IntechOpen

Protein-Protein Interactions

Computational and Experimental Tools

Edited by Weibo Cai and Hao Hong

PROTEIN-PROTEIN INTERACTIONS – COMPUTATIONAL AND EXPERIMENTAL TOOLS

Edited by **Weibo Cai** and **Hao Hong**

Protein-Protein Interactions - Computational and Experimental Tools
http://dx.doi.org/10.5772/2679
Edited by Weibo Cai and Hao Hong

Contributors

Amirhossein Sakhteman, Hamid Nadri, Alireza Moradi, Nicolas Ferey, Patrick Bourdot, Alex Tek, Marc Baaden, Olivier Delalande, Matthieu Chavent, Yoshihito Niimura, Takeshi Hase, Jian Huang, Beibei Ru, Ping Dai, Jarek Meller, Alexey Porollo, José Campos-Terán, Jaime Mas, Rolando Castillo, Paola Mendoza Espinosa, Narcis Fernandez-Fuentes, Joan Segura, Jose Ramon Blas Pastor, Shuichi Hirose, Hamid Ravaee, Pavol Jancura, Elena Marchiori, Maruthi Krishna Mohan Poluri, Baldo Oliva, Joan Planas-Iglesias, Jaume Bonet, Manuel A. Marín-López, Elisenda Feliu, Attila Gursoy, Zelmina Lubovac, Jifeng Zhang, Lusheng Wang, Tien-Hao Chang, KiYoung Lee, Woojin Jung, Hyun-Hwan Jeong, Takatoshi Fujiki, Takuya Yoshihiro, Jianhua Xing, Fan Bai, Zhanghan Wu, Philip Hochendoner, Jianshi Jin, Ernesto Iacucci, Samuel Xavier De Souza, Yves Moreau, Jurisica, Weibo Cai

Notice

First published in Croatia, 2012 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

Protein-Protein Interactions - Computational and Experimental Tools
Edited by Weibo Cai and Hao Hong

p. cm.

ISBN 978-953-51-0397-4

eBook (PDF) ISBN 978-953-51-4312-3

Meet the editors

Dr Weibo Cai received his BS degree from Nanjing University in 1995 and a PhD degree in Chemistry from UCSD in 2004. After three years of post-doctoral training at the Molecular Imaging Program at Stanford University, he joined the University of Wisconsin - Madison as a Biomedical Engineering Cluster Hire in February 2008 and is currently an Assistant Professor with joint appointment in the Departments of Radiology and Medical Physics. Prof Cai has authored > 90 peer-reviewed articles which have been cited > 3,100 times (H-index: 29). Prof Cai has won many prestigious awards and served on the Editorial Board of 18 scientific journals. He is currently the Executive Editor of the American Journal of Nuclear Medicine and Molecular Imaging.

Dr Hao Hong received his PhD degree in Biochemistry and Molecular Biology from Nanjing University in 2008. He is currently a Research Associate under the supervision of Prof Weibo Cai in the Department of Radiology at University of Wisconsin - Madison. Dr Hong has published > 30 peer-review articles and won many awards, including the Susan G. Komen Postdoctoral Fellowship and multiple Travel Awards to attend international conferences. Dr Hong is currently a member of the Society of Nuclear Medicine (SNM) and the World Molecular Imaging Society (WMIS). His research interests involve the design and optimization of novel molecular imaging and therapy agents based on peptides, antibodies, and various nanomaterials for the diagnosis and treatment of cancer and other diseases.

Contents

Preface

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. To provide a centralized resource for scientists who are either new to or working in the area of PPIs, we have organized this book. An international ensemble of experts in the field were invited to contribute a total of 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

The section of "Computational Approaches" contains 14 chapters. In the first chapter, Dr. Porollo and Dr. Meller gave an excellent review of the computational methods for the prediction of protein interaction sites, which were mainly focused on structure-based approaches. Next, an international team of experts from France, United Kingdom, and USA summarized the recent advances that are related to interactive molecular simulation approaches. Simulation design, software architectures, and applications in protein-protein docking were all discussed in exquisite detail. The following chapter, written by Jung et al. from the Republic of Korea, reviewed the PPI data available through public databases. Both non-network-based and network-based approaches were discussed, along with computational prediction methods of protein subcellular localization by exploiting the PPI data. Dr. Lubovac-Pilav from Sweden focused on defining the similarity between protein interactions based on an integrated score. The SWEMODE (Semantic WEights for MODule Elucidation) algorithm was discussed in detail in this chapter.

Next, Dr. Wang from Hong Kong, China introduced the use of quasi-bicliques for finding interacting protein group pairs and proposed approximation and heuristic algorithms for finding large quasi-bicliques in PPI networks. In the following chapter, Fujiki et al. from Japan focused on the interactions among three proteins. The

combinatorial effect level, which emerges only when those three proteins gather, was derived and estimated in a fully statistical manner. Dr. Hirose provided an excellent review on PPI prediction by computational techniques. The concepts and applications of several methods for inferring PPIs were covered, along with the databases and prediction methods that deal with protein flexibility, as well as the possibility of inferring PPIs from protein dynamics.

Prof. Xing and co-workers presented a unified mathematical formalism describing both conformational change and chemical reactions of proteins. The implications of slow conformational changes in protein allostery and network dynamics were also discussed in this chapter. Next, Prof. Huang and colleagues reviewed the methods for prediction of PPI sites using mimotope analysis. The current status, as well as the challenges and future directions of the field, were summarized. Prof. Oliva from Spain covered the strategies for modeling the interaction between two proteins from sequence data and reviewed the existing techniques to model large cellular protein complexes. In the next chapter, Dr. Chang focused on the concept of co-occurrence pattern and implementation details of methods in PPI prediction based on this concept.

Sakhteman et al. from Iran gave an overview on the biochemistry details of G-Protein coupled receptors (GPCRs) and provided information on homology modeling and molecular dynamic simulation methods for studying interactions involving GRPRs. Next, Dr. Iacucci and Dr. Moreau from Belgium evaluated the application of least square support vector machines (LS-SVM) to receptor-ligand interaction prediction and discussed various other methods to study PPIs, most of which relying on the phylogenetic profile analysis of candidate interactors. In the last chapter of this section, Ravaee et al. from Iran introduced the fuzzy learning vector quantization (FLVQ) as a high tolerant method for clustering PPI network to find protein complexes, which is less vulnerable to false-negative and false-positive interactions in PPI data than other techniques.

Although computational simulation is a powerful tool for studying PPIs, novel experimental approaches for investigating PPIs that can overcome the limitations of existing techniques are continuously been developed. Such techniques represent a vibrant area of research on PPIs. In the section of “Experimental Approaches”, the current state-of-the-art experimental strategies to study PPIs are presented in four chapters.

Molecular imaging, an extremely powerful tool to study molecular events in living subjects, can provide invaluable information and insight in elucidating the process of various PPIs. In the first chapter of this section, we summarized the current status of in vivo imaging of PPIs with various techniques, including fluorescence, bioluminescence, and positron emission tomography imaging. Next, Dr. Mohan illustrated the theoretical aspects of non-linear behavior of amide proton chemical shifts. In this chapter, he demonstrated the residue level nuclear magnetic resonance

(NMR) description of the low energy excited states representing locally different alternative conformations in different complex protein systems. Mendoza-Espinosa et al. described the physics and chemistry behind the disorder-to-order transitions in proteins and introduced different experimental measures to study the structure and function of multiple types of apolipoproteins. The last chapter of this section, contributed by Dr. Zhang, focused on the specific modulation of electrostatic interactions between proteins by salt.

The third section of this book contains four chapters that do not readily fall into either of the abovementioned categories. In the first chapter of this section, Prof. Fernandez-Fuentes and colleagues presented the theoretical basis of computational tools designed to predict PPIs, and then focused on the computational methods developed to predict protein interfaces. Dr. Hase and Dr. Niimura summarized the current knowledge of the statistical properties of PPI networks. They also reviewed the studies related to drug discovery and the possibilities of medical studies as an integration of network and evolutionary biology. The next chapter written by Dr. Jancura and Dr. Marchiori gave a general overview of the relevant literature and advances in the analysis and application of evolution in PPI networks. Lastly, Otasek et al. described pathway-centric analysis and the analysis of networks generated from protein-target interactions, which can elucidate the role of these proteins.

The research field of PPIs is highly dynamic and constantly evolving. We are truly grateful to the exceptional team of authors for their tremendous effort, all of whom have many responsibilities and yet they spent countless hours in these 22 chapters to make this book possible. With such whole-hearted support and participation from international experts/leaders of the field, we are confident that this endeavor will serve as a comprehensive reference book and help moving the field forward.

Weibo Cai, PhD
Assistant Professor
Departments of Radiology and Medical Physics
University of Wisconsin - Madison
USA

Hao Hong, PhD
Research Associate
Department of Radiology
University of Wisconsin - Madison
USA

Part 1

Computational Approaches

1

Computational Methods for Prediction of Protein-Protein Interaction Sites

Aleksey Porollo and Jaroslaw Meller
University of Cincinnati
USA

1. Introduction

Studies of protein-protein interactions play a central role in understanding protein function in biological systems, closing the gap between large-scale sequencing efforts and medically relevant outcomes. Increasingly, protein interaction interfaces that mediate communication between proteins are becoming targets for therapeutics, offering a possibility to disrupt critical interactions and specifically attenuate function (Fletcher and Hamilton, 2007; Fry, 2006).

Efforts to catalog, characterize, and link protein interactions with disease states and other phenotypes are ongoing, building on improvements in experimental techniques, such as high throughput two-hybrid assays or chip-based proteomics. Significant progress has also been achieved in structural genomics, providing detailed information for a growing number of macromolecular complexes and interaction interfaces by means of X-ray crystallography, NMR spectroscopy and other methods.(Aloy et al., 2005; Slabinski et al., 2007)

Despite impressive progress, existing experimental methods for mapping protein interactions suffer from many limitations. High throughput methods, such as two-hybrid or chip-based essays, are characterized by high rates of false positives and false negatives (Bader and Chant, 2006; Han et al., 2005), requiring further validation and detailed characterization of individual interactions. Obtaining detailed high-resolution information about protein interaction interfaces can also be challenging in many instances.

For example, some complexes may not crystallize, or crystallize in a different than biologically relevant conformation. X-ray crystallography may also fail when multiple and incompletely mapped interactions or membrane domains are involved.(Lacapere et al., 2007) This is exacerbated by the fact that each protein has been estimated to have around 9 distinct interacting partners (and some are estimated to have hundreds interactants), with majority of the implied complexes unlikely to be resolved experimentally in the foreseeable future.(Aloy and Russell, 2004; Ritchie, 2008)

Limitations of experimental techniques and attempts to circumvent the problem by focusing directly on protein interactions create an opportunity for computational approaches to complement and facilitate experimental efforts in that regard. In particular,

statistical and machine learning-based approaches are being increasingly used to facilitate identification of protein interfaces. There are a growing number of methods for protein interaction sites prediction that vary in terms of principles of the recognition of interaction interfaces, descriptors used to identify interacting sites (feature space) and learning algorithms used.

From the point of view of a representation used to capture characteristics of interaction interfaces, one may distinguish two main groups of methods. The first group attempts to predict interaction sites using sequence information only.(Gallet et al., 2000; Ofran and Rost, 2007) The second group of methods, takes available structural information into account (Fariselli et al., 2002; Lichtarge et al., 1996), typically involving the identification of sites on the surface of a monomeric structure that are either evolutionarily conserved (as for example in the pioneering evolutionary trace method by Lichtarge and colleagues (Lichtarge et al., 1996)), or have a propensity for interaction interfaces (see, e.g., (Jones and Thornton, 1997)).

Although evolutionary trace methods are relatively insensitive to structural detail and can identify conserved "hot spots", their overall accuracy is limited.(Caffrey et al., 2004; Porollo and Meller, 2007) On the other hand, detailed structural information can be used to characterize patches on the surface of a protein in terms of their geometric and other properties (see, e.g., (Bordner and Abagyan, 2005; Koike and Takagi, 2004; Neuvirth et al., 2004)). Structural conservation can also be taken into account when multiple structures within families are available.(Chung et al., 2006; Ma et al., 2003)

While structural information improves prediction accuracies (with the risk of increasing the sensitivity to the choice of a specific structure), challenges remain and new insights are required to improve state-of-the-art in the field.(de Vries and Bonvin, 2008; Zhou and Qin, 2007) Further progress also requires continued systematic evaluation of new methods. In this regard, the lack of standard definitions and consistent evaluation criteria adds to the challenge and often makes direct comparison of existing methods impossible.

One problem that contributes to the difficulty of fair evaluation and objective comparison of different methods is related to the uncertainty concerning the definition of the negative class. The assignment to the "non-interacting" class is at best tentative, given the incompleteness of information regarding all possible interactions and interacting partners. Despite the growing number of resolved structures of protein-protein complexes, another challenge is the relative paucity of carefully curated and properly stratified (to represent different types of complexes) benchmarks.

This chapter reviews computational methods for the prediction of protein interaction sites, with a primary focus on structure-based approaches. The goal is to help the reader better understand the underlying concepts and limitations pertaining to current methods in the field. A number of methodological issues related to the training and validation of such methods are discussed as well. The benchmarks and assessment included in this chapter should also help making an informed decision as to when computational predictions can be regarded as sufficiently confident for a particular system of interest to warrant further experimental validation.

2. Definition of protein-protein interaction site

The recognition of protein-protein interaction sites can be cast as a classification problem, i.e., each amino acid residue is assigned to one of the two classes: interacting or non-interacting residues. Consequently, the problem may be solved using statistical and machine learning techniques, such as neural networks (Ofran and Rost, 2003b; Zhou and Shan, 2001) or Support Vector Machines (Bock and Gough, 2001; Yan et al., 2004).

A clear definition of interacting residues is obviously required in order to predict whether a given amino acid residue is involved in protein-protein interactions. However, many alternative definitions are being used in the field. As the definition of an interaction site varies from one prediction method to another, it becomes difficult to directly compare their performance.

2.1 Commonly used definitions

If available, high resolution structural data readily provides a basis for atom or residue based definition of interaction sites. In fact, prediction methods discussed in this chapter primarily use information from resolved protein complexes to define the positive ("interacting") and negative ("non-interacting") classes. Protein quaternary structures are typically resolved by X-ray crystallography, and less frequently by NMR-spectroscopy or other techniques (Protein Data Bank, PDB – http://www.pdb.org/). While providing a high resolution structure, crystallographic data often remains inconclusive regarding the nature of the observed intermolecular contacts between protein chains. In particular, some of the observed contacts (and the resulting putative interaction interfaces) may be the result of crystal packing, rather than representing biologically relevant interactions.

A number of methods have been introduced to facilitate the process of filtering out crystal packing artefacts. Here, we used the approach adopted by the PISA server (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html). PISA discriminates crystal packing contacts from the functional protein–protein interaction using the size of solvent exposed area buried during association, as well as the number of residues constituting the interface, the number of salt and disulphide bridges at the interface, and the difference in approximate solvation energy upon complex formation.(Henrick and Thornton, 1998; Krissinel and Henrick, 2007)

Two different approaches are commonly used to define an interaction site based on 3D structural data: (i) interatomic distance and (ii) change in accessible surface area (ASA) upon complex formation. Following the first approach, interaction sites can be defined based on the distance between non-hydrogen atoms of different protein chains. For example, distance cutoffs of 4Å (Bordner and Abagyan, 2005); 4.5Å (Hamer et al., 2010); 5Å (Chen and Zhou, 2005); or 6Å (Ofran and Rost, 2003b) are used. This way of defining interaction sites is likely to miss some interchain contacts when water molecules are involved. A polar solvent, such as water, may bridge the interaction between two charged groups of amino acids that are too far apart to form a direct hydrogen bond.(Janin, 1999) In this regard, Neuvirth *et al.* introduced the Connolly interface index (CII) that is computed for circles of radius 10 Å around anchoring dots on the surface of monomeric structures. Atoms with CII above certain threshold are assigned to be interaction sites.(Neuvirth et al., 2004)

The second approach defines an interaction interface by using the concept of solvent accessibility or ASA. Specifically, ASA or the solvent accessibility of an amino acid residue in an unbound protein chain is contrasted with the corresponding ASA value for the same residue in a complex. Residues with a significant difference in ASA between the isolated chain and complex structures are then classified as "interacting". The following cutoffs for ASA were used: the loss of > 99% ASA for a given atom (Bradford and Westhead, 2005); a residue loses > 1Å^2 ASA in the complex (Chen and Jeong, 2009; Jones and Thornton, 1995; Liang et al., 2006); a residue ASA change by more than 20Å^2 (Kufareva et al., 2007); relative solvent accessibility (RSA) of a given residue decreased by more than 4% and its ASA decreased greater than 5Å^2 (Porollo and Meller, 2007). The latter definition uses relative ASA to address the considerable difference in size of amino acids, e.g. between glycine and tryptophan.

Both approaches require high resolution structural data. However, the interatomic distance based approach seems to be more sensitive to problems with missing atoms or atoms with multiple occupancies. Table 1 illustrates the difference in the protein interface recognition resulting from alternative definitions. As can be seen from the table, the same protein quaternary structure may yield different subsets of residues deemed to be interaction sites, therefore leading to different prediction models and their reported performances.

In what follows, we will refer to protein interfaces derived using our own ASA-based definition, dRSA > 4% and dASA > 5Å^2 (Porollo and Meller, 2007), unless stated otherwise. This definition takes into account both relative and absolute change in ASA, and it attempts to filter out noise related to variation in RSA observed in structures resolved under different conditions, or for closely related homologs.

Definition	Chain	Residues at the interface	Interface ASA, Å^2
dASA > 1Å^2	I	Y35 T41 C42 H57 C58 D60 R61 N95 T96 D97 D98 V99 A99A L143 L151 W172 T175 C191 Q192 G193 S195 T213 S214 F215 V216 S217 R217A L218 K224	830
	E	I18 I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 G47 S48 C49 A52 C53 F54	994
dRSA > 4% and dASA > 5Å^2	I	Y35 T41 H57 D60 R61 T96 D97 V99 A99A L143 L151 W172 T175 C191 Q192 G193 S195 S214 F215 V216 S217 R217A L218	810
	E	I18 I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 G47 S48 C49 A52 F54	989
dASA > 20Å^2	I	Y35 H57 R61 T96 D97 V99 W172 Q192 S195 F215 V216 R217A L218	692
	E	I19 L20 I21 R22 C23 A24 M25 L26 N27 P29 R31 E46 S48 C49	938

Table 1. The effects of using alternative definitions of protein interaction interfaces for a specific hetero-dimeric complex (PDB ID 1fle); dASA is the total loss of ASA for a given protein chain upon complex formation.

It should be noted that information on protein interaction sites may be also derived from the alanine scanning mutagenesis (ASM). Systematic replacement of the residues at the protein interface with alanine enables the evaluation of individual contribution of each interaction site to the binding energy. In this regard, the Alanine Scanning Energetics database (ASEdb, http://www.asedb.org/) provides ASM data on a number of protein-protein, as well as on some protein-DNA and protein-ligand interactions (Thorn and Bogan, 2001)

However, ASM approach is very costly and laborious, thus considerably limiting the number of comprehensively studied proteins. A protein interface needs to be approximately defined beforehand to limit the number of alanine mutants to evaluate. Results of ASM may not necessarily indicate the contribution to the binding energy, as some alanine mutants may cause an adverse protein conformational change and therefore indirectly decrease the efficacy of the protein-protein binding. Moreover, some protein-protein interactions are allosterically regulated, and ASM may not reflect the actual driving forces for a given protein complex. Nevertheless, such data is of great value and may be used as an additional validation of prediction methods. For example, it was used to evaluate ability of the methods ISIS (Ofran and Rost, 2007) and APIS (Xia et al., 2010) to identify hot spots.

2.2 Mapping interaction sites

Methods that do not require information about the interacting partner(s) are the primary focus of this chapter. These methods aim at the recognition of either individual residues, surface patches, or whole interaction interfaces using only sequence, structure and other information about an individual target protein, assuming that it is involved in some sufficiently stable interactions.

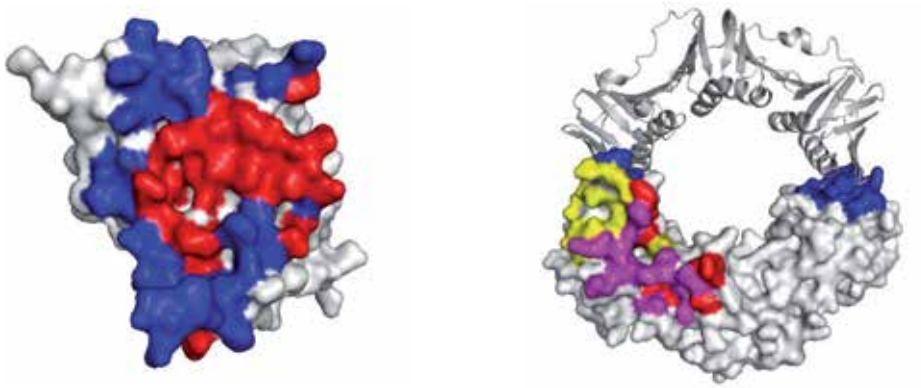
In light of the above, an important part of defining the residues as interaction sites is to retrieve as much information as possible on physical interactions for a given protein. Published studies on methods for the prediction of protein-protein interaction quite often ignore the fact that most proteins have multiple interaction partners that are mediated by alternative or overlapping interfaces. Therefore, using just one particular complex to identify the interaction interface and to derive the corresponding definition of the positive class, while ignoring all other complexes and interactions involving the same target protein chain (or its close homolog), may result in highly biased estimates of both false positive and false negative rates.

With the significant growth of structural data, the problem can be addressed by taking into account interaction sites from alternative complexes that contain the same protein chain or its close homologs. Interaction sites identified in such homologs can be mapped to a representative sequence in order to enable more sensitive prediction and perform its fair accuracy evaluation. Figure 1 illustrates this issue for two proteins resolved in complexes with different partners.

The protein shown in the left panel, caspase-9, utilizes overlapping interfaces for homo-oligomerization (PDB ID 1jxq), and for its interaction with ecotin (PDB ID 1nw9). However, the former protein-protein interaction involves many more residues than the latter interaction (affected ASA 1954Å^2 and 1019Å^2, respectively). If the definition of the positive ("interacting") class in caspase-9 were to be derived from the complex with ecotin (1nw9),

the accuracy of any method predicting correctly also the more extensive interface would have been wrongly underestimated. This problem can be addressed by mapping the interface from the homooligomer into the target structure, leading to the union of homo-dimerization and caspase-9/ecotin interfaces to be taken as the true positive class.

The second example on the right illustrates the mapping of the known interfaces into the beta subunit of *E. coli* DNA polymerase III. In addition to homodimerization interface (PDB ID 2pol), physical interactions with the delta subunit of the gamma complex (PDB IDs 1jqj, 1jql) and DNA polymerase Pol IV (PDB ID 1unn) are mapped. Again, without this additional mapping step, prediction of these alternative interfaces would be considered as false positives during the evaluation process.



A B

Fig. 1. Mapping interfaces from alternative protein complexes: A. Interaction interfaces in caspase-9, derived from the complex with ecotin (PDB ID 1nw9, chains B-A, shown in red) and caspase-9 homooligomer (PDB ID 1jxq, chains A-B), which includes both red and blue patches; B. Interaction interfaces mapped into DNA Pol III from the homodimer of the beta subunit of DNA Pol III (PDB ID 2pol, blue), delta subunit (PDB IDs 1jqj and 1jql, red), and DNA Pol IV (PDB ID 1unn, yellow), with the overlap of the latter two shown in magenta. Interfaces identified by using the SPPIDER server (http://sppider.cchmc.org/) and mapped into the target structure by using POLYVIEW-3D (http://polyview.cchmc.org/polyview3d.html).

The mapping, though, needs to be performed carefully, keeping in mind some important caveats. Sequence homology-based approach assumes that similar protein sequences adopt the same 3D fold and carry the same function, which is not always true. For example, paralogs may evolve to have distinct interaction partners and therefore perform different functions while having high sequence homology. Mapping interaction sites from such homologs might then result in incorrect expansion of the positive class to include patches utilized by other proteins with sequence similarity but distinct functions. In this context, one should comment that many methods for the prediction of interaction sites incorporate information about evolutionary profiles of protein families (e.g., obtained using PSI-BLAST to generate PSSM (Altschul et al., 1997)). Therefore, at least in some cases such methods arguably identify sites with a propensity to interact within the whole family, rather than just for the target protein.

Interactions specific to only some (or even only one) family members may require the identification of distinct interaction patches, rather than considering the problem of predicting the union of alternative interaction interfaces. Thus, mapping interaction interfaces might not be appropriate for evaluation of methods that attempt to predict such individual interaction patches. On the other hand, if ANY interaction patch that corresponds to a stable protein complex is to be found, then the union of all known interfaces constitutes the best approximation of the positive class and should be used for evaluation of the overall accuracy. As indicated above, this issue is often ignored altogether, even though it highlights the difficulty with a proper definition of a classification problem that best captures biologically relevant information while providing sufficiently "accurate" predictions.

Conversely, some protein domains with conserved 3D structure and specific function may be very divergent in terms of amino acid sequence, and only structure alignment might be able to detect such distant similarity. For example, PB1 domain displays low sequence homology between proteins, but it has a highly conserved secondary structure pattern and the overall 3D fold.(Lamark et al., 2003) While having just a few conserved residues playing a role of hot spots, this domain is widely utilized in various biological systems for interactions between the PB1-containing proteins to conduct cell signaling.(Moscat et al., 2006)

A PDB-wide structure alignment remains a computationally challenging task when it comes to a large protein set compiled for training or benchmarking a method for protein-protein interaction prediction. However, some current efforts, including for example the Dali database (http://ekhidna.biocenter.helsinki.fi/dali/start) (Holm et al., 2008), provide valuable resources in this regard. There have been also a number of studies published on the structure-based mapping of interaction sites, utilizing different schemes of hit weighting and homology recognition.(Albou et al., 2011; Oldfield, 2002; Park et al., 2001; Xu and Dunbrack, 2011)

However, it remains to be seen how structure-based mapping methods can deal with situations when a protein undergoes a significant conformational change upon complex formation (e.g., in case of calmodulin), and a structure alignment is likely to fail to identify similarity between apo- and holo-forms. Most likely, the future methods will utilize a balanced combination of sequence- and structure-based homology in order to more accurately map interaction sites from the known physical interactions. In this work, in order to test the effects of mapping interaction sites from multiple resolved complexes, we used a sequence homology-based mapping with conservative thresholds for homology hits: 70 or 90% of sequence identity. The interaction sites mapping process was automated through the SCORPPION web-server (http://scorppion.cchmc.org/).

3. Types of protein complexes

Biological diversity is very well represented at molecular level, in particular showing broad versatility in protein-protein interactions. Protein complexes can be classified into a number of broad categories, for example as homo- and hetero-oligomers; transient and obligatory (permanent), rigid and flexible complexes. Homo-oligomers are complexes consisting of two or more protein chains with identical amino acid sequence. Accordingly, assemblies of chains with different sequences are hetero-oligomeres. The number of chains participating in the assembly dictates the distinction on dimers, trimers, tetramers, and so forth.

Obligatory complexes (sometimes called obligomers) are considered to be protein assemblies that perform function only in the coupled state, whereas transient complexes are formed by proteins that were found to exists as monomers and to function separately as well. Rigid complexes may be considered as products of interaction between stable rigid-body domains. Flexible complexes, on the other hand, are formed when one or more constituting proteins undergo significant conformational changes.

Systematic analysis of the known protein complexes by several studies resulted in a number of observations that have significantly influenced the field of protein-protein interaction sites prediction. Ofran and Rost suggested that there are at least 6 types of contacts in proteins that display distinct amino acids compositions and contact preferences.(Ofran and Rost, 2003a) Thus, methods utilizing statistical contact propensities in their prediction models have to take into account different types of interactions. Another study found that even within a single interface the composition of amino acids varies depending on where the interacting amino acids are located, in the core of the interface or at its rim.(Chakrabarti and Janin, 2002)

A closer look at transient complexes was presented in (Nooren and Thornton, 2003). The study distinguished "weak" and "strong" homodimers, and it found that weak transient homodimers demonstrate smaller, more planar and polar interfaces compared to permanent homodimers, whereas strong transient homodimers undergo large conformational changes upon complex formation, and demonstrate larger, less planar, and more hydrophobic interfaces. Interestingly, only weak transient homodimers were found to have residues at interfaces more conserved than other surface residues, whereas other proteins with different oligomeric states showed no pronounced amino acid conservation.

These findings were further supported by the study on a larger set of protein complexes.(Caffrey et al., 2004) Comparing the conservation scores derived from multiple sequence alignments to orthologs vs. paralogs, the study demonstrated that residues at the interfaces are rarely more conserved than other residues on the protein surface. This observation implies that prediction models solely based on evolutionary profiles are likely to have limited overall accuracy.

Another large scale study has recently reported the results of PDB-wide analysis of protein-protein interactions. Both sequence and structure based characteristics of protein interfaces were characterized, with special focus on proteins with multiple interaction partners.(Kim et al., 2006) This analysis showed that, while there are ancient interfaces conserved across archea, bacteria, and eukaryotes (attributed primarily to symmetric homodimers), by and large interfaces are not conserved and vary in shape and amino acid composition due to broad diversity of interactions and interaction partners. The suggested classification introduced as many as 6000 different types of interfaces that are available for search and matching from the SCOPPI database (http://www.scoppi.org/).

4. Benchmarks of protein complexes

Benchmarks specifically designed for the training and evaluation of methods for the recognition of protein-protein interaction sites are critical for further progress in the field. Such benchmarks should allow an unbiased and fair evaluation of prediction methods. Consequently, benchmark sets used for comparison of different methods should comprise a

diverse representative set of protein-protein interactions and contain no redundancy to the training sets used by individual methods.

The uncertainty of the negative class assignment further complicates the choice of appropriate benchmarks. Designing a dataset that includes only carefully curated and well-studied proteins, or their domains, with all known physical interactions mapped, may result in a very limited number of data points for training and validation. As a more feasible alternative one could consider assembling several diverse and non-redundant training and validation data sets that include complexes of different type and are characterized by some level of completeness of information regarding interactions and interaction sites.

As a result of these difficulties, there is no established gold standard in the field. Most of the published methods refer to their own compilation of protein complexes derived from PDB. Here, we consider three protein sets used in the literature. The first compilation of protein complexes is a benchmark set for protein-protein docking, current version 3.(Hwang et al., 2008) For this set, proteins in bound and unbound state were retrieved from PDB in a semi-automated manner. Current version contains the total of 124 test cases; among those 88 are rigid-body cases, 19 of medium difficulty, and 17 difficult cases, which are classified by the degree of conformational change at the interface upon complex formation.

While the primary purpose of Hwang *et al*. benchmark was to evaluate the protein docking methods, many protein interface prediction methods used it for their own and comparative evaluation.(de Vries and Bonvin, 2011; de Vries et al., 2006; Fiorucci and Zacharias, 2010; Guharoy and Chakrabarti, 2010; Li et al., 2008; Liu and Zhou, 2009; Qin and Zhou, 2007; Zhou and Qin, 2007) However, a thorough analysis of this benchmark set led us to conclusion that it is not suitable for evaluation of the methods predicting protein-protein interaction sites. For example, it contains 25 antibody-antigen cases (PDB IDs: 1fc2, 1ahw, 1bvk, 1dqj, 1e6j, 1jps, 1mlc, 1vfb, 1wej, 2fd6, 2i25, 2vis, 1bj1, 1fsk, 1i9r, 1iqd, 1k4c, 1kxq, 1nca, 1nsn, 1qfw, 2jel, 1bgx, 1e4k, 2hmi), which are asymmetrical functional protein-protein interactions, i.e. while one partner (in general: antibody, protease, or major histocompatibility complex) is evolved to bind its substrate, the second partner is not (except for the protease inhibitors).

Therefore, all antibody-antigen complexes were removed from the set. In addition, protein chains no longer available in PDB (PDBID_ChainID: 1cd8_B, 1ml0_B, 2pab_C, 2pab_D, 2viu_C, 2viu_E, 1aly_B, 1aly_C, 1jb1_B, 1jb1_C), difficult to interpret in terms of protein chains (1hia_A, 1hia_B, 1n8o_B, 1n8o_C) or too short (1n8o_A, 1k74_B, 1mzn_B, 1zgy_B) were removed. Finally, before using this benchmark set for evaluation of protein interface prediction methods, redundant chains were also removed.

The second benchmark set represents 85 cases of proteins found in PDB both in bound and unbound state.(Albou et al., 2009) No complexes with asymmetrical function are included, such as antibody-antigen cases and others listed above. This set represents diverse protein-protein interactions and allows the evaluators to estimate the role of conformational change on the accuracy of the methods, when predictions using bound structures *versus* unbound are compared. However, the set contains two cases, when only α-carbon coordinates are available (PDBID_ChainID: 3dpa_A and 2tld_I). These cases may be challenging to prediction methods that rely on high resolution data with all atoms resolved.

The last benchmark set to be used in this work is the control set of the SPPIDER method.(Porollo and Meller, 2007) It was compiled based on the protein complexes

deposited in PDB after the compilation of the training set for the same prediction method. This manually curated and non-redundant (to the training set and within itself) set includes 149 protein chains, deemed to be sufficiently diverse and representative enough to be used for cross-validation studies. The only update to the set involved replacing the chain 1r72_A by 1xcb_A, as the PDB entry 1xcb now supersedes 1r72. In what follows, this set is referred to as SPPIDER149.

Table 2 and Figure 2 summarize the three datasets described above, after removing problematic cases from the first set, and redundant proteins from the first two sets. Redundancy was defined in terms of sequence homology: BLAST e-value < 0.001 when the alignment covers at least 70% of the query sequence (derived from the ATOM section of a PDB file). 150 chains derived from complexes in the first set and 78 chains in the second set were found non-redundant, and these (sub-) sets will be referred to as Hwang150B and Albou78B, respectively. The corresponding sets of chains that were retrieved from their unbound structures will be referred to as Hwang150U and Albou78U, respectively.

Dataset	Total chains	Families	Domains
Hwang150B	150	42	107
Albou78B	78	16	44
SPPIDER149	149	76	75

Table 2. Protein families and domains represented in non-redundant chains of the three benchmark sets used in this work. Families and domains defined according to the Pfam database (http://pfam.sanger.ac.uk/) (Finn et al., 2008) and mapped using sequence based search as implemented in SCORPPION (http://scorppion.cchmc.org/).

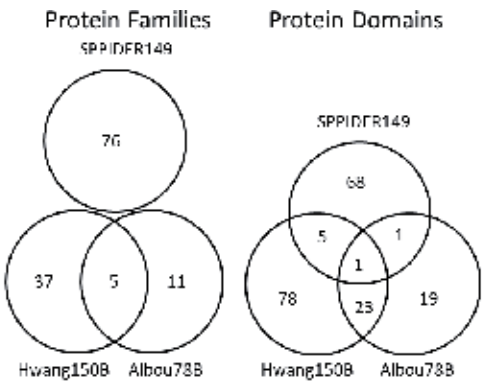


Fig. 2. Overlap between protein families (left) and domains (right) identified within the three benchmark sets used here.

Low to no overlap between the datasets discussed here is observed in terms of protein families and domains, suggesting a broad coverage of protein-protein interactions. This bodes well for estimates of the performance on different types of protein interfaces. On the other hand, the training sets for tested methods might partially overlap with the benchmark sets used here, leading to potentially overestimated accuracy.

Mapping of known interaction interfaces from alternative complexes was performed for each set using different approaches discussed in Section 2.2. Table 3 shows the number and

fraction of interacting residues for each protein set. Interaction sites were derived from (i) asymmetric units defined in the original PDB files, (ii) biological units (BUs) as defined by Protein Quaternary Structure (PQS) database, and (iii) BUs as defined by the PISA database. In addition, interaction sites were mapped from the PISA-based BUs of their close homologs using sequence identity 90 and 70% as a cutoff (Table 4). The estimates of accuracy for methods compared here were overall quite similar, and only the results for the latter threshold are reported in the following sections of the chapter.

PDB also provides its own definition of biological units that differs from PISA.(Xu and Dunbrack, 2011) PDB defines biological units as separate models in the same PDB file. In addition, both PISA and PDB may rename chain labels starting from 'A' within each BU. This all makes it difficult sometimes to trace back the chains from the asymmetrical unit in automated manner. To be consistent, we will map interaction sites from BUs as defined by PISA. However, when no information can be mapped for a given chain, due to technical difficulties or inconsistency in BU definition, we will use a PDB-based asymmetric unit for the mapping of interaction sites.

Dataset	Total residues / On the surface	PDB-based, %	PQS-based, %	PISA-based, %
Hwang150B	31208 / 24687	19	21	19
Albou78B	17412 / 13375	16	16	15
SPPIDER149	25883 / 20885	29	28	28

Table 3. Summary of the benchmarks used in this work with regards to the total number of residues, residues on the surface, and percentage of the surface residues found to be at protein interfaces derived from the asymmetric unit (PDB-based), and biological units (PQS-based and PISA-based), respectively.

Dataset	Total / Surface residues	SI70	SI90
Hwang150B	31208 / 24687	10011	9674
Hwang150U	32471 / 24595	10201	9661
Albou78B	17412 / 13375	5819	5506
Albou78U	16838 / 12342	5572	5294
SPPIDER149	25883 / 20885	7863	7668

Table 4. Summary of the benchmarks used in this work with regards to the total number of residues, residues on the surface, and interacting residues on the surface mapped to representative protein chains using BUs derived from the PISA database and 70 or 90% sequence identity cutoffs (SI70 and SI90), respectively.

5. Prediction methods

All prediction methods can be broadly classified by the type of data they use as an input. Sequence-based methods rely on some combination of the following protein features: amino acid hydrophobicity, evolutionary profile (e.g., similarity scores or Shannon entropy), amino acid composition or propensity to be at the interface, predicted structural features (e.g., secondary structure, solvent accessibility, order/disorder region, *etc.*), or their derivatives like mean or weighted average over a sequence window.

The structure-based methods, on the other hand, also utilize features derived from a 3D protein structure, such as solvent accessibility and secondary structure states, local topology (e.g., protrusions and cavities), hydrophobic and polar surface patches, temperature or B-factors (for X-ray based structures), etc. In addition, there are a number of methods built using a consensus of the individual predictors with reportedly improved accuracy.(de Vries and Bonvin, 2011; Huang and Schroeder, 2008; Qin and Zhou, 2007) However, consensus-based methods are not discussed here in detail, as the goal is to evaluate the discriminating power of the underlying principal features for each representative method.

Described below are selected structure-based methods with at least somewhat orthogonal feature spaces that were available as web-servers at the time of data preparation for this work. Methods are listed in the order of the publication year of the original work.

Evolutionary trace (ET) method (Lichtarge et al., 1996) identifies evolutionary conserved residues and maps them onto a protein 3D structure. Conserved residues in the core of a protein are deemed to be structurally important, whereas those on the surface are assumed to be functionally important. The method starts from constructing a multiple sequence alignments, and partitions the aligned sequences into groups by using their mutual sequence similarity. For each group, a consensus sequence is defined highlighting the positions with invariant amino acids. Consensus sequences are further aligned to identify (i) conserved residues across the entire protein family; (ii) class-specific residues that are invariant in some groups; and (iii) neutral residues that are not preserved in any single sequence group. Conserved and class-specific residues are then mapped onto 3D structure. Clusters of such residues on the surface of a protein structure are predicted to be functional. The ET method is available at http://mammoth.bcm.tmc.edu/ETserver.html

ConSurf (Glaser et al., 2003) follows a similar approach by mapping the evolutionary conserved residues on 3D protein structure. The difference lies in computing the conservation scores that are relative with respect to other residues in a given protein. In addition, the outcome of the method is sensitive to the quality of multiple sequence alignment and to the overall length of a query sequence. For example, two 3D structures of the same protein, but with different sequence length representing its resolved part, may result in different location of the most conserved residues. The ConSurf method is available at http://consurf.tau.ac.il/, whereas its pre-computed results for the PDB deposited proteins are available from the ConSurfDB database (http://consurfdb.tau.ac.il/).

It should be noted that the two methods described above were not designed to identify specifically protein-protein interaction sites, but rather to reveal any functional residues, e.g. involved in protein-DNA or protein-ligand interactions. However, since the authors of these methods refer to identification of protein interfaces as examples in their original publications, we chose these methods to serve as a separate group of predictors that rely primarily on evolutionary information, and can be contrasted with structure-based methods.

PROMATE (Neuvirth et al., 2004) considers residues on the surface of a protein structure within 10Å circles around a given point. Spatially neighboring residues provide the following descriptors: (i) statistically derived chemical composition of binding sites, such as

propensity of individual amino acids, atom types, pairs of amino acids, and collective chemical properties (positively and negatively charged, polar, hydrophobic, and aromatic residues); (ii) evolutionary conservation in terms of diagonal elements of the PSI-BLAST-derived position specific scoring matrix (PSSM); (iii) distance in the sequence between residues in the circle; (iv) secondary structure states, including extent of the loops. Additionally, temperature factors (B-factors) and bound waters are incorporated into the model whenever available. These descriptors are combined to yield a cumulative score that allows the circles to be classified as Interface, Non-interface, or Boundary. The neighboring circles are further clustered to define predicted interface patches. PROMATE is available at http://bioinfo.weizmann.ac.il/promate/

Cons-PPISP (Chen and Zhou, 2005) employs a consensus of neural networks trained on (i) the position specific similarity scores derived from the PSI-BLAST multiple sequence alignment and (ii) observed (in the target structure provided as input) solvent accessibility for spatially neighboring residues. In addition to validation on crystal structures, cons-PPISP was shown to provide accurate prediction of protein interfaces for a set of 8 NMR-derived complexes, non-redundant to its training set. The web-server is available at http://pipe.scs.fsu.edu/ppisp.html

WHISCY (de Vries et al., 2006) introduces prediction scores that are based on evolutionary and structural information. Conservation of residues on the surface is computed as the corrected sum of similarity scores between amino acids at a given position by pairwise comparison of a query sequence and sequences from a multiple alignment. Similarity scores are taken from the Dayhoff mutation matrix. ASA is the only structural information used. WHISCY is available at http://nmr.chem.uu.nl/Software/whiscy/index.html

PIER (Kufareva et al., 2007) combines (i) statistically derived interatomic contact potentials, (ii) physical descriptors, such as observed solvent accessibility for separate atomic groups within amino acids, and (iii) sequence alignment based features, in particular, three different conservation scores (frequency-based, similarity matrix-based, and entropy-based). The surface of a protein structure is divided on individual patches. Using the descriptors listed above, all patches obtain a set of cumulative scores that further fed to a partial least squares (PLS) based regression model to predict protein interfaces. Since the PIER scoring heavily relies on atomic resolution, it may have difficulties with incomplete or of low resolution crystal structures. The corresponding prediction server is available at http://abagyan.ucsd.edu/PIER/

SPPIDER (Porollo and Meller, 2007) is a neural network-based method that uses the difference between predicted from sequence and observed in an unbound structure RSA of amino acid residue as a novel and highly informative signal of interaction sites. Solvent accessibility prediction methods tend to predict residues at protein interfaces as buried, which is consistent with the fact that they are indeed getting buried upon complex formation, even though they are exposed in an unbound structure. The SABLE (Adamczak et al., 2004) method for RSA prediction was used to generate the input for SPPIDER. Additional features include averaged over spatially neighboring residues of (i) RSA predicted by SABLE; (ii) evolutionary conservation (in terms of Shannon entropy) of amino acid type, charge, hydrophobicity, and side chain size; (iii) amino acid contact numbers and hydropathy constants. The server is available at http://sppider.cchmc.org/

6. Evaluation

6.1 Accuracy measures

Prediction of protein interaction sites is typically cast as a classification problem. Therefore, a number of commonly used measures for two class classification problems can be employed to evaluate the accuracy. These measures include the two-class classification accuracy (Q_2), recall or sensitivity (R), and precision or specificity (P), all expressed as percentage.

$$Q_2 = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \quad (1)$$

$$R = \frac{TP}{TP+FN} \cdot 100\% \quad (2)$$

$$P = \frac{TP}{TP+FP} \cdot 100\%, \quad (3)$$

where TP are true positives, TN – true negatives, FP – false positives, and FN – false negatives.

However, since the number of interaction sites can be much smaller than the number of non-interacting residues, the classification problem at hand may be highly unbalanced. As a result, the measures listed above may be difficult to interpret and compare for different benchmarks. For example, with 90% of data points assigned to the negative class, a baseline classifier that predicts all residues as non-interacting achieves numerically high 90% classification accuracy. To provide a measure that balances sensitivity and specificity of predictions, the Matthews correlation coefficient (MCC) is often used (4) together with other measures. MCC ranges from -1, indicating an inverse prediction, through 0, which corresponds to a random classifier, to +1 for perfect prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (4)$$

Other measures that can be used to assess and compare classification methods are area under the receiver operating characteristic (ROC) curve and F-measure.

6.2 Performance of selected methods

The performance of several representative methods discussed in the previous section is assessed here in order to compare more systematically individual methods, and to quantify the effects of mapping additional interaction interfaces and using truly unbound structures. Different aspects of the performance are evaluated using benchmark datasets described in section 4 (SPPIDER149, Hwang150B/U, and Albou78B/U).

For all evaluations, only residues with RSA of at least 5% were considered, thus excluding all fully buried residues in a given protein conformation. For methods providing a real valued score, multiple thresholds were tested as a basis for projection into two classes. The results for the best performing threshold in terms of MCC are reported in Tables 5 through 9. The following values were found to be optimal for each method: ET with residues being ranked 1 (out of top 1, 5, and 10 rankings evaluated), ConSurf with evolutionary rank ≥ 5 (5,

7, 9 evaluated), WHISCY with threshold ≥ 0 (0, 0.18 evaluated), PIER with threshold ≥ 15 (0, 15, 30 evaluated), and SPPIDER with threshold ≥ 0.3 (0.3, 0.5, 0.7 evaluated).

Method	SPPIDER149	Hwang150B	Albou78B
ET	0.08	0.04	0.01
ConSurf	0.12	0.07	0.02
PROMATE	0.10	0.10	0.09
Cons-PPISP	0.30	0.22	0.17
WHISCY	0.19	0.11	0.08
PIER	0.37	0.27	0.22
SPPIDER	0.41	0.28	0.20

Table 5. The performance of representative methods measured using MCC on three different sets, with only the original PDB complexes used to define the positive class.

As can be seen from Table 5, the overall accuracy of the methods evaluated here is rather limited. The two best performing methods, i.e., PIER and SPPIDER achieve MCC of about 0.4 for SPPIDER149 set, 0.3 for Hwang150B, and 0.2 for Albou78B, respectively. Similar relative drop in accuracy is also observed for other methods, indicating that Hwang150B and Albou78B sets are more difficult to classify. This can be explained in part due to a larger imbalance between positive and negative classes in these benchmarks, especially in the Albou78B dataset (see Table 3).

Method	SPPIDER149		Hwang150B		Albou78B	
	R, %	P, %	R, %	P, %	R, %	P, %
ET	7.03	43.92	3.99	28.18	2.84	17.55
	6.39	51.73	3.44	48.89	3.57	60.60
ConSurf	65.27	32.87	61.42	22.18	55.17	16.40
	63.00	40.97	55.91	41.07	53.19	41.66
PROMATE	3.91	60.71	4.06	48.98	3.69	43.43
	3.22	64.29	2.56	63.78	1.85	58.29
Cons-PPISP	33.40	60.59	26.25	42.42	22.46	34.80
	29.39	69.12	19.35	67.62	15.33	64.40
WHISCY	29.38	45.42	21.15	29.77	20.49	21.83
	26.66	54.32	17.21	51.71	16.53	48.38
PIER	61.10	52.62	49.66	37.46	45.43	30.61
	54.38	60.31	38.64	60.86	31.20	56.99
SPPIDER	80.36	48.47	63.15	34.11	56.22	26.49
	73.14	56.81	53.04	59.82	43.48	55.52

Table 6. The effect of mapping interaction sites from homologous protein complexes on recall (R) and precision (P): the first line in each row shows R and P using original PDB complexes, whereas the second line indicates accuracy derived after mapping interaction sites using PISA BUs and homologous chains with 70% sequence identity.

It should be noted that due to a sufficiently large number of data points (surface residues, see Table 3) included in each benchmarks, each of the correlation coefficients reported above

is statistically significantly different from 0 with a p-value < 0.05. Nevertheless, practical applicability of methods that achieve correlations of 0.2 and lower has to be judged using also other criteria and specific examples. In particular, evolutionary methods achieve very limited accuracy in this test, even though they may provide biologically valuable insights, as discussed later.

The effects of mapping interaction residues from alternative complexes are illustrated in Table 6 using measures of sensitivity and specificity. The accuracy using the assignment of the positive class (interaction sites) derived from the original complexes is compared to the accuracy obtained re-labeling the "non-interacting" residues in mapped interfaces as "interacting" sites. Due to largely canceling effects of decreased rates of false positives and increased rates of false negatives, the mapping of interaction sites from PISA biological units does not affect significantly the performance of the prediction methods in terms of MCC, although a systematic small drop in accuracy is observed in most cases (data not shown).

However, as can be seen from Table 6, all methods show a drop in recall while precision improves when mapping is applied. These results also allow one to trace how the trade-off between sensitivity and specificity was optimized for different methods. One striking example is ConSurf vs. ET comparison. On the other hand, most structure-based methods provide fairly well balanced predictions. In particular, precision improves considerably, with only a relatively limited drop in recall for the best performing SPPIDER method, followed by PIER and Cons-PPISP. The observed ranking could reflect the fact that SPPIDER was trained (although on a different set without homology to SPPIDER149 set) using mapping from alternative complexes to reduce the noise in learning from data and to provide a more balanced classification problem.

Method	Hwang150B SI70	Hwang150U SI70	Albou78B SI70	Albou78U SI70
ET	0.03	0.00	0.06	0.08
ConSurf	0.03	0.05	0.00	0.00
PROMATE	0.06	0.05	0.04	0.01
Cons-PPISP	0.20	0.18	0.14	0.13
WHISCY	0.09	0.16	0.06	0.08
PIER	0.24	0.23	0.15	0.11
SPPIDER	0.29	0.29	0.17	0.14

Table 7. The effect of the bound versus unbound state of the protein structures used as an input in terms of MCC. In all cases, interacting residues were mapped using homology to PISA BUs with 70% sequence identity.

The impact of conformational change and the use of structures in bound as opposed to unbound state as an input is assessed in Table 7. For that purpose, the overall accuracy in terms of MCC is compared using two pairs of sets of bound (taken from a complex by simply ignoring other chains) and truly unbound structures: Hwang150B *vs.* Hwang150U and Albou78B *vs.* Albou78U, respectively. Slight decrease in performance is observed for all but one structure-based method, the exception being WHISCY. The latter method starts from a low level, though. In addition, the WHISCY server did not generate results for a number of more difficult cases, suggesting that this trend might not hold on other data sets.

While the drop in accuracy is limited for other methods tested, it should be emphasized that benchmarks included here sample relatively small conformational changes due to induced fit. Therefore, further systematic studies will be required to better delineate the range of applicability of structure-based method for the recognition of protein interaction sites.

Table 8 demonstrates how the performance estimates can be inflated when accuracy measures are computed based on all residues as opposed to computing the accuracy for each protein and then averaging over all proteins. Per protein averages, together with measures of variance (here we report standard deviations), allow one to assess better the range of expected accuracies for individual proteins. As can be seen from Table 8, the observed large standard deviations suggest large protein to protein variation and indicate that all tested methods fail dramatically for at least some proteins. It should be also noted that using per protein measures PIER is the top performing method, followed by SPPIDER and Cons-PPISP.

Method	MCC	Q_2, %	R, %	P, %
ET	0.06±0.12 0.08	65.64±17.83 71.21	9.60±16.07 7.03	29.35±35.01 43.92
ConSurf	0.12±0.15 0.12	54.44±8.16 52.80	64.54±14.06 65.27	39.61±22.69 32.87
PROMATE	0.07±0.13 0.10	64.01±19.63 71.16	5.72±8.93 3.91	28.30±39.31 60.71
Cons-PPISP	0.23±0.23 0.30	69.52±13.23 74.15	37.50±22.11 33.40	58.99±29.71 60.59
WHISCY	0.14±0.20 0.19	67.39±13.14 71.03	26.58±19.79 29.38	42.64±28.00 45.42
PIER	0.30±0.23 0.37	71.18±11.47 72.54	58.73±24.80 61.10	55.22±27.09 52.62
SPPIDER	0.29±0.20 0.41	66.94±13.82 69.39	79.16±24.79 80.36	49.19±21.69 48.47

Table 8. Comparison of the accuracy measures calculated per residue by merging data from all chains (the bottom line in each row) and per protein averages and standard deviations (the top line in each row), using the SPPIDER149 set (similar effect is observed on other benchmarks).

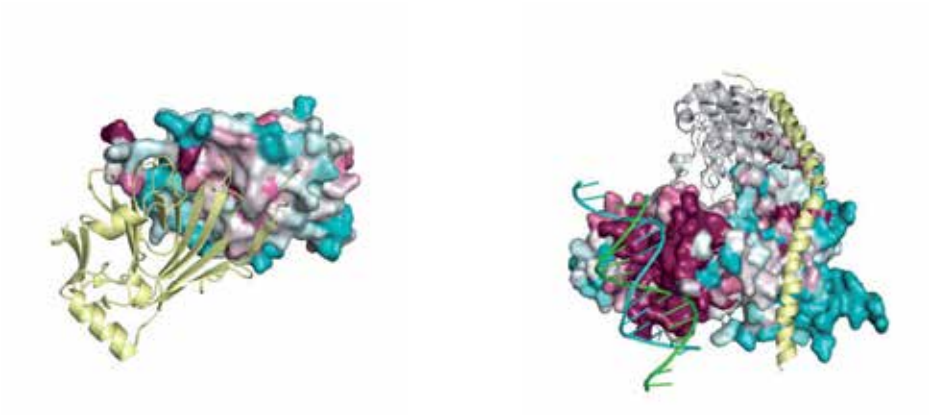
Not all web-based implementations of the methods are reliable. While requesting and retrieving predictions from the evaluated servers, we faced multiple failures. Table 9 illustrates the reliability of the corresponding servers from the user`s point of view by presenting the numbers of proteins failed to be processes within each benchmark set. The most reliable web-servers appear to be PIER and SPPIDER, whereas ET, ConSurf, and WHISCY are quite unreliable, which makes it more difficult to evaluate servers on a large scale.

Prediction methods that seemingly perform poorly according to some evaluation criteria can still greatly facilitate further experimental and computational studies on protein interactions. One might argue that predicting possible interaction interfaces should be directed at the recognition of the sites that contribute most to the binding energy. Such hot

spots also represent the most natural target for further validation, e.g., using mutagenesis, or as targets for therapeutics.

Method	SPPIDER149	Hwang150B	Hwang150U	Albou78B	Albou78U
ET	14	14	28	8	8
ConSurf	21	12	13	4	4
PROMATE	1	3	8	3	12
Cons-PPISP	7	3	1	0	4
WHISCY	34	15	17	8	9
PIER	0	0	0	0	0
SPPIDER	0	0	0	0	0

Table 9. The number of proteins not included in each benchmark due to problems with the retrieval of the results as an indicator of the reliability of web-servers tested.



A **B**

Fig. 3. Examples of protein interaction sites predicted by ConSurf: **A.** A successful identification of the protein interface for the homodimer of phosphoglucose isomerase (PDB ID 1qxr, chain A); **B.** A multi-interface protein (CSL transcription factor) illustrates possible confusion with DNA binding sites that are the most slowly evolving residues at the surface of the protein in this case (PDB ID 2fo1, chain A). Residues in magenta are the most conserved, whereas variable sites are colored using cyan (see the ConSurf documentation).

In this context, a special note needs to be made on the performance of evolutionary methods, such as ET and ConSurf. As we mentioned before, these methods were not designed specifically to predict protein-protein interaction sites, but rather to identify evolutionary conserved residues. Therefore, these methods may not able to discriminate between protein-protein, protein-ligand (e.g., co-factor or substrate), and protein-DNA/RNA binding sites. An example of such a case is shown in Figure 3.

On the other hand, highly conserved residues that are exposed on the surface of a protein are very likely functionally relevant, irrespective of the actual involvement in interaction. Despite all the limitations, evolutionary methods for the prediction of interaction sites have significantly contributed to the mapping of protein interactions and other functional

annotations, see e.g., (Kniazeff et al., 2002; Shenoy et al., 2006) and (He et al., 2003; Lietha et al., 2007), for ET and ConSurf, respectively.

7. Discussion and conclusions

Protein-protein interactions are essential for enzymatic functions, signal transduction, cell cycle regulation and other fundamental biological processes. In addition to addressing the fundamental questions of molecular biology, identification of residues involved in protein-protein interactions has important medical relevance. Combined with recent advances in genome sequencing it facilitates delineating natural functional variants from pathological mutants, and conducting 'molecular diagnostics' as part of personalized medicine.(Su et al., 2011) Detailed structural information on thousands of protein complexes also stimulates growth in the field of rational drug design by providing a new class of targets that include known protein interaction interfaces.(White et al., 2008)

However, experimental identification and validation of a protein interface remains a challenging task, both in terms of labor and cost. Therefore, efforts to map and characterize protein interactions can considerably benefit from computational biology and structural bioinformatics. In particular, methods that integrate sequence and structure information achieved accuracies that are useful in selecting and prioritizing targets for mutagenesis and other experimental studies.

In this chapter, we reviewed state-of-the-art in the field of computational prediction of protein-protein interaction sites. We evaluated some representative methods using several published benchmarks of protein complexes. The overall accuracy of existing methods, in accord with other recent evaluations, was found to be limited (the Matthews correlation coefficient between the predicted and true class assignment of up to 0.4). Therefore, further concerted efforts will be required to improve state-of-the-art in the field. To that end, we discussed the need for standard definition of protein interaction sites, developing more comprehensive benchmark protein sets, and appropriate ways of measuring/reporting the accuracy of predictions.

We quantified the effects of taking into account multiple interaction interfaces and using as an input unbound structures that were resolved without interacting partners. Both of these issues are often ignored when evaluating the performance of interaction sites prediction methods. Yet, they are shown to impact significantly the estimates of performance. These two issues also highlight more fundamental difficulties with the definition of the negative class and current attempts to cast the problem in a computationally feasible way.

Casting the prediction of interaction sites in terms of a two-class classification problem requires that examples of the negative ("non-interacting") class be used for the training. With data points representing both "interacting" and "non-interacting" residues, a decision boundary separating the two classes can be optimized. These negative examples are defined in most cases by simply taking the complement of the positive class, i.e., all other (surface exposed) residues that are not known to be involved in interactions.

Consequently, without mapping known interfaces alternative complexes, residues within such interfaces are incorrectly regarded as "non-interacting". This could introduce problems in training, as misclassified vectors from the negative class may coincide with the bulk of the

density for the positive class. One strategy to address this issue is to filter out such difficult cases. As an alternative, one could also consider one-class approaches, in which only the positive class examples are used to learn a predictor. On the other hand, if residues from multiple complexes are systematically mapped, as advocated here, the negative class assignment as a source of noise should be gradually reduced with the progress in experimental mapping of interaction sites.

Conformational changes upon complex formation pose another problem for the methods considered here. Protein flexibility and the induced fit effects upon complex formation are assumed to be limited. Obviously, this assumption does not hold in many instances of protein-protein interactions (and sometimes it breaks spectacularly, e.g., when the co-folding of otherwise disordered interacting domains occurs). Therefore, methods presented here are of limited applicability when large conformational changes or flexible domains are involved.

It should be also stressed that even a limited induced fit can pose significant challenges for structure-based methods. Simply ignoring all but one chain in a protein complex, and thus taking a *de facto* bound conformation as input, may lead to spurious effects in training and overly optimistic estimates of accuracy. For example, low B-factors of surface residues, which can be "locked" in a specific conformation by interactions with a co-factor, may not be a true signal of interaction sites (in many cases the opposite can actually be observed). Features that are capable of identifying interaction sites starting from a truly unbound structure should be emphasized.

Reliable identification of residues that participate in binding to other proteins can help direct and streamline mutagenesis and other experimental studies, and to facilitate efforts to map entire interactomes. It can also reduce the levels of false positives (by assessing compatibility between predicted interfaces), and false negatives (by helping identify novel interactions) observed for experimental approaches that are used to map protein interactions. Another promising application is protein docking, in which predicted interfaces can be used for evaluating and ranking potential complex structures (de Vries and Bonvin, 2011), in analogy to docking methods that utilize limited NMR data. (Dominguez et al., 2003; Kohlbache et al., 2001)

Further progress in the field will require new insights to overcome current limitations, as well as careful assessment of the accuracy in order to address possible biases in training and validation. Constant improvements in experimental techniques and a growing number of resolved macromolecular complexes, from which to learn better predictors, bode well for future efforts in this regard.

8. References

Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks-based regression. Proteins *56*, 753-767.

Albou, L. P., Poch, O., and Moras, D. (2011). M-ORBIS: mapping of molecular binding sites and surfaces. Nucleic Acids Res *39*, 30-43.

Albou, L. P., Schwarz, B., Poch, O., Wurtz, J. M., and Moras, D. (2009). Defining and characterizing protein surface using alpha shapes. Proteins *76*, 1-12.

Aloy, P., Pichaud, M., and Russell, R. B. (2005). Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol *15*, 15-22.

Aloy, P., and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. Nat Biotechnol *22*, 1317-1321.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Bader, J. S., and Chant, J. (2006). Systems biology. When proteomes collide. Science *311*, 187-188.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Bock, J. R., and Gough, D. A. (2001). Predicting protein--protein interactions from primary structure. Bioinformatics *17*, 455-460.

Bordner, A. J., and Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. Proteins *60*, 353-366.

Bradford, J. R., and Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics *21*, 1487-1494.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci *13*, 190-202.

Chakrabarti, P., and Janin, J. (2002). Dissecting protein-protein recognition sites. Proteins *47*, 334-343.

Chen, H., and Zhou, H. X. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins *61*, 21-35.

Chen, X. W., and Jeong, J. C. (2009). Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics *25*, 585-591.

Chung, J. L., Wang, W., and Bourne, P. E. (2006). Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins *62*, 630-640.

de Vries, S. J., and Bonvin, A. M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One *6*, e17695.

de Vries, S. J., and Bonvin, A. M. J. J. (2008). How proteins get in touch: Interface prediction in the study of biomolecular complexes. Curr Protein Pept Sc *9*, 394-406.

de Vries, S. J., van Dijk, A. D., and Bonvin, A. M. (2006). WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins *63*, 479-489.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc *125*, 1731-1737.

Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein--protein interaction sites in heterocomplexes with neural networks. Eur J Biochem *269*, 1356-1361.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., and Bateman, A. (2008). The Pfam protein families database. Nucleic Acids Res *36*, D281-288.

Fiorucci, S., and Zacharias, M. (2010). Prediction of protein-protein interaction sites using electrostatic desolvation profiles. Biophys J *98*, 1921-1930.

Fletcher, S., and Hamilton, A. D. (2007). Protein-protein interaction inhibitors: small molecules from screening techniques. Curr Top Med Chem *7*, 922-927.

Fry, D. C. (2006). Protein-protein interactions as targets for small molecule drug discovery. Biopolymers *84*, 535-552.

Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. J Mol Biol *302*, 917-926.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163-164.

Guharoy, M., and Chakrabarti, P. (2010). Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics *11*, 286.

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., and Deane, C. M. (2010). i-Patch: interprotein contact prediction using local network information. Proteins *78*, 2781-2797.

Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol *23*, 839-844.

He, X. L., Bazan, J. F., McDermott, G., Park, J. B., Wang, K., Tessier-Lavigne, M., He, Z., and Garcia, K. C. (2003). Structure of the Nogo receptor ectodomain: a recognition module implicated in myelin inhibition. Neuron *38*, 177-185.

Henrick, K., and Thornton, J. M. (1998). PQS: a protein quaternary structure file server. Trends Biochem Sci *23*, 358-361.

Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics *24*, 2780-2781.

Huang, B., and Schroeder, M. (2008). Using protein binding site prediction to improve protein docking. Gene *422*, 14-21.

Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein-protein docking benchmark version 3.0. Proteins *73*, 705-709.

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Structure *7*, R277-279.

Jones, S., and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol *63*, 31-65.

Jones, S., and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. J Mol Biol *272*, 121-132.

Kim, W. K., Henschel, A., Winter, C., and Schroeder, M. (2006). The many faces of protein-protein interactions: A compendium of interface geometry. PLoS Comput Biol *2*, e124.

Kniazeff, J., Galvez, T., Labesse, G., and Pin, J. P. (2002). No ligand binding in the GB2 subunit of the GABA(B) receptor is required for activation and allosteric interaction between the subunits. J Neurosci *22*, 7352-7361.

Kohlbache, O., Burchardt, A., Moll, A., Hildebrandt, A., Bayer, P., and Lenhof, H. P. (2001). Structure prediction of protein complexes by an NMR-based protein docking algorithm. J Biomol NMR *20*, 15-21.

Koike, A., and Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel *17*, 165-173.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*, 774-797.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). PIER: protein interface recognition for structural proteomics. Proteins *67*, 400-417.

Lacapere, J. J., Pebay-Peyroula, E., Neumann, J. M., and Etchebest, C. (2007). Determining membrane protein structures: still a challenge! Trends Biochem Sci *32*, 259-270.

Lamark, T., Perander, M., Outzen, H., Kristiansen, K., Overvatn, A., Michaelsen, E., Bjorkoy, G., and Johansen, T. (2003). Interaction codes within the family of mammalian Phox and Bem1p domain-containing proteins. J Biol Chem *278*, 34568-34581.

Li, N., Sun, Z., and Jiang, F. (2008). Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinformatics *9*, 553.

Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. Nucleic Acids Res *34*, 3698-3707.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lietha, D., Cai, X., Ceccarelli, D. F., Li, Y., Schaller, M. D., and Eck, M. J. (2007). Structural basis for the autoinhibition of focal adhesion kinase. Cell *129*, 1177-1187.

Liu, R., and Zhou, Y. (2009). Using support vector machine combined with post-processing procedure to improve prediction of interface residues in transient complexes. Protein J *28*, 369-374.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A *100*, 5772-5777.

Moscat, J., Diaz-Meco, M. T., Albert, A., and Campuzano, S. (2006). Cell signaling and function organized by PB1 domain interactions. Mol Cell *23*, 631-640.

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol *338*, 181-199.

Nooren, I. M., and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol *325*, 991-1018.

Ofran, Y., and Rost, B. (2003a). Analysing six types of protein-protein interfaces. J Mol Biol *325*, 377-387.

Ofran, Y., and Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information. FEBS Lett *544*, 236-239.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics *23*, e13-16.

Oldfield, T. J. (2002). Data mining the protein data bank: residue interactions. Proteins *49*, 510-528.

Park, J., Lappe, M., and Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol *307*, 929-938.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630-645.

Qin, S., and Zhou, H. X. (2007). meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics *23*, 3386-3387.

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci *9*, 1-15.

Shenoy, S. K., Drake, M. T., Nelson, C. D., Houtz, D. A., Xiao, K., Madabushi, S., Reiter, E., Premont, R. T., Lichtarge, O., and Lefkowitz, R. J. (2006). beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. J Biol Chem *281*, 1261-1273.

Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., and Godzik, A. (2007). The challenge of protein structure determination--lessons from structural genomics. Protein Sci *16*, 2472-2482.

Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., and Shi, L. (2011). Next-generation sequencing and its applications in molecular diagnostics. Expert Rev Mol Diagn *11*, 333-343.

Thorn, K. S., and Bogan, A. A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics *17*, 284-285.

White, A. W., Westwell, A. D., and Brahemi, G. (2008). Protein-protein interactions as targets for small-molecule therapeutics in cancer. Expert Rev Mol Med *10*, e8.

Xia, J. F., Zhao, X. M., Song, J., and Huang, D. S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics *11*, 174.

Xu, Q., and Dunbrack, R. L., Jr. (2011). The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res *39*, D761-770.

Yan, C., Honavar, V., and Dobbs, D. (2004). Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. Neural Comput Appl *13*, 123-129.

Zhou, H. X., and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics *23*, 2203-2209.

Zhou, H. X., and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins *44*, 336-343.

2

Advances in Human-Protein Interaction - *Interactive and Immersive Molecular Simulations*

Nicolas Férey, et al.*
CNRS - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - Université Paris XI Bâtiment 508, 512 et 502 bis, 91403 Orsay Cedex France

1. Introduction

Molecular simulations allow researchers to obtain complementary data with respect to experimental studies and to overcome some of their limitations. Current experimental techniques do not allow to observe the full dynamics of a protein at atomic detail. In return, experiments provide the structures, i.e. the spatial atomic positions, for numerous biomolecular systems, which are often used as starting point for simulation studies. In order to predict, to explain and to understand experimental results, researchers have developed a variety of biomolecular representations and algorithms. They allow to simulate the dynamic behavior of macromolecules at different scales, ranging from detailed models using quantum mechanics or classical molecular mechanics to more approximate representations. These simulations are often controlled *a priori* by complex and empirical settings. Most researchers visualise the result of their simulation once the computation is finished. Such post-simulation analysis often makes use of specific molecular user interfaces, by reading and visualising the molecular 3D configuration at each step of the simulation. This approach makes it difficult to interact with a simulation in progress. When a problem occurs, or when the researcher does not achieve to observe the predicted behavior, the simulation must be restarted with other settings or constraints. This can result in the waste of an important number of compute cycles, as some simulations last for a long time: several days to weeks may be required to reproduce a short timespan, a few nanoseconds, of molecular reality. Moreover, several biomolecular processes, like folding or large conformational changes of proteins, occur on even longer timescales that are inaccessible to current simulation techniques. It can thus be necessary to impose empirical constraints in order to accelerate a simulation and to reproduce

*Alex Tek, Benoist Laurent, Marc Piuzzi, Zhihan Lu and Marc Baaden (CNRS - Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, 13, rue Pierre et Marie Curie, 75005 Paris, France)
Olivier Delalande (CNRS - Interactions Cellulaires et Moléculaires - Université de Rennes 1, Avenue du Professeur Léon Bernard, 35065 Rennes cedex, France)
Matthieu Chavent (Structural Bioinformatics and Computational Biochemistry Unit, Dept. of Biochemistry, University of Oxford, United Kingdom)
Christine Martin, Lorenzo Piccinali, Brian Katz and Patrick Bourdot (CNRS - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - Université Paris XI Bâtiment 508, 512 et 502 bis, 91403 Orsay Cedex, France)
Ludovic Autin(Molecular Graphics Laboratory Department of Molecular Biology, MB-5 - The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037-1000, USA)

an experimental result in MD. These constraints have to be defined *a priori*, rendering it difficult to explore all possibilities in order to examine various biological hypotheses.

A new approach allowing to address these problems has emerged recently: Interactive Molecular Simulation (IMS). IMS consists in visualising and interacting with a simulation in progress, and provides the user with control over simulation settings in interactive time. With the recent advances in human computer interaction and the impressive increase of available computing power, the IMS approach allows a user to interact in 3D space in real time with a molecular simulation in progress. This approach provides quality control features by visualizing results of a simulation in progress and supplies interactive features, such as feeling forces involved in the simulation as well as triggering specific events by applying custom forces during the simulation in progress. These advances led to a new generation of scientific tools to better understand life science phenomena, which place the human expertise at the centre of the analysis process, complementarily to automatic computational methods.

The IMS approach emerged from the breakthrough initiated by the *Sculpt* precursor program proposed by Surles et al. (1994). Since then, the interactive molecular simulations field has been developing continuously. Initial interactive experiments using molecular mechanics techniques gave quickly rise to "guided" dynamics simulations [Wu & Wang (2002)] or *Steered Molecular Dynamics* (SMD) [Israelewitz et al. (2001)] [Leech et al. (1996)]. The interest for these methods increased with the enhancement of simulation accuracy and thanks to the exciting new possibilities for dynamic structural exploration of very large and complex biological systems. In the Interactive Molecular Dynamics (IMD) approach, steering forces are applied interactively with a chosen amplitude, direction and application point. This enables the user to explore the simulation system while receiving instant feedback information from real-time visualisation or haptic devices [Leech et al. (1997)]. Schulten's group has carried out several applications of IMS simulations to macromolecular structures [Grayson et al. (n.d.)] [Stone et al. (2001)]. This effort lead to the design of two efficient software tools facilitating the process of setting up an IMS : *NAMD* and *VMD* [Phillips et al. (2005)] [Nelson et al. (1995))]. The underlying exchange protocol is also supported by ProtoMol [Matthey et al. (2004)], LAMMPS [Plimpton (1995)], HOOMD-blue [Anderson et al. (2008)] and any software using the MDDriver library [Delalande et al. (2009)]. Similar projects proposing an interactive display for molecular simulations exist, such as the *Java3D* interface proposed in Knoll & Mirzaei (2003) and Vormoor (2001), or the *Protein Interactive Theater* [Prins et al. (1999)].

With fast generalization of new computer hardware devices and increasing accessibility to powerful computational infrastructures, IMS showes a fast and promising evolution, even for very large molecular systems (over 100.000 atoms). Such applications are now in the reach of state-of-art desktop computing. This evolution was possible given the strong increase in raw computing power leading to faster and bigger processing units (multi-processors, multi-core architectures). Currently ongoing technological developments such as GPU computing and the spread of parallelized entertainment devices (PS3, Cell) with specific graphic and processing capabilities open exciting new opportunities for interactive calculations. These approaches could provide even more processing power for highly parallelizable computational problems, for instance by differentiating the parallelisation of molecular calculations and graphical display functionalities. Given these developments, the range of accessible computational methods and representations is bound to grow. It may soon be possible to extend the IMS approach to *ab initio* or QM/MM calculations. Indeed, the precision achieved in the description of a system can be improved by switching to a more

accurate physical model and/or by improving the representation of the molecular context simulated. Thus, multi-scale simulations [Baaden & Lavery (2007)] would indeed benefit from an interactive approach leading to important advantages with respect to the study of complex biological systems. However, the raw increase in computer speed alone is not sufficient to grant a successful future evolution of the IMS approach. In addition, it is necessary to develop adapted software solutions, which are generally more efficient [Grayson et al. (n.d.)], as it is commonly admitted in the numeric simulation field. Finally, the most recent and famous work illustrating the revolution of this approach is the "Fold It" serious game, which allows a user to interactively propose a protein folding solution [Cooper et al. (2010)].

We will describe in this chapter the recent advances relating to these IMS approaches previously described. As IMS implies to efficiently combine simulation and interaction features, we will explain how we designed specific simulation, visualisation, and interaction techniques to solve the real time constraint, to study complex biomolecular systems, and to address a larger simulation timescale. Then we will discuss software architectures to efficiently put the different building blocks together. Finally, we will explain how we apply IMS to different fields of research including various topics such as protein-protein docking in a virtual reality and multimodal context, an ion substitution study using an haptic device, and a study about the opening and closure of the Guanylate Kinase enzyme.

2. Multiscale and multiphysics protein simulation models

In structural biology, recent advances in experimental techniques allow us to solve larger and larger protein 3D structures. However, even if structure is known to be strongly linked to biological function, static states often lack in providing dynamical informations that are crucial for the understanding of the subtle mechanisms occuring at the molecular level. Thus, molecular simulations are nowadays used to complete experimental biostructural studies, especially to better understand the dynamic behaviour and the fundamental mechanisms involved in a protein complex. In spite of the increasing computational resources, classical simulation tools are not well adapted to quickly obtain insight into the global biomechanical properties, because of the limited timescale covered by all-atom or coarse-grained simulations. For these reasons, it is necessary to develop new modeling approaches at a larger scale, complementary to all-atom and coarse-grained models, especially designed to interactively study protein complex formation and biomechanical properties of large biomolecular structures. We present in this part unconventional approaches that could address these requirements. The first one, based on a rigid body model of a protein, was especially designed to study protein-protein interactions for an interactive rigid docking application. The second one, based on a spring netwok model, takes into account protein flexibility in order to study biomechanical behavior of large protein structures in interactive time.

2.1 A rigid body simulation model to interactively study protein-protein interactions

At a larger scale, it is sometimes not necessary to model and simulate the flexibility of a protein, but sufficient to consider the protein as a rigid body. Using a simple but accurate model at the macroscopic scale allows us to overcome the main constraint to provide an interactive time biophysical simulation as required for IMS: taking into account the user interaction during a simulation in progress. To present our rigid body simulation model dedicated to IMS, especially interactive rigid docking, we have to focus on the main phenomena that are involved in the protein interactions.

2.1.1 Geometry and surface

Proteins can be viewed as both the building blocks and the workforce of cells. They are synthesized based on portions of DNA (Deoxyribonucleic Acid) called coding sequences or genes. Genes are transcribed in the form of mRNA (Messenger RiboNucleic Acid), which is then translated by ribosomes in the form of a protein, following a specific coding scheme (figure 1A). Each triplet of mRNA bases corresponds to one AA (Amino Acid) or residue, of which there are twenty basic types. The various physicochemical properties of AAs give rise to interactions at the atomic level, inducing protein folding which contributes in turn to protein stability (figure 1B). These properties also play a crucial part in protein-protein interactions.

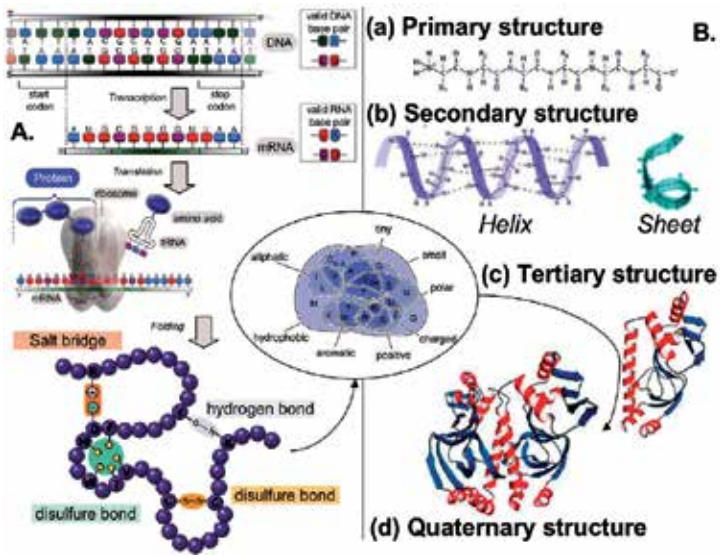


Fig. 1. (**A)** - Overall view of protein synthesis: transcription of DNA to messenger RNA (mRNA) and translation of mRNA to amino acid sequences chosen from 20 possible varieties, here shown according to their physicochemical properties (using a Venn diagram). **(B)** - Based on the chemical nature of component amino acids, resulting interactions cause the protein to fold to a favorable arrangement in space. This 3D shape can be described according to four levels: the (a) primary, (b) secondary, (c) tertiary and (d) quaternary structure.

Proteins, therefore, can be seen as long chains composed of successive amino acids folded in space, which are the product of the expression of an organism's genetic makeup. But in order to execute their functions within cells, proteins must undergo folding and take a specific 3D form. This form may be characterized following four levels of structure (see figure 1B). The order in which residues are linearly arranged, *i.e.* their sequence, constitutes the protein's primary structure. (see figure 1B-a). Some of the structure's segments organize themselves into sequences of specific substructures called secondary structures (see figure 1B-b). These structures, stabilized by hydrogen bonds, can be divided into two groups: regular secondary structures, called alpha helices and beta sheets, which are linked together by irregular structures called loops. The arrangement of these secondary structures thus

constitutes the 3D, or tertiary structure of the protein (see figure 1B-c), which determines protein function within the cell.

Once folded, proteins carry out various functions within the cell, such as transporting molecules to and from various components of the organism (*e.g.* hemoglobin, chaperone proteins), inter- and intracellular signaling and communications (*e.g.* hormones, neurotransmitters, ions), immune defense functions (immunoglobulins, adhesion molecules) or cellular metabolism (chlorophyll, apoptosis proteins, transcription factors, ATP synthesis). These cellular functions are closely linked to the protein's tertiary structure, but also to its interactions with other proteins.

In short, better understanding of protein-protein interactions is a major stake for biomedical research. Indeed, designing new drugs increasingly involves targeting specific protein-protein interactions [Villoutreix et al. (2008)], or alternatively, involves synthesizing recombinant proteins meant to emulate interactions with the original native protein [Pipe (2008)]. It becomes more and more necessary, therefore, to identify the 3D structure of protein complexes. Two main experimental methods are currently used to determine the 3D structure of a protein complex. These are X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. All known publicly available protein structures are currently housed on the website of the Protein Data Bank (PDB) [Berman et al. (2000)]. This database contains now several hundred thousand protein structures for many organisms. However, this number remains small in comparison to estimates of the number of existing proteins in the natural world. This is because experimental determination of protein structure is often difficult, and in some cases impossible. Indeed, solving a problem of this kind involves mass production and purification of the protein, and in the case of crystallography, production of diffractive crystals. In determining the structure of a protein complex, difficulties in production and purification are all the more critical, because partner proteins must be produced at the same time for complexes to form. Furthermore, the time necessary for crystallization may be incompatible with the lifespan of some complexes. For all these reasons, many scientists have attempted to predict the structure of such complexes using computational tools through methods and algorithms for molecular docking.

Current techniques for the experimental study of the 3D structure of protein complexes (crystallography, NMR, electron cryomicroscopy, SAXS, etc.) have several limitations (in terms of size and type of proteins) and are costly in terms of time and money. For that reason, computer-based (*in silico*) docking methods have been developed in the past, to deduce the functional 3D structure of a complex based on single molecules, which turns out to be considerably easier and cheaper than experimental *in vitro* methods. Current approaches are strictly computational and results are evaluated using visualization tools. These approaches can be divided into 4-5 successive stages (figure 2): (1) choice of the representation mode for proteins (atomic view, pseudo-atoms, grid, etc.); (2) conformational exploration (taking into account position, orientation, and shape of the ligand); (3) minimization of the function used to evaluate binding energy (i.e. *score*) for conformations derived from the exploration; (4) grouping by similarity and classification through evaluation or fine-tuning of the *scores*, augmented with a manual stage of visualization when the score alone doesn't allow native conformations (i.e. the ones present in nature) to be discriminated from other generated conformations; (5) an optional stage for fine-tuning selected complexes, through energy minimization or molecular dynamics.

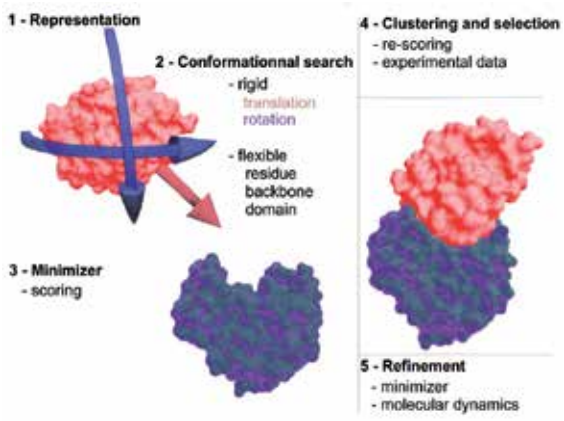


Fig. 2. The 5 stages of the docking task

A large number of fully automatic computational docking algorithms depend on an comprehensive approach of conformational exploration, the main problem being combinatorial explosion of the number of possible solutions. These approaches can be sorted into three categories: those based on systematic sampling, on molecular dynamics techniques and on classification interaction modes between proteins. An ideal function would yield, for a given mode of interaction, the binding energy of two proteins involved in a complex (see section 2.1.3). Such functions aim to reproduce experimental values of free binding energy, and through minimization, to reach the overall minimum energy in the set of all possible protein-protein complexes.

Consequently, in real life cases, automatic docking algorithms, such as *ClusPro* [Comeau et al. (2004)] or *Hex* [Ritchie (2003)], must manage two difficulties in order to reach a relevant result. The first is to process a space of potential solutions which increases in size along with the number of degrees of freedom in describing protein position and conformation, thus running the risk of not beeing processed in an acceptable amount of time. The second problem is that search algorithms produce local minima, and cannot easily find the global minimum that is associated to the native form of the complex [Wang et al. (2003)].

To finalize a docking simulation, experts rely upon a manual stage of visualization to analyse the generated complexes. This task consists in a detailed analysis of residues and atoms involved in the interface each complex, through the identification of hydrogen bonds, salt bridges, and especially the presence of hotspots, *i.e.* amino acids at the interface, known from experimental studies to be an essential part of this interface. However, it can be difficult to manipulate two 3D structures at the same time to observe the interface with traditional interaction tools, since one protein usually hides the other. Therefore docking assisted by user interaction is a useful alternative to improve the work of experts in this field. Such techniques might allow a more intuitive interaction with 3D protein structures.

Finally, two approaches are used to "thin the herd" of selected complexes. One consists in minimizing the rigid bodies and lateral chains of amino acids present at the interface. This approach is implemented in several applications such as *ICM-DISCO* [Fernandez-Recio et al. (2003)], *MMTK* [Hinsen (2000)], *FireDock* [Andrusier et al. (2007)], *PELE* [Borrelli et al. (2005)], *ATTRACT* [Zacharias (2005)], etc. The other approach involves studying the dynamic behavior of the selected complex. The software program *Gromacs* [Hess et al. (2008)], for

example, allows evaluation of atomic positions over time based on their physicochemical properties. This approach allows first to evaluate the complex stability, as well as possible conformational changes induced by the interaction, *e.g.* loop deformation. We should add, however, that this approach remains very costly in terms of processing time, compared to minimizers which allow users to process a given configuration very quickly.

As the automatic docking software programs previously presented did not respond to the interactive time constraint, we developed a new simulation tool dedicated to interactive protein-protein rigid docking. Our protein docking method is essentially based on two sets of criteria: geometric/topological criteria, and biophysical criteria.

2.1.2 Interactive time evaluation of geometry and surface complementarity

One of the earliest criteria identified in protein-protein interaction is surface topology of the proteins involved. In most known structures of 3D complexes, partners exhibit good surface complementarity. Studies have also shown that the surface of the protein-protein interface generally covers between 1000 and 2500 square Angstroms. This criteria allowed the development of first-generation docking software, based solely on shape recognition [Connolly (1983)] (*i.e.* complementarity of molecular surfaces). This approach is well adapted to "hard" rigid protein docking. We used these geometric/topological criteria in our multimodal immersive environment in two ways:

Surface collision. For each protein, a surface mesh is computed using the *MSMS* software before interactive docking occurs [Sanner et al. (1996)]. The resolution of this mesh can be adjusted using parameters. Collision detection during interaction then uses the *RAPID* library [Gottschalk et al. (1996)], which allows real-time computation of a list of colliding triangles among the two protein surface meshes during docking. This set of triangles can be used to generate feedback based on triangle normals and on the intersection volume of the two protein surfaces.

Atomic surface complementarity. Atomic surface complementarity is estimated essentially as a calculation of the variance of the inter-atomic distances on the two protein surfaces. We use this overall atomic surface complementarity score in audio or visual feedback.

2.1.3 Interactive time computation of physicochemical properties and energies

However, geometric criteria turned out to be insufficient to predict the structure of a complex. Thus, we had to rely on methods including energy criteria. Protein-protein complexes seem to follow the rule of thumb that the active configuration is the one whose level of free energy is lowest [Wang et al. (2003)]. In order to evaluate free energies between two proteins, we rely on molecular mechanics methods. For this purpose, atoms are viewed as spheres, and interactions between atoms can be computed using the van der Waals and electrostatic potentials. The free energy for protein-protein interaction can then be approximated by the sum of these potentials, which is known as the *score*. In the context of real-time immersive docking, the choice of equations and methods to evaluate the energy of a complex and hence its score is a crucial issue [Wang et al. (2003)].

Van der Waals interactions. Van der Waals interactions are an empirical approximation of atomic interactions. The van der Waals force, obtained by constructing a gradient of the potential field, is defined by the Lennard-Jones potential equation (equation 1). In this

equation, r_{ij} is the distance between two atoms i and j, σ the interatomic distance for which the potential becomes zero, and ϵ the depth of the potential well. ϵ and σ are determined empirically and depend on what pair of atoms is considered. The van der Waals potential includes an attractive component when atoms are bound, and a repulsive component when atoms are too close to each other. It prevents two proteins from penetrating into each other during interactive docking, through calculation of interatomic forces at the protein-protein interface.

$$U_{vw}(r) = 4\epsilon[(\frac{\sigma}{r_{ij}})^{12} - (\frac{\sigma}{r_{ij}})^{6}] \tag{1}$$

These forces apply only to very short distances and mostly concern surface atoms. As computing distances between all pairs of atoms has a quadratic complexity, we apply specific filtering rules to keep only surface atoms and opposite atoms from each protein (see figure 3). The resultant translational and rotational components of van der Waals's forces on each atom are calculated and applied to the barycenters of the proteins.

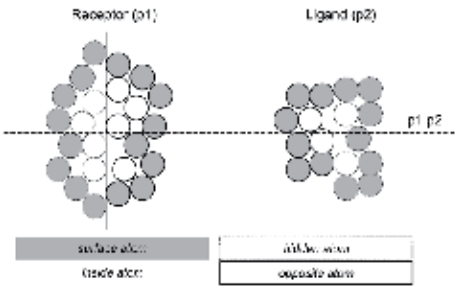


Fig. 3. Dynamic and static atom filtering for optimized computing of van der Waals interactions

Electrostatic interactions Unlike van der Waals interactions, electrostatic interactions even operate when "long" distances (about 10 Angstrom) separate groups of electrically charged atoms. Indeed some amino acids or atoms may present a positive or negative electric charge, which gives rise to electrostatic phenomena allowing formation of a protein-protein complex. Two approaches have been implemented to compute electrostatic phenomena.

We consider the interaction between two point charges in vacuum, and we use Coulomb's law (equation 2) with r_{ij} being the distance between the barycenters of charges q_i and q_j of the atoms considered, and ϵ_0 is the constant of the permittivity of vacuum. This potential can be translated to a force (F_{el}) usable for haptic interaction for example. This first approach involves calculating the forces to apply to each electrically charged particle considering only pairs of charged particles. This computation has quadratic complexity, because all distances between atoms must be computed. But it remains relevant in the case of medium-sized proteins, since the number of charged particles in a protein is limited in several models.

$$U_{el}(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}} \tag{2}$$

In the second approach (see figure 4, designed for a more efficient optimised calculation, the overall field of the electrostatic potential of the target protein (receptor) is computed beforehand using the *APBS* software [Baker et al. (2001)]. It allows to generate a 3D electrostatic potential grid, which can be used as a 3D texture. The gradient of the electrostatic

potential allows computation of force field vectors for each point of the grid. Atoms from the ligand protein are then "immersed" in this 3D force field surrounding the receptor. This method allows us to compute electrostatic forces for each atom in linear time, depending only on the number of charged atoms in the ligand. In both cases, we are able to obtain overall electrostatic energy and electrostatic forces on each atom.

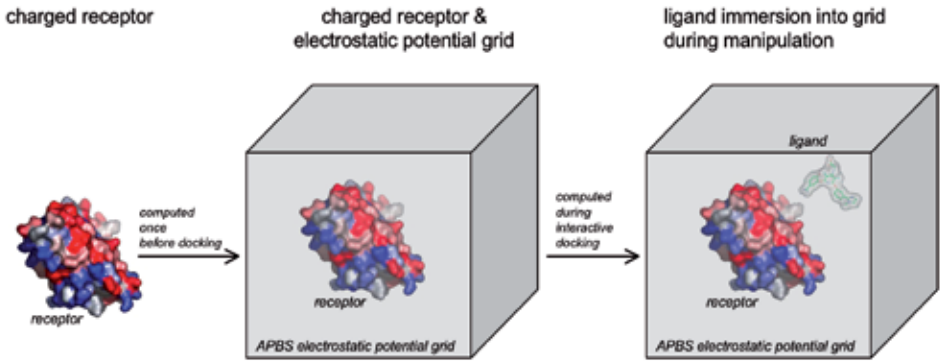


Fig. 4. Ligand immersion in the electrostatic potential grid of the receptor

2.1.4 Other criteria

In order to reach a finer description of protein-protein interactions, other criteria, based on energy, can be taken into account. To geometric/topological and biophysical criteria, one can add other criteria of utmost importance to protein-protein interactions, such as hydrogen bonds or the hydrophobic effects.

Hydrogen bonds. Hydrogen bonds (*e.g.* figure 1 in the bottom left corner) may strongly contribute to the favorable interactions of the complex binding energy. On average, there are 5-6 hydrogen bonds per protein-protein interface. In our application, when several atoms (nitrogen and oxygen) on the surface of each protein are close enough, closer than a distance of 3 Angstroms, and when their chemical environment is adequate, hydrogen bonds are created between these atoms. We use the same methods as described above for van der Waals interactions to filter surface atoms in order to decrease the complexity of calculating distances between atoms.

Hotspots at the interface. The number of "hotspots" at the complex interface refers to the list of amino acids present within the current interface region and previously identified using experimental methods as being important actors to stabilize the protein-protein complex.

2.1.5 Conclusion

This simulation model, based on rigid body docking, including optimisation to efficiently compute geometrical as well as biophysical properties, allows us to present these properties in real time during the interactive building of a protein complex performed by the user. Designing real time simulations is a first step of the IMS approach, providing a user with interactive control on the simulated object in real time.

2.2 A multiphysics and multiscale approach based on elastic networks

Simulation models dedicated to IMS must also deal with the intrinsic flexibility of proteins, and especially take into account local moves as well as large conformational changes. The representation used for our interactive simulation approach is quite simple, yet innovative, and has proven efficient on large biomolecular structures. Our method is based on a spring network simulation, inspired by the success of the Normal Mode Analysis method (NMA), known to accurately reproduce the elastic behavior [Cui & Bahar (2006)]. Moreover NMA is not sensitive to the scale of representation used for modeling. Hence this method can be applied to all-atom, coarsed-grain and residue/CA representations. We augmented this spring network model with non-bonded interactions and we propose to surround the charged spring network by an electrostatic field, allowing us to study conformational changes guided by electrostatic constraints. We called our implementation of this method *BioSpring* and describe it in this section. *BioSpring* allows us to simulate large structures in real time, fulfulling the most important constraint to provide IMS features.

2.2.1 BioSpring : an enhanced interactive spring network model

The first step of our approach is to build the spring network according to the 3D structure of the biomolecular system [Berman et al. (2000)]. At this stage, the user needs to choose a scale, targeting for example an all atom (AA), a coarse-grained (CG), or an alpha-carbon representation (CA). In this context, individual atoms can be considered as separate particles (AA model), or can be grouped into a single pseudo-particle, according to rules defined by the user (e.g. at the CG or CA level). In this way, we can adapt our approach to most commonly-used modeling approaches in theorical biochemistry.

The second step, is to connect the particles by springs obtained in the previous step. For this purpose, we define a distance cut-off, and we add a spring between two particles if the distance between them is less than the cut-off distance. This cut-off will depend on the scale and the representation mode (all-atom, coarse-grained, alpha-carbon, ...). For example, a cut-off between 7 and 15 angstrom is classically used for the CA representation [Cui & Bahar (2006)]. This process can be computationnally very time-consuming, especially on large structures. For each particle, we need to test if any other particles are closer than the cut-off distance. This approach has a quadratic complexity according to the number of particles. In order to deal with large structures and to decrease the complexity of the previous approach, we use a classical technique based on a regular 3D grid to partition a three dimensional space into cubes, also named "voxels". The grid covers the entire space occupied by the particles. The size of each voxel is the cut-off distance. According to its coordinates in space, each particle is projected into its voxel on the grid. In order to determine for a given particle p, all particles near p within the cut-off distance, we have to test the particles in the same and in the direct neighbouring voxels. This method has linear complexity according to the number of particles, allowing us to address very large structures. We will see in the next part of this paper how we can use the same grid to efficiently compute non-bonded interactions between particles.

After building the initial spring network, interactive manipulation of this molecular structure is provided using a classical newtonian particle-based simulation, taking into account spring forces (see equation 3) between particles and external forces $\vec{F}_{control}(p)$ (see 5 equation) provided by the user on a particles p through a specific graphical user interface. In the

equation 3, $k_{stiffness}$ is a global stiffness for all the springs and the force between two particles p and p' linked by a spring depends on the distance between these particles. If this distance between p and p' is equal to the equilibrium spring length $e_{pp'}$, which is the distance between these particles in the initial structure used to compute the network, this force is null by definition. In the other cases, when the distance $d_{pp'}$ between p and p' changes because of external forces, the generated spring forces tend to bring the structure back to its equilibrium conformation at $e_{pp'}$. Damping forces are used to stabilize the system (see equation 4). This is necessary because the user injects energy into the system by adding external arbitrary forces. It should be noted that some experimental and theorical studies provide estimates for $k_{stiffness}$, which allows us to work with magnitudes of forces in the simulation that are relevant from a biophysical point of view.

$$\vec{F}_{spring}(p) = \sum_{p' \in Springs(p)} k_{stiffness} \vec{u}_{pp'} (d_{pp'} - e_{pp'}) \tag{3}$$

$$\vec{F}_{damping}(p) = -k_{damping} \vec{V}(p) \tag{4}$$

$$\vec{F}(p) = \vec{F}_{spring}(p) + \vec{F}_{damping}(p) + \vec{F}_{control}(p) \tag{5}$$

At each time step of the simulation, to compute new positions and velocities according to the spring and external forces applied on the particles, we use a velocity verlet integrator described in equation 6.

$$\begin{aligned} \vec{P}(t + \Delta t) &= \vec{P}(t) + \vec{P}(t)\Delta t + \tfrac{1}{2}\vec{A}(t)\Delta t^2 \\ \vec{V}(t + \Delta t) &= \vec{V}(t) + \tfrac{\vec{A}(t) + \vec{A}(t+\Delta t)}{2}\Delta t \end{aligned} \tag{6}$$

Finally, the graphical user interface must provide interactive simulation features, allowing a user to visualise the spring network simulation in progress and to interactively apply external forces $\vec{F}_{control}(p)$ on particles. Combining these interactive simulation features and our interactive spring network simulation approach, a user can manipulate some parts of a large biomolecular system, and interactively observe the effects of this manipulation highlighting biomechanical properties such as rigid vs. flexible areas or allosteric effects.

However, even if the spring network model embeds an approximation of bonded and non-bonded interactions at the local scale, this model is not able to deal with long range and steric interactions during a simulation, because the spring 'particles' (atom, coarse grain, or residue-level) are considered as points. For example, domain interpenetration is allowed in the default spring network model. This is not a problem when the goal is to highlight local flexibility or rigid areas, but it is a critical issue for our objective of interactive modeling of large biomolecular systems. Similarly, it is also necessary to take into account electrostatic interactions during interactive modeling. For these reasons, in addition to spring forces, we introduce classical non-bonded forces to take into account steric and electrostatic interactions between particles in our model.

In order to meet the specific needs of the user, BioSpring provides a variety of terms to represent steric interactions. The most simple term is the linear steric model (see equation 7) which can be used to avoid atom or pseudo-atom collisions and take into account the 3D shape of the biomolecular model. This avoids domain interpenetration during the interactive manipulation, without taking into account complex realistic steric energy considerations and

in particular attractive terms. More classical models such as Lennard-Jones (see equation 8) are also avalaible, in order to take into account both attractive and repulsive interactions between atoms or pseudo-atoms, and to compute a more relevant steric energy during an interactive simulation. Atom radius r, epsilon ϵ and sigma σ parameters can be set up using configuration files, allowing us to use many of the currently available forcefields. However, this method has a complexity in $O(n^2)$ which is quadratic according to the number n of particles, because for each particle we need to compute the distance with respect to all other particles in the simulation. To address larger biomolecular systems, we necessarily have to decrease the complexity. We can remark that beyond a certain distance, several pairwise interactions become null or negligible. This is especially the case for linear or Lennard-Jones steric interactions. In this case, according to this distance cutoff, we can use the same optimisation techniques as in section 2.2.1, projecting particles into a 3D grid to accelerate the distance computation between particles at each time step of the simulation, by reducing quadratic complexity to linear complexity according to the number of particles in the simulation. The complexity is in $D.O(n)$ which is linear according to the number n of particles and the mean number D of particles in a voxel, which can be considered as a constant because it is related to the mean density of particles at a molecular scale.

We also use another way to optimize our simulation method as in many cases some part of the biomolecular complex can be considered as a rigid component. In this case, we can consider that these particles are static, *i.e.* have a constant position in space, because they belong to a rigid component. Hence it is useless to compute their interactions and to apply a positional integration on these static particles. We only have to take into account interactions between the dynamic particles belonging to the flexible part, and the interactions originating from the static particles and acting on the dynamic ones. This is a simple way to decrease complexity.

The following equations 7 to 9 describe the last two optimisations. *Dynamic* is the dynamic particle set, which contains all the particles belonging to the flexible part. The complexity is in $D(|Dynamic|)$ which is linear according to the number $|Dynamic|$ of particles in the flexible part.

$$s_{pp'} = (r_p + r_{p'}) - d_{pp'}$$

$$\vec{F}_{linearsteric}(p \in Dynamic) = \begin{cases} \vec{0} & \text{if} \quad s_{pp'} \leq 0 \\ \sum_{p' \in Neighbors(p)} -k_{steric} \vec{u}_{pp'} s_{pp'} & \text{else} \end{cases} \tag{7}$$

$$\vec{F}_{lennardjonessteric}(p \in Dynamic) = \sum_{p' \in Neighbors(P)} \vec{u}_{pp'} 4\epsilon_{pp'} \left[\left(\frac{\sigma_{pp'}}{9 d_{pp'}} \right)^9 + \left(\frac{\sigma_{pp'}}{7 d_{pp'}} \right)^7 \right] \tag{8}$$

$$\vec{F}_{coulomb}(p \in Dynamic) = \sum_{p' \in Neighbors(p)} -\vec{u}_{pp'} \frac{q_p q_{p'}}{4\pi\epsilon_0 d_{pp'}^2} \tag{9}$$

We can highlight another important fact: the previous approach is well-adapted to efficiently compute steric interactions by defining a distance cut-off. For long range interactions such as electrostatic ones, we must be extremely careful with this cut-off. It is preferable to avoid the use of cut-offs to stay biophysically relevant, but in this case, we fall down to quadratic complexity. We thus propose an efficient alternative to take into account long range

electrostatic interactions, considering that some parts of a biomolecular complex are rigid. A charge distribution can be translated into an electrostatic potential map, using the APBS tools [Baker et al. (2001)] for example. Charged particles belonging to the rigid components of our complex can be considered as a charge distribution, and are used as an input for APBS. The results of APBS can be interpreted as a 3D grid, and each voxel $V_{i,j,k}$ of this grid contains an electrostatic potential $E_{i-1,j,k}$. For a dynamic particle belonging to $V_{i,j,k}$, using this potential, we can compute an electrostatic force $\vec{F}_{map}$ using the charge of the particle V_p, by spatial derivation of this electrostatic potential (see equation 10). The potential forces $\vec{F}_{map}$ act on the flexible part and originate from the electrostatic potential map. They are defined by computing of the electrostatic potential gradient using the finite central difference method. In equation 10, we consider particle p belonging to the voxel $V_{i,j,k}$ of the electrostatic potential grid, and $E_{i,j,k}$ the value of the potential in this voxel. We define the gradient as the mean of the difference between the $E_{i,j,k}$ potential and the potentials of the six adjacent voxels, two for each axis. This method of computing the gradient reduces the bias related to the discretization of the grid. As regular grids are usually provided by tools such as APBS, Δ_x, Δ_y and Δ_z is the size of the voxel.

$$\vec{F}_{map}(p \in V_{i,j,k}) \simeq \begin{bmatrix} \frac{E_{i+1,j,k} - E_{i-1,j,k}}{2\Delta_x} \\ \frac{E_{i,j+1,k} - E_{i,j-1,k}}{2\Delta_y} \\ \frac{E_{i,j,k+1} - E_{i,j,k-1}}{2\Delta_z} \end{bmatrix} \tag{10}$$

To summarize, this last optimization technique is particularly well-adapted to study the behaviour of flexible biomolecules interacting with a large rigid biomolecular complex. Flexible parts are immersed into a grid and guided by a potential field induced by the rigid component, computed before the simulation. We have combined Eulerian (particle-based) and Lagrangian (grid-based) representations for molecular simulations, inspired by Joe Stam's works on Computational Fluid Dynamics [Stam (1999)]. This approach is also called semi-lagrangian or semi-eulerian method.

During the simulation, forces are computed and applied on the dynamic particle set ($P_dynamic$). We explicitly consider potential, van der Waals, Coulomb and external forces. Finally, these new forces are summed with an external force $\vec{F}_{control}(p)$ provided by the user through the graphical interface during the simulation.

$$\vec{F}(p) = \vec{F}_{spring}(p) + \vec{F}_{damping}(p) + \vec{F}_{map}(p) + \vec{F}_{steric}(p) + \vec{F}_{coulomb}(p) + \vec{F}_{control}(p) \tag{11}$$

2.2.2 Conclusion

BioSpring allows a user to quickly study the biomechanical properties, by interactively highlighting rigidity, flexibility, and allosteric effects, in order to provide new hypotheses about a biomolecular system. Moreover, our approach is also designed to help the user in the complex task of modeling large biomolecular complexes before using more classical (and more time-consuming) simulation tools.

3. Multimodal interaction models

In order to interact with biomolecular complexes during a particle-based interactive simulation such as *Gromacs* [Hess et al. (2008)], *NAMD* [Phillips et al. (2005)] or *BioSpring* (see section 2.2.1) in progress, it is common to use a mouse for adding force constraints on

particles, providing two degrees of freedom (2DoF), e.g. the x- and y-axes, for the interaction. Using a 3DoF device such as a 3D mouse or a 3D haptic device is even better adapted to this task, in particular for selecting and moving particles in 3D space. Such a device with three instead of two degrees of freedom is more intuitive and efficient for interacting with a complex three-dimensional object, especially when stereoscopic features are used to improve the spatial perception. Furthermore, the immediate force feedback using a haptic device when a particle is actually picked significantly improves the user experience and greatly helps to immerse the user in the molecular scene. If visual feedback is essential especially during the selection and picking task of a particle, the user often asks for additional explanations before getting started. With force feedback, this barrier is lifted, as the interactive simulation becomes more intuitive and is comparable to intuitive dextrous manipulations such as those carried out in daily life. Hardware requirements are modest. In our experience, this approach is viable using a small and affordable haptic device, providing 3D positions and handling 3D directional force feedback. Such an entry-level solution designed for a desktop use is targeted at a large user community and is very easy to set up.

3.1 Pick and pull particle interaction models

The haptic device is used in order to control the direction of the forces applied to selected particles and to adjust the amplitude of these forces. This interaction method contains two stages. The first stage comprises the selection of a single probe particle or a set of particles that we will name $P_{selection}$, using a 3D tool attached to a haptic device and its buttons. In a second stage, the model described by equation 12 is used in order to compute the forces $F_{control}$ applied to the selected particles and sent to the BioSpring simulation as control force (see section 2.2.1). $F_{control}$ is proportional to the distance between the geometric centre of the particle set and the tracker position $P(tool)$. For computing force feedback, the main idea of this approach is to link the selected atoms and the 3D haptic tool with a spring. Instead of providing direct haptic rendering of forces computed in the simulation, the force feedback $F_{feedback}$ only depends on the spring length according to equation 13, which in turn is influenced by the way the simulation reacts to the applied force.

$$\vec{F}_{control}(p \in Selection) = -k_{control}[\vec{P}(tool) - \frac{1}{|Selection|} \sum_{p' \in Selection} \vec{P}(p')] \tag{12}$$

$$\vec{F}_{feedback}(tool) = -k_{feedback}[\vec{P}(tool) - \frac{1}{|Selection|} \sum_{p' \in Selection} \vec{P}(p')] \tag{13}$$

The resulting forces are rendered by haptic feedback if a haptic device is used, and by visual feedback such as the blue arrows shown in the Molecular User Interface (MUI), top left part of Figure 5. These forces are simultaneously sent to the interactive simulation. It will take these forces applied by the user on the selected atoms set into account as a control force as described in section 2.2.1.

We emphasize that the haptic loop computation frequency must be between at least 300 to 1000 Hz in order to provide a haptic rendering of good quality. A strong point of the approach described above is that a low physical simulation framerate does not cause instabilities and does not affect the quality of the haptic feedback. With this decoupled spring model, force feedback can be computed at a very high frequency required by the haptic device.

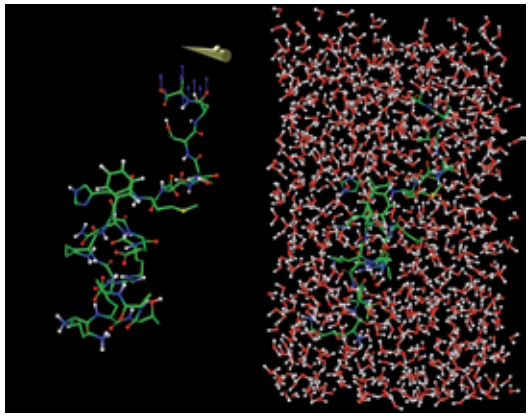


Fig. 5. Dynamic haptic control of a simple polypeptide with (right) or without (left) solvent - "ball and stick" representation

3.2 Interaction models for manipulating proteins as rigid body

The interaction method is more complicated when we want to provide controls and feedback during a rigid body based simulation, comparing to pick and pull a particle set as described in the last section. In order to manipulate both individual proteins and attempt to interactively study interactions between two proteins, the user may rely on various devices and interaction paradigms. A first paradigm associates the position and orientation of the protein with a 6DoF (6 degrees of freedom, 3 for translation, 3 for rotation) devices, such as a 3D mouse or a haptic device. Commonly used in the Virtual Reality domain, haptic devices are specifically used for manipulation and assembly tasks. Collision feedback rendered by 6 Degrees of Freedom (6DoF) haptic devices helps users to assemble 3D objects (see figure 6).

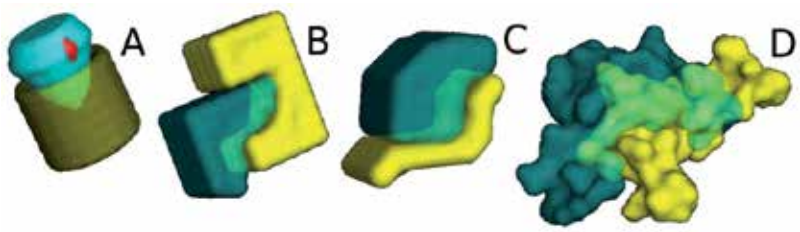


Fig. 6. Different kinds of assembly

3.2.1 Related works dealing with device workspace limitations

All devices have a limited workspace, a limited precision, and limited rotational movements. In order to overcome these limitations, a basic manipulation control is the clutching/unclutching interaction technique, which is however time consuming and does not allow a user to focus on his task. When the user is physically stopped in his movement, by reaching either a boundary of the device or an uncomfortable wrist position, he can press the clutching button to find a better position without moving the virtual object. When releasing, the object is re-attached with the same position and orientation. In this technique, the position and orientation of the 3D virtual object is an isomophic mapping of the position and orientation of the device.

Other solutions avoiding clutching/declutching, such as the *Bubble* technique, [Dominjon et al. (2005)] propose, to perceive (via haptic and visual feedback) the hardware limitations of the device, and provide a rate control based on an isomorphic mapping when the device is far from its workspace boundaries, and on a non-isomorphic mapping near the boundaries, also proposed LaViola & Katzourin (2007).

3.2.2 Related works that deal with high precision assembly

Morover protein-protein docking tasks require high precision. Haptic-guidance based approaches are often used in order to help the user reach a precise and predefined assembly goal. However, there are a few haptic interaction techniques designed to facilitate microassembly tasks for which haptic guidance is unsuitable, such as protein docking. The objective is to find an optimal but precise 3D configuration by interactive exploration.

3.2.3 A haptic interaction paradigm for rigid body based biophysical simulation

We propose an innovative technique to both overcome the physical limitations of the device and to reach the high accuracy required by micromanipulation tasks without a predefined goal, as it is the case during interactive docking simulations. This approach is based on a non-isomorphic mapping around a neutral referential retrieved by an elastic haptic feedback, in addition to external haptic feedback computed by the biophysical rigid body simulation.

In contrast to the method based on the haptic workspace boundaries, our approach is based on a *neutral referential*. Our solution is an implementation of a rate control technique with a 6DoF feedback device, based on a neutral referential, inspired by Bourdot & Touraine (2002).

Our contribution is to use the elastic force feedback to help the user return to the position/orientation of the neutral referential. First, we define a neutral referential with an origin corresponding to the most convenient position/orientation for the user holding the device. For each movement, we calculate a feedback force and torque to bring the user back to this neutral orientation/position (see figure 7 A). In order to compensate the inherent imprecisions of the device, we define a "dead zone" near the neutral referential, in which no movement occurs. The device is then physically restrained inside a comfortable workspace, while the virtual objects have an infinite motion space.

The rate control is based on the difference between the position/orientation of neutral referential initially chosen by the user and the position/orientation of the device during manipulation. The interpolation of movements is obtained by a downscale factor for translation and by a quaternion interpolation for rotation. The level of interpolation varies according to the distance of the two objects to be assembled. Concerning translational motion, the interpolation is done by rescaling the distance vector representing the position of the device from the origin of the neutral referential. In the following equation, $\vec{S}$ is the interpolated translation, $\vec{p}$ the current device's vector position and i the scaling factor. Concerning rotation, at any time of the manipulation, the rotational motion of the device controls the angular velocity of the object. The orientations of the device q_d and the object q_o are represented by quaternions. The rotational motion q_s of the object is then given by the multiplication of the two quaternions (see figure 7 B). The *SLERP* interpolation [Shoemake (1985)] is traditionally used to calculate intermediate frames between two quaternions (start and end orientions) in order to produce smooth rotation. Here, we use the *SLERP* interpolation to calculate a range of

quaternion orientations, between the current object's orientation and the one it would adopt after the motion. Then, an orientation at the time t can be picked according to the desired level of attenuation. In the following equation, q_s is the quaternion representing the rotational motion applied to the object without interpolation, q_d is the quaternion of the device and q_o the quaternion of the object. q_a is the softened speed using the *SLERP* function applied between q_o and q_s at the time t of the interpolation. Here, t and i are functions of the distance between the manipulated object and the area of assembly. The attenuation increases when this distance decreases.

$$\vec{S} = \vec{p} \cdot i \tag{14}$$

$$q_s = q_d \cdot q_o \tag{15}$$

$$q_a = \text{slerp}(q_o, q_s, t) \tag{16}$$

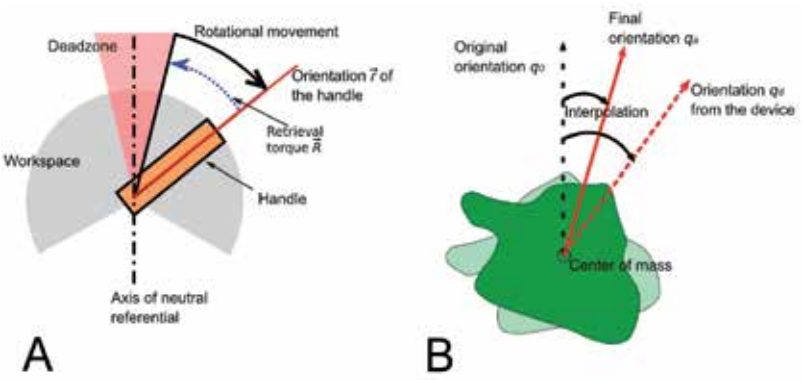


Fig. 7. **Results of a device rotation.** A, the device is rotated from the axis of the neutral referential to the orientation $\vec{r}$, taking into account the deadzone. The retrieval torque $\vec{R}$ is thus equal to $-\vec{r}$. B, the motion of the device is mapped on the object, q_o representing the neutral referential. The final orientation q_a is obtained by applying the *SLERP* interpolation

3.2.4 External feedback

The interaction between the manipulated protein and the other one during the docking task produces biophysical interactions provided by the rigid body based biophysical simulation. These forces are not directly applied to the 3D object in the scene, but are rendered by a haptic force-feedback summed with the elastic feedback used to retrieve the neutral referential.

3.2.5 Conclusion

One of the challenges lies in the protein interaction paradigm simulated by rigid body biophysical simulation. We developed a new method to provide a fine control of the protein with a 6DoF force-feedback device, providing simultaneously biophysical feedback coming from rigid body based simulation. According to the results of an ergonomic study, our technique provides at least the same precision (RMSD) and performance (task time) as direct manipulation with clutching/declutching and successfully overcomes the physical limitations of the device. Moreover subjective results show that users feel more comfortable with our method which avoids the clutching mechanism. We suspect that these results come from the fact that the user is more focused on the assembly task, instead of spending time in

clutching/declutching. Further evaluation must be lead in this way. Participants found our technique less disturbing than clutching, appreciating the fact that there is no button to press to manipulate the object. Furthermore, their arm was never in an uncomfortable posture. They furthermore liked the adaptive interpolation. The slowness of the interaction when the two objects are very close was judged pertinent in order to accurately assemble the objects. Another interesting observation is that the most negative comments were not about the manipulation technique itself, but concerned difficulties with the 3D visual perception of a complex protein surface (see figure 6 D). Our approach could thus be an alternative to classical ones and provide at least the same efficiency. We are working on improving the precision of our approach by dynamically tuning the scaling factor used to control rotational and translational velocity. This could be done using the minimal distance between the two objects during the assembly. Finally, we highlight the fact that our approach addresses most problems of the physical limitations of haptic devices (workspace size, precision, mechanical constraints), avoids the use of a clutching/declutching mechanism, is well-adpated to both manipulation and navigation, and could be applied to other 6DoF devices, and does not require complementary visual feedback.

4. Multimodal rendering models

Given the large quantity of biophysical or geometrical information provided by IMS and conveyed to the user in real time, it seems relevant to supplement visual feedback with audio and haptic feedback. Haptic rendering is known to improve the quality of operator interactivity in an immersive environment, as well as his perception of the objects handled or data analyzed [Seeger & Chen (1997)]. Likewise, audio renderings may improve communication of complex information [Barass & Zehner (2000)]. Furthermore, substitutions and redundancy between these channels of communication may have beneficial results on user performance, as long as the choice of modalities is relevant to the task at hand. Richard et al. (2006) and Kitagawa et al. (2005) showed that specific audio and visual renderings can effectively convey information that is presented using haptic modalities. In this part, we provide some examples of haptic audio rendering especially dedicated to study protein interactions using rigid body based biophysical simulation.

4.1 Visual rendering

To represent protein structures, the community of biologists uses standard representations that any specialist can understand 8. They range from a per-atom representation 8(A, B, C) to molecular surfaces 8(H). Some high-level metaphors with ribbons and arrows 8(E,F,G) can describe the secondary structure in a schematic way. Color schemes for atoms respect different standards to simplify the distinction between the different elements of the molecule.

IMS can act at different scales, from whole proteins to precise atomistic interactions, sometimes in the same simulation run. The visuals must then follow the user needs. Three main features have to be fulfilled : interactive frame rate, display of potentially huge molecules and coherent visual information.

For rigid-body docking, pre-computed triangulated surfaces of the proteins and secondary structure representations can be used. But if the atoms are allowed to move inside the structure, computing their surface in real-time is too time consuming in most of the cases.

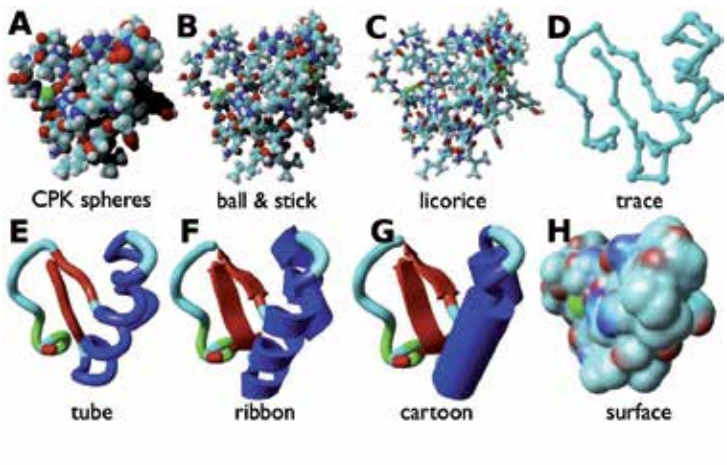


Fig. 8. Standard representation of protein structures. A-D : atomistic representation with spheres and bonds. Colors depend on the atom element. D : Backbone of the protein. E-G : secondary structure visualisation with high-level metaphor objects. Beta-sheets are in red, helices in blue and turns in green

So when it comes to more precise interaction and soft docking, spheres and bonds are more tractable.

Representation of proteins by spheres and bonds using common graphical primitives is easy to implement but generally not appropriate to reach an interactive frame rate. Each primitive is composed of many triangles, then displaying spheres or cylinders consumes a lot of computation time.

Other methods use Graphics Processing Unit (GPU) programming capabilities to draw the spheres and bonds directly on the GPU with no other information than the size and position of the particles.

Different textures and effects can be applied to emphasize interesting locations, collisions or other physical properties.

4.1.1 GPU shaders and HyperBalls

The computer visualization field evolves very quickly due to continuously renewed graphics hardware capabilities. So, the latest contributions from this domain of research has clearly helped scientists to display more and more complex systems. The latest graphics techniques can provide an improved visual perception which could drastically impact the way to visualize molecular structures [Chavent, Lévy, Krone, Bidmon, Nominé, Ertl & Baaden (2011)]. For example, using GPU shaders, i.e. code used to directly program the GPU, it is possible to accelerate and enhance the quality of well known molecular representations such as Molecular Surfaces (figure 9 A), Ball & Stick (figure 9 B), Van der Waals (figure 9 D and E) or protein Secondary Structure (figure 9 C). It is also possible to add lighting effects in real time in order to improve the perception of molecular shape or highlight molecular contours (figure 9 D and E). Furthermore, one can add effects such as blur to depict protein flexibility (figure 9 B). All these graphics techniques, available in real time, will be a great help for the users to interact in a wiser manner with their molecular structures.

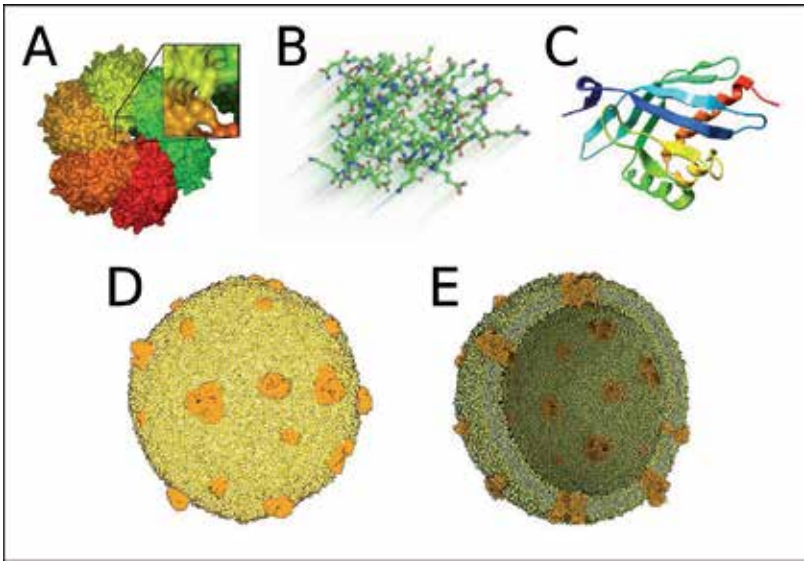


Fig. 9. New and revived molecular metaphors. (A-D) several molecular surface representations: (E) illustrate lighting effect to enhance molecular structure perception.

We have recently developed our implementation of molecular representations on the GPU [Chavent, Vanel, Tek, Lévy, Robert, Raffin & Baaden (2011)]. In this work, we introduced a visual molecular model, the HyperBalls representation, that offers a continuous representation smoothly connecting between classical representations such as licorice or ball-and-stick (figure 10). This representation takes benefit of a GPU ray-casting implementation to visualize molecular systems efficiently. The proposed implementation of the HyperBalls method is efficient for both static and dynamic visualizations of a large number of molecules and is particularly well adapted to visualize huge molecular systems. At present, without further optimization, we can smoothly and interactively render systems with more than 560,000 atoms, reaching some limits for systems comprising a few million spheres. We can expect that our implementation will benefit from the future GPU architectures, where performance increases drastically from a generation to another. This HyperBalls implementation is clearly well suited for an interactive and immersive approach due to the quality rendering and the display efficiency. Furthermore, it is possible to see in real time atomic bond evolution that can be beneficial for interactive docking (see figure 10).

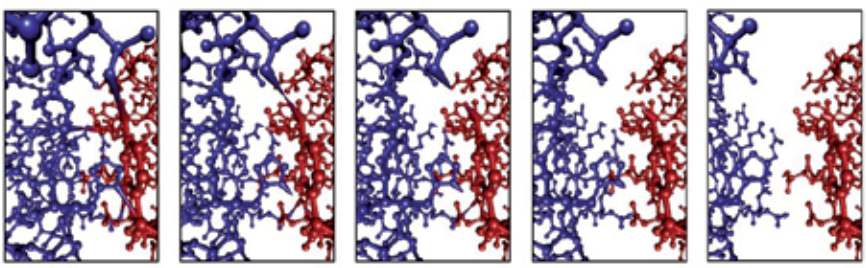


Fig. 10. HyperBalls representation to depict hydrogen bond disruption at a protein interface

4.1.2 Point-sprites

A simple way to represent molecular structure is to depict it as a collection of spheres. To represent spheres, it is possible to use only one square per sphere, always oriented perpendicular to the screen plane. Then an image (also called sprite) of a sphere is pasted on this square (figure 11). This method is usually used to depict visual effects such as flames, smoke or dust where one needs to display a big amount of animated particles. This method is really efficient and commonly implemented in 3D graphics libraries. The main drawback is that sprites are superimposed on each other, so there is no intersection between the individual spheres and it implies to sort the particles along the depth axis.

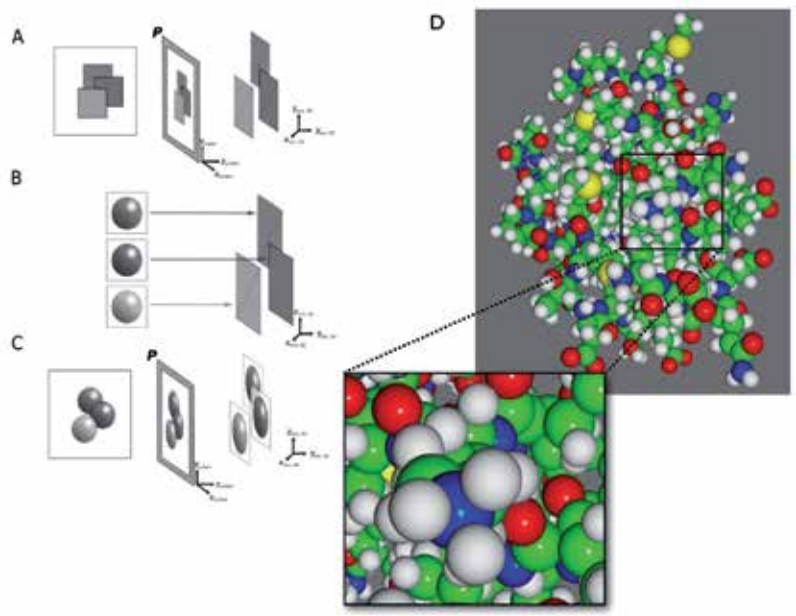


Fig. 11. Point-Sprites method to represent atoms of a protein

4.1.3 Benchmarks

These methods, as well as HyperBalls, have been implemented in an Unity3D (http://unity3d.com/) application and evaluated in terms of frame rate. The HyperBalls GPU shaders were adapted to fit the constraints of Unity3D but the performance was not as good as the initial implementation. The benchmarks show that the point-sprites method is far more efficient than the others. However, the frame rate is not constant when the camera is moving. In fact, the particles must be sorted to be displayed correctly which takes some time when there is a huge number of particles. Domain decomposition can be used to reduce this effect but then some visual glitches at the frontiers of the domains can occur.

The visual result is quite different depending on the methods. Point-sprites can be confusing as the bonds are missing and the spheres are superimposed (figure 11 D). But from a far point of view, the general form of big proteins is kept and using a good color scheme helps to distinguish the interesting areas of the molecules. The traditional primitives can be used

along with visual effects such as ambient occlusion, shading or texturing that are often pre-implemented for triangulated objects.

So what we suggest is to combine those methods in an interactive way, according to the size of the system and the user actions. Triangulated primitives are easy to implement for quick development and nice visual effects. For big systems and when the proteins are far away from the camera, particles are particularly suited. When the user zooms into a specific area, a more precise representation such as HyperBalls is adapted, especially to depict interactions.

4.1.4 Visual effects

To allow an accurate interaction with the particles, the position of each one in space must be easily discriminated by the user. The best option is to use stereoscopy but 3D display devices are not common yet. However it is possible to add some visual effects on the objects to overcome this problem. Shading, depth-cueing and ambient-occlusion are commonly used to add realistic lighting and depth perception to a 2D image (9 E). Also, texturing and contouring can help to highlight particular areas and particles and blurring effects can be used to emphasize some movements (9 B).

4.2 Haptic rendering

Currently, there are very few IMS frameworks that include large-scale haptic feedback (force/tactile feedback). This is mainly due to the complexity of computing operations of molecular simulation, which makes it difficult to comply with constraints in terms of refresh rates for real time haptic feedback (from 200 Hz to 1 kHz). Another difficulty is to render various kinds of physico-chemical interactions such as steric or electrostatic interaction. In order to obtain a consistent haptic feedback, only one type of rendering is provided to the user at a time. However one should note that at the perceptual level, steric interaction rendered using haptic feedback are similar to surface collision renderings since it prevents molecular interpenetration.

Most of haptic feedback presented in this section are computed using the rigid body simulation model described in section 2.1. In all rendering, the haptic-device controlled protein, which we will call ligand, can be considered as a big probe against the other protein, which we will call receptor.

4.2.1 Steric and electrostatic interactions

This rendering is used to provide haptic feedback of non-bonded interaction. Haptic rendering of physicochemical interactions consists in feeding the haptic device with the resultant forces computed as described in section 2.1. Forces can be computed and rendered independently or summed up to obtain a total resultant force. Exploration of the receptor by the ligand thus aims at finding stable areas. When the two proteins are in an unstable conformation it renders an unsteady feedback, thus leading the user to drag the ligand towards the surface of the receptor to find a better position and orientation. However the complexity of the force fields induce very irregular directional forces affecting the precision of the manipulation. It appears especially with steric interactions because of the non-linearity in the *Lennard-Jones* potential used to model these forces.

4.2.2 Surface collision

Two approaches were explored to render collisions between both proteins considering their surfaces. The first consists in computing a repulsive force. The direction of this force is the opposite of the direction provided and the module is proportional to the number of colliding triangles determined by the *RAPID* computation as explained in section 2.1. This force can also be weighed by a distance or a volume of interpenetration. Therefore the feedback is more relevant, but the complexity of the computation induces lower refresh rates which could lead to lags in feedback. Rather than repulse the two molecules from each other, the second approach, also based on distance computation, aims to prevent collisions locally by modeling contacts points as springs. The method is introduced by Johnson & Willemsen (2003) and allows fast computation of local minimum distances based on the geometry of the model as well as resulting force and torque. Interestingly the spring model described can be easily adapted to model atomic clashes, such as steric ones in our case. Instead of using the complex Lennard-Jones potential to render the resulting force, interactions are modeled through this more simple spring model with realistic cutoffs (2.5 Angstroms). As the atomic distance computation is already optimized to take into account only surface and opposite atoms, the refresh rate is sufficient and allows a very precise rendering of the contacts, allowing users to feel holes and bumps at the surface. Hence computation speed and consistent feedback constraints are observed ensuring a biological relevance. Current research aims to determine how the size of the proteins affects computing time. It will also be interesting to compare this atomic clashes-based approach with the geometric one which could provide faster computation.

4.3 Audio rendering

Sonification is the use of non-speech audio to convey information. Due to the high temporal resolution and wide bandwidth, the use of auditory stimuli seems highly suitable for time-varying parameters (very high temporal definition when compared to other modalities such as video and haptics), concurrent streams (overlapping of multiple audio renderings for various parameters is possible and easily understandable if these are properly designed), and spatial information (lower definition if compared to visual stimuli, but perceptible over the 360 degree sphere, therefore allowing true 3D rendering).

A large variety of sonification techniques exist and are used in various applications [Walker & Lane (1994)]. One sonification technique is referred to as "parameter mapping" [Hermann & Ritter (1999)], and it is this technique we used to study protein interaction. Parameter mapping sonification is based on creating a link between the data to be rendered and the parameters of a synthesizer (or of any other device which generates or plays back sound). In this particular sonification typology, three elements need to be carefully considered [Walker & Lane (1994)]:

- The nature of the mapping: which data dimension (*i.e.* temperature, pressure, velocity...) is mapped onto, or represented by, each acoustic parameter (*i.e.* frequency, loudness, tempo...). As an example, for a sonification task the temperature might be linked with the frequency of a sound, therefore as the temperature increases, the frequency of the corresponding sonification increases.
- Mapping polarity: in the event of an increase in the sonified data, the sonification parameter can decrease or increase. In the case of temperature-frequency mapping, it

is common to use an increasing-TO-increasing (up-up) polarity. An alternate example could be the size of an object being mapped to frequency: the polarity would likely be increasing-TO-decreasing such that large objects are linked to low sounds and vice versa.

- Mapping scale: in response to a specific increase of the data to be sonified, how much should the sonification parameter increase or decrease. One must take into account the possible range of the data, and the percentage of the usable audible range which is to be exploited. Human hearing is more sensitive to small frequency changes at low frequencies, rather than higher, following an exponential scale. In the case of temperature-frequency mapping the temperature could be exponentially linked to the frequency.

In our application, sound spatialization is used in two different ways: firstly, for local parameters the sonification is spatialised in the specific position where the parameter is calculated, in accordance with visual or haptic rendering, to provide additional information in the protein coordinate system (*i.e.* if the task is to sonify the collision between two different atoms on both proteins, sonification is spatialised at the position of the collision). Then, multiple concurrent sonifications can be spatially distributed in order to give a better intelligibility of the sonifications themselves (*i.e.* stream segregation, selective attention in auditory perception, cocktail party effect studied by Moore (2003)). In 2007, a set up a test for the validation of different sonification methods for object manipulation. Within this test, the subject was asked to change the orientation of a simplified 3D chemical compound in order to be the same as that of a given reference. To do this, the subject used an orientation tracking device. Three approaches for data parameter sonification were tested to improve the speed and accuracy of this manipulation: manipulation speed, angular distance from the reference configuration, and guidance towards the reference position.

Regarding the protein-protein docking task, the following biophysical information has been selected for the sonification:

4.3.1 Molecular surface collision and complementarity

Atomic surface complementarity is estimated essentially as a calculation of the variance of the inter-atomic distances on the two proteins surfaces. This parameter is used to control the variance of a randomly applied pitch to different grains of a granular synthesis process. Granular synthesis has been applied using a spoken word as audio sample (for this particular application, the french word "complementaire" has been recorded and used), repeated cyclically within the granular engine. In this instance, the word is unintelligible if the geometrical complementarity parameter is low, becoming more intelligible as the parameter increases. The rendered audio stream is doubled and associated to each of the two proteins, in preparation for further processing.

The collision parameter represents the number of collisions computed between the two surfaces. The employed method for atomic collision sonification uses a modulation of the phase of a sinusoidal wave whose parameters (carrier and modulator) are controlled by the global number of collisions. Starting with a continuous 400 Hz sinusoidal wave modulated by a 1 Hz signal, the frequency of the modulation increases as the global collision score gets higher, and with it the number of modulating waves, going from 1 to 4, when the two proteins are completely superposed. A second method developed is based on the individual association of every collision with a broadband noise processed with subtractive synthesis (the result is similar to wind noise). The noise is specifically filtered for every collision, adding a

controlled randomization of the filtering parameters, so that every "noise generator" sounds different from the others, and spatialised according to its proper position in space. Both of these sonification methods are based on the principle that the signal produced becomes more and more annoying as the number of collisions increases, encouraging the user to change the position and distance of the proteins in order to reduce the number of collisions, and as such stopping the annoying sound. Regarding the second sonification method, sound spatialization helps the listener to localize the part of the protein surface where the collision is taking place, and to guide him/her towards an orientation of the protein for which no collisions are present.

4.3.2 Electrostatic energy

This electrostatic parameter is computed from electrostatic interaction energies between charged particles. Electrostatic energy sonification is performed through the alternation of two sounds, generated using additive synthesis, whose pitch and timbre vary as a function of the global value of this specific force (scalar value). The electrostatic force value is highly variable, and there is not a direct linear relationship between this parameter and a quality judgement of it being good or bad for the docking condition. The link between the parameter and the quality of its specific value has therefore been traced in a two dimensional Cartesian diagram, with the value of the parameter on the X axis, and the quality (being good or bad) on the Y axis. At a given electrostatic force value, the correspondent value on the Y axis has been sonified with the method previously described. For good values, the frequencies of the two sounds are coincident, and their spectra are perfectly harmonic, whilst as the value worsens, the two frequencies become more distant, and the spectra more inharmonic.

4.3.3 Steric energy

This parameter is computed from van der Waals interaction energies between particles. The sonification of the van der Waals energy is based on the principle of the beatings between two sounds frequentially close. As with the electrostatic force, for the van der Waals force value there is not a linear relationship between the parameter and a quality judgement (being good or bad). A mapping similar to the one described for the previous sonification method (electrostatic force) has been employed, with the Y axis value being sonified. Two intermittent sinusoidal pulses are played back simultaneously: if the quality value for the van der Waals force is good, then the two waves have the same frequency, whilst as it becomes worse, one of the two pulses reduces in frequency by up to 20 Hz from the other. This processing results in the creation of beatings between the two frequencies. If there are no beatings, then the score can be considered to be good. In contrast, if the beatings become more frequent (more rapid beat frequency indicates greater frequency separation between the two pulses) the score is becoming worse.

4.3.4 Hotspots at the interface

The number of "hotspots" at the complex interface refers to the list of amino acids present within the current interface region, previously identified using experimental methods as being important actors for protein-protein interaction. Finding hotspots at the protein-protein interface is an important part in judging the quality of solutions. In a second stage, the two audio streams are processed with a low-pass filter with the cutoff frequency controlled by the percentage of protein hotspots which are situated on the interface region. If none of the

hotspots are present on the interface the low-pass filter frequency is set at 200 Hz, making the sound nearly inaudible. The cutoff frequency of the filter increases with the number of hotspots present at the interface, making the sound clearer and brighter until, in the optimal position, the frequency filtering is completely deactivated. The two audio streams are rendered stereophonically, associating the left and right channels respectively to the first and second protein.

5. Coupling simulation and interaction codes

Molecular simulation engines previously described provide time-dependent atomic or particle positions, velocities and system energies according to biophysical models at different scale. These models can now compute a molecular dynamics trajectory of interesting biological systems in interactive time. This progress allows to control and visualise a molecular simulation in progress. We have developed a generic library, called *MDDriver*, in order to facilitate the implementation of such interactive simulations. It allows to easily create a network connection between a molecular user interface and a physically-based simulation. We use this library in order to study a real biomolecular system, simulated by various interaction-enabled molecular engines and models. We use a classical molecular visualisation tool and a haptic device to control the dynamic behavior of the molecule. This approach provides encouraging results for interacting with a biomolecule and understanding its dynamics. Our goal is to extend IMS approach to a broader range of simulation engines, as the use of a specific simulation sofware or model often depends on the studied biological system. We have thus developed a generic and independent library, called *MDDriver*, which allows us to easily interface molecular simulation engines with molecular visualisation tools through a network connection. As a first step, we have rendered the calculation modules easily interchangeable while keeping the existing *VMD* user interface as MUI.

5.1 *MDDriver* : a library to coupling molecular simulations codes and molecular user interfaces

In the *VMD/NAMD* architecture, the IMD network protocol [Stone et al. (2001)] was developed in order to interface the Molecular User Interface (MUI) with the MD engine. However, the use of a specific simulation engine and MUI strongly depends on the studied biological system and on user habits. Adding IMD capabilities to other simulation engines and molecular models as well as to a variety of MUIs in addition to *VMD* and *NAMD* enables a whole range of new possibilities in interactive molecular simulations. This approach allows us to address a larger user community working on molecular modeling and simulations, sometimes based on their own home-made simulation engines. Following these motivations, we developed a generic and independent library, called *MDDriver*, inspired by the *VMD/NAMD* approach.

5.1.1 Software architecture

We have thus encapsulated the IMD protocol in the *MDDriver* library, allowing a developer to easily adapt MUI code and MD code in order to extend them with IMD features. This interface provides functions for the exchange of specific data structures over a network: atom positions and system energies, computed for each simulation step by the MD engine (server part), and user-applied forces on a selected atom set.

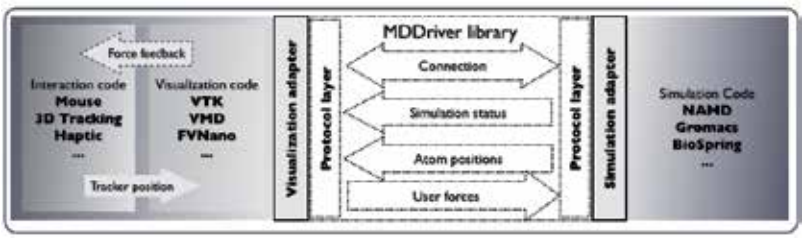


Fig. 12. *MDDriver* library for interfacing a Molecular Dynamics simulation with a Molecular User Interface (Interaction and Visualization code)

5.1.2 Molecular simulation *MDDriver* wrapper

This approach was tested, applied and improved by integrating calls to the *MDDriver* library into the *GROMACS* simulation engine [Hess et al. (2008)], thus rendering the simulation interactive via a MUI. We have used *VMD* as MUI in order to study the molecular behavior of Guanylate Kinase (GK) using an all-atom model and a coarse-grained representation [Baaden & Lavery (2007)] with *GROMACS*. Then we have tested a home-made simulation engine dedicated to molecular docking, which was also IMD-enabled.

MDDriver module offers a simple, modular and generic solution to combine any coordinates-based calculation code with various visualization programs. IMD simulation, this powerful tool for exploration of biomolecules structure in large biological system, is now accessible in a easier way to desktop or virtual reality computational environment. We insist on the fact that the *MDDriver* library was designed for easy integration into any molecular simulation engine providing time series of particle positions. Indeed there are many approaches capable of simulating the dynamic behavior of biomolecules, such as lattice simulations, elastic networks, coarse grain models or even quantum mechanical and semi-empirical methods.

5.1.3 Performances

We will only briefly comment here the desktop use performances obtained for the *MDDriver* library implementation to the *GROMACS* code. The data (coordinates, status and forces) transfer rate between calculation and visualization modules essentially depends on the size of the simulated system.Force application component alters slightly more IMD performances for small systems, depending essentially on selected/total particles ratio (increasing data exchange). In the context of large computing infrastructure deployment for *GROMACS* IMD using *MDDriver*, similar performances have been observed. This confirming robustness of the *MDDriver* library coupled to parallelized applications, performance of the display/interaction installation being the main limitation for IMD simulations of large molecular systems.

6. Applications

We propose in this last section to illustrate the previous simulation, interaction and rendering concepts especially designed for IMS, with several applications. In the first application, these concepts was used to designed new approach and methodology for docking. In the three next ones, these concepts was used in a research context to study some biostructural phenomena. In the two last one, we present two cutting edge scientific sofware that used and included all the innovative concepts presented in this chapter.

6.1 CoRSAIRe : a multimodal and immersive molecular docking project

6.1.1 Main focus

The main focus of the *CoRSAIRe* project [**?**] is to design a new methodology in that field based on advanced interaction and rendering possibilities, that Virtual Reality (VR) technologies may offer. With respect to other works on docking, we are specifically studying multi-sensorimotor rendering during an interactive docking task.

Usually user participation during the computationnal docking process was very limited, since it only involved configuring docking scripts and choosing one result amongst the computer-generated solutions to the studied problem. Indeed classic approaches to docking provide large numbers of complex configuration based on 3D data describing partner proteins. These algorithms take a long time to produce results, since they test all possible geometric configurations to dock the two proteins. These configurations are then filtered according to energy and physicochemical criteria. Finally, the scientist selects, in this set of results, a smaller set of possible solutions that can be tested against each other experimentally.

Relying on user expertise before applying automatic docking algorithms interactive context allows the user to use natural abilities for the detection of surface complementarity, as well as prior implicit or literature based knowledge regarding for example the nature of the protein-protein interface, what hotspots are present, etc.

It seems thus relevant to develop complementary or alternative approaches to docking. In project *CoRSAIRe* (Combination of Sensorimotor Renderings for the Immersive Analysis of Results) our hypothesis is that using Virtual Reality (VR) technologies and related interactions, which rely on multiple sensory and motor channels, may help experts in this docking task. There are several reasons for this. Firstly, stereoscopy, especially when it is adaptative, may improve perception of 3D protein models. Furthermore, direct manipulation of several proteins at the same time, as afforded by peripherals commonly used today for such tasks (*e.g.* 3D mouse, force-feedback interfaces, etc.) may be more intuitive and efficient than traditional, desktop WIMP[1]-type interfaces. Additionally, multimodal management of sensorimotor feedbacks (based on an approach aiming to dynamically specify adaptation of visual, haptic and audio renderings to the characteristics of the information in use) is one possible answer to the problems related to the simultaneous presentation of large amounts of data. Finally, a strongly interactive approach of VR docking allows the docking expert to be placed on the forefront of the work, rather than giving an automatic algorithm a complete control over the generation of possible sets of solutions. We believe our approach, which combines benefits of multimodal interaction with the capitalization of docking experts' occupational skills (in biology, crystallography, bioinformatics) in modelling will allow improvements in the speed of predictions for the structure of protein-protein complexes, as well as in overall search efficiency and in the quality of results obtained when analyzing possible solutions.

6.1.2 Discussion and results

This project allow us to define the multimodal allocation space ("modal allocation" [André (2000)]) that refers to the specific use of one or more sensory modalities to display an information. It is preferable for users to use optimal modal allocation considering both technical constraint, task (*e.g.* characteristics of information relevant to scientists) and

[1] Acronym of Window, Icon, Menu, and Pointing device.

	Visual	Auditive	Haptic
Surface representation	ok		
Surface collisions	ok	ok	ok
Surface complementarity	ok	ok	ok
Electrostatic interactions			ok
Electrostatic energy	ok	ok	ok
Steric interactions			ok
Steric energy	ok	ok	ok
Hydrophobic patchs	ok		
Hotspots at the interface	ok	ok	

Table 1. Restricting the modal allocation space

operator-related constraints (*e.g.* characteristics of perception, of expertise, etc.). The results of the project allow us to formulate the following principles for the design of a multimodal application for molecular docking, summed up in table 1.

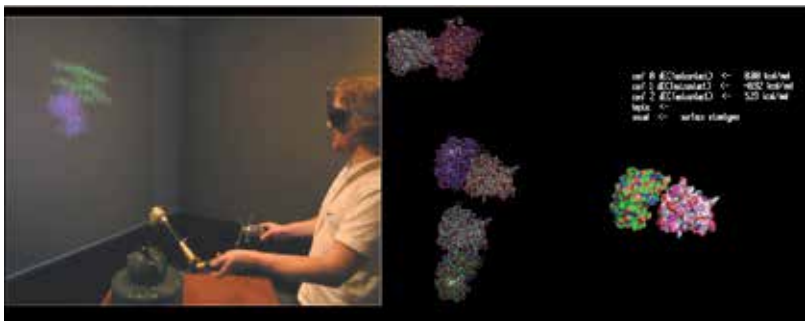


Fig. 13. A user immersed in the docking application (left). On the right, a sample screen capture following selection of three conformations by the user.

The interactive process dedicated to protein protein docking designed into the CoRSAIre project, allows significant reduction of the number of configurations to be tested by algorithms used afterwards, and we maximize the use of the user expertise. Our approach could also be reused in the design of future docking software, integrating factors such as protein flexibility, based on the premise that many docking problems involve flexible partners. Furthermore, this work should also focus on defining future situations of use of such tools. Indeed, our interactions with future users identified several possible avenues for the use of docking tools, *e.g.* teaching, scientific discovery, collaborative work, etc.

6.2 Interactively locating ion binding sites by steering particles into electrostatic potential maps

Interactively locating ion binding sites by steering particles into electrostatic potential maps Metal ions drive important parts of biology, yet it remains experimentally challenging to locate their binding sites into biomolecules (protein, DNA). With the MyPal method (Molecular scrutinY of PotentiALs), implemented in the *BioSpring* program, we use interactive steering of charged ions in an electrostatic potential map in order to identify their potential binding sites [Delalande et al. (2010)]. We use this method in order to facilitate the discovery of new relevant

ion binding sites by successfully retrieving the location of cation binding sites in DNase I enzyme and assessing their selectivity, combining atomic and coarse-grained resolutions.

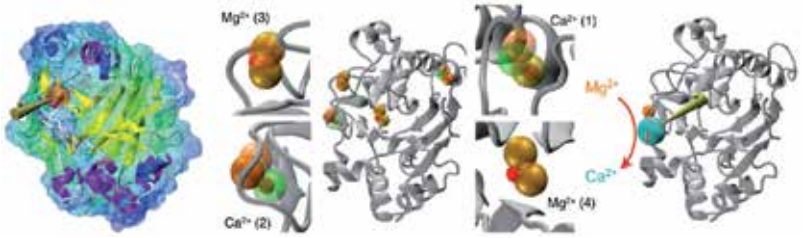


Fig. 14. Visual summary of the interactive experiments. On the left, interactive potential exploration of the DNase I enzyme using the MyPal method. On the middle, results of experiments for detecting a priori unknown ion binding sites. The reference position of each binding pocket is shown as a red sphere and MyPal predictions for a potential map with (orange) or without (green) ionic strength are displayed by transparent spheres. On the right, an ion substitution experiment ("molecular-billiard") at site 2 is depicted. Such an experiment probes the selectivity of a given ionic pocket for different ions

We interactively scanned the electrostatic potential of DNase I by using Na+, Ca2+ and Mg2+ as ionic probes. For the binding sites detection, Mg2+ cation was chosen as its double charge facilitates long-range electrostatic steering towards the binding pockets and its small size increases the accuracy for sensing the rough and detailed molecular surface at atomic resolution. Taking into account ionic strength for calculating the electrostatic potential leads to more accurate maps. However, without considering ionic strength we achieved comparable predictions and more easily detect binding sites thanks long-range driving forces (Figure 16). All four ion binding sites identified were retrieved by the MyPal approach. To assess the selectivity of identified binding sites, we start with different ions (Na+, Ca2+, Mg2+ and Cl-) at a given site and tried to interactively substitute this initial ion by another. Figure 16 and Table 2 illustrate and summarize the results for these ion substitution "molecular-billiard" simulations. As might be expected, chloride as an anion cannot be stabilized within any of the four cation binding pockets, nor can it displace a bound cation. Sites that are magnesium selective are well characterized by our approach. Less efficient substitution experiments may be related to the simplicity of our model in which selectivity depends on the shape of the pocket itself and the pathway for accessing it. Generally speaking, buried and narrow sites are unreachable for large ions, whereas sites localized at the enzyme surface are readily subject to ion exchange. In the latter case, haptic feedback helps the user to distinguish between favourable and unfavourable substitutions.

The current implementation of MyPal/BioSpring was not designed in order to provide precise quantitative binding affinity estimates, but to be capable of distinguishing in real time between non-existing, weak and strong ion binding sites and assess the relative selectivity of significantly different ionic probes. Despite the approximations made in the choice of the model representation it should remain possible to quantify the strength of binding by calculating the work required by the user to extract an ion from its binding site.

6.3 Interactive study of Guanylate Kynase opening and closure

In this study, we have worked on an intensive studied biomolecular system, the Guanylate Kinase (GK) enzyme. Structures for this molecule are provided by experimental methods

Probe (Site)	Ca^{2+} (1)	Ca^{2+} (2)	Mg^{2+} or Ca^{2+}(3)	Mg^{2+} (4)
Mg^{2+}/Ca^{2+}	$Ca \mapsto Mg$	$Ca \mapsto Mg$	$Mg \dashv Ca$	$Mg \dashv Ca$
	$Mg \dashv Ca$	$Mg \mapsto Ca$	$Ca \mapsto Mg$	-
Na^{+}	$Ca \dashv Na$	$Ca \mapsto Na$	$Mg \dashv Na$	$Mg \dashv Na$
	$Na \mapsto Ca$	$Na \mapsto Ca$	$Na \mapsto Mg$	-
Cl^{-}	$Ca \dashv Cl$	$Ca \dashv Cl$	$Mg \dashv Cl$	$Mg \dashv Cl$
	-	-	-	-

Table 2. **Ion substitution interactive simulation results.** The table indicates whether exchange from X to Y is possible ($\mapsto$) or impossible ($\dashv$). For instance, $Ca \dashv Cl$ means that Ca^{2+} cannot be displaced by Cl^{-}. A minus sign indicates that initial positioning of the chosen probe ion at the given binding pocket was not possible *via* our approach.

such as Nuclear Magnetic Resonance or X Ray cristallography. The molecule has a U shape with either a closed or an open conformation (see 15). The closure mechanism of GK consists in increasing the proximity of two substrate binding sites, for GMP and ATP, both essential for the enzymatic reaction. The goal of our study is to understand which parts of this system are involved in the closure mechanism. This mechanism has been investigated using our *MDDriver* framework (VMD/MDDriver/GROMACS) at two levels of detail. The first level corresponds to an all-atom model (18098 atoms), the second to a lower resolution coarse-grain model (1900 beads), and the third to a augmented spring network model. Prospective tests using coarse-grain simulations allowed for the efficient exploration of a broad range of possibilities to close the enzyme, trying to reach a closed conformation similar to the available experimental structures.

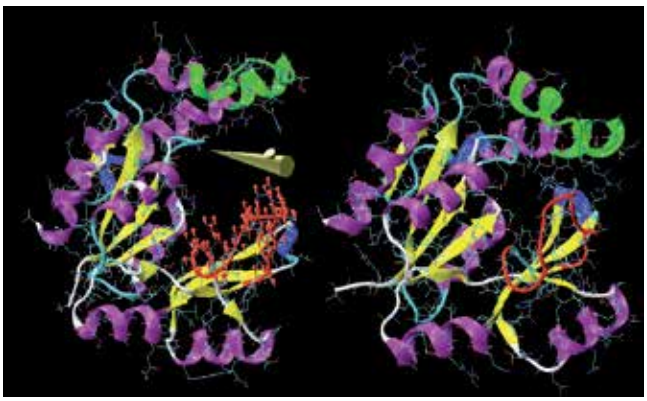


Fig. 15. Haptic control (red arrows in the red loop) of Guanylate Kinase closure. Secondary structure cartoon representation of the open state (left) and the closed state (right)

Figure 15 shows a secondary structure representation of the protein, considering specific architectural units such as the loops (white tubes), the helices (purple ribbons) and the beta sheets (yellow arrows). The crucial role of one loop (highlighted in red in Figure 15) in the initiation of GK's closure could thus be identified. It was then confirmed in a second phase using more detailed all-atom simulations. Understanding the features of this early intermediate state occurring as an impulse for the closure mechanism allows us to propose a novel mechanistic hypothesis. The loop move could be initiated by GMP docking, which may drive this loop via long range electrostatic interactions. When the loop draws closer to the

other side of the enzyme, conformational changes could be triggered, subsequently inducing a global closure of the enzyme. The interactive exploration of the simulation using the haptic modality lead us to this theoretical hypothesis. It also suggests that electrostatic interactions could be the main driving force for closure.

6.3.1 Modelling a transient stage of DNA repair by flexible docking of double stranded DNA to RecA nucleoprotein filaments

Homologous recombination is a fundamental process enabling the repair of double-strand breaks with a high degree of fidelity. In prokaryotes, it is carried out by RecA nucleofilaments formed on single-stranded DNA (ssDNA). These filaments incorporate genomic sequences that are homologous to the ssDNA and exchange the homologous strands. Due to the highly dynamic character of this process and its rapid propagation along the filament, the sequence recognition and strand exchange mechanism remains unknown at the structural level. By the interactive and flexible approach available from the BioSpring program, we investigated the possible geometries of association of the early encounter complex between RecA/ssDNA filament and double-stranded DNA (dsDNA) [Saladin et al. (2010)]. Due to the huge size of the system and its dense packing, we used a reduced representation for both protein and DNA. In this study, a systematic docking was also performed to associate dsDNA and the RecA/ssDNA complex, but this approach didn't enable the consideration of flexible regions of the nucleofilament RecA. BioSpring approach promoted to easily build a hybrid rigid-flexible representation of the molecular system by combining an Augmented Spring Network model (ASN) and a static molecular shape, and finally enabled to include very flexible L2 loops in the structure of RecA/ssDNA receptor. These flexible L2 loops constituted the only interactively controlled protein region, the rest of the RecA nucleofilament and the ssDNA were considered as static. Incoming dsDNA (ligand) was the second molecular fragment described by a flexible ASN model. During the interactive docking simulation, L2 loops moves were obtained by pulling user-selected atoms, while position and orientation of the dsDNA were controlled by acting on a fixed particles group (central nucleobases). Each single interactive simulation consisted in (i) moving L2 loops and simultaneously (ii) pulling the dsDNA toward the ssDNA, then (iii) allowing the relaxation of the system and finally (iv) saving the ligand and receptor positions.

Docking of curved dsDNA structures permitted to reach a more stable molecular complex than the one obtained from B-type DNA ligands. These simulations also demonstrate that it is possible for the double-stranded DNA to access the RecA-bound ssDNA while initially retaining its Watson-Crick pairing and emphasize the importance of RecA L2 loop mobility for both recognition and strand exchange.

6.4 ePMV : embedding molecular modeling software directly inside of professional 3D animation applications

ePMV, the Embedded Python Molecular Viewer [Johnson & Autin (2011)] is an open-source plug-in, that runs the molecular modeling software PMV [Sanner (1999)] directly inside of numerous professional 3D animation applications (hosts), to provide seamless access the capabilities of both systems and to simultaneously link the host software to other scientific algorithms. ePMV currently plugs into Maxon Cinema4D, Autodesk Maya, and Blender. Uniting host and scientific algorithms into a single interface allows users from

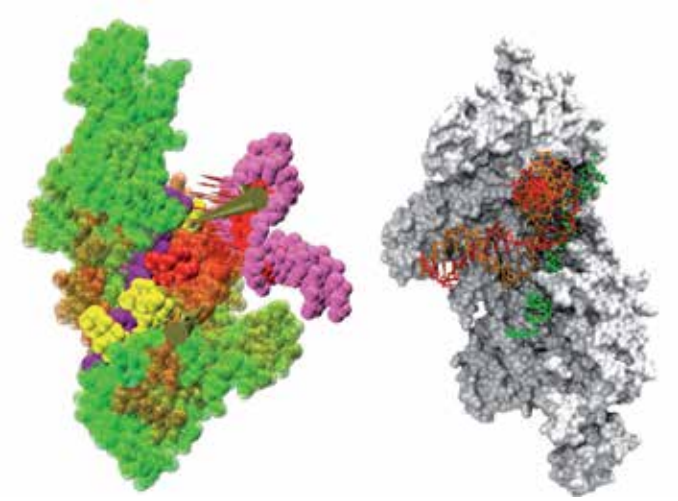


Fig. 16. Visual summary of the interactive experiments. On the left, interactive and flexible low-resolution docking of dsDNA to the RecA/ssDNA complex, using BioSpring. The two trackers enable the user to move at the same time (i) protein L2 loops (yellow or red, if selected) and (ii) dsDNA (pink and red, for selected nucleobases). Static fragments of RecA protein and ssDNA are shown in green/brown and purple spheres, respectively. On the right, all-atom model obtained after reconstruction from one of the best BioSpring prediction.

varied backgrounds to assemble professional quality visuals and to perform computational experiments with relative ease. The hybrid provides:

- high quality rendering with shadows, global illumination, ambient occlusion, etc
- intuitive GUI workflows that help users set up animations ranging from easy turntable rotations to sophisticated mechanism-of-action movies
- mesoscale modeling that allows users to illustrate or animate complex cell events in molecular detail by positioning objects with intuitive controls
- a common Python Platform that allows users to initiate sophisticated algorithms like molecular dynamics or docking energy calculations on the fly and to interoperate these algorithms with each other and with the host

The Interactive Molecular Driver [Stone et al. (2001)] and the callback action from Modeller [Eswar & Sali (2008)] enable real-time interactive molecular simulations with additional forces provided by the user. This interactive steering can operate at different levels, from selected atoms or residues, to selected curve points associated with molecular backbones. Mouse gestures and animated key frames can transmit forces or new coordinates to the simulation calculator that is linked to the host GUI via ePMV. Sophisticated host algorithms like inverse kinematics and efficient collision detection algorithms can operate on the same data as well. With this setup, a ligand can be hand-guided into a binding site with real-time docking scores provided by the Python modules of Autodock [Huey et al. (2007)]. Host-provided physics shortcuts (e.g., soft-body springs for bonds) enable interactive flexible docking with real-time scoring. At the cutting edge of molecular Augmented Reality, a user can interact with data via handheld markers tracked by a camera [Gillet et al. (2005)] to perform an interactive Rigid-body docking with intuitive midair hand gestures (see Figure 17 and http://epmv.scripps.edu/videos/structure2011).

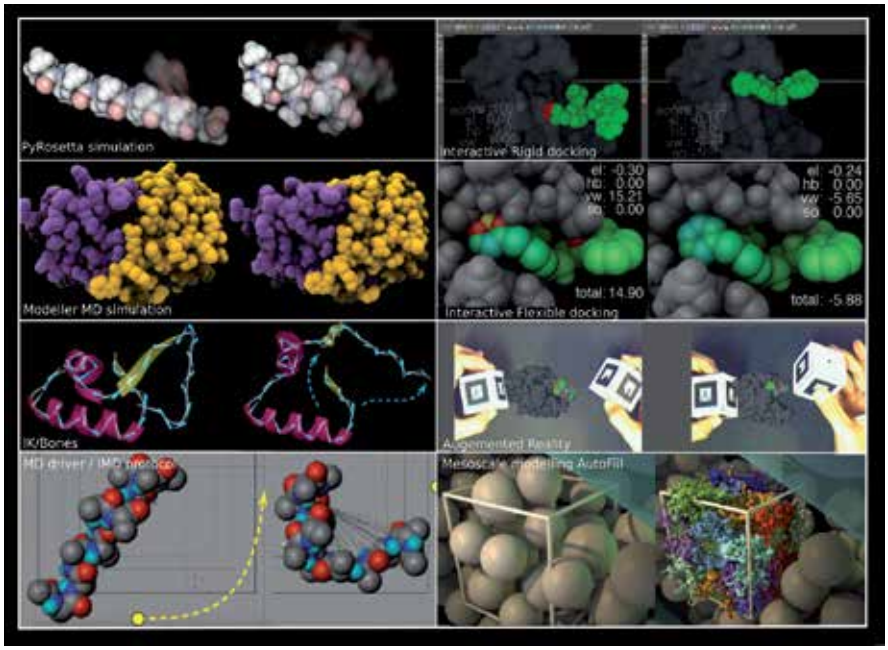


Fig. 17. ePMV features through 8 examples

6.5 FvNano: A virtual laboratory to manipulate and visualize molecular systems

The main goal of FvNano is to provide an easy to use program to manipulate molecular structures in "real time" on regular or high-performance computing (HPC) platforms. The idea is to combine molecular dynamics (MD) software with modern human-computer interaction (HCI) peripherals and GPU rendering. As previously told, combining MD with user interaction is crucial for a better understanding of the molecular motions inside a protein structure when a particular solvent is used or with an important number of active compounds. MD simulations require a lot of computing power. Hence, the ability to use high-performance computing platforms is mandatory when studying complex macromolecular systems. However, the difference between regular and massively parallel architectures can make the program hard to optimize for both platforms. This is solved by using a modular architecture based on the Flow-VR middleware **?**, http://flowvr.sourceforge.net/. In that case, MD simulation, interaction and visualization can be represented as modular blocks linked together by Flow-VR. Each of these blocks can then run on single or multiple threads according to the user's choice. Manipulating objects in a 3D environment with a 2D screen can be challenging, for that purpose, FvNano currently implements two types of HCI peripherals: SpaceBalls and haptic arms. SpaceBalls are used to move the viewpoint and haptic arms to manipulate the molecular structures with force feedback support. The visualization part is also modular, as of now two renderers are available: VMD and OpenGL. The OpenGL renderer uses the HyperBalls GPU shaders previously described. FvNano can also be used to visualize molecular trajectories computed by MD softwares with a simple user interface inspired by video players. Related work within the VMD software has recently been discussed [Stone et al. (2001)].

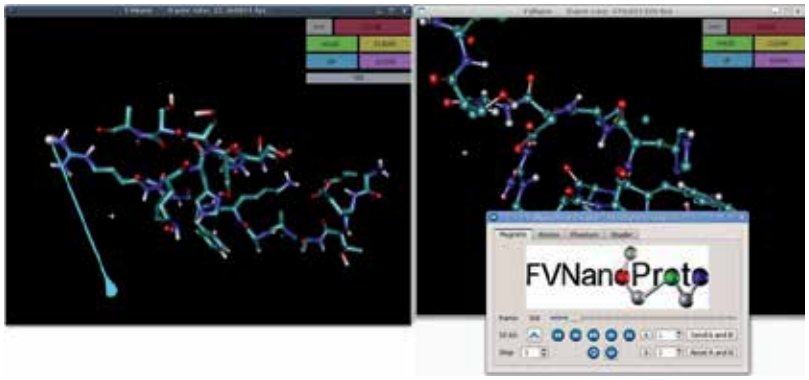


Fig. 18. Screenshot of the interactive molecular dynamics application (left). The cyan cone is the haptic arm avatar and the line shows the atom movement in progress. On the right a screenshot of the molecular trajectory reader.

7. Conclusion

Protein interactions are now routinely studied via computer simulation to understand aspects that cannot currently be studied by experiments. Recently, the Fold It! project [Cooper et al. (2010)] - a 3D-puzzle desktop game, in which the user's task is to fold proteins interactively and without any knowledge prerequisites - showed that using interactivity and insight of human minds can lead to more accurate results than pure computation. Removing false-positive results is done implicitly by users that intuitively avoid erroneous ways of molecule assembly using their experience and logic mind. More generally, molecular simulations can now benefit from this approach to reduce computation and analysis time.

In this document, we presented an interactive approach to assist scientists in their study of protein-protein docking phenomena using some advanced interaction and rendering features offered by a Virtual Reality or advanced human-computer-interaction environment. In such a context, it is important to take into account existing practices of domain experts used in their everyday work. By formalizing user needs and tasks in order to propose a limited set of design principles leading towards an appropriate tool such practices can be further improved, whilst leaving some room for them to evolve in new directions.

Through different examples, we have seen in this chapter that this goal requires efforts from many scientific domains. Experimental biologists describe the needs and validate the results that bioinformaticians extract from the analysis of simulations. Computer science experts are needed to provide efficient codes and graphics. Also, cognitive science helps to design suitable interaction paradigms and user interfaces. Interaction can be used in many applications, from rigid-body docking to accurate atomistic simulations, allowing the user to obtain a wide range of results. The novelty of our approach is that it strives to ensure continuous user participation in the process through direct manipulation of the protein models. In proposing such an approach in which users are involved both upstream and downstream from automatic docking procedures, we hope to maximize the use of their expertise. Hence, the interactive approach efficiently introduces a human element in the process and benefits from the user's experience and insight. The resemblance of this kind of applications with video games should not delude scientists to underestimate the scientific value of such techniques.

8. References

Anderson, J., Lorenz, C. & Travesset, A. (2008). General purpose molecular dynamics simulations fully implemented on graphics processing units, *J. Comp. Phys.* 227: 5342.

André, E. (2000). *Handbook of natural language processing*, chapter The generation of multimedia presentations, pp. 305–327.

Andrusier, N., Nussinov, R. & Wolfson, H. J. (2007). Firedock: fast interaction refinement in molecular docking, *Proteins* 69(1): 139–59.

Baaden, M. & Lavery, R. (2007). There's plenty of room in the middle: Multiscale Modelling of Biological Systems, *Recent Advances in Structural Bioinformatics* pp. 173–196.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome, *Proceedings of the National Academy of Science*, Vol. 98, pp. 10037–10041.

Barass, S. & Zehner, B. (2000). Responsive sonification of well-logs, *Proceedings of the International Conference on Auditory Display (ICAD'00)*.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). The protein data bank, *Nucleic Acids Research* 1(28): 235–242.

Borrelli, K., Vitalis, A., Raul Alcantara, R. & Guallar, V. (2005). Pele: Protein energy landscape exploration. a novel monte carlo based technique, *Journal of Chemical Theory and Computation* 6(1): 1304–1311.

Bourdot, P. & Touraine, D. (2002). Polyvalent display framework to control virtual navigations by 6DoF tracking, *Proceedings of the IEEE Virtual Reality International Conference (IEEE-VR'02)*, pp. 277–278.

Chavent, M., Lévy, B., Krone, M., Bidmon, K., Nominé, J., Ertl, T. & Baaden, M. (2011). Gpu-powered tools boost molecular visualization, *Briefings in Bioinformatics* .

Chavent, M., Vanel, A., Tek, A., Lévy, B., Robert, S., Raffin, B. & Baaden, M. (2011). Gpu-accelerated atom and dynamic bond visualization using hyperballs: A unified algorithm for balls, sticks, and hyperboloids, *Journal of Computational Chemistry* 32(13): 2924–2935.

Comeau, S., Gatchell, W., Vajda, S. & Camacho, C. (2004). Cluspro: an automated docking and discrimination method for the prediction of protein complexes, *Bioinformatics* 20(1): 45–50.

Connolly, M. (1983). Analytical molecular surface calculation, *Journal of Applied Crystallography* 16: 548–558.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. & Players, F. (2010). *Nature* 466(7307): 756–60.

Cui, Q. & Bahar, I. (2006).

Delalande, O., Ferey, N., Grasseau, G. & Baaden, M. (2009). Complex molecular assemblies at hand via interactive molecular simulations, *Journal of Computationnal Chemistry* 30(15): 2375–2387.

Delalande, O., Férey, N., Laurent, B., Gueroult, M., Hartmann, B. & Baaden, M. (2010). Multi-resolution approach for interactively locating functionally linked ion binding sites by steering small molecules into electrostatic potential maps using a haptic device, *Pacific Symposium on Biocomputing*, pp. 205–215.

Dominjon, L., Lécuyer, A., Burkhardt, J., Andrade-Barroso, G. & Richir, S. (2005). The "Bubble" Technique: Interacting with Large Virtual Environments Using Haptic Devices with Limited Workspace, *Proceedings of the World Haptics Conference (joint Eurohaptics Conference and Haptics Symposium)*.

Eswar, N., E.-D. W. B. S. M. & Sali, A. (2008). Protein structure modeling with modeller, *Methods Mol. Biol.* 426: 145?159.

Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2003). Icm-disco docking by global energy optimization with fully flexible side-chains, *Cambridge, MA: Bradford Books / MIT Press* 1(52): 113–117.

Gillet, A., Sanner, M., Stoffler, D. & Olson, A. (2005). Tangible interfaces for structural molecular biology, *Structure* 13(3): 483–91.

Gottschalk, S., Lin, M. & Manocha, D. (1996). Obbtree: A hierarchical structure for rapid interference detection, *Proceedings of the 23rd Conference on Computer graphics and interactive techniques*, Vol. 30, pp. 171–180.

Grayson, P., Tajkhorshid, E. & Schulten, K. (n.d.).

Hermann, T. & Ritter, H. (1999). *Listen to your Data: Model-Based Sonification for Data Analysis*, pp. 189–194.

Hess, B., Kutzner, C. Vanderspoel, D. & Lindahl, E. (2008). Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation., *Journal of Chemical Theory and Computation* 4(3): 435–447.

Hinsen, K. (2000). The molecular modeling toolkit: A new approach to molecular simulation, *Journal of Computational Chemistry* 21: 79–85.

Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation, *Journal of computational chemistry* 28(6): 1145–52.

Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D. & Schulten, K. (2001). Steered molecular dynamics investigations of protein function, *J. Mol. Graph.* 19: 13–25.

Johnson, D. & Willemsen, P. (2003). Six degree-of-freedom haptic rendering of complex polygonal models, *Proceedings of the 11th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS'03)*.

Johnson, G. & Autin, L., G. D. S. M. O. A. (2011). ePMV Embeds Molecular Modeling into Professional Animation Software Environments, *Structure* 19: 293–303.

Kitagawa, M., Dokko, D., Okamura, A. & Yuh, D. (2005). Effect of sensory substitution on suture-manipulation forces for robotic surgical systems, *Journal of Thoracic and Cardiovascular Surgery* 129(1): 151–158.

Knoll, P. & Mirzaei, S. (2003). Development of an interactive molecular dynamics simulation software package, *Rev. Sci. Instrum.* 74: 2483–2487.

LaViola, J. & Katzourin, M. (2007). An exploration of non-isomorphic 3d rotation in surround screen virtual environments, *Proceedings of the IEEE International Conference on 3D User Interfaces (IEEE-3DUI'07)*, pp. 49–54.

Leech, J., Prins, J. F. & Hermans, J. (1996). SMD : visual steering of molecular dynamics for protein design, *Computational Science and Engineering* 3: 38–45.

Leech, J., Prins, J. F. & Hermans, J. (1997). , *Physica A.* 240: 246–254.

Matthey, T., Cickovski, T., Hampton, S., Ko, A., Ma, Q., Nyerges, M., Raeder, T., Slabach, T. & Izaguirre, J. A. (2004). Protomol, an object-oriented framework for prototyping novel algorithms for molecular dynamics, *ACM Trans. Math. Softw.* 30-265(3): 237–265.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing*.

Nelson, M., Humphrey, W., Kufrin, R., Gursoy, A., Dalke, A., Kale, L., Skeel, R. & Schulten, K. (1995). MDscope - a visual computing environment for structural biology, *Comp. Phys. Comm.* 91: 111–133.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. & Schulten, K. (2005). Scalable Molecular Dynamics with NAMD, *Journal of Computational Chemistry* 26: 1781–1802.

Pipe, S. (2008). Recombinant clotting factors, *Thromb Haemost* 99(5): 840–850.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics, *J. Comp. Phys.* 117: 1–19.

Prins, J. F., Hermans, J., Mann, G., Nyland, L. S. & Simons, M. (1999). A virtual environment for steered molecular dynamics., *Fut. Gen. Comp. Sys.* 15: 485–495.

Richard, P., Chamaret, D., Inglese, F.-X., Lucidarme, P. & Ferrier, J.-L. (2006). Human-scale haptic virtual environment for product design: Effect of sensory substitution, *International Journal of Virtual Reality* 5(2): 37–44.

Ritchie, D. (2003). Evaluation of protein docking predictions using hex 3.1 in capri rounds 1 and 2, *Proteins* 52(1): 98–106.

Saladin, A., Amourda, C., Poulain, P., Ferey, N., Baaden, M., Zacharias, M., Delalande, O. & Prevost, C. (2010). Modeling the early stage of dna sequence recognition within reca nucleoprotein filaments, *Nucleic Acid Research* 38(19): 6313–6323.

Sanner, M. F. (1999). Python: a programming language for software integration and development, *J. Mol. Graph. Model.* 17(1): 57–61.

Sanner, M., Olson, A. & Spehner, J.-C. (1996). Reduced surface: An efficient way to compute molecular surfaces, *Biopolymers* 38: 305–320.

Seeger, A. & Chen, J. (1997). Controlling force feedback over a network, *Proceedings of the Second PHANToM User's Group Workshop*.

Shoemake, K. (1985). Animating rotation with quaternion curves, *Proceedings of the 12th annual conference on Computer graphics and interactive techniques (SIGGRAPH'85)*, pp. 245–254.

Stam, J. (1999). Stable fluids, *In SIGGRAPH 99 Conference Proceedings, Annual Conference Series* 38: 121–128.

Stone, J. E., Gullingsrud, J. & Schulten, K. (2001). A system for interactive molecular dynamics simulation, *Proceedings of Interactive 3D Graphics* pp. 191–194.

Surles, M. C., Richardson, J. S., Richardson, D. C. & Brooks-JR., F. P. (1994). Sculpting proteins interactively : Continual energy minimization embedded in a graphical modelling system, *Protein Science* 3: 198–210.

Villoutreix, B., Bastard, K., Sperandio, O., Fahraeus, R., Poyet, J., Calvo, F., Deprez, B. & Miteva, M. (2008). In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics, *Current Pharmaceutical Biotechnology* 9(2): 103–22.

Vormoor, O. (2001). Quick and easy interactive molecular dynamics using Java3D, *Comp. Sci. Eng.* 3: 98–104.

Walker, B. N. & Lane, D. M. (1994). *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Westview Press.

Wang, R., Lu, Y. & Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking, *Journal of Medicinal Chemistry* 46: 2287–2303.

Wu, X. & Wang, S. (2002). Direct observation of the folding and unfolding of b-hairpin in explicit water through computer simulation, *J. Am. Chem. Soc.* 124: 5282–5283.

Zacharias, M. (2005). Attract: protein-protein docking in capri using a reduced protein model., *Proteins* 60(2): 252–6.

3

Protein Interactome and Its Application to Protein Function Prediction

Woojin Jung[1], Hyun-Hwan Jeong[1,2], and KiYoung Lee[1]
[1]*Department of Biomedical Informatics, Ajou University School of Medicine*
[2]*Department of Computer Engineering, Ajou University*
Republic of Korea

1. Introduction

Diverse molecules interact with proteins to produce a biological function. Proteins exhibit many interactions with other molecules including other proteins, nucleic acids, carbohydrates, lipids, minerals, metabolites, and chemical compounds, resulting in diverse roles within and/or between cells. Some of these proteins locate in subcellular organelles, where they modulate biochemical reactions, and some other proteins locate in membranes mediating various stimuli to signaling pathways. Cellular systems can be represented as complex networks. We may consider the molecules as nodes and the associations among the molecules as edges in the network. In this network, all kinds of the molecular interactions can be referred to as an interactome. Even though all kinds of the interactome are important, we here focus on protein-protein interactions (PPIs) since they are fundamental in cellular systems. To function correctly, a protein should interact with other proteins in the context of complex formation, signalling pathways and biochemical reactions. To perform a specific biological function, these interactions need to be specifically formed with proper interacting partners at the right time and locations.

Given the knowledge of genome sequencing on model organisms including human, we have elucidated a large number of unknown molecular structures and interactions within nucleic acids. In the post-genomic era, functional genomics is an emerging area of research that seeks to annotate every bit of information of the genome structure with relevant biological function. Still, many proteins (or genes) remain functionally unannotated (Apweiler et al, 2004; Sharan et al, 2007). These missing links between structures and functions need to be resolved to understand complex biological phenomena including human diseases, development and aging.

Protein function is widely defined in several different ways. It is highly context- and condition-dependent, which means that proteins participate in most biological processes. There have been various attempts to categorize the protein functions (Bork et al, 1998). One of them categorized the protein function into three parts: molecular function, cellular function and phenotypic function. First, the molecular function is defined as biochemical reactions performed by proteins. Second, the cellular function is defined as various pathways associated with proteins. Lastly, the phenotypic function is defined as an integration of all physiological subsystems to environmental stimuli.

Aside from the conceptual definition, many annotation efforts on protein function have been undertaken (Table 1). One of these efforts, the Gene Ontology (GO) consortium (Ashburner et al, 2000), made a standard and multi-labelled hierarchical annotation on proteins in the category of biological process, molecular function and cellular component. The GO consortium is regularly accumulating annotations on proteins according to GO category in open databases. In this chapter, we consider the three kinds of GO terms in annotation of protein function.

Many experimental techniques are available for discovering the protein function, such as gene knockout and transcript knockdown, but these approaches are low-throughput and time-consuming. In recent decades novel high-throughput techniques have been developed, and we are now able to analysis genome-wide data, which is broadening our biological insights. Computational methods are necessary for analysing the massive quantity of data and they are complementary with the low- and high-throughput experimental methods.

In this chapter, we first introduce PPI data available through public databases and compare the contents of major databases. We also describe PPI detection methods by experimental and computational approaches. Next, network- and non-network-based computational methods for the identification of protein function are described. Finally, computational prediction methods of protein subcellular localization, especially by exploiting PPI data, are shown.

Databases	Description
GO	The Gene Ontology project/consortium
COGs	Clusters of Orthologous Groups of proteins
ENZYME	A repository of information relative to the nomenclature of enzymes
Pfam	A database of protein families that includes their annotations and multiple sequence alignments
PROSITE	Database of protein domains, families and functional sites
HAMAP	High-quality Automated and Manual Annotation of microbial Proteomes
UniProt	The Universal Protein Resource
FunCat	MIPS (Munich Information Center for Protein Sequences) Functional Catalogue
DAVID	The Database for Annotation, Visualization and Integrated Discovery
FANTOM	A database for functional annotation of the mammalian genome
ANNOVAR	Functional annotation of genetic variants from high-throughput sequencing data
EFICAz	A genome-wide enzyme function annotation database
KEGG	Kyoto Encyclopedia of Genes and Genomes

Table 1. Databases for the functional annotation of genes and proteins.

2. PPI data

PPI can be considered as one kind of protein interactome. Proteins mutually interact in the biological context for specific functions. Given the knowledge of a single gene, expressing

distinct transcripts and protein isoforms, a protein also interacts with other proteins including itself to give specific function. PPIs are defined as physical interactions between protein pairs (Bonetta, 2010). There are also non-physical interactions such as genetic and functional interactions. Genetic interaction is typically defined as when two genes are simultaneously perturbed, with the quantitative phenotype being more or less than expected (Mani et al, 2008). Functional interaction between two proteins is a much broader concept than other experiment-derived interactions. It may include any functionally associated gene/protein pairs which are integrated and predicted from heterogeneous data. We will explain these computational prediction methods later in this section.

The physical interactions between protein pairs also can be either direct or indirect. Binary interaction is an example of a direct interaction while indirect interaction includes subunits of protein complex. To give a specific function, proteins often form a large complex including direct and indirect interaction among the participant proteins. These interactions are also separable according to their binding lifetime. Some interactions between protein pairs are transient, with the interactions associating and dissociating under particular physiological conditions. On the other hand, some of proteins form stable complexes where the participants in the complexes permanently interact with each other. Various PPI types are defined in standard and annotated across many PPI databases (Cote et al, 2010; Kerrien et al, 2007).

2.1 PPI databases

Currently, there are 132 PPI databases indexed by the Pathguide (Bader et al, 2006; accessed 23 Dec 2011). The quantity of physical interactions to date is 386,495 across all species when integrated among major 11 databases by the iRefWeb (Turner et al, 2010; accessed 23 Dec 2011). The PPI data derived from both high- and low-throughput experiments are altogether deposited into any of primary databases which manually curate experimental results. These primary databases include not only physical interactions but also genetic interactions and annotate standard minimal information about a molecular interaction (MIMIx) (Orchard et al, 2007). There is an inconsistency problem related to the literature curation across different databases (Turinsky et al, 2010). Turinsky et al. confirmed that the agreement between curated interactions from 15,471 papers shared across nine databases was only 42% for interactions and 62% for proteins. This result was averaged between any two databases curated from the same publication. Some of the primary databases altogether formed a consortium called IMEx (The International Molecular Exchange) to enhance the quality of literature curation efforts.

Since we have plenty of primary databases, comprehensive integration of those primary databases has become an intriguing research field. Such meta-databases minimize redundancy and inconsistency that are limitations of the primary databases (Turinsky et al, 2010). Moreover, functional interaction databases consist of both experimentally-detected and computationally-predicted data. Sometimes, these predicted and experimental PPIs need to be distinguished for the degree of confidence. They both give useful information but should be separated according to the relevant evidence codes. There are also species-specific functional interaction databases (Lee et al, 2011; Lee et al, 2010a).

Type	Name	Description	URL
Primary databases	BioGRID	Physical and genetic interaction	http://thebiogrid.org
	MINT	Physical interaction	http://mint.bio.uniroma2.it
	IntAct	Physical interaction	http://www.ebi.ac.uk/intact
	DIP	Physical interaction	http://dip.doe-mbi.ucla.edu
	BIND	Physical and genetic interaction	http://bond.unleashedinformatics.com
	Phospho-POINT	A human kinase interactome resource	http://kinase.bioinformatics.tw
	PIG	Host-Pathogen interactome	http://pig.vbi.vt.edu
	SPIKE	A database of highly curated human signaling pathways	http://www.cs.tau.ac.il/~spike
	MPPI	The MIPS mammalian PPI database	http://mips.helmholtz-muenchen.de/proj/ppi
	HPRD	Human physical interaction	http://www.hprd.org
	CORUM	Mammalian protein complexes	http://mips.helmholtz-muenchen.de /proj/corum
	APID	Agile Protein Interaction DataAnalyzer	http://bioinfow.dep.usal.es/apid
	MiMi	Michigan Molecular Interactions	http://mimi.ncibi.org
	UniHI	Unified Human Interactome	http://www.mdc-berlin.de/unihi
Meta-databases	iRefWeb	Interaction Reference Index	http://wodaklab.org/iRefWeb
	DASMI	Distributed Annotation System for Molecular Interactions	http://dasmi.de/dasmiweb.php
	HIPPIE	Human Integrated Protein-Protein Interaction rEference	http://cbdm.mdc-berlin.de/tools/hippie
	HAPPI	Human Annotated and Predicted Protein Interaction database	http://bio.informatics.iupui.edu/HAPPI
Functional databases	STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org
	Gene-MANIA	Multiple Association Network Integration Algorithm	http://genemania.org
	Functional-Net	Species-specific functional gene networks	http://www.functionalnet.org

Table 2. List of PPI databases.

Contents	BIND	BioGRID	DIP	HPRD	IntAct	MINT	MPPI
Biological role (PSI-MI)					O	O	
Experimental role (PSI-MI)					O	O	
Taxonomy ID					O	O	
Interaction category					O	O	
Interaction title							
Interaction type (Text)		O					
Interaction type (PSI-MI)			O		O	O	
Interactor type (peptide, protein)					O	O	
Detection method (Text)				O			O
Detection method (PSI-MI)		O	O		O	O	
Evidence (PMID or doi-number)						O	
PubMed ID	O	O	O	O	O		O
BioGRID ID		O					
HPRD ID				O			
NCBI Gene ID	O	O		O			
Protein ID					O	O	O
ID type					O	O	O
Protein accession number	O			O			
UniProt ID			O				
Link to source ID	O	O	O			O	
Description	O				O		
Confidence score		O				O	

Table 3. Contents of primary PPI databases. Available contents are colored in grey with "O" shape.

We have listed some of the major primary databases, meta-databases, and functional databases in Table 2. Comparisons among the primary databases are shown in Table 3. We compared various features including interaction types, detection methods, references, and biological and experimental roles. This information would be valuable for researchers when they need to select and integrate various PPI data bases.

2.2 Methods for PPI identification

There are two major ways to determine PPIs. One is an experimental detection and the other is computational prediction. The former method is more reliable and well-established in both small and large scales while the latter method is based on the characteristics of accumulated protein interactions. In this section, we will briefly describe both approaches.

2.2.1 Experimental detection methods for PPIs

Experimental detection of interactions between protein pairs is achieved by various methods. Here, we describe only two representative methods: yeast two-hybrid (Y2H) (Suter et al, 2008) and mass spectrometry (MS) (Berggard et al, 2007). These methods both detect physical PPIs but the type of PPIs is different. As previously stated, direct and binary PPIs are distinct from protein groups in a complex and this type of PPI is detected only by Y2H method. This method uses a transcription factor found in yeast which consists of two other domains. Y2H method relies on an artificial insertion of a protein coding sequence to one of the domains and another protein inserted on the other domain using a plasmid. PPI can be assessed by confirming phenotype of the target gene of the transcription factor. The Y2H method can detect PPIs in large-scale and the sensitivity is high, enabling detection of even weak transient PPIs. But, since the experiment is done only in the nucleus, the real location information of such PPIs is hard to annotate, which obscures the detailed biological interpretation. Moreover, Y2H detects only binary interactions and results in a high rate of false positive, which are noteworthy limitations.

Another method in this category is based on mass-spectrometry (MS). The MS analyzes the mass of molecules rapidly and accurately. If the weight of all proteins in question is known, this information can be linked to the specific protein. This method is powerful when protein co-complexes are examined. Although it cannot provide details on the direct-level of interactions, the grouping of the proteins in a complex can be revealed. For this method, one protein ("bait") and all of interacting partners in a complex are pulled out and separated by electrophoresis. Finally, all the constructs derived from electrophoresis are used for MS. This method yields many false positive results when the sampling strategy is thoroughly different. This sampling might include fake interactions resulting in a high rate of false readings. There are many strategies related to this problem (Bousquet-Dubouch et al, 2011; Gingras et al, 2007). The experimental results obtained with MS-based methods are different from those obtained with binary methods (Y2H). Data derived from co-complex experiments cannot directly assign a binary interpretation. An algorithm is needed to translate group-based observations into pairwise interactions.

2.2.2 Computational prediction methods for PPIs

While recent reviews (Lees et al, 2011; Pitre et al, 2008; Shoemaker & Panchenko, 2007; Skrabanek et al, 2008; Xia et al, 2010) have discussed computational prediction methods for PPIs in details, we here briefly introduce some of approaches that are widely used. Although the amounts of experimental resources of PPIs are growing rapidly, proteome-wide PPIs information is still lacking and mostly limited on several model organisms. Given

wide types of indirect but genome-wide resources, we can enhance our understanding of overall protein interactome. Methods in prediction of direct physical PPIs are less investigated than those of functional association between protein pairs. These functional association methods of PPIs can give information of which protein pairs have same biological process and potential physical interactions.

The first data used in these prediction methods is genomic sequences. Co-occurrence-based methods use assumption that if gene pairs are co-inherited across evolutionary processes (i.e. species), they are considered as functionally associated (Barker & Pagel, 2005; Bowers et al, 2004; Pellegrini et al, 1999). These methods applied to microorganisms and successfully discovered novel participants of known pathway (Carlson et al, 2004; Luttgen et al, 2000). Other similar methods based on this genomic sequence use the information of gene fusion events (Marcotte et al, 1999; Reid et al, 2010; Zhang et al, 2006) and gene neighbourhood (Ferrer et al, 2010; Itoh et al, 1999; Koonin et al, 2001). Another type of data used is amino acid (AA) sequences and the interface of interacting protein pairs are composed of specific AA residues (Tuncbag et al, 2008; Tuncbag et al, 2009). This knowledge is reflected in the co-evolution of specific interface residues between interacting proteins and by alignments of multiple sequences, the results are highly correlated with physical PPIs (Pazos et al, 2005). Commonly occurring domain pairs are also considered in this context (Eddy, 2009; Finn et al, 2010; Stein et al, 2009; Yeats et al, 2011) and simple AA sequence such as 3-mers of interacting residues can be used (Ben-Hur & Noble, 2005). Another well-known information is homology of PPIs across different species. Methods on this information simply find PPIs which are conserved across species, called interologs (Matthews et al, 2001). Here, any known PPIs regarded as query to find conserved interactions across species using an ortholog database. There are many algorithms which follow this approach (Kemmer et al, 2005; Persico et al, 2005). Aside from the sequence-level data, structural information is also a valuable resource to predict PPIs, especially a protein 3D structure. (Aloy & Russell, 2003; Ezkurdia et al, 2009; Hosur et al, 2011; Shoemaker et al, 2010; Singh et al, 2010; Zhang et al, 2010). A huge amount of genome-wide gene expression profiles are another useful data to predict PPIs and they are investigated to define gene co-expression patterns of any pairs and consider higher correlation degree as higher probability of PPIs (Grigoriev, 2001; Lukk et al, 2010; Stuart et al, 2003). As shown in the earlier section, there are many literature-curated PPI databases. While those approaches are based on the manual inspection, such PPIs information can be automatically extracted using a text-mining algorithm (Blaschke et al, 2001; Szklarczyk et al, 2011; Tikk et al, 2010).

3. Computational prediction methods for protein function

Even before the prevalence of genome sequencing technologies, typical experimental identification on a protein function has been executed. Such identification has focused on a specific target gene or protein, or a small set of protein complexes. Gene knockout, knockdown of gene expression, and targeted mutations are some methods for protein function identification (Recillas-Targa, 2006; Skarnes et al, 2011). Such low-throughput experiments were replaced by high-throughput experiments including genome sequencing and determination of the protein interactome. Computational methods followed by massively archived data have been developed for better analysis. Based on the assumption that structural

similarity correlates with functional similarity, homology-based functional annotation across organisms has now become a trivial approach (Aloy et al, 2001; Gaudet et al, 2011).

3.1 Non-network based approaches

Classical computational methods use features from only a single protein in prediction of protein function (Bork et al, 1998). These approaches use a set of features like amino acid sequences, genome sequences, protein structures (2D and 3D), phylogenetic data, and gene expression data. PSI-BLAST (Altschul et al, 1997) and FASTA (Mount, 2007) are popular sequence alignment tools used to reveal homologous proteins between known and unknown (query) proteins. Proteins with similar sequences are assumed to have similar functions. Moreover, protein folding patterns are also preserved enough to identify homologs (Huynen et al, 1998; Sanchez-Chapado et al, 1997). The comparative genomics across different species is a powerful approach for analysing functional annotation of proteins. In fact, it has been suggested that correlation of sequence-structure is much stronger than that of sequence-function (Smith et al, 2000; Whisstock & Lesk, 2003). So many approaches take the sequence to structure to function route for protein function prediction (Fetrow & Skolnick, 1998).

Likewise, these data are showing only single aspect of functional features conserved during evolution. Data derived from different sources can be inter-connected it should be integrated to analyse simultaneously (Kemmeren & Holstege, 2003). We next show that PPI networks potentially enrich functional relationship between protein pairs that may not be detectable from other genomic data such as primary or higher level sequence structure.

3.2 Network-based approaches

As we mentioned in the Introduction, biological function is never achieved by a single protein. Rather, proteins dynamically interact with each other and the interacting partners adopt similar performances for specific functions. With a plethora of data being generated by high-throughput proteomic experiments, it has become possible to use proteome-wide PPI patterns in protein function prediction. Among a broad type of protein interactome, a PPI network generates well-known data that is invaluable in prediction of protein function. It is possible to annotate the function of undefined proteins according to its neighbours that are functionally annotated. This assumption is based on simple idea called "guilt-by-association", and we consider an association by possible physical interaction in any condition and, sometimes, functional association are given with relevant evidence score.

Here, we review the general network-based approaches in predicting protein functions. These approaches are categorized into two methods for better description. The first one is a straightforward method of inferring protein function based on the topological structure of a PPI network. The other method first identifies distinct sub-networks from a whole PPI network. These sub-networks are also referred to as functional modules since they perform specific biological functions such as protein complexes, and metabolic and signalling pathways. Functional modules are detected by a broad variety of clustering

algorithms and, thereafter, each module is annotated with appropriate functional association. In this section, basic concepts and pioneering studies on this corresponding approaches are introduced.

3.2.1 Direct annotation of protein function using PPI network

3.2.1.1 Neighbourhood approaches

Direct functional annotation considers the correlation of the network distance between two proteins, which means the closer the two proteins are in the network the more similar are their functions. One of the earliest studies extrapolated only adjacent neighbours within an entire PPI network. This simple approach used information of the immediate neighbourhood and took the most common functions up to three among its neighbours. In spite of the effectiveness, accuracy was achieved by 72% (Schwikowski et al, 2000). However, this method lacked significance values for each association and the full network topology was not considered in the annotation process. A strategy was proposed to tackle the first problem of assigning statistical significance (Hishigaki et al, 2001). This was done by using χ^2 .-like scores and, instead of using the immediate neighbours, the *n*-neighbourhood of a protein that consists of proteins with distance of *k*-links to the protein is considered. Simply put, the neighbours of adjacent neighbours are taken into account with the frequencies of all the distance of in this neighbourhood. For an unknown protein, the functional enrichment in its *n*-neighbourhood in identified with χ^2 test, and the top ranking functions are assigned to the unknown protein. In another approach, the shared neighbourhood of a pair of proteins are considered besides from the neighbourhood of the protein of interest. Chua et al. investigated the correlation between functional similarity and network distance (Chua et al, 2006). They developed a functional similarity score, called the FS-weight measure, which gives different weights to proteins depending on their network distance from the query protein. This approach showed higher accuracy when employing indirect interactions and its functional association.

3.2.1.2 Global optimization approaches

Although the neighbourhood approach is very attractive and effective by its simplicity, shortcomings arise when there is not enough number of protein neighbours and sufficiently annotated proteins. To overcome this issue, several approaches that utilize the entire topology of the network have been proposed. These global approaches attempt to optimize annotation of function-unknown protein using the topology of a whole network. One of the first studies that took this approach used the theory of Markov random fields, which determines the probability of a protein having a certain function (Deng et al, 2004). This theory is then used to determine the joint probability of the whole interaction network regarding to a certain function. This formulation is transformed to that of the conditional probability of a protein having a certain function given the annotations of its interaction partners. After that, the Gibbs sampling technique is iteratively applied to determine the stable values of this probability for each protein. This approach resulted in higher performance than those of neighbourhood-based approaches (Chua et al, 2006; Hishigaki et al, 2001; Schwikowski et al, 2000) when utilized to the yeast PPI data.

Additional attempts according to this approach had been followed. Here, the objective function is defined for the whole network, which is a sum of the following variables (Vazquez et al, 2003).

1. The number of neighbours of a protein having the same function as itself.
2. The number of neighbours of a protein having the function under consideration.

Thus, this function estimates the number of pairs of interacting proteins with no common functional annotation. Since a high value of this function is biologically undesirable, it is minimized using a simulated annealing procedure. As expected, this approach outperformed the majority rule-based strategy on the *Saccharomyces cerevisiae* interaction data (Schwikowski et al, 2000), since the latter tried to optimize only the second factor above. An additional advantage of this approach was that multiple annotations of all proteins were obtained in one shot, unlike earlier approaches which ran independent optimization procedures for different functions.

The above discussion shows that a wide variety of approaches based on principles of global optimization have been proposed in the literature and many more are in the pipeline. The most accurate results in the field of function prediction from PPI networks have also been achieved by these approaches, which is intuitively acceptable since they extract the maximum benefit from the knowledge of the structure of the entire network.

3.2.2 Indirect annotation of protein function

This approach uses a protein interaction network, not directly for annotation, but identifies functional modules first and then assigns functions to unknown proteins based on their membership in the functional modules. This is based on the assumption that most biological networks are organized as distinct sub-networks to give specific functions (Hartwell et al, 1999). We assume that proteins in the same module participate in a similar biological process. Modular patterns and dense regions are found in the PPI network (Gavin et al, 2006).

3.2.2.1 Distance-based clustering approaches

To find biologically significant modules, clustering algorithms can be applied efficiently. Clustering is a popular unsupervised learning algorithm that does not use any prior information about the class label. There are two widely-used ways of clustering: topology-based or distance-based. The key procedure in distance-based clustering is to select the similarity measure between two proteins to detect modules. The distance between two proteins (also called as nodes) in a network is usually defined as the number of interactions (also called as edges) on the shortest path between them. However, there is a serious problem in this hierarchical clustering, known as the 'ties in proximity' problem (Arnau et al, 2005). This means that the distance between many protein pairs are identical.

To solve this problem, a network clustering method was developed to identify modules in the biological network based on the fact that each node has a unique pattern of shortest path lengths to every other node. But for a specific module in the network, the nodes/members of the module shared similar pattern of shortest path lengths (Rives & Galitski, 2003). Another study used the hierarchical clustering method with the shortest path length

between proteins as a distance measure to overcome the 'ties in proximity'. This was achieved by exploiting equally valid hierarchical clustering solution with a random select when ties are met (Arnau et al, 2005). Although many methods in the similarity measures have been proposed, a single validation for such methods is insufficient. For this, two evaluation schema are suggested, which are based on the depth of a hierarchical tree and width of the ordered adjacency matrix (Lu et al, 2004). Furthermore, there are various types of cellular network with distinct modular patterns, and so network-specific methods should be investigated in the future.

3.2.2.2 Graph-based clustering approaches

Dissecting functional modules in a large PPI network is the same problem of graph partitioning and clustering. One of the pioneering method using this network topology-based concept was the MCODE (molecular complex detection algorithm) (Bader & Hogue, 2003). This method predicts complexes in a large PPI network consisting of three processes. First, the nodes of the network are weighted by their core clustering coefficients (the density of the largest *k*-core of its adjacent neighbourhood), and then densely connected modules are identified in a greedy fashion. The use of this coefficient instead of a standard clustering coefficient was proposed, as it increases the weights of densely interconnected graph regions while giving small weights to the less connected nodes. The next step is to filter or add proteins based on the connectivity criteria. This method was applied to large-scale PPI networks and given as a plug-in for the Cytoscape (Kohl et al, 2011).

Another similar study to find complexes and functional modules is based on super paramagnetic clustering. This method used an analogy to the physical features of a heterogeneous ferromagnetic model to detect densely connected clusters in a large graph (Spirin & Mirny, 2003). There is also an algorithm called the restricted neighbourhood search clustering (RNSC), which starts with an initial random cluster assignment and then proceeds by reassigning nodes to maximize the partition's score. Here, the score represents an intra-connectivity in the cluster, not an inter-connectivity across other clusters. The RNSC algorithm is known to perform better than the MCODE algorithm (King et al, 2004). The Markov clustering algorithm (MCL) is another fast and scalable clustering algorithm based on simulation of random walks on the underlying graph (Pereira-Leal et al, 2004). This algorithm has an assumption that a random walker in natural clusters (i.e. dense region of the graph) sparsely goes from one to another natural cluster. Such clusters in a whole graph are structurally identified by the MCL algorithm. It starts by measuring the probabilities of random walks through the graph to build a stochastic "Markov" matrix, by alternating two operations: expansion and inflation. The expansion takes the squared power of the matrix while the inflation takes the Hadamard power of a matrix, followed by a re-scaling. Therefore the resulting matrix is remained as stochastic. Clusters are detected by alternation of expansion and inflation until the graph is partitioned into distinct subsets where no paths between these subsets are available. This algorithm can be efficiently implemented to weighted and large dense graphs. Various PPI networks were applied using the MCL algorithm to find functional modules such as protein complex (Krogan et al, 2006).

It is true that a protein might have multiple functions and this characteristics of a protein leads to overlap of different modules. That means graph partitioning in a strict manner

might not be reasonable for the PPI network. However, most current methods are based on the hard-partition algorithms, meaning that each protein can belong to only one specific module. To handle this limitation, a clustering algorithm based on the information flow was suggested. This algorithm efficiently identified the overlapping clusters in weighted PPI network by integrating semantic similarity between GO function terms (Cho et al, 2007). Since the common proteins in the overlapping modules are interpreted as a connecting bridge across the different modules, biologically significant and functional sub-networks could be identified. Still, there are few clustering methods identifying such overlapping modules. Novel clustering methods for this theme are required with enhancement of prediction accuracy.

4. Prediction of protein subcellular localization

4.1 Introduction

Proteins should move to specific locations after synthesis to work in our body correctly. Thus, knowing subcellular localization of proteins is important to understand their own functions. Unicellular organisms like budding and fission yeasts can find systematic protein localization by experimental studies. However, such studies could not be performed well in higher eukaryotes such as *Caenorhabditis elegans*, *Drosophila melanogaster*, or mammals because of large-scale proteome sizes and technical difficulties associated with protein tagging.

Therefore, bioinformatical approaches to develop efficient methods are required instead of wet experiments. Actually, many computational methods to predict subcellular localization of protein have been proposed over several decades. A considerable number of computational classification methods have been developed for this purpose. Typically these algorithms input list of features and output subcellular localizations of target proteins. The features contain various characteristics of the proteins. Molecular weight, amino acid content and codon bias can be the features. Input features for prediction of subcellular localization can be broadly categorized into four categories: protein sorting signals, empirically correlated characteristics, sequence homology with known answer sets, and other sources (Imai & Nakai, 2010).

During the training phase, in the methods, learning utilizes a set of gold-standard proteins whose localizations are well known. This set consists of the feature vectors. After the training phase, a model is constructed to recognize those features or patterns of features that are useful and then predicts the subcellular localization of proteins whose localization is unknown. Various algorithms have been used to construct a model for prediction of sub-cellular localization.

In the field of bioinformatics, there are several problems to resolve for predicting subcellular localization of proteins. First, there are generally too many classes (localization). According to Huh et al, 22 distinct localizations exist in budding yeast. Next, one protein may have multiple different localizations (Huh et al, 2003). This is referred to a multi-label classification problem and traditional classification algorithms have a limit on handling the multi-label problem well. Another problem is that there may be a higher dimensional feature space for prediction. More than tens of thousands features exist in some cases.

Another issue is that data for each localization is too imbalanced. All these characteristics make the prediction difficult. More importantly, the localization prediction is sometimes difficult to achieve sufficient performance when we use information of single proteins only. Recently, large-scale protein-protein interaction networks have been elucidated in yeast, fly, worm, and human. To interact physically, two proteins should localize to the same or adjacent subcellular localization. That means we can get useful information of a protein from its interacting neighbours. Thus, we can improve the localization prediction performance particularly using PPI networks.

4.2 Computational prediction of protein subcellular localization

4.2.1 Single-protein feature based localization predictions

Table 4 summarizes previous studies that have used the features of single proteins. The studies for prediction of subcellular localization have the following trends. The first is an increase in the number of predicting localizations. At first, Nakashima & Nishikawa predicted localization of a protein that is inter-cellular or extra-cellular using Amino Acid (AA) and Pair coupled Amino Acid (PairAA) (Nakashima & Nishikawa, 1994). After their study, many studies tried to increase the number of distinct localizations to predict. For example, Gardy et al predicted five distinct subcellular localization including 'cytoplasmic', 'inner membrane', 'periplasmic', 'outer membrane' and 'extra-cellular' (Gardy et al, 2003). Nair & Rost predicted ten distinct subcellular localizations (Nair & Rost, 2003). Also, Chou & Cai predicted 22 distinct subcellular localizations that experimentally identified localization of Huh et al. (Chou & Cai, 2003).

The second trend is handling of a multi-label problem. A protein can localize to several sub-cellular locations. However, most of these studies did not consider multiple localization property, but rather assumed that a protein has a single representative localization. Also, the accuracy of prediction is lower when the number of distinct localizations for a protein is increased. Some researchers have been tried to address this issue (Lee et al, 2006).

Another tendency is the development of a classification algorithm for an elaborate and efficient model construction. Least distance algorithm, artificial neural network, a nearest neighbour approach, a Markov model, a Bayesian network approach, and support vector machine (SVM) were used to archive the goal. Some studies mixed several algorithms. Lee et al. developed an algorithm that reflects of property of the prediction task (Lee et al, 2006). They developed an extended Density-induced Support Vector Data Description (D-SVDD) classification algorithm to handle well the issues related to class imbalance, higher dimensionality, multi-label, and many distinct classes. The classical D-SVDD algorithm can handle only one-class classification tasks. Thus, Lee et al. extended it to handle multi-label classification tasks.

4.2.2 Network-based localization prediction

As mentioned earlier, two proteins that localize to same or adjacent subcellular localization have a tendency to interact with each other. That means two proteins can be a tag protein to one other for subcellular localization. Therefore, if a molecular network such as PPIs is available, we may take advantage of the PPI network for the prediction. Several studies

tried to predict subcellular localization using network data. This section consists two parts: first one is a brief explanation of the study by Lee et al. (Lee et al, 2008), which is the cornerstone of the network-based approach for location prediction using PPI network. We describe a methodology to generate of feature vectors for a protein in the aforementioned study and introduce a DC-kNN classifier for the prediction. The second part is a summary of the network-based approaches from the work of Lee et al. to the present.

Author(s)	Method(s)	Feature(s)	# Classes	Multi-label	Imbalanced
(Nakai & Kanehisa, 1991)	Expert Systems	SignalMotif	4	X	X
(Nakai & Kanehisa, 1992)	Expert Systems	AA, SingalMotif	14	X	X
(Nakashima & Nishikawa, 1994)	Scoring System	AA, diAA	2	X	X
(Cedano et al, 1997)	LDA using Mahalanobis distance	AA	5	X	X
(Reinhardt & Hubbard, 1998)	ANN Approach	AA	3, 4	X	X
(Chou & Elrod, 1999)	CDA	AA	12	X	X
(Yuan, 1999)	Markov Model	AA	3, 4	X	X
(Nakai & Horton, 1999)	k-NN approach	SignalMotif	11	X	X
(Emanuelsson et al, 2000)	Neural network	SignalMotif	4	X	X
(Drawid & Gerstein, 2000)	CDA	Gene Expression Pattern	8	X	X
(Drawid & Gerstein, 2000)	Bayesian Approach	SignalMotif, HDEL motif	5, 6	X	X
(Cai et al, 2000)	SVM	AA	12	X	X
(Chou, 2000)	Augumented CDA	AA, SOC factor	5, 7, 12	X	X
(Chou, 2001)	LDA using various distance measures	pseuAA	5, 9, 12	X	X
(Hua & Sun, 2001)	SVM	AA	4	X	X
(Chou & Cai, 2002)	SVM	SBASE-FunD	12	X	X
(Nair & Rost, 2002)	Nearest Neighbor Approach	functional annotation	10	X	X
(Cai et al, 2003)	SVM	SBASE-FunD, pseuAA	5	X	X
(Cai & Chou, 2003)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	3, 4	X	X
(Chou & Cai, 2003)	LDA using various distance measures	pseuAA	14	X	X
(Pan et al, 2003)	Augumented CDA	pseuAA with filler	12	X	X
(Park & Kanehisa, 2003)	SVM	AA, diAA, gapAA	12	X	X
(Zhou & Doctor, 2003)	Covariant discrinant algorithm	AA	4	X	X
(Cai et al, 2003)	SVM	SBASE-FunD, pseuAA	5	X	X

Author(s)	Method(s)	Feature(s)	# Classes	Multi-label	Imbalanced
(Gardy et al, 2003)	SVM, HMM, Baysian	AA, motif, homlogy analysis	5	X	X
(Reczko & Hatzigerrorgiou, 2004)	ANN Approach	AA, SingalMotif	3	X	X
(Huang & Li, 2004)	fuzzy k-NN	diAA	11	X	X
(Cai & Chou, 2004)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	3, 4	X	X
(Chou & Cai, 2005)	Nearest Neighbor Approach	FunDC(5875D), pseuAA	3, 4	X	X
(Bhasin & Raghava, 2004)	SVM	AA, diAA	4	X	X
(Lee et al, 2006)	PLPD	AA, diAA, gapAA, InterProFunD	22	O	O
(Chou & Shen, 2007)	Nearest Neighbor Approach	GO, InterProFunD, pseuAA	22	O	X
(Shatkay et al, 2007)	SVM	SignalMotif, AA, text-based feature	11	X	X
(Garg et al, 2009)	k-NN, PNN	AA, sequence order, physicochemical properties	11	X	X
(Zhu et al, 2009)	SVM	AA, PSSM	14	O	X
(Shen & Burger, 2010)	SVM	AA, groupedAA, gapAA,, GO	4	X	X
(Mei et al, 2011)	SVM	AA, diAA, gapAA, GO	10	O	X
(Wang et al, 2011)	Frequent Pattern Tree	Motif, Overall-sequence	12	X	X
(Mooney et al, 2011)	N-to-1 Neural Network	BLAST	5	X	O
(Tian et al, 2011)	PCA, WSVM	PesAA	20	X	X
(Pierleoni et al, 2011)	SVM	AA, ChemAA, protein length, GO	3	X	X

Table 4. Summary of previous methods for prediction of protein subcellular location.

4.2.2.1 Generation of feature vectors

Lee et al. used three types of feature to predict the localization and integrated these features (Lee et al, 2008). These are single protein features (***S***) and two kinds of network neighbourhood features (***N*** and ***L***).

Seven ***S*** features were based on a protein's primary sequence and its chemical properties. Amino acid composition frequencies (AA), adjacent pair amino acid frequencies (diAA) and pair-wise amino acid frequencies with a gap which is length of 1 (gapAA) from a protein's

primary sequence were used. Also, three kinds of chemical amino acid compositions (chemAA) were generated from normalized hydrophobicity (HPo), hydrophilicity (HPil), or side-chain mass (SCM). Also, they combined these chemical properties into pseudo-amino acid composition (pseuAA), which is another ***S*** feature vector. Occurrences of known signalling motifs in the primary protein sequence (Motif) are also used as one of the ***S*** features. The last ***S*** feature encoded functional annotations of the protein from Gene Ontology (GO) (Ashburner et al, 2000). Figure 1 provides an example.

N network features are summary of ***S*** features from neighbourhood of a protein. Knowledge for neighbours of a protein comes from PPI data, which are pooled from various databases such as BioGRID (Stark et al, 2011), DIP (Salwinski et al, 2004) and SGD (Engel et al, 2010). ***L*** network features are summary of location distribution of interacting neighbours. Figure 2A shows a relationship among the three PPI databases. It shows that a single protein interaction database covers a different part of the whole reported interactions. The diagonal pattern in Figures 2B-D shows that interacting protein pairs share similar localization information. For example, a protein in an "ER to Golgi" tends to interact with other proteins which localized in the "ER to Golgi" more than other localizations.

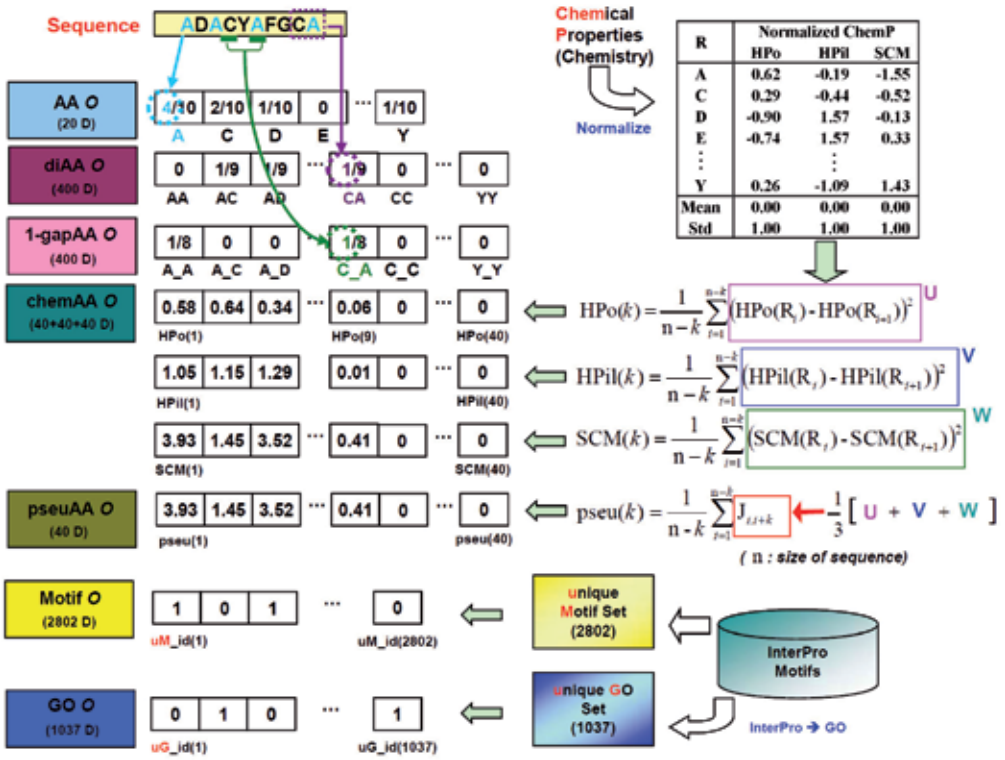


Fig. 1. Summary of feature generation scheme for a single protein (adapted from Lee et al, 2008).

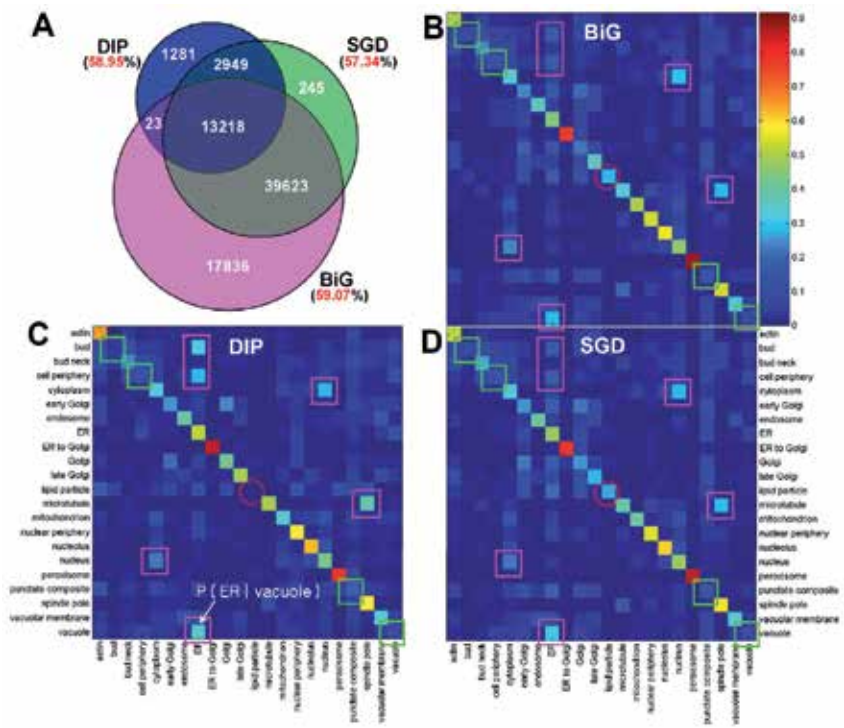


Fig. 2. Correlation between known localizations and protein interactions of yeast proteins. (A) The number of interactions (inside the circles) and the fraction of interactions whose proteins share localization information (outside the circles) of three interaction databases: BiG, DIP and SGD. (B-D) They show that interacting protein pairs have similar localization information in DIP, BiG and SGD (adapted from Lee et al, 2008).

4.2.2.2 Divide-and-Conquer k-Nearest Neighbour (DC-kNN) Classifier

After generating feature vectors, large-scale feature vectors with a high order may generate. A high dimensional feature vectors generally cause some problems like *curse-of-dimensionality*. In other words, data from higher dimensional feature vectors usually require a corresponding amount of inputs and it, sometimes, causes an over-fitting problem to a given dataset (Guyon et al, 2002). Also some feature vectors may be useless in constructing a model for a specific localization. Thus, individual model for different subcellular localizations may require different sets of useful feature sets. Therefore, extraction for feasible feature vectors for individual localizations may be needed to construct robust and reliable prediction models.

To construct a prediction model, Lee et al. proposed a DC-kNN classifier which is a variety of a k-Nearest Neighbours classification algorithm. A DC-kNN classifier tackles high-dimensional features in a divide-and-conquer manner. Briefly mentioning, a DC-kNN has three main steps (Figure 3): dividing, choosing, and synthesizing. In the dividing step, the full feature vector is divided into $\boldsymbol{m}$ meaningful subsets. After the dividing step, the k-nearest neighbours are chosen for each protein and for each subvector. In the synthesizing step, results of kNNs of individual $\boldsymbol{m}$ sets are synthesized to produce confidence scores

using an average of Area under the ROC curve (AUC) for each localization. DC-kNN finds a feasible combination of feature sub-vectors for each label (localization) based on a feature forward selection approach.

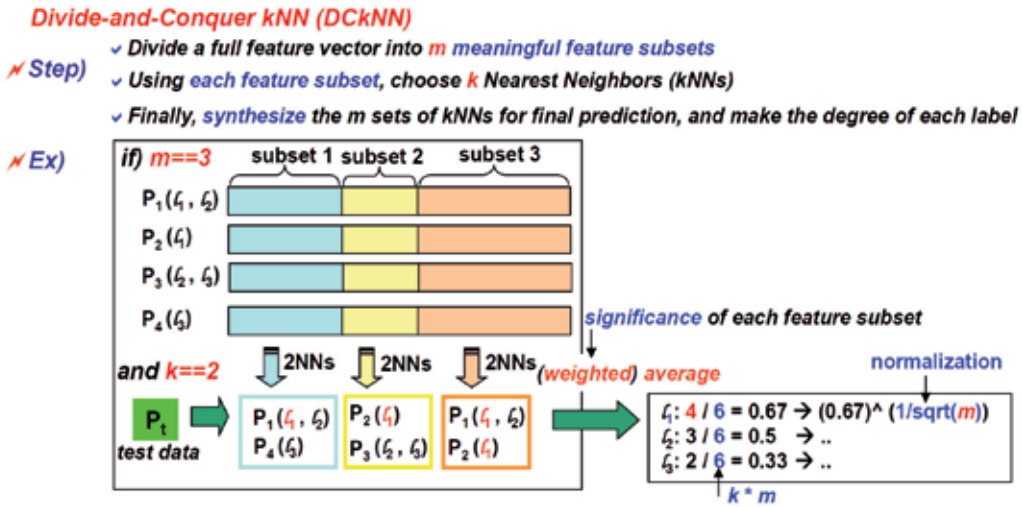


Fig. 3. Brief description of a DC-kNN (adapted from Lee et al, 2008).

4.2.3 Results of location prediction

Lee et al. first compared prediction performance of a DC-kNN for localization prediction with different feature sets: ***S*** features only, ***N*** features only, ***L*** features only, all features together (***S***+***N***+***L***), and random guesses. ***N*** and ***L*** features are generated using DIP (Salwinski et al, 2004). Performance of each case was evaluated by the technique of leave-one-out cross-validation (LOOCV). Proteins of *Saccharomyces cerevisiae* (n=3914) (Huh et al, 2003) were used for the LOOCV. They used three different performance metrics: Top-K, Total, and Balanced. These metrics were used to summarize the results of 3914 LOOCV runs. Top-K measurement considers as correct if at least one of the real localization of a protein is in the top-K predictions. Total measurement counts all the correctly predicted localizations based on the number of real localizations of test data. Balanced measure calculates the averaged fraction of correctly predicted proteins in each localization. As a result, every classifier showed clearly better performance than random guess (Figure 4A), and combination of ***S***, ***N***, and ***L*** features showed the highest performance.

Figures 4A and 4B inform that information of neighbourhood acquired from a PPI database improves prediction performance. However, Figure 4C illustrates that acquiring more information does not always contribute to an improvement of performance. On the contrary, additional information can decrease prediction performance. To find the necessary feature vectors for each localization, Lee et al. used a DC-kNN and found feasible subsets using the prepared feature vectors for individual localizations (Figures 4C and 4D). Using the selected features for individual localizations, the average of the AUC values was 0.94.

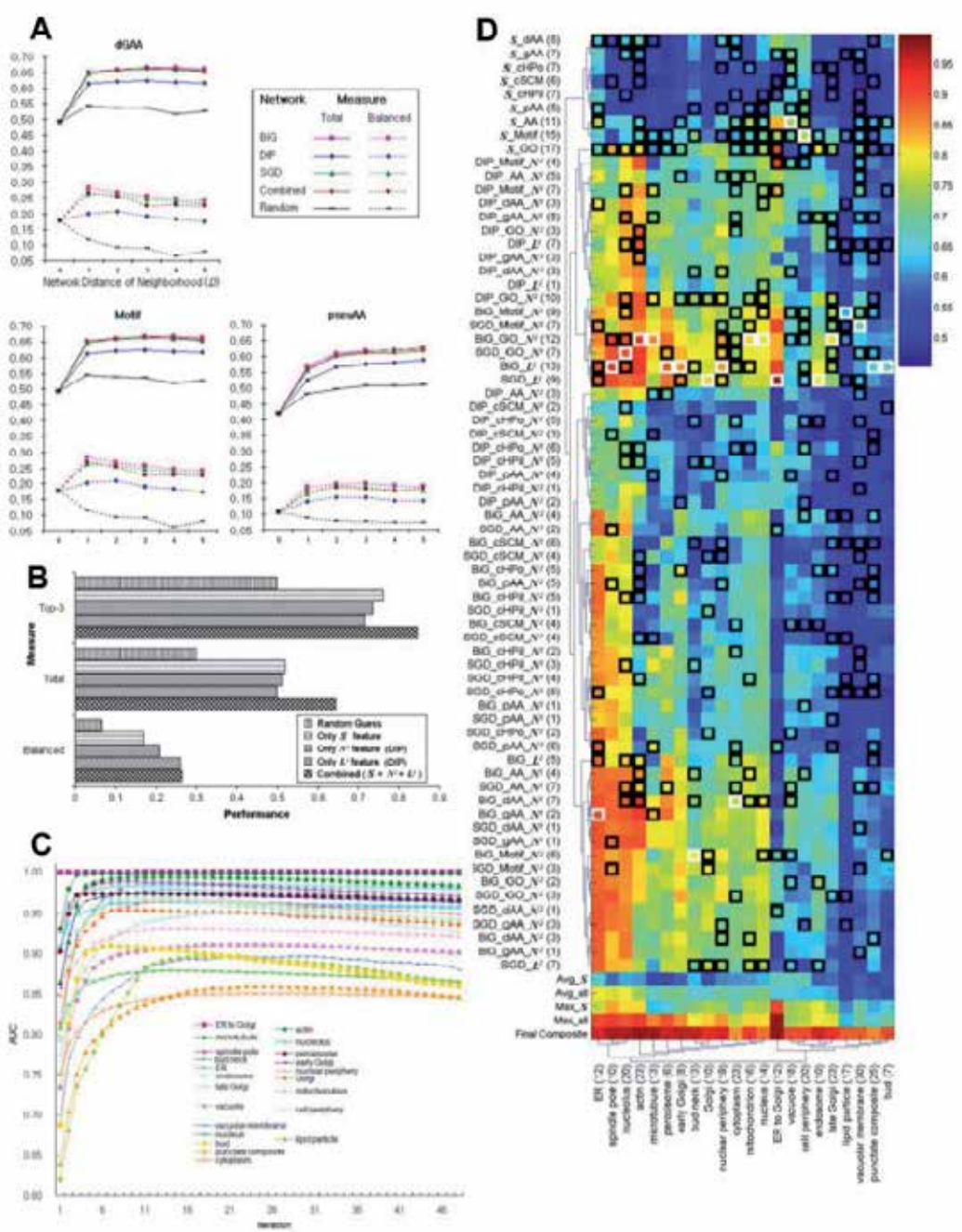


Fig. 4. (A) Shows performance of the classifiers by input from various kinds of feature. (B) shows performance for combination of feature vectors. (C) shows averaged AUC of the classifier for each localization based on feature selection using a DC-kNN. (D) shows selected feature sets for each of 22 localizations in yeast (adapted from Lee et al, 2008).

Based on the methodology, Lee et al. applied their method to the prediction of the localizations of genome-wide yeast proteins. Surprisingly, they also validated novel localizations of 61 proteins. For example, Huh et al. reported that Noc4/Ypr144c and Utp21/Ylr409c were localized in the nucleus (Huh et al, 2003). However, the proposed method developed by Lee et al. predicted the localization of the both proteins as the nucleolus. They revaluated for both proteins using new experiments and finally confirmed the previous results of Huh et al. had errors (Figures 5A and 5B). The correct prediction mainly owes to the fact that Lee et al. combined evidence from multiple interacting partners. For example, Noc4 interacts with many other proteins known to exist in the nucleolus, so we can assume that Noc4 localizes nearby or directly in the nucleolus. They confirmed the assumption by the network neighbours (Lee et al, 2008) (Figure 5C).

The number of localizations and known PPIs for yeast proteins are larger than those for other organisms. In other words, some organisms have less information on known localization and protein interaction, which might make the location prediction difficult based on a PPI network. Lee et al. evaluated their method using yeast data with some random missing information (Lee et al, 2008). As a result relatively robust results were obtained with less information. For example, the average number of neighbours of a protein in yeast is 27 and the number in worm is three. Decrement in the number of neighbours from yeast to worm was 9-fold. However, the average of AUC value decreased from 0.94 (yeast) to 0.87 (worm) (Figure 6). In other words, their method can be easily applied, not only to yeast but to other species with less known localization and/or interaction information. Actually they predicted subcellular localization of fly, human, and Arabidopsis (Lee et al, 2008; Lee et al, 2010b) using protein interactions. The results of both works showed that the prediction worked well for the other organisms and could find real localizations of some unknown proteins (Figures 6-7).

They also compared a DC-kNN with two previous popular methods, ISort (Chou & Cai, 2005) and PSLT2 (Scott et al, 2005). ISort is a comprehensive sequence-based machine learning method. ISort can predict more than 15 compartments. PSLT2 is a previous method that used a protein interaction network to predict subcellular localizations. They compared to DC-kNN with ISort and PSLT2 using both total and balanced measures. As illustrated in Figure 8, DC-kNN outperformed both methods in total and balanced measurement.

4.2.4 Other network-based methods

After the study of Lee et al. in 2008, several studies based on network-based approaches tried to predict subcellular localization. Mintz-Oron et al. used a constraint-based method for predicting subcellular localization of enzymes based on their embedding metabolic network, relying on a parsimony principle of a minimal number of cross-membrane metabolite transporters (Mintz-Oron et al, 2009). They showed that their method outperformed pathway enrichment-base methods. Another group constructed a decision tree-based meta-classifier for identification of essential genes (Acencio & Lemke, 2009). Their method relied on network topological features, cellular localization and biological process information for prediction of essential genes. Tung & Lee integrated various biological data sources to get information of neighbour proteins in a probabilistic gene-network (Tung & Lee, 2009). They predicted the subcellular localization using a Fuzzy k-nearest neighbour classifier. Lee et al. curated IntAct *Arabidopsis thaliana* PPI dataset

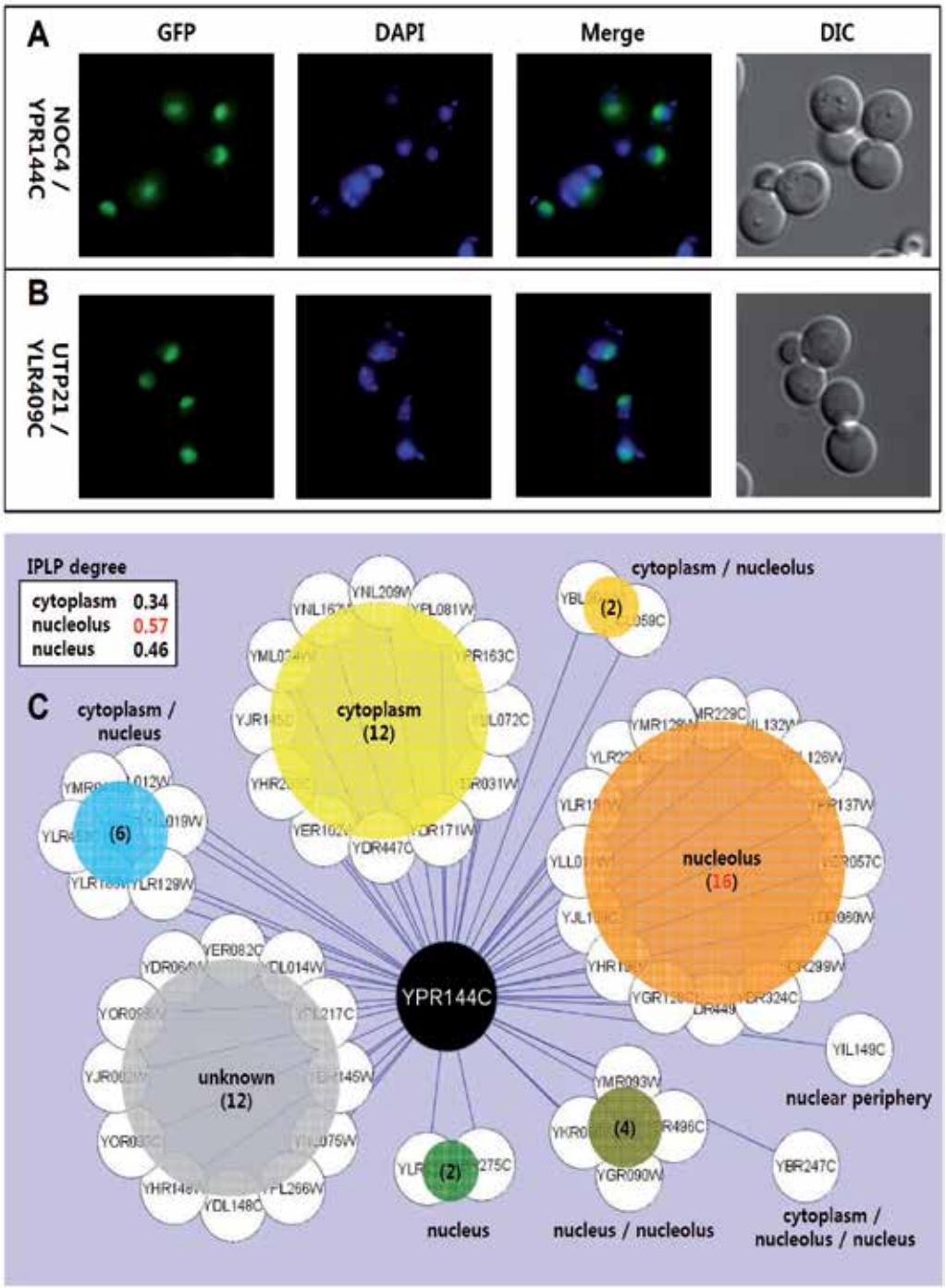


Fig. 5. (A, B) represent results of new experiments for Noc4/Ypr144c and Utp21/Ylr409c. (C) shows the interacting neighbours of Ypr144c (adapted from Lee et al, 2008).

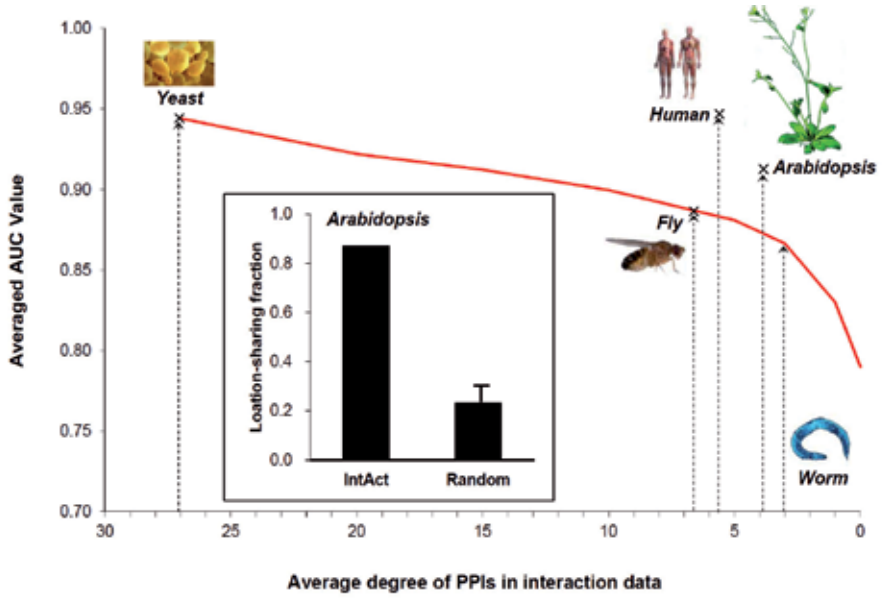


Fig. 6. Averaged AUC values across different organisms (adapted from Lee et al, 2010b).

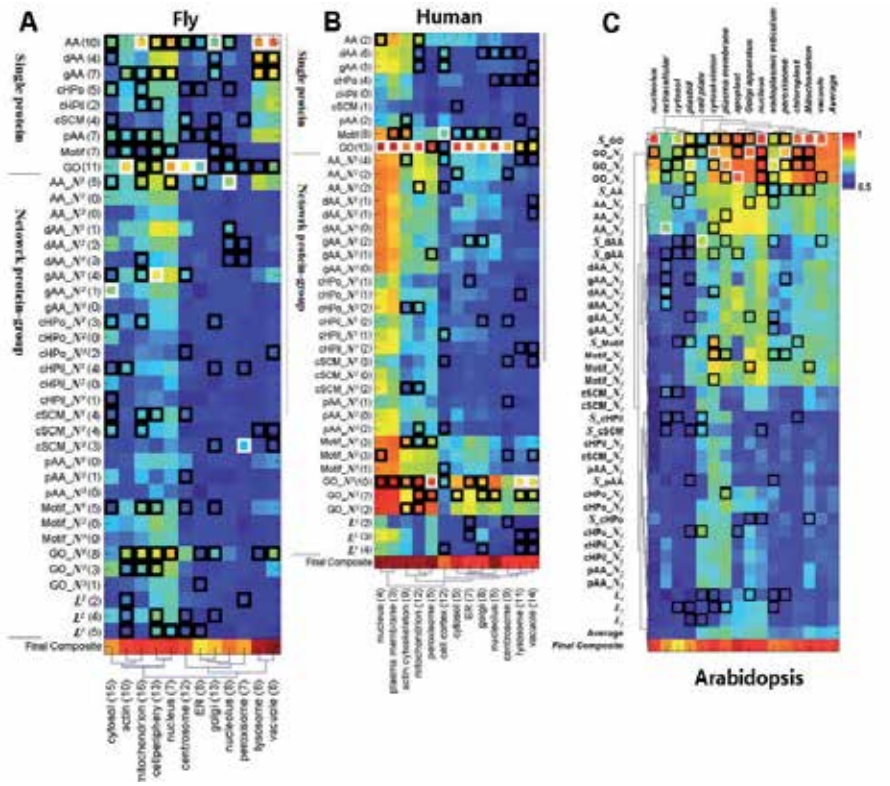


Fig. 7. Generated models for the location prediction for Fly (A), Human (B), and Arabidopsis (C) (adapted from Lee et al, 2008 and Lee et al, 2010b).

Lee et al. curated IntAct *Arabidopsis thaliana* PPI dataset (Aranda et al, 2010) using the DC-kNN method, which was proposed before and which showed good performance (Lee et al, 2010b). They also showed that the DC-kNN is applicable to other organisms. Kourmpetis et al. predicted a function of proteins in *Saccharomyces cerevisiae* based on network data, such as PPI data (Kourmpetis et al, 2010). They took a Bayesian Markov Random field analysis method for prediction and predicted the functions of 1170 un-annotated *Saccharomyces cerevisiae* proteins.

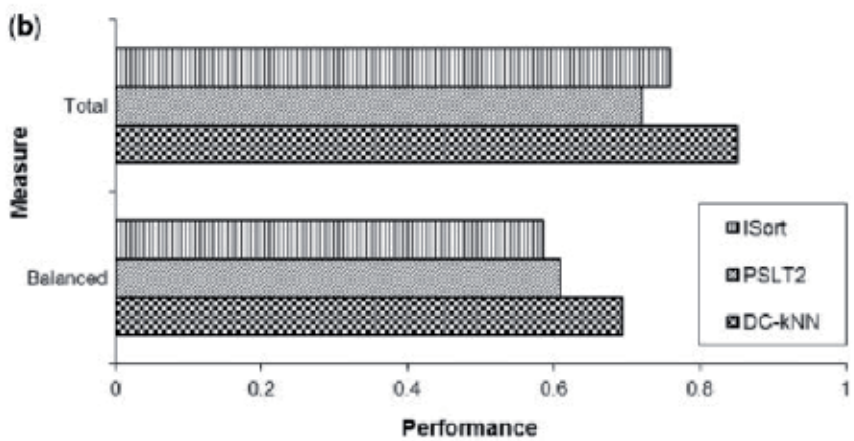


Fig. 8. Performance comparison of Isort, PSLT2 and DC-kNN (adapted from Lee et al, 2008).

5. Conclusions

We reviewed on PPI databases and the methods for detection of PPIs. Then, the computational methods of protein function prediction were briefly reviewed. We finally discussed that the prediction of protein function, especially the subcellular localization, shows outstanding performance when using PPIs data. This is because real biological functions are maintaining through a cascade of PPIs. Moreover, the computational approaches are very much promising when compared to the experimental identification especially for the false reading corrections. Functional genomics is an ongoing field in systems biology and this must be done well to drive further progress. We are facing other issues concerning the lack of conditional protein interactomes. We have identified and accumulated only static information at the molecular level in cells to make a scaffold of cellular systems. Computational methods should be applied to this conditional analysis when sufficient data become available and the next field of utilization would be personalized medicines, such as the early diagnosis with specific markers and treatments with specific drug targets.

6. Acknowledgement

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0022887 and 2010-0018258).

7. References

Acencio ML, Lemke N (2009) Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC bioinformatics* 10: 290

Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of molecular biology* 311: 395-408

Aloy P, Russell RB (2003) InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 19: 161-162

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25: 3389-3402

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic acids research* 32: D115-119

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K et al (2010) The IntAct molecular interaction database in 2010. *Nucleic acids research* 38: D525-531

Arnau V, Mars S, Marin I (2005) Iterative cluster analysis of protein interaction data. *Bioinformatics* 21: 364-378

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25: 25-29

Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic acids research* 34: D504-506

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics* 4: 2

Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS computational biology* 1: e3

Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics* 21 Suppl 1: i38-46

Berggard T, Linse S, James P (2007) Methods for the detection and analysis of protein-protein interactions. *Proteomics* 7: 2833-2842

Bhasin M, Raghava GP (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic acids research* 32: W414-419

Blaschke C, Hoffmann R, Oliveros JC, Valencia A (2001) Extracting information automatically from biological literature. *Comparative and functional genomics* 2: 310-313

Bonetta L (2010) Protein-protein interactions: Interactome under construction. *Nature* 468: 851-854

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. *Journal of molecular biology* 283: 707-725

Bousquet-Dubouch MP, Fabre B, Monsarrat B, Burlet-Schiltz O (2011) Proteomics to study the diversity and dynamics of proteasome complexes: from fundamentals to the clinic. *Expert review of proteomics* 8: 459-481

Bowers PM, Cokus SJ, Eisenberg D, Yeates TO (2004) Use of logic relationships to decipher protein network organization. *Science* 306: 2246-2249

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and biophysical research communications* 305: 407-411

Cai YD, Chou KC (2004) Predicting 22 protein localizations in budding yeast. *Biochemical and biophysical research communications* 323: 425-428

Cai YD, Liu XJ, Xu XB, Chou KC (2000) Support vector machines for prediction of protein subcellular location. *Molecular cell biology research communications : MCBRC* 4: 230-233

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical journal* 84: 3257-3263

Carlson BA, Xu XM, Kryukov GV, Rao M, Berry MJ, Gladyshev VN, Hatfield DL (2004) Identification and characterization of phosphoseryl-tRNA[Ser]Sec kinase. *Proceedings of the National Academy of Sciences of the United States of America* 101: 12848-12853

Cedano J, Aloy P, Perez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. *Journal of molecular biology* 266: 594-600

Cho YR, Hwang W, Ramanathan M, Zhang A (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC bioinformatics* 8: 265

Chou KC (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and biophysical research communications* 278: 477-483

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43: 246-255

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *The Journal of biological chemistry* 277: 45765-45769

Chou KC, Cai YD (2003) Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *Journal of cellular biochemistry* 90: 1250-1260

Chou KC, Cai YD (2005) Predicting protein localization in budding yeast. *Bioinformatics (Oxford, England)* 21: 944-950

Chou KC, Elrod DW (1999) Protein subcellular location prediction. *Protein engineering* 12: 107-118

Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. *Analytical biochemistry* 370: 1-16

Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics* 22: 1623-1630

Cote R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H (2010) The Ontology Lookup Service: bigger and better. *Nucleic acids research* 38: W155-160

Deng M, Chen T, Sun F (2004) An integrated probabilistic model for functional prediction of proteins. *Journal of computational biology : a journal of computational molecular cell biology* 11: 463-475

Drawid A, Gerstein M (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of molecular biology* 301: 1059-1075

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome informatics International Conference on Genome Informatics* 23: 205-211

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of molecular biology* 300: 1005-1016

Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Krieger CJ, Livstone MS, Miyasato SR, Nash R, Oughtred R, Park J, Skrzypek MS, Weng S, Wong ED, Dolinski K et al (2010) Saccharomyces Genome Database provides mutant phenotype data. *Nucleic acids research* 38: D433-436

Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML (2009) Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics* 10: 233-246

Ferrer L, Dale JM, Karp PD (2010) A systematic study of genome context methods: calibration, normalization and combination. *BMC bioinformatics* 11: 493

Fetrow JS, Skolnick J (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *Journal of molecular biology* 281: 949-968

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. *Nucleic acids research* 38: D211-222

Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic acids research* 31: 3613-3617

Garg P, Sharma V, Chaudhari P, Roy N (2009) SubCellProt: predicting protein subcellular localization using machine learning approaches. *In silico biology* 9: 35-44

Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in bioinformatics* 12: 449-462

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M et al (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631-636

Gingras AC, Gstaiger M, Raught B, Aebersold R (2007) Analysis of protein complexes using mass spectrometry. *Nature reviews Molecular cell biology* 8: 645-654

Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic acids research* 29: 3513-3519

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46: 389-422

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* 402: C47-52

Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein--protein interaction data. *Yeast* 18: 523-531

Hosur R, Xu J, Bienkowska J, Berger B (2011) iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. *Journal of molecular biology* 405: 1295-1310

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics (Oxford, England)* 17: 721-728

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics (Oxford, England)* 20: 21-28

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. *Nature* 425: 686-691

Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P (1998) Homology-based fold predictions for Mycoplasma genitalium proteins. *Journal of molecular biology* 280: 323-326

Imai K, Nakai K (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10: 3970-3983

Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular biology and evolution* 16: 332-346

Kemmer D, Huang Y, Shah SP, Lim J, Brumm J, Yuen MM, Ling J, Xu T, Wasserman WW, Ouellette BF (2005) Ulysses - an application for the projection of molecular interactions across species. *Genome biology* 6: R106

Kemmeren P, Holstege FC (2003) Integrating functional genomics data. *Biochemical Society transactions* 31: 1484-1487

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ, Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stumpflen V, Salwinski L, Nerothin J, Cerami E et al (2007) Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC biology* 5: 44

King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* 20: 3013-3020

Kohl M, Wiese S, Warscheid B (2011) Cytoscape: software for visualization and analysis of biological networks. *Methods Mol Biol* 696: 291-303

Koonin EV, Wolf YI, Aravind L (2001) Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome research* 11: 240-252

Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ (2010) Bayesian Markov Random Field analysis for protein function prediction based on network data. *PloS one* 5: e9293

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440: 637-643

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* 21: 1109-1121

Lee I, Lehner B, Vavouri T, Shin J, Fraser AG, Marcotte EM (2010a) Predicting genetic modifier loci using functional gene networks. *Genome research* 20: 1143-1153

Lee K, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research* 36: e136

Lee K, Kim DW, Na D, Lee KH, Lee D (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic acids research* 34: 4655-4666

Lee K, Thorneycroft D, Achuthan P, Hermjakob H, Ideker T (2010b) Mapping plant interactomes using literature curated and predicted protein-protein interaction data sets. *The Plant cell* 22: 997-1005

Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA (2011) Systematic computational prediction of protein interaction networks. *Physical biology* 8: 035008

Lu H, Zhu X, Liu H, Skogerbo G, Zhang J, Zhang Y, Cai L, Zhao Y, Sun S, Xu J, Bu D, Chen R (2004) The interactome as a tree--an attempt to visualize the protein-protein interaction network in yeast. *Nucleic acids research* 32: 4804-4811

Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A (2010) A global map of human gene expression. *Nature biotechnology* 28: 322-324

Luttgen H, Rohdich F, Herz S, Wungsintaweekul J, Hecht S, Schuhr CA, Fellermeier M, Sagner S, Zenk MH, Bacher A, Eisenreich W (2000) Biosynthesis of terpenoids: YchB protein of Escherichia coli phosphorylates the 2-hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol. *Proceedings of the National Academy of Sciences of the United States of America* 97: 1062-1067

Mani R, St Onge RP, Hartman JLt, Giaever G, Roth FP (2008) Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America* 105: 3461-3466

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-753

Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M (2001) Identification of potential interaction networks using sequence-based searches for

conserved protein-protein interactions or "interologs". *Genome research* 11: 2120-2126

Mei S, Fei W (2010) Amino acid classification based spectrum kernel fusion for protein subnuclear localization. *BMC bioinformatics* 11 Suppl 1: S17

Mei S, Fei W, Zhou S (2011) Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics* 12: 44

Mintz-Oron S, Aharoni A, Ruppin E, Shlomi T (2009) Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics (Oxford, England)* 25: i247-252

Mooney C, Wang YH, Pollastri G (2011) SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics (Oxford, England)* 27: 2812-2819

Mount DW (2007) Using a FASTA Sequence Database Similarity Search. *CSH protocols* 2007: pdb top16

Nair R, Rost B (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics (Oxford, England)* 18 Suppl 1: S78-86

Nair R, Rost B (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins* 53: 917-930

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in biochemical sciences* 24: 34-36

Nakai K, Kanehisa M (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* 11: 95-110

Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897-911

Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of molecular biology* 238: 54-61

Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, Ceol A, Chatr-aryamontri A, Armstrong J, Woollard P, Salama JJ, Moore S, Wojcik J, Bader GD, Vidal M, Cusick ME, Gerstein M, Gavin AC, Superti-Furga G, Greenblatt J et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature biotechnology* 25: 894-898

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *Journal of protein chemistry* 22: 395-402

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics (Oxford, England)* 19: 1656-1663

Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of molecular biology* 352: 1002-1015

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4285-4288

Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. *Proteins* 54: 49-57

Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC bioinformatics* 6 Suppl 4: S21

Pierleoni A, Martelli PL, Casadio R (2011) MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics (Oxford, England)* 27: 1224-1230

Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, Golshani A (2008) Computational methods for predicting protein-protein interactions. *Advances in biochemical engineering/biotechnology* 110: 247-267

Recillas-Targa F (2006) Multiple strategies for gene transfer, expression, knockdown, and chromatin influence in mammalian cell lines and transgenic animals. *Molecular biotechnology* 34: 337-354

Reczko M, Hatzigerrorgiou A (2004) Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. *Proteomics* 4: 1591-1596

Reid AJ, Ranea JA, Clegg AB, Orengo CA (2010) CODA: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion. *PloS one* 5: e10908

Reinhardt A, Hubbard T (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic acids research* 26: 2230-2236

Rives AW, Galitski T (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America* 100: 1128-1133

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. *Nucleic acids research* 32: D449-451

Sanchez-Chapado M, Angulo JC, Ibarburen C, Aguado F, Ruiz A, Viano J, Garcia-Segura JM, Gonzalez-Esteban J, Rodriquez-Vallejo JM (1997) Comparison of digital rectal examination, transrectal ultrasonography, and multicoil magnetic resonance imaging for preoperative evaluation of prostate cancer. *European urology* 32: 140-149

Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. *Nature biotechnology* 18: 1257-1261

Scott MS, Calafell SJ, Thomas DY, Hallett MT (2005) Refining protein subcellular localization. *PLoS computational biology* 1: e66

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Molecular systems biology* 3: 88

Shatkay H, Hoglund A, Brady S, Blum T, Donnes P, Kohlbacher O (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics (Oxford, England)* 23: 1410-1417

Shen YQ, Burger G (2010) TESTLoc: protein subcellular localization prediction from EST data. *BMC bioinformatics* 11: 563

Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational biology* 3: e43

Shoemaker BA, Zhang D, Thangudu RR, Tyagi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR (2010) Inferred Biomolecular Interaction Server--a web

server to analyze and predict protein interacting partners and binding sites. *Nucleic acids research* 38: D518-524

Singh R, Park D, Xu J, Hosur R, Berger B (2010) Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. *Nucleic acids research* 38: W508-515

Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, Mujica AO, Thomas M, Harrow J, Cox T, Jackson D, Severin J, Biggs P, Fu J, Nefedov M, de Jong PJ, Stewart AF, Bradley A (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* 474: 337-342

Skrabanek L, Saini HK, Bader GD, Enright AJ (2008) Computational prediction of protein-protein interactions. *Molecular biotechnology* 38: 1-17

Smith KL, DeVos V, Bryden H, Price LB, Hugh-Jones ME, Keim P (2000) Bacillus anthracis diversity in Kruger National Park. *Journal of clinical microbiology* 38: 3780-3784

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America* 100: 12123-12128

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M (2011) The BioGRID Interaction Database: 2011 update. *Nucleic acids research* 39: D698-704

Stein A, Panjkovich A, Aloy P (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic acids research* 37: D300-304

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302: 249-255

Suter B, Kittanakom S, Stagljar I (2008) Two-hybrid technologies in proteomics research. *Current opinion in biotechnology* 19: 316-323

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* 39: D561-568

Tian J, Gu H, Liu W, Gao C (2011) Robust prediction of protein subcellular localization combining PCA and WSVMs. *Computers in biology and medicine* 41: 648-652

Tikk D, Thomas P, Palaga P, Hakenberg J, Leser U (2010) A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS computational biology* 6: e1000837

Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O (2008) Architectures and functional coverage of protein-protein interfaces. *Journal of molecular biology* 381: 785-802

Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in bioinformatics* 10: 217-232

Tung TQ, Lee D (2009) A method to improve protein subcellular localization prediction by integrating various biological data sources. *BMC bioinformatics* 10 Suppl 1: S43

Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database : the journal of biological databases and curation* 2010: baq026

Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database : the journal of biological databases and curation* 2010: baq023

Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. *Nature biotechnology* 21: 697-700

Wang J, Li C, Wang E, Wang X (2011) An FPT approach for predicting protein localization from yeast genomic data. *PloS one* 6: e14449

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics* 36: 307-340

Xia JF, Wang SL, Lei YK (2010) Computational methods for the prediction of protein-protein interactions. *Protein and peptide letters* 17: 1069-1078

Yeats C, Lees J, Carter P, Sillitoe I, Orengo C (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic acids research* 39: W546-550

Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS letters* 451: 23-26

Zhang QC, Petrey D, Norel R, Honig BH (2010) Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences of the United States of America* 107: 10896-10901

Zhang Z, Sun H, Zhang Y, Zhao Y, Shi B, Sun S, Lu H, Bu D, Ling L, Chen R (2006) Genome-wide analysis of mammalian DNA segment fusion/fission. *Journal of theoretical biology* 240: 200-208

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. *Proteins* 50: 44-48

Zhu L, Yang J, Shen HB (2009) Multi label learning for prediction of human protein subcellular localizations. *The protein journal* 28: 384-390

4

Integrative Approach for Detection of Functional Modules from Protein-Protein Interaction Networks

Zelmina Lubovac-Pilav
University of Skövde, Systems Biology Research Centre
Sweden

1. Introduction

Advances in large scale technologies in proteomics, such as yeast two-hybrid (Y2H) screening and mass spectrometry (MS) have enabled us to generate large protein-protein interaction (PPI) networks. The structure of such networks has been frequently analysed to identify the modules, which constitute the basic "building blocks" of molecular networks. One of the challenges that systems biology is facing consists of explaining biological organisation in the light of the existence of modules in networks (Han et al., 2004; Pereira-Leal et al., 2004; Petti and Church, 2005; Rives and Galitski, 2003). A series of studies attempting to reveal the modules in cellular networks, ranging from metabolic (Ravasz et al., 2002), to protein networks (Spirin and Mirny, 2003; Yook et al., 2004), support the proposal that modular architecture is one of the principles underlying biological organisation.

Several key issues are being addressed in current research in systems biology, as a result of our post-genomic view that has expanded the role of the protein into an element of a network in which it has contextual functions within functional modules (Eisenberg et al., 2000; Jeong et al., 2001). How do modules interact to achieve a certain functionality (Han et al., 2004; Rives and Galitski, 2003)? How can we evaluate the biological relevance of modules (Pereira-Leal et al., 2004; Poyatos and Hurst, 2004)? Answering those questions may contribute to better understanding of the relationships between structure, function and regulation of molecular networks, which is an important aim of systems biology (Qi and Ge, 2006; Stelling et al., 2002).

From the structural perspective, modules are often associated with highly connected clusters of proteins. Many efforts in this area have been directed towards analysing structural properties of the protein interaction graph, measured by clustering coefficient and shortest path distance for example, to derive modular formations. The main focus presented in this chapter is on defining similarity between protein interactions based on an integrated score that takes into consideration topology of PPI network along with the functional knowledge determined by semantic similarity. An important reason for considering knowledge represented in annotations a valuable complement to topological characteristics is

encompassed in the concept of functional modules themselves. A functional module consists of proteins that cooperate towards achieving a particular function or participate in similar processes. Hence, considering annotation that describes molecular functions and biological processes should enrich the protein-protein interactions. Functional information can be retrieved from Gene Ontology (GO), which is a structured vocabulary used to annotate proteins with information about their molecular function, participation in biological processes or localization in cellular components. A module-identifying algorithm proposed earlier (Lubovac et al., 2006), SWEMODE (Semantic WEights for MODule Elucidation), that relies on an integrated measure, called semantic cohesiveness, corresponds to one of the successful approaches that contributes to achieve the important aims of systems biology. This method will be the focus of attention in this chapter.

2. Background

Molecular biology is becoming a highly modular science where functional modules are considered to be a critical level of biological organization. The term "module", as understood in molecular biology, was originally defined as a discrete unit with a function that is separable from those of other modules (Hartwell et al., 1999). Furthermore, modularity refers to clusters of elements that work in a co-operative fashion to achieve some defined function. Protein complexes constitute one example type of module, since the proteins within a complex interact functionally and physically to form a robust unit, which in its turn carries out some biological function (Yook et al., 2004).

One of the key issues to be solved with help of bioinformatics is the deciphering of the complex architecture of biological networks.

2.1 Climbing life's complexity pyramid

Biological networks are often modular and compound, and involve connections between groups of genes and proteins as well as between individual elements. A simple complexity pyramid (see Fig. 1) suggested by Oltvai and Barabasi (2002), illustrates different levels of cellular organisation.

Living systems are organised at both logical and physical levels. The individual nucleotides are elementary building blocks of DNA and RNA molecules, which, in turn, are organised into higher level structures such as regulatory elements, and genes. DNA is physically organised into larger structures such as chromatin and chromosomes. Groups of genes, proteins, RNAs (the bottom level of the pyramid in Fig. 1) may be organised into pathways in metabolism, and motifs in genetic regulatory networks (see level 2). Regulatory motifs may in turn serve as building blocks of functional modules (level 3). There is a growing body of evidence that the modules are then organised in a hierarchical manner (Barabasi and Oltvai, 2004; Oltvai and Barabasi, 2002; Ravasz et al., 2002), defining the large-scale functional organisation of the cell (level 4 in Fig. 1).

The way these various structures interact with each other determines the machinery of a cell. Cells and the extracellular matrix, which surrounds and supports cells, build up the tissues that in turn are organised into organs, and so forth.

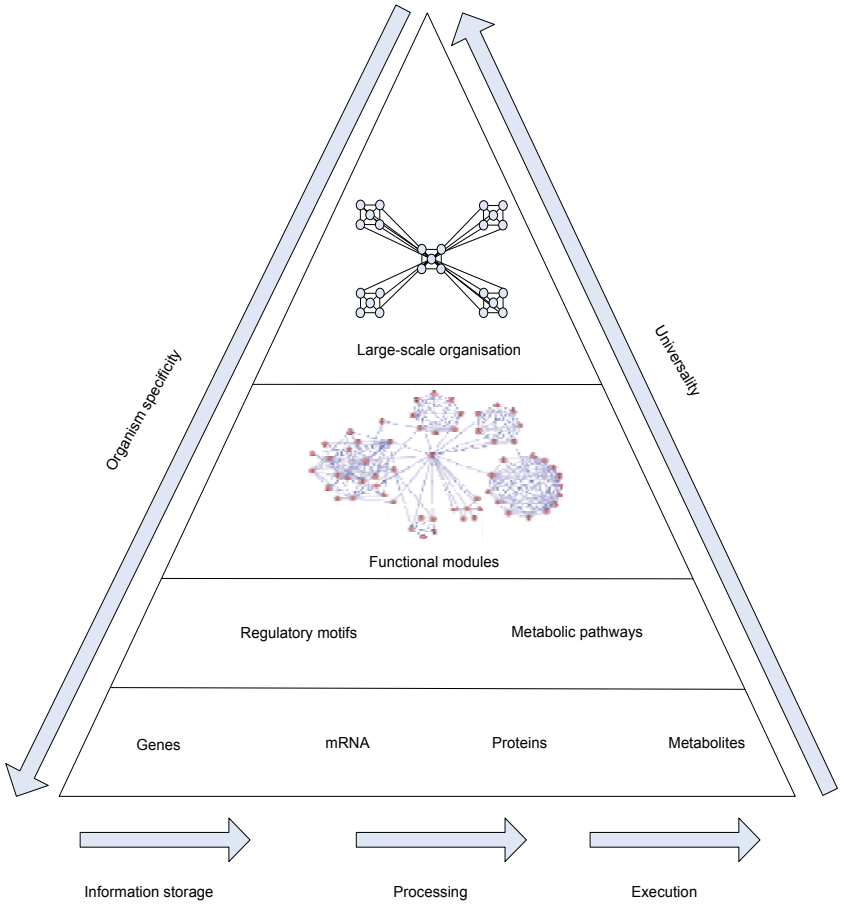


Fig. 1. Life's complexity pyramid redrawn from (Oltvai and Barabasi, 2002).

The integration of different layers in the pyramid to achieve a better understanding of system-level rules that govern cell function is one of the challenges in systems biology. Computational analysis tools and methods are needed at each level but also across different levels. Here, the integrative approach for deriving modules at the third level in the pyramid is described, which also make it possible to climb to the top, and provide means for revealing large-scale organisation.

2.2 Modularity in cellular networks

"Modularity is a fundamental design principle whereby components are partitioned according to common physical, regulatory, or functional properties" (Petti and Church, 2005). Modules can be found in many systems, for example, food webs, networks of web pages describing related subjects (Flake et al., 2002), networks of friends in sociology (Newman, 2003), or scientific collaboration networks (Newman, 2001). A usual synonym for the term module in other scientific disciplines, like sociology for example, is community or community structure. In a study by Flake et al., (2002), the term web community is for

example defined as "a collection of web pages such that each member page has more hyperlinks within the community than outside of the community". This definition may be adjusted further, according to Flake et al., (2002), to identify communities of varying sizes and levels of cohesiveness (clustering).

Furthermore, modularity involves groups of elements that work in a co-operative fashion to achieve some well-defined function. In a general network representation, a module appears as a highly interconnected group of nodes (Barabasi and Oltvai, 2004). Modules can be interpreted as separated substructures of a network or pathway, e.g. a protein complex is a module of a protein interaction network. Protein complexes are well-defined examples of modularity since they consist of proteins that interact functionally and physically to form a tightly connected unit, which, in turn, carries out some biological function (Yook et al., 2004). Another example of modular organisation can be found in genetic regulatory networks where several transcription factor binding sites, organised into functional units, i.e. modules, play a crucial role in gene transcription.

The members that constitute modules are more strongly related to each other than to members of other modules, which is reflected in the network topology. The modular nature of PPI networks is reflected by a high degree of clustering, measured by the clustering coefficient. The clustering coefficient measures the local cohesiveness around a node, and it is defined, for any node i, as the fraction of neighbours of i that are connected to each other (Watts and Strogatz, 1998). Simply stated, the clustering coefficient ci measures the presence of 'triangles' which have a corner at *i* (see the triangles with dashed sides in Fig. 2). The high degree of clustering is based on local sub-graphs with a high density of internal connections, while being less tightly connected to the rest of the network (Uhrig, 2006).

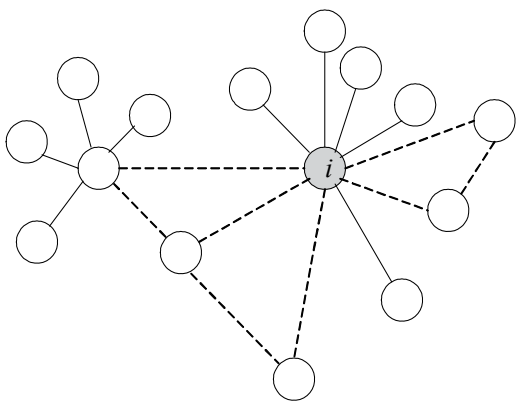


Fig. 2. Example of a protein sub-graph with triangle-forming proteins.

As pointed out by Barabasi and Oltvai (2004), each module may be reduced to a set of triangles, and a high density of such triangles is highly characteristic for PPI networks, pointing at the modular nature of such networks. By averaging the clustering coefficient over all nodes we can obtain a global measure of the cohesiveness of the network, where a high average clustering coefficient indicates the presence of modularity. It has been confirmed in many studies that most real large-scale networks tend to contain dense

clusters, in the sense that the average clustering coefficient of such networks is much greater than for random networks. In contrast, if modularity is absent in the network, the average clustering coefficient is comparable to that of a randomised network.

The exact meaning of modularity in biological networks depends on the network under consideration. For example, modules in protein networks are often seen as static molecular complexes (such as the ribosome) or as dynamic signalling pathways (such as the MAPK cascade). There are also examples of large modular molecule complexes that are in turn organised in modules. One of such complexes is yeast Mediator, which transmits regulatory signals from DNA-binding transcription factors to RNA polymerase II. The Mediator complex is thought to be composed of 24 subunits organised in four modules, named the head, middle, tail and Cdk8 modules. In gene regulatory networks, modules are often seen as sets of genes controlled by the same set of transcription factors under certain conditions (Segal et al., 2003).

Modules should not be seen as isolated components, since it has been shown that some crosstalk and overlap exists between them (Han et al., 2004; Schwikowski et al., 2000). Instead, modules should be considered as components that have dense intra-connectivity but sparse inter-connectivity. In a study analysing protein interaction networks in the yeast *Saccharomyces cerevisiae*, Schwikowski et al., (2000) reported global patterns of interactions of proteins within functional classes or subcellular compartments, as well as many possible cross-connections. It is further pointed out by Qi and Ge (2006) that the existence of the links between modules emphasises the coordination of the cellular processes. For example, Petti and Church (2005) investigated possible transcriptional coordination between glycolysis and lipid metabolism modules.

A growing body of work supports the idea that such modules underlie much of cellular functioning (Gavin et al., 2006; Han et al., 2004; Pereira-Leal et al., 2004; Qi and Ge, 2006; Rives and Galitski, 2003), and that functional modules are the most relevant organisational units of a cell from the perspective of systems biology (Hartwell et al., 1999).

2.3 Integrating functional knowledge in module discovery

Although topology-based network measures, such as clustering coefficient, play an important role in module discovery, there are some reasons why we should integrate functional knowledge as well when deriving modular formations. High-throughput protein interaction data that is often used to identify modules is very noisy (Titz et al., 2004). Technologies such as Y2H often result in many false positives that may cause false conclusions in the analysis. A possible approach to decrease the number of false interactions may be to focus on the "high confidence" data sets, where all interactions have been confirmed by several experiments. However, in this way the majority of the existing interactions would be discarded from further analysis. A better approach should imply incorporating the functional knowledge associated with available interactions into the analysis. This has also been pointed out in previous studies that focus on deriving protein complexes by using topological information. In (Przulj et al., 2004), it has been observed that the increasing size of PPI networks (by including medium and low confidence interactions) has resulted in a decreasing number of highly connected sub-graphs or clusters which may correspond to protein complexes. As Przulj, et al., (2004) state, the reason for this may be the

increasing noise in the data, and a possible solution to this problem is the integration of PPI networks with annotation or gene expression data. In sub-chapter 2.4 a possible general framework for such integrative approach for module identification is described.

2.4 A general framework for integrative module identification

There are many ways of measuring similarity between proteins. The main proposal presented here considers protein similarity based on an integrated score that takes into consideration protein interaction data (as a topology source) and functional information based on semantic similarity. As pointed out previously, an ideal approach should take into consideration both temporal and spatial data, to be able to reflect the true dynamics of the cellular networks. It is therefore worthwhile to discuss how the methods presented here may be generalised to cope with several sources of information. Our module-identifying framework may be generalised by:

1. considering several sources of topological information
2. considering several sources of functional information

Topological information may refer to, for example, protein-protein interactions obtained from different experimental sources, such as Y2H and MS. However, this information may also be derived from different topological properties like clustering coefficient, edge betweenness, etc.

Besides semantic similarity values based on protein GO terms that we used in this work, there are many other sources of functional information that may be useful for predicting membership in protein complexes. One of the most prominent sources is gene expression data generated using various high-throughput platforms, such as microarrays. Expression profile correlation coefficients may, for example, be used to assign similarity scores to pairwise interactions. Other sources of functional information are essentiality, phylogenetic profiles, localisation, the MIPS functional catalogue, etc.

In this study, as in the majority of others, protein interactions are treated as binary, i.e. the edges in a network are either present or absent. Bearing in mind the fact that large-scale methods, although offering vast improvements in efficiency, still have much higher error rates than small-scale methods, a step towards generalisation of the proposed algorithms would be to treat protein interaction networks probabilistically. By treating the edges as binary (indicating presence/absence of interaction), we cannot distinguish edges supported by multiple evidence types, from edges supported by evidence of differing quality. There are several ways of assigning probabilities to individual pairs of proteins based on the amount and type of supporting evidence (Asthana et al., 2004; Jansen et al., 2002; Jansen et al., 2003). When dealing with several data sources that need to be combined in order to improve the prediction, a usual way of combining these consists of overlapping different interactomes. This approach, in turn, gives rise to the question whether it is more beneficial to consider the union of the disparate datasets or their intersection. One of the extremes that may be envisaged is that each one of the networks that are to be integrated has a low rate of false positives (FP) but a high rate of false negatives (FN). In this case, the union of the two sets of interactions would be advantageous. At the other extreme, when dealing with networks with high FP rates and low FN rates, the intersection between the different networks is preferable.

The problem of finding an optimal combination of unions and intersections among the different networks may be defined, as described in (Jansen et al., 2002), as finding a trade-off between the highest possible coverage (TP/(TP+FN)) and the lowest possible error rate (FP/(TP+FP)). Determining the error rate is still an open question, as pointed out in (Jansen et al., 2002).

A hypothetical example of integrating different data sources that may be useful in generalising the proposed approaches is given in Fig. 3. The top part of the figure shows four possible data sources that may be useful for module identification. Two of them are topological sources, denoted as t_1 and t_2, and are usually treated as binary networks. The other two sources, denoted as f_1 and f_2, may be used to assign functional weights to the edges. For example, when using gene expression as a possible source for weighting the edges, the probability of finding two proteins in a complex, given a certain correlation between their expression profiles, may be a possible way to assign weights (Jansen et al., 2002). Gene ontology sub-graphs as a possible source of functional information is visualised in the third square in Fig. 3, where semantic similarity between ontology terms may be used to reflect the functional similarity between the proteins, as assumed in this work. These functional weights may also be transformed into binary values, by setting different thresholds, where the level of the threshold determines the sensitivity and specificity of the experiment.

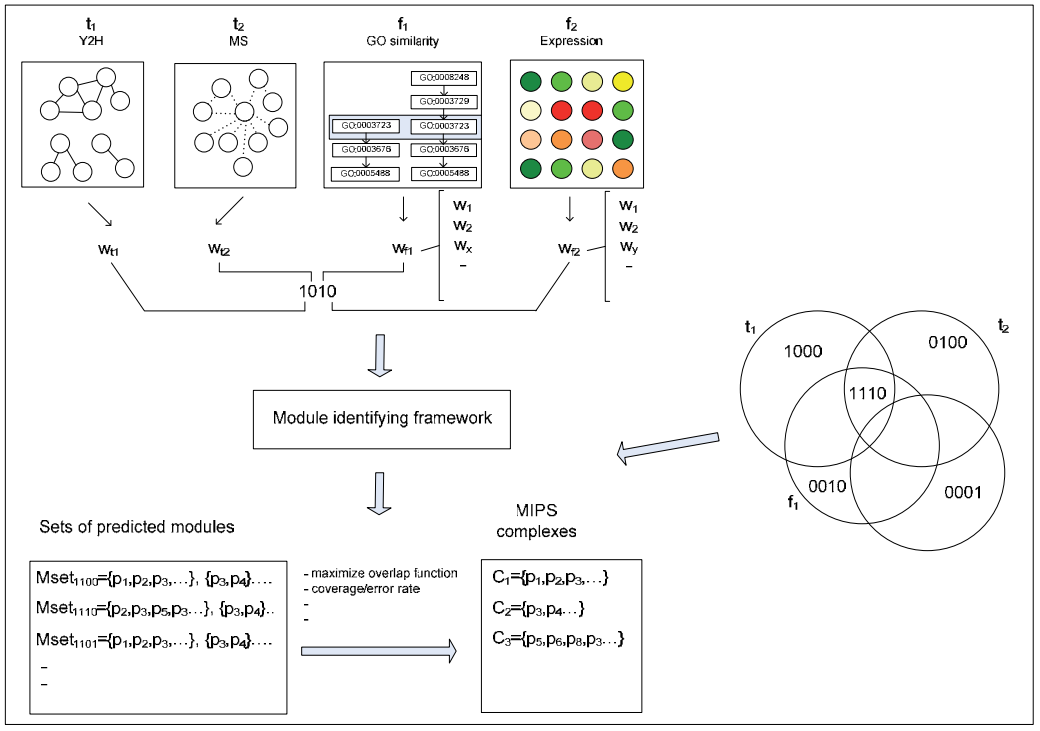


Fig. 3. Hypothetical integration of four data sources for module identification.

The bottom part of Fig. 3 shows the hypothetical module sets generated with different combinations of data sets. The Venn diagram to the right in the figure shows binary subset

profiles, where profile 1110 includes all data points that are present in data sets t_1, t_2, and f_1. Mset1110, for example, denotes the set of modules derived from the combination of MS, Y2H, and GO semantic similarity weights, where p_x denotes a protein x belonging the module.

3. Module identification based on an integrated approach

The algorithm described in previous work (Lubovac et al., 2006), SWEMODE (Semantic WEeights for MODule Elucidation), is an example of a method that employs an integrated approach for deriving functional modules, based on the functional and topological cohesiveness of the sub-graphs. Here, an integrated weighting score, called weighted clustering coefficient, that forms the bases for this method will be described. The reason for focusing on description of the integrative score here is that it can be applied as a part of node weighting procedure in other methods for deriving modules of PPI networks.

3.1 Weighted clustering coefficient

As depicted in earlier work, the separate edge weights do not provide an overall picture of the network's complexity. Therefore, we here consider the sum of all weights between a particular node and its neighbours, also referred to as the node strength. The strength s_i of the node i is defined as:

$$s_i = \sum_{\forall j, j \in N(i)} ss_{ij} \tag{1}$$

Given two proteins, i and j, with T_i and T_j containing m and n terms, respectively, the protein-protein semantic similarity ss_{ij} based on GO terms, is defined as the average inter-set similarity between terms from the given term sets (see Equation 2).

$$ss_{ij} = \frac{1}{m \times n} \sum_{t_k \in T_i, t_l \in T_j} sim(t_k, t_l) \tag{2}$$

Determining the similarity between two proteins i and j, is preceded by calculation of the similarity between the terms belonging to the term sets T_i and T_j that are used to annotate these proteins. Given the ontology terms $t_k \in T_i$ and $t_l \in T_j$, the semantic similarity measure proposed by (Lin, 1998) is defined as:

$$sim(t_k, t_l) = \frac{2 \ln p_{ms}(t_k, t_l)}{\ln p(t_k) + \ln p(t_l)} \tag{3}$$

Where $p(t_x)$ is the probability of term t_x and $p_{ms}(t_k,t_l)$ is the probability of the minimum subsumer of t_k and t_l, which is defined as the lowest probability found among the parent terms shared by t_k and t_l (Lord et al., 2003).

In previous work, some extensions of the topological clustering coefficient have been developed for weighted networks. In (Barrat et al., 2004), two scores that integrate topological and weighted features of the nodes – weighted clustering coefficient c^w and weighted average nearest-neighbours degree nn^w are introduced. These scores have

previously been applied to two types of complex weighted networks, namely, the world-wide airport network and the scientist collaboration network. A first attempt to apply these integrated scores on PPI networks was described in (Lubovac et al., 2006). A weighted measure that uses semantic similarity weights was introduced. Weighted clustering coefficient c^w is defined as:

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{\forall j,h \mid \{j,h\} \in K(i)} (ss_{ij} + ss_{ih}) \tag{4}$$

Where s_i is the functional strength of node i (see Equation 1) and ss_{ij} is the semantic similarity reflecting the functional weight of the interaction (see Equation 2). For each triangle formed in the neighbourhood of node i, involving nodes j and h, the semantic similarities ss_{ij} and ss_{ih} are calculated. Hence, not only the number of triangles in the neighbourhood of the node i is considered but also the relative functional similarity between the nodes that form those triangles, with regard to the total functional strength of the node. The normalisation factor $s_i(k_i\text{-}1)$ represents the summed weight of all edges connected from node i, multiplied by the maximum possible number of triangles in which each edge may participate. It also ensures that $0 \le c^w \le 1$. This measure can be involve any of the three aspects of Gene Ontology - molecular function, biological process and cellular component, or the combination of these.

4. Comparison with topology-based methods for module identification

The aim of this sub-chapter is to demonstrate the performance of the approach called SWEMODE (Lubovac et al., 2006), based on an integrative score described in 3.1, by comparing it to two purely topological approaches. One of the topology-based method for detecting modules from a PPI networks has been developed by Luo and Scheuerman (2006) and further analysed in (Luo et al., 2007). The module notion proposed was based on the degree definition of the sub-graphs. Unlike the approach described in Section 3, this method is based solely on topological properties of the protein sub-graph.

Modules generated with SWEMODE were also compared with the modules derived in (Przulj et al., 2004), based on HCS (Highly Connected Subgraphs) clustering algorithm (Hartuv and Shamir, 2000). This method aims to find disjoint subsets (clusters) that should satisfy following criteria: homogeneity – members of the same cluster are highly similar to each other; and separation: members of different clusters have low similarity to each other.

4.1 Protein-protein interaction data

For the evaluation purpose, two different PPI networks have been used. The first one was derived from the Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu), which is a database that stores and organises experimentally determined PPI (Xenarios et al., 2000). There is the subset of PPI from Yeast *S. cerevisiae*, denoted as CORE, which is the result of assessment with the Expression Profile Reliability Index (ERP Index) and the Paralogous Verification Method (PVM) (for further details, see (Deane et al., 2002)). The CORE subset contained 6379 interactions.

The second data set of PPI is obtained from the study by (von Mering et al., 2002). In that study, a quality assessment of large-scale data sets of protein-protein interactions in yeast was performed. A critical evaluation of the accuracy of high-throughput data is needed, because of the high rate of false interactions in these data sets. In (von Mering et al., 2002), data sets from yeast two-hybrid (Y2H) systems, protein complex purification techniques that rely on mass-spectroscopy (TAP and HMS-PCI), correlated mRNA expression profiles, genetic interactions, and *in silico* interaction predictions were analysed. As stated further in this study, each of these methods can be used to predict protein interactions, even though their goals are slightly different.

The authors integrated about 80 000 interactions between yeast proteins and found that only 2 455 were supported by more than one method. This low overlap between sets of protein interactions obtained from different methods may be due to the high fraction of false positives, but may also be caused by the difficulties for some methods to capture certain types of interactions. All interactions are classified by the level of confidence (low, medium, high), based on the evidence that supports them. In our study, we have used the interaction set with high level of confidence, meaning that all interactions are confirmed by several methods. This data set will be referred to as "von Mering". The data set contains 2 455 interactions between 988 proteins.

4.2 Evaluation against MIPS functional categories

The Munich Information Center for Protein Sequences (MIPS) provides high quality curated genome-related information, such as protein-protein interactions, protein complexes, protein functional categories, etc., spanning over several organisms.

The MIPS functional catalogue database consists of different fields, such as functional catalogue (FunCat) number, EC number, GO number, keywords etc. FunCat is an annotation scheme that provides functional descriptions of proteins (Ruepp et al., 2004). There are in total 28 main functional categories that are hierarchically structured. These categories cover functional fields such as metabolism, signal transduction, cellular transport etc.

The MIPS Comprehensive Yeast Genome Database (CYGD) provides information on the molecular structure and functional network of *S. cerevisiae*. The information used here for the evaluation purposes is the protein complex catalogue that contains a manually curated set of protein complexes that serve as an example of a type of module. There is another data set containing protein complexes obtained from (Gavin et al., 2002). This data set was produced by using a single experimental method, whereas the complex data set from MIPS has been derived from experiments from many labs using different techniques. Therefore, MIPS database is more realistic and appropriate to use for evaluation.

To evaluate and compare the performance of SWEMODE with two other methods for module identification, overlap score is used. In previous work, a similar evaluation has been applied to the clustering algorithm MCODE (Bader and Hogue, 2003), with respect to the number of matched complexes, but here slightly different definition of overlap score is used (see Equation 5).

The overlap score Ol (Poyatos and Hurst, 2004), is defined as:

$$Ol_{ij} = \left|M_i \cap M_j\right| \Big/ \sqrt{\left|M_i\right|\left|M_j\right|} \tag{5}$$

where M_i is the predicted module, and M_j is a module from the MIPS complex data set. The Ol measure assigns a score of 0 to modules that have no intersection with any known protein complex, whereas modules that exactly matches a known complex get the score 1.

4.3 Results

A total of 99 modules were detected in (Luo and Scheuermann, 2006). A new agglomerative algorithm was developed to identify modules from the network by combining the new module definition with the relative edge order generated by the Girvan-Newman algorithm. A JAVA program, MoNet, was developed to implement the algorithm Luo et al. (2007). Applying MoNet to the yeast core protein interaction network from the database of interacting proteins (DIP) identified 86 simple modules with sizes larger than 3 proteins. For convenience, those modules will be referred to as MoNet modules.

Evaluation of the MoNet modules with the overlap score threshold has been performed, and the results are compared with the resulting modules from SWEMODE, generated across approximately 400 different parameter settings (for parameter settings, see (Lubovac et al., 2006). We found that the modules derived from the latter show higher agreement with MIPS complexes (see Fig. 4). This comparison also indicates that introducing knowledge in terms

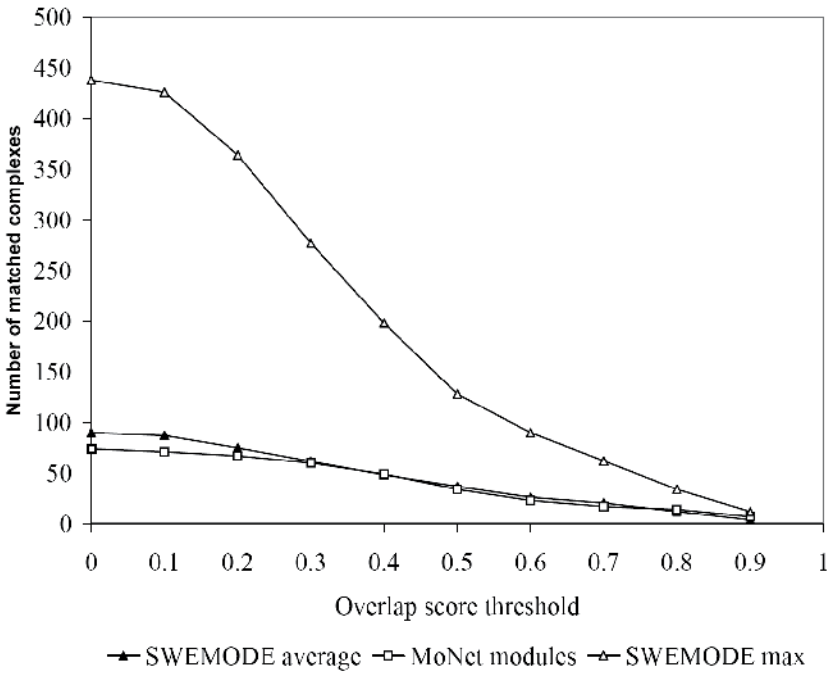


Fig. 4. Comparison between MoNet modules and SWEMODE modules.

of semantic similarity into the network topology seems to be advantageous over using only topology information. Furthermore, this method produces one single partition of the network, which does not seem biologically plausible, as many proteins may be involved in different processes.

We also compared our SWEMODE modules obtained from von Mering data with the modules derived in (Przulj et al., 2004), based on HCS. The modules generated with SWEMODE showed also here higher overlap with MIPS complexes (see Fig. 5). A more detailed analysis shows that both algorithms resulted in 39 identical modules. However, as HCS only discern the complexes that are highly interconnected, it discards many clusters that correspond to known complexes.

Another disadvantage of both methods that are here compared to SWEMODE is that they do not allow any overlap between modules, i.e. they produce disjoint clusters.

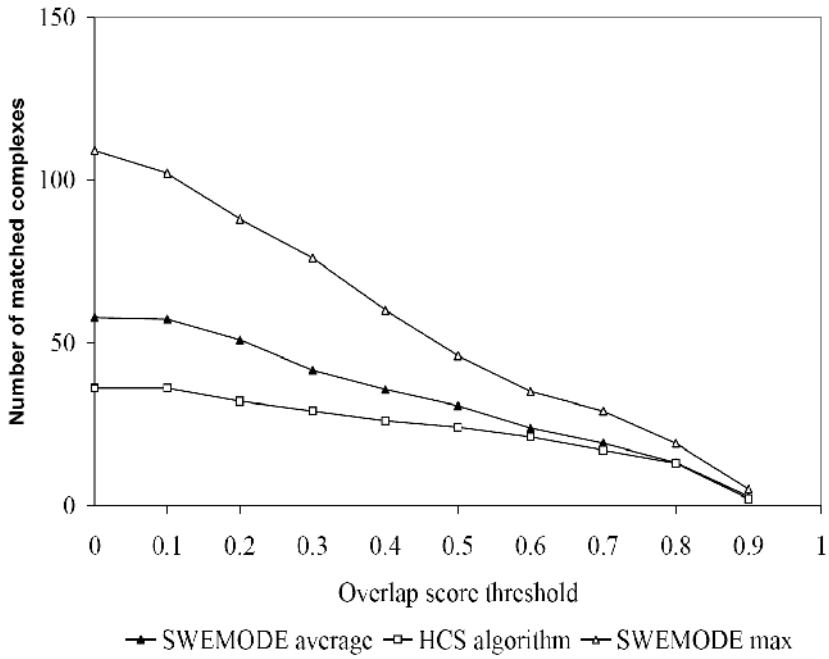


Fig. 5. Comparison between SWEMODE modules and modules generated with HCS clustering method.

5. Conclusion

The focus of attention in this chapter is the knowledge-based method that integrates domain specific knowledge, in this case functional information from Gene Ontology, with topological information, to derive modular structures from PPI networks. There are clear

disadvantages with the approaches that only rely on topological information, as previously described. In contrast to these methods that often suffer from lack of biological plausibility, the approach described here takes into consideration the functional knowledge about the experimental interactions, and in this way strengthen the validity of the obtained modular structures. Modules obtained in this way serve as models for studying interconnectivity, which is a step towards reconstruction of the higher order hierarchy of cellular networks.

Three different biological aspects – molecular function, biological process and cellular component, have been employed and tested for their suitability for deriving modules. The identification of protein complexes may become more challenging as additional PPI data becomes available, because the interactions are noisy, and the integration of PPI data with annotation might prove a useful solution to this problem. The integrated approaches contribute to this solution, by increasing the confidence in high-throughput Y2H data. The approach also provides means for an increased understanding of the higher-order structures underlying cellular function. As annotations become more complete, the increased biological relevance of our module predictions with integrated approaches is expected to be even more evident.

One of the biggest issues in this type of study is the difficulty to clearly characterise modules. There is no generally accepted definition of modules. A pioneering work in this area, performed by Hartwell et al. (1999) provides a wide definition, which leaves space for different authors to define different more specific criteria. This is, as also pointed out in (Schlosser and Wagner, 2004), unavoidable, and "retaining a pragmatic pluralism of different modularity concepts is probably a fruitful strategy for broadening our perspective and illuminating the importance of modularity at many different levels of organization".

A possible future application of the method described in this chapter is identification of modules of genes and proteins involved in various diseases, such as cancer. This module-level knowledge can contribute to the understanding of cancer on system-level, which may be useful for developing new drugs. Cancer-related networks for a specific type of cancer may be derived from, for example, gene expression data. Deriving gene networks makes it possible to apply network theoretic approaches on the interconnected genes that are potentially related to cancer development. Furthermore, a comparative analysis of the cancer-related networks derived from different types of cancer could be performed to identify modules that are shared among different types, but also to identify the specific processes that characterize a certain type of cancer.

Modular analysis may also be applied to identify general properties of the interrelated genes that are involved in the origin of cancer cells. A suitable model for this analysis is a gene fusion network in human neoplasia (Hoglund et al., 2006). By investigating topological properties of the cancer nodes in the network, such as node betweenness centrality, the cancer-related genes that act as "bridges" or communication points between various modules that correspond to cancer related processes may be identified.

Explaining the relationships between structure, function and regulation of molecular networks at different levels of the complexity pyramid of life is one of the main goals in systems biology. By integrating the topology, i.e. various structural properties of the

networks with the functional knowledge encoded in protein annotations, and also analysing the interconnectivity between modules at different levels of the hierarchy, we aim to contribute to this goal. With the increasing availability of protein interaction data and more fine-grained GO annotations, this will help constructing a more complete view of interconnected modules to better understand the organisation of cells.

6. References

Asthana, S., King, O. D., Gibbons, F. D., & Roth, F. P. (2004). Predicting protein complex membership using probabilistic network reliability. *Genome Res* 14, 1170-1175.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101-113.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proc Natl Acad Sci U S A* 101, 3747-3752.

Deane, C. M., Salwinski, L., Xenarios, I., & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 1, 349-356.

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405, 823-826.

Flake, G. W., Lawrence, C., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer* 35, 66-71.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88-93.

Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters* 76, 175-181.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47-52.

Hoglund, M., Frigyesi, A., & Mitelman, F. (2006). A gene fusion network in human neoplasia. *Oncogene* 25, 2674-2678.

Jansen, R., Lan, N., Qian, J., & Gerstein, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics* 2, 71-81.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-453.

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.

Lin, D. (1998). An information-theoretic definition of similarity. *The 15th International Conference on Mashine Learning* (Madison, WI).

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275-1283.

Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins* 64, 948-959.

Luo, F., & Scheuermann, R. H. (2006). Detecting Functional Modules from Protein Interaction Networks. *Proceeding of the First International Multi-Symposiums on Computer and Computational Scineces* (IMSCCS'06) (IEEE Computer Society).

Luo, F., Yang, Y., Chen, C. F., Chang, R., Zhou, J., & Scheuermann, R. H. (2007). Modular organization of protein interaction networks. *Bioinformatics* 23, 207-214.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A* 98, 404-409.

Newman, M. E. J. (2003). Ego-centered networks and the ripple effect. *Socal networks* 25, 83-95.

Oltvai, Z. N., & Barabasi, A. L. (2002). Systems biology. Life's complexity pyramid. *Science* 298, 763-764.

Pereira-Leal, J. B., Enright, A. J., & Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins* 54, 49-57.

Petti, A. A., & Church, G. M. (2005). A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. *Genome Res* 15, 1298-1306.

Poyatos, J. F., & Hurst, L. D. (2004). How biologically relevant are interaction-based modules in protein networks? *Genome Biol* 5, R93.

Przulj, N., Wigle, D. A., & Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics* 20, 340-348.

Qi, Y., & Ge, H. (2006). Modularity and dynamics of cellular networks. *PLoS Comput Biol* 2, e174.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551-1555.

Rives, A. W., & Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A* 100, 1128-1133.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., & Mewes, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32, 5539-5545.

Schlosser, G., & Wagner, G. P. (2004). Modularity in development and evolution: The University of Chicago Press).

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol* 18, 1257-1261.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34, 166-176.

Spirin, V., & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 100, 12123-12128.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., & Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190-193.

Titz, B., Schlesner, M., & Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics* 1, 111-121.

Uhrig, J. F. (2006). Protein interaction networks in plants. Planta 224, 771-781.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440-442.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res* 28, 289-291.

Yook, S. H., Oltvai, Z. N., & Barabasi, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* 4, 928-942.

5

Mining Protein Interaction Groups

Lusheng Wang
Department of Computer Science
City University of Hong Kong
Hong Kong

1. Introduction

Proteins with interactions carry out most biological functions within living cells such as gene expression, enzymatic reactions, signal transduction, inter-cellular communications and immunoreactions. As the interactions are mediated by short sequence of residues among the long stretches of interacting sequences, these interacting residues or so-called interaction (binding) sites are at the central spot of proteome research. Although many imaging wet-lab techniques like X-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy and mass spectrometry have been developed to determine protein interaction sites, the solved amount of protein interaction sites constitute only a tiny proportion among the whole population due to high cost and low throughput. Computational methods are still considered as the major approaches for the deep understanding of protein binding sites, especially for their subtle 3-dimensional structure properties that are not accessible by experimental methods.

The classical graph concept—maximal biclique subgraph (also known as maximal complete bipartite subgraph)—has been emerged recently for bioinformatics research closely related to topological structures of protein interaction networks and biomolecular binding sites. For example, Thomas *et al.* introduced complementary domains in (Thomas et al., 2003), and they showed that the complementary domains can form near complete bipartite subgraphs in PPI networks. A lock-and-key model has been proposed by Morrison *et al.* which is also based on the concept of maximal complete bipartite subgraphs (Morrison et al., 2006). Very recently, Andreopoulos *et al.* used clusters in PPI networks for identifying locally significant protein mediators (Andreopoulos et al., 2007). Their idea is to cluster common-friend proteins, which are in fact complete-bipartite proteins, based on their similarity to their direct neighborhoods in PPI networks. Other computational methods studying bipartite structures of PPI networks include (Bu et al., 2003; Hishigaki et al., 2001) which are focused on protein function prediction.

To identify motif pairs at protein interaction sites, Li *et al.* introduced a novel method with the core idea related to the concept of complete bipartite subgraphs from PPI networks (Li et al., 2006). The first step of the algorithm in (Li et al., 2006) finds large subnetworks with all-versus-all interactions (complete bipartite subgraphs) between a pair of protein groups. As the proteins within these protein groups have similar protein interactions and may share the same interaction sites, the second step of Li's algorithm is to compute conserved motifs

(possible interaction sites) by multiple sequence alignments within each protein group. Thus, those conserved motifs can be paired with motifs identified from other protein groups to model protein interaction sites. One of the novel aspects of the algorithm in (Li et al., 2006) is that it combines two types of data: the PPI data and the associated sequence data for modeling binding motif pairs.

Each protein in the above PPI networks is represented by a vertex and every interaction between two proteins is represented by an edge. Discovering complete bipartite subgraphs in PPI networks can thus be formulated as the following biclique problem: Given a graph, the biclique problem is to find a subgraph which is bipartite and complete. The objective is to maximize the number of vertices or edges in the bipartite complete subgraph. We note that the maximum vertex biclique problem is polynomial time solvable (Yannakakis, 1981). This problem is also equivalent to the maximum independent set problem on bipartite graphs which is known to be solvable by a minimum cut algorithm. However, the maximum vertex balanced biclique problem is NP-hard (Garey & Johnson, 1979). The maximum edge biclique problem is proved to be NP-hard as well (Peeters, 2003).

In this paper, we consider incompleteness of biological data, as the interaction data of PPI networks is usually not fully available. On the other hand, within an interacting protein group pair, some proteins in one group may only interact with a proportion of the proteins in the other group. Therefore, many subgraphs formed by interacting protein group pairs are not perfect bicliques. They are more often near complete bipartite subgraphs. Therefore, methods of finding bicliques may miss many useful interacting protein group pairs. To deal with this problem, we use quasi-bicliques instead of bicliques to find interacting protein group pairs. With the quasi-biclique, even though some interactions are missing in a protein interaction subnetwork, we can still find the two interacting protein groups. In this paper, we introduce and investigate the maximum vertex quasi-biclique problem. We show that the problem is NP-hard. We also propose approximation and heuristic algorithms for finding large quasi-bicliques in PPI networks. The applications for finding protein-protein binding sites are illustrated.

2. Bicliques and quasi-bicliques

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, where each vertex represents a protein and there is an edge connecting two vertices if the two proteins have an interaction. Since $\mathcal{G}$ is an undirected graph, any edge $(u, v) \in \mathcal{E}$ implies $(v, u) \in \mathcal{E}$. For a selected edge (u, v) in $\mathcal{G}$, in order to find the two groups of proteins having the similar pairs of binding sites, we translate the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into a bipartite graph. Let $X = \{x|(x, v) \in \mathcal{E}\}$, $Y_1 = \{y|(u, y) \in \mathcal{E} \& y \notin X\}$ and $Y_2 = \{w|(u, w) \in \mathcal{E} \& w \in X\}$. For a vertex $w \in Y_2$, w is incident to both u and v in $\mathcal{G}$. Thus both X and Y_2 contain w. We keep w in X and replace w in Y_2 with a new virtual vertex $\overline{w}$. After replacing all vertices w in Y_2 with $\overline{w}$, we get a new vertex set $\overline{Y_2}$. Let $Y = Y_1 \cup \overline{Y_2}$ and $E = \{(x, y)|(x, y) \in \mathcal{E} \& x \in X \& y \in Y_1\} \cup \{(x, \overline{w})|(x, w) \in \mathcal{E} \& x \in X \& \overline{w} \in \overline{Y_2}\}$. In this way, we have a bipartite graph $G = (X \cup Y, E)$. A biclique in G corresponds to two subsets of vertices, say, subset A and subset B, in $\mathcal{G}$. In $\mathcal{G}$, every vertex in A is adjacent to all the vertices in B, and every vertex in B is adjacent to all the vertices in A. Moreover, $A \cap B$ may not be empty. In this case, for any vertex $w \in A \cap B$, $(w, w) \in \mathcal{E}$. This is the case, where the protein has a self-loop. Self-loops are very common in practice. When a self-loop appears, one protein

molecule interacts with another identical protein molecule. For example, two identical protein subunits can assemble together to form a homodimeric protein.

In the following, we focus on the bipartite graph $G = (X \cup Y, E)$. For a vertex $x \in X$ and a vertex set $Y' \subseteq Y$, the degree of x in Y' is the number of vertices in Y' that are adjacent to x, denoted by $d(x, Y') = |\{y | y \in Y' \& (x, y) \in E\}|$. Similarly, for a vertex $y \in Y$ and $X' \subseteq X$, we use $d(y, X')$ to denote $|\{x | x \in X' \& (x, y) \in E\}|$. Now, we are ready to define the *δ-quasi-biclique*.

Definition 1. *For a bipartite graph $G = (X \cup Y, E)$ and a parameter $0 < \delta \leq \frac{1}{2}$, G is called a δ-quasi-biclique if for each $x \in X$, $d(x, Y) \geq (1 - \delta)|Y|$ and for each $y \in Y$, $d(y, X) \geq (1 - \delta)|X|$.*

Similarly, a δ-quasi-biclique in G corresponds to two subsets of vertices, say, subset A and subset B, in $\mathcal{G}$. In $\mathcal{G}$, every vertex in A is adjacent to at least $(1 - \delta)|B|$ vertices in B, and every vertex in B is adjacent to at least $(1 - \delta)|A|$ vertices in A. Moreover, according to the translation and the definition, $A \cap B$ may not be empty. Again, if a protein appears in both sides of a δ-quasi-biclique and there is an edge between the two corresponding vertices, the protein has a self-loop. In our experiments, we observe that about 22% of the δ-quasi-bicliques produced by our program contain self-loop proteins.

In many applications, due to various reasons, some edges in a clique/biclique may be missing and a clique/biclique becomes a quasi-clique/quasi-biclique. Thus, finding quasi-cliques/quasi-bicliques is more important in practice. Here we show that large quasi-bicliques may not contain any large bicliques.

Theorem 1. *Let $G = (X \cup Y, E)$ be a random graph with $|X| = |Y| = n$, where for each pair of vertices $x \in X$ and $y \in Y$, (x, y) is chosen, randomly and independently, to be an edge in E with probability $\frac{2}{3}$. When $n \to \infty$, with high probability, G is a $\frac{1}{2}$-quasi-biclique, and G does not contain any biclique $G' = (X' \cup Y', E')$ with $|X'| \geq 2 \log n$ and $|Y'| \geq 2 \log n$.*

In the biological context, Theorem 1 indicates that it is possible that some large interacting protein groups cannot be obtained by simply finding a maximal biclique if a few (interaction) edges are missing. As large interacting protein groups are more useful, according to this theorem, we have to develop new computational algorithms to extract from PPI networks large interacting protein groups which form quasi-bicliques.

In terms of false positive edges, both quasi-biclique and biclique can handle spurious edges very well. If very few spurious edges are added, in most cases, an irrelative protein will not be included in the quasi-bicliques or biclique unless $(1 - \delta)|A|$ spurious edges are simultaneously added to the protein that has no interaction with any of the proteins in A, where A is one of the two interaction groups.

The maximum vertex quasi-biclique problem is defined as follows.

Definition 2. *Given a bipartite graph $G = (X \cup Y, E)$ and $0 < \delta \leq \frac{1}{2}$, the maximum vertex δ-quasi-biclique problem is to find $X' \subseteq X$ and $Y' \subseteq Y$ such that the $X' \cup Y'$ induced subgraph is a δ-quasi-biclique and $|X'| + |Y'|$ is maximized.*

The maximum vertex biclique problem, where $\delta = 0$, can be solved in polynomial time (Yannakakis, 1981). Here we show that the maximum vertex δ-quasi-biclique problem

when $\delta > 0$ is NP-hard. The reduction is from X3C (Exact Cover by 3-Sets), which is known to be NP-hard (Karp, 1972).

Theorem 2. *For any constant integers $p > 0$ and $q > 0$ such that $0 < \frac{p}{q} \leq \frac{1}{2}$, the maximum vertex $\frac{p}{q}$-quasi-biclique problem is NP-hard.*

3. A polynomial time approximation scheme

The following lemma that is originally from (Li et al., 2002) will be repeatedly used in our proofs.

Lemma 1. *Let $X_1, X_2, \ldots, X_n$ be n independent random 0-1 variables, where X_i takes 1 with probability p_i, $0 < p_i < 1$. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = E[X]$. Then for any $0 < \epsilon \leq 1$,*

$$\mathbf{Pr}(X > \mu + \epsilon\, n) < exp(-\frac{1}{3}n\epsilon^2),$$
$$\mathbf{Pr}(X < \mu - \epsilon\, n) \leq exp(-\frac{1}{2}n\epsilon^2).$$

The Main Ideas and Techniques: The problem can be formulated as a quadratic programming problem. We use a random sampling technique and a randomized rounding method to get a good approximate solution for the quadratic programming problem under the conditions that $|X_{opt}| = \Omega(|X|)$ and $|Y_{opt}| = \Omega(|Y|)$. The random sampling technique involves to randomly select $r1 = \Omega(\log |X_{opt}|))$ vertices from X_{opt} when X_{opt} is not known. This can be done when $|X_{opt}| = \Omega(|X|)$ and $|Y_{opt}| = \Omega(|Y|)$.

In order to make sure that $|X_{opt}| = \Omega(|X|)$ and $|Y_{opt}| = \Omega(|Y|)$, we design a combinatorial approach to find a subset $X' \subseteq X$ and a subset $Y' \subseteq Y$ such that $|X'| = \Omega(|X_{opt}| + |Y_{opt}|)$, $|X' \cap X_{opt}| \geq (1-\epsilon)|X_{opt}$, $|Y'| = \Omega(|X_{opt}| + |Y_{opt}|)$ and $|Y' \cap Y_{opt}| \geq (1-\epsilon)|Y_{opt}|$. See Lemma 2. Thus, we can work on a bipartite graph induced by X' and Y'. Without loss of generality, we can assume that $|Y_{opt}| \geq |X_{opt}|$. Now, two subcases arise: Case 1: $|X_{opt}| \leq \epsilon|Y_{opt}|$, and Case 2: $|X_{opt}| > \epsilon|Y_{opt}|$. For case 1, we can use linear programming approach and a brute-force approach to solve the problem. For case 2, we can use the quadratic programming approach to solve the problem.

Let $G = (X \cup Y, E)$ be the input bipartite graph. Let $X_{opt} \subseteq X$ and $Y_{opt} \subseteq Y$ be the optimal biclique for the maximum quasi-biclique problem. Without loss of generality, we can assume that

Assumption 1: $|Y_{opt}| \geq |X_{opt}|$.

The basic idea of our algorithm is to (1) formulate the problem into a quadratic programming problem and (2) use a random sampling approach to approximately solve the problem. In order to make the random sampling approach work, we have to make sure that

$$|X_{opt}| = \Omega(|X|) \tag{1}$$

and

$$|Y_{opt}| = \Omega(|Y|). \tag{2}$$

However, for any input bipartite graph $G = (X \cup Y, E)$, there is no guarantee that (1) and (2) hold. Here we propose a method to find a subset X' of X and Y' of Y such that for any $t > 0$, $|X_{opt}| = \Omega(|X'|)$, $|X_{opt} \cap X'| \geq \frac{t-1}{t}|X_{opt}|$, $|Y_{opt}| = \Omega(|Y'|)$, and $|Y_{opt} \cap Y'| \geq \frac{t-1}{t}|Y_{opt}|$. If we can obtain this kind of X' and Y', then we can work on the induced bipartite graph $G' = (X' \cup Y', E')$, where $E' = \{(u,v)|u \in X', v \in Y' \text{ and } (u,v) \in E\}$. Obviously, any good approximate solution of G' is also a good approximate solution of G.

Let x_i be a vertex in the bipartite graph $G = (X \cup Y, E)$. Define $D(x_i, Y)$ to be the set of vertices in Y that are incident to x_i. The following lemma tells us how to obtain X' and Y'.

Lemma 2. *For any $t > 0$, there exist k vertices $x_1, x_2, \ldots, x_k$ in X for $k = \lceil \delta t \rceil$ such that $|\bigcup_{i=1}^{k} D(x_i, Y)| \leq k(|Y_{opt}| + |X_{opt}|)$ and $|Y_{opt} \cap \bigcup_{i=1}^{k} D(x_i, Y)| \geq \frac{t-1}{t}|Y_{opt}|$. Similarly, there exists k vertices $y_1, y_2, \ldots, y_k$ in Y for $k = \lceil \delta t - 1 \rceil$ such that $|\bigcup_{i=1}^{k} D(y_i, X)| \leq k(|Y_{opt}| + |X_{opt}|)$ and $|X_{opt} \cap \bigcup_{i=1}^{k} D(y_i, X)| \geq \frac{t-1}{t}|X_{opt}|$.*

Though we do not know which k vertices in X we should choose, we can try all possible size k subsets of X in $O(|X|^k)$ time for constant k. The value of k is $\lceil \delta t \rceil$ and is determined by t later. Thus, from now on, we assume that the k vertices $x_1, x_2, \ldots, x_k$ are known. Let $X' = \bigcup_{i=1}^{k} D(y_i, X)$ and $Y' = \bigcup_{i=1}^{k} D(x_i, Y)$. We will focus on finding a quasi-biclique in the sub-graph $G' = (X' \cup Y', E')$ of G induced by X' and Y'.

Let $X'_{opt} \subseteq X'$ and $Y'_{opt} \subseteq Y'$ be a quasi-$(\delta + \frac{1}{t})$-biclique with maximum number of vertices in G'. From Lemma 2, $|X'_{opt}| + |Y'_{opt}| \geq (1 - \frac{1}{t})(|X_{opt}| + |Y_{opt}|)$ since $X' \cap X_{opt}$ and $Y' \cap Y_{opt}$ also form a quasi-$\delta + \frac{1}{t}$-biclique of size $(1 - \frac{1}{t})(|X_{opt}| + |Y_{opt}|)$. From now on, we will try to find a good approximate solution for X'_{opt} and Y'_{opt}.

If $|X'_{opt}|$ and $|Y'_{opt}|$ are approximately the same, then we have $|X'_{opt}| = \Omega(|X'|)$ and $|Y'_{opt}| = \Omega(|Y'|)$. That is, (1) and (2) hold for graph G'. Therefore, we can use quadratic programming approach to solve the problem. Nevertheless, there is no guarantee that $|X'_{opt}|$ and $|Y'_{opt}|$ are approximately the same. For any $\epsilon > 0$, we consider two cases.

Case 1: $|X'_{opt}| < \epsilon |Y'_{opt}|$. In this case, the number of vertices in Y'_{opt} will dominate the size of the whole quasi-biclique. If we select a vertex $x \in X'_{opt}$, then x and $D(x, Y')$ form a biclique of size at least $1 + (1 - \delta)|d(x, Y')| \geq 1 + (1 - \delta)|Y'_{opt}|$. When the value of δ is big with respect to ϵ, we do not have the desired quasi-biclique. If we try to add more vertices from Y', we have to guarantee that for every selected vertex y in Y', y is incident to at least $(1 - \delta)|X'|$ selected vertices in X'. This is impossible if x is the only selected vertex from X'. Therefore, we have to consider to add more vertices from both X' and Y'. It is clear that the task here is non-trivial.

In the following lemma, we will show that there exists a subset of r vertices (for some constant r) $X_r \subseteq X'$ and a subset $Y''_{opt} \subseteq Y'_{opt}$ such that X_r and Y''_{opt} form a quasi-$(\delta + \epsilon'')$-biclique with $|Y''_{opt}| \geq (1 - \epsilon'')|Y_{opt}|$ for some $\epsilon'' > 0$. Here r and ϵ'' are closely related.

Lemma 3. *Let $\frac{1}{t} = \epsilon'$. There exists a subset X'_r of X'_{opt} containing $r = \frac{2}{\epsilon'^2} \log(\frac{1}{\epsilon'})$ elements and a subset Y''_{opt} of Y'_{opt} with $|Y''_{opt}| \geq (1 - \frac{r(r-1)}{2|X_{opt}|} - 2\epsilon')|Y'_{opt}|$ such that X'_r and Y''_{opt} form a quasi-$(\delta + \frac{r(r-1)}{2|X'_{opt}|} + 2\epsilon')$-biclique.*

Based on Lemma 3, we can design an algorithm that finds a quasi-$(\delta + 4\epsilon')$-biclique with size at least $(1 - 4\epsilon' - \epsilon)(|X'_{opt}| + |Y'_{opt}|)$. Let $G' = (X' \cup Y', E')$ be the sub-graph obtained from Lemma 2. For any $\epsilon' > 0$, define $r = \frac{2}{\epsilon'^2}\log(\frac{1}{\epsilon'})$.

Case 1.1. $|X'_{opt}| \geq \frac{r(r-1)}{\epsilon'}$: When $|X'_{opt}| \geq \frac{r(r-1)}{\epsilon'}$, $\frac{r(r-1)}{2|X'_{opt}|} \leq \epsilon'$. Thus, there exist a quasi-$(\delta + 3\epsilon')$-biclique $X_r \subset X'$ and Y''_{opt} as described in Lemma 3.

We select r vertices from X'. For each subset $X_r \subseteq X'$ of r vertices $\{v_1, v_2, \ldots, v_r\}$, we define the following integer linear programming. Let $c_{i,j}$ be a constant, where $c_{i,j} = 1$ if $(v_i, u_j) \in E'$; and $c_{i,j} = 0$ if $(v_i, u_j) \notin E'$. Let y_i be a $0/1$ variable, where $y_i = 1$ indicates that the vertex u_i in Y' is selected in the quasi-biclique and $y_i = 0$ otherwise.

$$y_i(\sum_{j=1}^{r} c_{i,j}) \geq (1 - \delta - \frac{1}{t} - \epsilon')r \tag{3}$$

$$\sum_{i=1}^{|Y'|} y_i c_{i,j} \geq (1 - \delta - 3\epsilon')|Y'_{opt}| \text{ for } j = 1, 2, \ldots, r, \tag{4}$$

Here we do not know $|Y'_{opt}|$. However, we can guess the value of $|Y'_{opt}|$ by trying $|Y'_{opt}| = 1, 2, \ldots, |Y'|$. The integer programming problem formulated by (3) and (4) has no objective function and we just want a feasible solution to fit (3) and (4). The integer programming problem is hard to solve. However, we can obtain a fractional solution $\bar{y}_i$ for (3) and (4) with $0 \leq \bar{y}_i \leq 1$ in polynomial time. After obtaining the fractional solution $\bar{y}_i$, we randomly set y_i to be 1 with probability $\bar{y}_i$.

Lemma 4. *Assume that* $\frac{1}{2}(1 - \delta - 3\epsilon')|Y'_{opt}|\epsilon'^2 \geq 2\log r$ *and* $\frac{1}{t} = \epsilon'$. *With probability at least* $1 - \frac{1}{r}$, *we can get a pair of subsets* $X_A \subseteq X'$ *and* $Y_A \subseteq Y'$ *(an integer solution) by randomized rounding according to the probability* $\bar{y}_i$ *such that* X_A *and* Y_A *form a quasi-*$(\delta + 4\epsilon')$*-biclique with* $|X_A| + |Y_A| \geq (1 - \delta - 4\epsilon')|Y'_{opt}|$.

A standard method in (Li et al., 2002) can give a de-randomized algorithm.

When $\frac{1}{2}(1 - \delta - 3\epsilon')|Y'_{opt}|)\epsilon'^2 < 2\log r$, we can enumerate all possible subsets of size $(1 - \delta - 3\epsilon')|Y'_{opt}|$ in Y' in polynomial time to get the desired solution.

Case 1.2. $|X'_{opt}| < \frac{r(r-1)}{\epsilon'}$: In this case, X'_{opt} and Y'_{opt} form the desired quasi-δ-biclique. Instead of selecting r vertices in X', we select $|X'_{opt}|$ vertices in X'. Though we do not know the value of $|x'_{opt}|$, we can guess the value for $|x'_{opt}| = 1, 2, \ldots, \frac{r(r-1)}{\epsilon'}$. We also solve the integer linear programming (3) and (4) in the same way as in Case 1.1. The algorithm for Case 1 is given in Fig. 1.

Theorem 3. *Assume* $|X'_{opt}| \leq \epsilon|Y'_{opt}|$. *We set* $\frac{1}{t} = \epsilon'$ *in the algorithm. With probability at least* $1 - \frac{1}{r}$, *Algorithm 1 finds a quasi-*$(\delta + 4\epsilon')$*-biclique* $X_A \subseteq X$ *and* $Y_A \subseteq Y$ *with* $|X_A| + |Y_A| \geq (1 - \delta - 4\epsilon')(|X_{opt}| + |Y_{opt}|)(1 - \epsilon')/(1 + \epsilon)$ *in time* $O((|X||Y|)^{\lceil \delta t \rceil}[|X||Y||Y'|^{\frac{4\log r}{\epsilon'^2}} + |X'|^{\frac{r(r-1)}{\epsilon'}}\frac{r(r-1)}{\epsilon'}(|X| + |Y|)^3)])$.

Algorithm 1: Algorithm for Solving Case 1: $|X'_{opt}| \leq \epsilon |Y'_{opt}|$.

Input: a bipartite graph $G = (X \cup Y, E)$, a real number $0 \leq \delta \leq 0.5$, a number $t > 0$, a number $\epsilon > 0$, and a number $\epsilon' > 0$.

0. Let $k = \lceil \delta t \rceil$.
1. **for** any $v_1, v_2, \ldots, v_k \in X$ and any $u_1, u_2, \ldots, u_k \in Y$ **do**
2. Set $X' = \cup_{i=1}^{k} D(v_i, Y)$ and $Y' = \cup_{i=1}^{k} D(u_i, X)$.
3 $r == \frac{2}{\epsilon'^2} \log(\frac{1}{\epsilon'})$
4 Guess $|X'_{opt}|$ and $|Y'_{opt}|$ assuming $|X'_{opt}| \leq \epsilon |Y'_{opt}|$.
5 **if** $\frac{1}{2}(1 - \delta - 3\epsilon')|Y'_{opt}|\epsilon'^2 < 2 \log r$ **then** enumerate all possible subsets of size $(1 - \delta - 3\epsilon')|Y'_{opt}|$ in Y' in polynomial time to get the desired solution.
6 **if** $\frac{1}{2}(1 - \delta - 3\epsilon')|Y'_{opt}|)\epsilon'^2 > 2 \log r$) **then**
7 **for** $i = r, r+1, \ldots \frac{r(r-1)}{\epsilon'}$ **do**
8 **for** every i-elements subset $X_i = \{x_1, x_2, \ldots, x_i\}$ **do**
9 give a fractinal solution $\bar{y}_i$ for (3) and (4).
10 randomly set $y_i = 1$ with probability $\bar{y}_i$.
8. Output a $\delta + \frac{1}{t} + 4\epsilon'$ quasi-biclique with the biggest $|X_A| + |Y_A|$.

Fig. 1. The algorithm for solving Case 1.

Case 2: $|X'_{opt}| \geq \epsilon |Y'_{opt}|$. In this case, we have $|X'_{opt}| = \Omega(|X'|)$ and $|Y'_{opt}| = \Omega(|Y'|)$. We will use a quadratic programming approach to solve the problem. We can formulate the quasi-biclique problem for the bipartite graph $G' = (X' \cup Y', E')$ into the following quadratic programming problem.

Quadratic programming formulation:

Let x_i and y_j be 0/1 variables, where $x_i = 1$ indicates that vertex v_i in X' is in the quasi-biclique and $y_j = 1$ indicates that vertex u_j in Y' is in the quasi-biclique. Define $e_{i,j} = 1$ if $(v_i, u_j) \in E'$ and $e_{i,j} = 0$ otherwise. Let c_1 and c_2 be two integers representing the sizes of X'_{opt} and Y'_{opt}, respectively. We can guess the values of c_1 and c_2 in polynomial time though we do not know c_1 and c_2. We have the following inequalities:

$$y_i(\sum_{j=1}^{|X'|} e_{i,j} x_j) \geq (1 - \delta - \frac{1}{t}) y_i c_1 \text{ for } i = 1, 2, \ldots, |Y'| \tag{5}$$

$$x_i(\sum_{j=1}^{|Y'|} e_{i,j} y_j) \geq (1 - \delta - \frac{1}{t}) x_i c_2 \text{ for } i = 1, 2, \ldots, |X'| \tag{6}$$

$$\sum_{i=1}^{|Y'|} y_i = c_1, \tag{7}$$

$$\sum_{i=1}^{|X'|} x_i = c_2. \tag{8}$$

(5) and (6) indicate that $x_i > 0$ and $y_i > 0$ imply that $\sum_{j=1}^{|X'|} e_{i,j} x_j \geq (1 - \delta - \frac{1}{t}) c_1$ and $\sum_{j=1}^{|Y'|} e_{i,j} y_j \geq (1 - \delta - \frac{1}{t}) c_2$, respectively.

Let $\hat{x}_i$ and $\hat{y}_j$ be the 0/1 integer solution for the quadratic programming problem (5)-(8). Let $\hat{r}_i = \sum_{j=1}^{|X'|} e_{i,j}\hat{x}_j$ and $\hat{s}_i = \sum_{j=1}^{|Y'|} e_{i,j}\hat{y}_j$. To deal with the quadratic programming problem, the key idea here is to estimate the values of $\hat{r}_i$ and $\hat{s}_j$. If we know the values of $\hat{r}_i$ and $\hat{y}_j$, then (5) and (6) become

$$y_i\hat{r}_i \geq y_i c_1(1-\delta-\frac{1}{t}) \text{ for } i = 1,2,\ldots,|Y'| \tag{9}$$

$$x_i\hat{s}_i \geq x_i c_2(1-\delta-\frac{1}{t}) \text{ for } i = 1,2,\ldots,|X'|, \tag{10}$$

where $\hat{r}_i$ and $\hat{s}_i$ in (9) and (10) are constants and the quadratic inequalities become linear inequalities.

Estimating $\hat{r}_i$ and $\hat{s}_i$.

The approach for giving a good estimation of $\hat{r}_i$ and $\hat{s}_i$ is to randomly and independently select a subset $B_{X'}$ of $O(\log(|X'_{opt}|))$ vertices and a subset $B_{Y'}$ of $O(\log(|Y'_{opt}|))$ vertices in X'_{opt} and Y'_{opt}, respectively. Let $c_1 = |X'_{opt}|$ and $c_2 = |Y'_{opt}|$. We do not know c_1 and c_2, but we can guess them in $O(|X'| \times |Y'|)$ time. Then we can use $\frac{c_1}{k}\sum_{v_j \in B_{X'}} e_{i,j}$ and $\frac{c_2}{k}\sum_{u_j \in B_{Y'}} e_{i,j}$ to estimate $\hat{r}_i$ and $\hat{s}_i$, respectively. Since we do not know X'_{opt} and Y'_{opt}, it is not easy to randomly and independently select vertices from X'_{opt} and Y'_{opt}. We develop a method to randomly select $p \times \log|Y'|$ vertices in Y'_{opt} from Y' when Y'_{opt} is not known. Here p is a constant to be determined later.

Finding $p\log|Y'|$ vertices in Y'_{opt} when Y'_{opt} is not known

Let $|Y'| = c|Y'_{opt}|$. The idea here is to randomly and independently select a subset B of $(c+1) \times p \times \log|Y'|$ vertices from Y' and enumerate all size $p \times \log|Y'|$ subsets of B in time $C^{p\log|Y'|}_{p(c+1)\log|Y'|} \leq O(|Y'|^{p(c+1)})$. We can show that with high probability, we can get a set of $p\log|Y'|$ vertices randomly and independently selected from Y'_{opt}.

Lemma 5. *With probability at least* $1 - |Y'|^{-\frac{p}{2c^2(c+1)}}$, *$B$ contains a size $p\log|Y'|$ subset of Y'_{opt}.*

Proof. Let us consider the probability that B contains less than $p\log|Y'|$ vertices in Y'_{opt}. Let b be the expected number of vertices in B that are also in Y'_{opt}. Recall that $|Y'| = c|Y'_{opt}|$. If we randomly select a vertex in Y', the probability that the vertex is in Y'_{opt} is $\frac{1}{c}$. Let μ be the expected number of vertices in B that are in Y'_{opt}. We have $\mu = \frac{|B|}{c} = \frac{1}{c}\lceil(c+1)p\log|Y'|\rceil$. Let $X_1, X_2, \ldots, X_{|B|}$ be $|B|$ independent random 0/1 variables, where $X_i = 1$ with probability $\frac{1}{c}$ indicating that the selected vertex is in Y_{opt}. Thus,

$$b = \sum_{i=1}^{|B|} X_i \tag{11}$$

and

$$\mu = E(\sum_{i=1}^{|B|} X_i) = \frac{1}{c}\lceil (c+1)p\log|Y'|\rceil. \quad (12)$$

Since we selected $(c+1)p\log|Y'|$ vertices,

$$|B| = \lceil (c+1)(p\log|Y'|)\rceil. \quad (13)$$

Based on Lemma 1, we have

$$\begin{aligned}
\mathbf{Pr}(b < p\log|Y'|) &\leq \mathbf{Pr}(b < (\frac{1}{c} - \frac{1}{c(c+1)})\lceil (c+1)(p\log|Y'|)\rceil) \\
&= \mathbf{Pr}(\sum_{i=1}^{|B|} X_i < \mu - \frac{1}{c(c+1)}|B|) \quad \text{(From (11), (12) and (13))} \\
&\leq exp(-\frac{1}{2c^2(c+1)^2}|B|) \\
&\leq exp(-\frac{1}{2c^2(c+1)^2}(c+1)(p\log|Y'|)) \\
&= exp(-\frac{p\log|Y'|}{2c^2(c+1)}) = |Y'|^{-\frac{p}{2c^2(c+1)}}.
\end{aligned}$$

Therefore, with probability at most $|Y'|^{-\frac{p}{2c^2(c+1)}}$, B does not contain any size $p\log|Y'|$ subset of Y'_{opt}. This completes the proof. □

Let $B_{X'}$ and $B_{Y'}$ be the sets of randomly and independently selected vertices in X'_{opt} and Y'_{opt}. Let $|B_{X'}| = p_1\log|X'|$ and $|B_{Y'}| = p_2\log|Y'|$. We define $\bar{r}_i = \sum_{v_j\in B_{X'}} e_{i,j}$ and $\bar{s}_i = \sum_{u_j\in B_{Y'}} e_{i,j}$. The following lemma shows that $\frac{c_1}{|B_{X'}|}\bar{r}_i$ and $\frac{c_2}{|B_{Y'}|}\bar{s}_i$ are good approximations of $\hat{r}_i$ and $\hat{s}_i$.

Lemma 6. *With probability at least* $1 - 2|Y'||X'|^{-\frac{\epsilon^2}{3}p_1} - 2|X'||Y'|^{-\frac{\epsilon^2}{3}p_2}$, *for any* $i = 1, 2, \ldots, |X'|$ *and* $j = 1, 2, \ldots, |Y'|$,

$$(1-\epsilon)\hat{r}_i \leq \frac{c_1}{|B_{X'}|}\bar{r}_i \leq (1+\epsilon)\hat{r}_i$$

and

$$(1-\epsilon)\hat{s}_j \leq \frac{c_2}{|B_{Y'}|}\bar{s}_j \leq (1+\epsilon)\hat{s}_j.$$

Now, we set $r_i = \frac{c_1}{|B_{X'}|}\bar{r}_i$ and $s_i = \frac{c_2}{|B_{Y'}|}\bar{s}_j$. We consider the following linear programming problem.

$$y_i r_i \geq y_i c_1(1-\epsilon)(1-\delta) \text{ for } i = 1, 2, \ldots, m, \tag{14}$$

$$x_i s_i \geq x_i c_2(1-\epsilon)(1-\delta) \text{ for } i = 1, 2, \ldots, m, \tag{15}$$

$$\sum_{i=1}^{|Y'|} y_i = c_1, \tag{16}$$

$$\sum_{i=1}^{|X'|} x_i = c_2 \tag{17}$$

$$\sum_{j=1}^{|X'|} e_{i,j} x_j \geq \frac{r_i}{1+\epsilon} \tag{18}$$

$$\sum_{j=1}^{|Y'|} e_{i,j} y_j \geq \frac{s_i}{1+\epsilon}. \tag{19}$$

The term $(1-\epsilon)$ in (14) and (15) ensures that the quadratic programming problem has a solution when the estimated values of r_i and s_i are smaller than $\hat{r_i}$ and $\hat{s_i}$. Similarly, the term $(1+\epsilon)$ in (18) and (19) ensures that the quadratic programming problem has a solution when the estimated values of r_i and s_i are bigger than $\hat{r_i}$ and $\hat{s_i}$.

Randomized rounding

Let x'_i and y'_j be a fractional solution for (14) -(19). In order to get a 0/1 solution, we randomly set x_i and y_j to be 1 using the fractional solution as the probability. That is, we randomly set x_i and y_j to be 1's with probability x'_i and y'_i, respectively. (Otherwise, x_i and y_j will be 0.)

Lemma 7. *With probability* $1 - 2exp(-\frac{1}{3}|X'|\epsilon^2) - 2exp(-\frac{1}{3}|Y'|\epsilon^2) - |Y'|exp(-\frac{1}{2}|X'|\epsilon^2) - |X'|exp(-\frac{1}{2}|Y'|\epsilon^2)$, *we can find a subset* $\hat{X} \subseteq X'$ *and a subset* $\hat{Y} \subseteq Y'$ *with* $(1-\epsilon)c_1 \leq |X'| \leq (1+\epsilon)c_1$ *and* $(1-\epsilon)c_2 \leq |Y'| \leq (1+\epsilon)c_2$ *such that for any* $x \in \hat{X}$, $d(x, Y') \geq (1-\delta-4\epsilon)|\hat{Y}|$ *and for any* $y \in \hat{Y}$, $d(y, X) \geq (1-\delta-4\epsilon)|\hat{X}|$.

The complete algorithm for Case 2 is given in Fig. 2. Let $k = \lceil \delta t \rceil$ as defined in Lemma 2. Here c_x , c_y are set to be $k(1+\frac{1}{\epsilon})$ and $2k$, respectively. $p_1 = p_2 = \frac{5}{\epsilon^2}$.

Theorem 4. *With probability at least* $1 - o(1)$, *Algorithm 2 finds a quasi-*$(\delta + 4\epsilon + \frac{1}{t})$*-biclique of size* $(1 - \frac{1}{t} - \epsilon)(|X_{opt}| + |Y_{opt}|)$ *in* $O((k \times \frac{1}{\epsilon^2}|X||Y|)^{\lceil \delta t \rceil}(|X|^{\frac{5}{\epsilon^2}k(1+\frac{1}{\epsilon})} + |Y|^{\frac{5}{\epsilon^2}2k})(|X| + |Y|^3))$ *time.*

We can derandomize the algorithm to get a polynomial time deterministic algorithm. Step 3 can be derandomized by using the standard method. For instance, instead of randomly and independently choosing $p_1 \log(|X'|)$ and $p_2 \log(|Y'|)$ vertices from X' and Y', we can pick the vertices encountered on a random walk of the same length on a constant degree expander. Obviously, the number of such random walks on a constant degree expander is polynomial. Thus, by enumerating all random walks of length $p_1 \log(|X'|)$ and $p_2 \log(|Y'|)$, we have a polynomial time deterministic algorithm.

Algorithm 2: Algorithm for Soving Case 2: $|X'_{opt}| > \epsilon|Y'_{opt}|$.

Input: a bipartite graph $G = (X \cup Y, E)$, a real number $0 \leq \delta \leq 0.5$, a number $t > 0$ and a number $\epsilon > 0$.

0. Let $k = \lceil \delta t \rceil$, $p_1 = p_2 = \frac{5}{\epsilon^2}$, $c_x = k(1 + \frac{1}{\epsilon})$ and $c_y = 2k$.
1. **for** any $v_1, v_2, \ldots, v_k \in X$ and any $u_1, u_2, \ldots, u_k \in Y$ **do**
2. Set $X' = \cup_{i=1}^{k} D(v_i, Y)$ and $Y' = \cup_{i=1}^{k} D(u_i, X)$.
3 Randomly and independently select a set $S_{X'}$ of $(c_x + 1)p_1 \log |X'|$ vertices in X' and a set $S_{Y'}$ of $(c_y + 1)p_2 \log |Y'|$ vertices in Y'.
4 **for** any size $p_1 \log |X'|$ subsset $B_{X'}$ of $S_{X'}$ and size $p_2 \log |X'|$ subset $B_{Y'}$ of $S_{Y'}$ **do**
   (a) $\bar{r_i} = \frac{c_1}{|B_{X'}|} \sum_{v_i \in |B|_{X'}} e_{i,j}$
   (b) $\bar{s_i} = \frac{c_2}{|B_{Y'}|} \sum_{u_i \in |B|_{Y'}} e_{i,j}$
   (c) Get a fractional solution x'_i and y'_i for $x_i \in X'$ and $y_i \in Y'$ of (11)-(16)
   (d) do randomrized rouding according to x'_i and y'_i
   (e) $X_A = \{v_i | x_i = 1\}$ and $Y_A = \{u_i | y_i = 1\}$
5. Output a $\delta + \frac{1}{t} + 4\epsilon$ quasi-biclique with the biggest $|X_A| + |Y_A|$.

Fig. 2. The algorithm for Case 2.

Step 4 (d) can be derandomized by using Raghavan's conditional probabilities method (Raghavan, 1988). From Case 1 and Case 2, we can immediately obtain the following theorem.

Theorem 5. *There exists a polynomial time approximation scheme that outputs a quasi-biclique $X_A \subseteq X$ and $Y_A \subseteq Y$ with $|X_A| + |Y_A| \geq (1 - \epsilon)(|X_{opt}| + |Y_{opt}|)$ such that any vertex $x \in X_A$ is incident to at least $(1 - \delta - \epsilon)|Y_A|$ vertices in Y_A and any vertex $y \in Y_A$ is incident to at least $(1 - \delta - \epsilon)|X_A|$ vertices in X_A for any $\epsilon > 0$, where X_{opt} and Y_{opt} form the optimal solution.*

4. The heuristic algorithm

In practice, we need to find large quasi-bicliques in PPI networks. Here, we propose a heuristic algorithm to find large quasi-bicliques. Consider a PPI network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Our heuristic algorithm has two steps. First, we construct a bipartite graph from the graph $\mathcal{G}$ based on a pair of interacting proteins (u, v). Using the method described at the beginning of Section 2, we can get a bipartite graph $G = (X \cup Y, E)$. Second, we find quasi-bicliques in G. The bipartite graph G contains all proteins that have interactions with u or v. So we can find large quasi-bicliques containing u and v in the bipartite graph.

In the algorithm for finding quasi-bicliques in G, we have two parameters δ and τ, which control the quality and sizes of the quasi-bicliques. We use a greedy method to get the seeds for finding large quasi-bicliques in G. At the beginning, we set $X' = \phi$ and $Y' = Y$. In each step, we find a vertex with the maximum degree in $X - X'$. The vertex is added into the biclique vertex set X', and we eliminate all vertices y in Y' such that $d(y, X') < (1 - \delta)|X'|$. We will continue this process until the size of Y' is less than τ. At each step, we get a seed for finding large quasi-bicliques.

The seeds may miss some possible vertices in the quasi-bicliques. We can extend the seeds to find larger quasi-bicliques. Let $X'' = X'$ and $Y'' = Y'$ be a pair of seed vertex sets. In the first step, we can find a vertex x in $X - X''$ with the largest degree $d(x, Y'')$ in $X - X''$. If

$d(x, Y'') \geq (1-\delta)|Y''|$, we add the vertex x to X''. In the second step, we can find a vertex y in $Y - Y''$ with the largest $d(y, X'')$ in $Y - Y''$. If $d(y, X'') \geq (1-\delta)|X''|$, we add the vertex y to Y''. We repeat the above two steps until no vertex can be added. The whole algorithm is shown in Fig. 3. We can also exchange the two vertex sets X and Y to find more quasi-bicliques using the algorithm.

Let n be the number of vertices in the bipartite graph G. In the greedy algorithm, the time complexity of Steps $3-5$ and Step 10 is $O(n)$, and the time complexity of Steps $6-9$ is $O(n^2)$. So the time complexity of Steps $3-10$ is dominated by $O(n^2)$. Since Steps $3-10$ is repeated $O(n)$ times, the time complexity of the whole algorithm is $O(n^3)$.

The Greedy Algorithm

Input A bipartite graph $(X \cup Y, E)$ and two parameters δ and τ.

Output A set of δ-quasi-bicliques $(X' \cup Y', E')$ with $|X'| \geq \tau$ and $|Y'| \geq \tau$.

1. Let $X' = \phi$ and $Y' = Y$.
2. **while** $|Y'| \geq \tau$ and $X' \neq X$ **do**
3. Find the vertex $x \in X - X'$ with the maximum degree $d(x, Y')$.
4. Add x into X', $X' = X' \cup \{x\}$, and delete from Y' all vertices $y \in Y'$ such that $d(y, X') < (1-\delta)|X'|$.
5. $X'' = X'$ and $Y'' = Y'$.
6. **repeat**
7. Find the vertex $x \in X - X''$ with the maximum degree $d(x, Y'')$. If $d(x, Y'') \geq (1-\delta)|Y''|$, add x to X'', $X'' = X'' \cup \{x\}$.
8. Find the vertex $y \in Y - Y''$ with the maximum degree $d(y, X'')$. If $d(y, X'') \geq (1-\delta)|X''|$, add y to Y'', $Y'' = Y'' \cup \{y\}$.
9. **until** no vertex is added in the steps 7 and 8.
10. **if** $|X''| \geq \tau$, $|Y''| \geq \tau$, for each $x \in X''$, $d(x, Y'') \geq (1-\delta)|Y''|$, for each $y \in Y''$, $d(y, X'') \geq (1-\delta)|X''|$, output $(X'' \cup Y'')$ as a quasi-biclique.

Fig. 3. The greedy algorithm.

5. Finding motifs from the multiple sequence alignment of computed δ-bicliques.

We implemented the heuristic algorithm described in the last section in JAVA. The software is called PPIExtend. In the implementation, we added a new parameter α to speed up the algorithm. In Step 3, instead of selecting one vertex with the best degree, we can select the best α vertices in $X - X'$ and add all the α vertices into X' in Step 4. As shown in the last step of the algorithm, some vertices in X'' may be adjacent to less than $(1-\delta)|Y''|$ vertices in Y'', but the average degree of the vertices in X'' is no less than $(1-\delta)|Y''|$. Similarly, some vertices in Y'' may be adjacent to less than $(1-\delta)|X''|$ vertices in $|X''|$, but the average degree of the vertices in Y'' is no less than $(1-\delta)|X''|$. In our experiments, these quasi-bicliques are still output to get more useful quasi-bicliques.

Our algorithm for PPIExtend consists of two steps: (i) find interacting protein group pairs (quasi-bicliques) using the greedy algorithm, (ii) find conserved motifs from multiple sequence alignments for each of the protein groups. (We use the existing multiple sequence alignment software PROTOMAT (Pietrokovski, 1996).)

The motifs found by PROTOMAT can be viewed as a *block*, that is a conserved region in a multiple sequence alignment of the proteins in a group. For each biclique X and Y obtained by the greedy algorithm, we use S_X and S_Y to denote the sets of motifs obtained by the multiple sequence alignments of protein sequences in X and Y, respectively. Any pair of motifs (m_1, m_2) with $m_1 \in S_X$ and $m_2 \in S_Y$ is a candidate protein-protein interaction motif pair. Thus, our algorithm can also output lots of motif pairs as candidate protein-protein interaction motif pairs.

We look at the numbers of motifs found by the programs PPIExtend and FPClose* that are also in the two block databases, BLOCKS (Pietrokovski, 1996) and PRINTS (Attwood & Beck, 1994). The LAMA program (Pietrokovski, 1996) is used to find the local optimal alignment of two blocks (the motif output by PPIExtend/FPClose* and a block in the databases), where the Z-score is computed to measure the alignments. The default threshold of Z-score was used in the experiments. The results are reported in Table 1. From this table, we can see that our method has more mappings to BLOCKS and PRINTS than FPClose* (Li et al., 2006; Grahne & Zhu, 2003).

	BLOCKS		PRINTS		BOTH	
	blocks	domains	blocks	domains	blocks	domains
FPClose*	6408/24294	3128/4944	2174/11170	1093/1850	24.1%	62.1%
PPIExtend	9325/29767	4191/6149	2423/11435	1160/1900	28.5%	66.4%

Table 1. The mappings between the motifs and the two databases: BLOCKS and PRINTS. FPClose* uses BLOCKS 14.0 and PRINTS 37.0. Our PPIExtend method uses BLOCKS 14.3 and PRINTS 38.0. Each entry a/b means the motifs are mapped to a blocks(domains) in all b blocks(domains) in the databases.

	BLOCKS	PRINTS	Pfam	*i*Pfam
Version	14.3	38.0	20.0	20.0
Number of domains	6149	1900	8296	2883
Number of entries	29767	11435	8296	3019

Table 2. Databases used in the experiments.

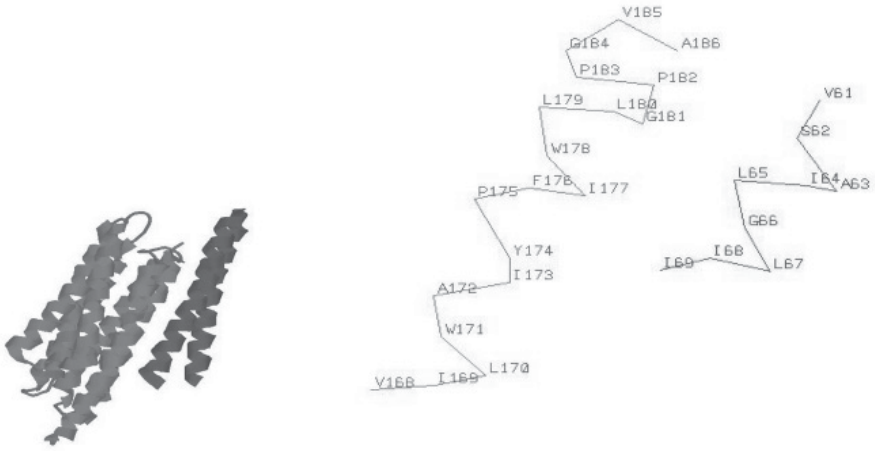
We look at the numbers of motif pairs found by the two programs PPIExtend and FPClose* that can be mapped into domain-domain interaction pairs in the domain-domain interaction database *i*Pfam (Finn et al., 2005). The versions of the databases are shown in Table 2. The *i*Pfam database is built on top of the Pfam database (Sonnhammer et al., 1997) which stores the information of protein domain-domain interactions. To examine whether the motif pairs found by PPIExtend and FPClose* can match some pairs of interacting domains in *i*Pfam, we map our motif pairs to domain pairs in *i*Pfam through the integrated protein family database InterPro (Apweiler et al., 2001) which integrates a number of databases. In fact, we strictly follow the procedure as suggested in (Li et al., 2006). (1) We map our motifs to domains (protein groups) in the database BLOCKS or PRINTS; (2) we map a protein group of BLOCKS to a protein group of InterPro based on the one-to-one mapping between an entry of BLOCKS

and an entry of InterPro; (Note that both PRINTS and Pfam are member databases of InterPro, and the mapping between PRINTS and Pfam is clear.) (3) we use existing cross-links between protein groups of InterPro and domains of Pfam to determine the crosslinks between the motifs found by PPIExtend/FPClose* and Pfam domains. In this way, we can map our motif pairs into domain pairs with Pfam domain entries. Note that the mapping between motif pairs and domain pairs is not one-to-one.

We observed that the motif pairs found by PPIExtend can map to 81 distinct domain pairs in *i*Pfam. However, only 18 domain pairs were reported in (Li et al., 2006). This is a significant improvement and the main reason is the use of quasi-bicliques. In the 81 domain pairs, 48 pairs are domain-domain interactions on one protein (self-loops) and 33 pairs are domain-domain interactions on different proteins. Although the self-loops form a large portion, we still find many other domain-domain interactions that are not self-loops.

6. Protein interaction sites: a case study

In this section, we present detailed information about binding motif pairs that can be mapped to interacting domain pairs. The first motif pair is derived from a protein group pair in which the left protein group contains 7 proteins and the right protein group contains 10 proteins. There are 66 interactions between the two groups of proteins. Using the hypergeometric probability model, the p-value of the protein group pair is less than 1.57×10^{-191}. PROTOMAT finds two left blocks and two right blocks in this protein group pair. The second left block contains 20 positions and the first right block contains 12 positions. By the mapping method, the positions $1-19$ of the second left block can be aligned with the positions $9-27$ of block IPB001425B in BLOCKS, and the positions $4-12$ of the first right block can be aligned with the positions $1-9$ of block IPB003660A in BLOCKS. Block IPB001425B is in the Bac_rhodopsin domain, and block IPB003660A is in the HAMP domain. See Table 3 for more details. Our binding motif pair can map into the domain pair (PF00672, PF01036) in *i*Pfam. *i*Pfam shows that the HAMP domain interacts with the Bac_rhodopsin domain in protein complexes such as lh2s. 1h2s is the complex of *Natronobacterium pharaonis* sensory rho-dopsin II (sRII) with receptor-binding domain of HtrII. The X-ray structure of 1h2s was obtained at 1.93 Å resolution (Gordeliy et al., 2002) and it provided an atomic picture of the first step of the signal transduction. The interactions in the sRII-HtrII complex have been intensively investigated to find the signal relay mechanism from the receptor to the transducer (Bergo et al., 2005; Inoue et al., 2007; Sudo et al., 2007). The 3D structure of the interactions is shown in Fig. 4(a) and 4(b), which are generated by Protein Explorer (Martz, 2002). The shortest residue-residue distance between the two motifs in a pair is also interesting. In protein complex 1h2s, there are two chains: chain A (1h2s_A) and chain B (1h2s_B). The left motif is located at positions $168-186$ of 1h2s_A, and the right motif is located at positions $61-69$ of 1h2s_B (Table 3). We downloaded the coordinate information of 1h2s from http://www.ebi.ac.uk/msd-srv/msdlite/atlas/summary/1h2s.html, and computed the residue-residue distances between the two motifs. The shortest residue-residue distance is 4.07 Å between atom 1346 of residue 177 in 1h2s_A and atom 2018 of residue 69 in protein 1h2s_B (Fig. 4(b)). The average shortest residue-residue distance is 9.17Å. From these



(a) The 3D structure of 1h2s (asymmetric unit).

(b) The backbone structure of the two motifs in 1h2s.

Fig. 4. (a) The 3D structure (best viewed in color) of the interactions between the Bac_rhodopsin domain and the HAMP domain in 1h2s. The left part is chain A and contains the Bac_rhodopsin domain. The right part is chain B and contains the HAMP domain. (b) The backbone structure of the interactions between segment [168V,186A] in 1h2s_A and segment [61V,69I] in 1h2s_B.

calculation and information, we may conclude that the positions $1-19$ of the second left block and the positions $4-12$ of the first right block are possibly interaction sites.

7. Prediction of binding sites

After obtaining candidate domains (conserved regions) in multiple sequence alignment, we can further verify if a pairs of predicted domains really interact with each other by using some tools for protein binding site prediction. Here we briefly introduce a method originally in (Guo & Wang, 2011). This method assumes that the 3D structures of the two given proteins are known.

Given two complete protein structures, the task is to find the binding sites between the two proteins. The method contains three steps. Firstly, we do local sequence alignment at the atom level to get the alignments of conserved regions. Those alignments of conserved regions may contain some gaps. Secondly, among the conserved regions obtained in Step 1, we use the 3D structure information to identify the surface segments. Finally, for any pair of the surface segments identified in Step 2, we compute a rigid transformation to compare the similarity of the two substructures in 3D space and output the qualified pairs as binding sites. When computing the rigid transformations, we treat each protein as a molecule with some volume and introduce a method to ensure that the two whole protein 3D structures have no overlap under such a rigid transformation in 3D space. The software package is available at http://sites.google.com/site/guofeics/bsfinder.

```
AC  l18493xB;
    distance from previous block=(4,396)
DE  none BL  IIK motif=[6,0,17] motomat=[1,1,-10]
    width=20 seqs=7
 DIP:8095N  ( 206) VIGILIISYTKATCDMLAGK
 DIP:4973N  ( 536) MILILIAQFWVAIAPIGEGK
 DIP:5150N  ( 417) LIKDEINNDKKDNADDKYIK
 DIP:5371N  ( 384) IILALIVTILWFMLRGNTAK
 DIP:676N   ( 402) VIVAWIFFVVSFVTTSSVGK
 ...
 pdb 1h2s_A ( 168) VILWAIYPFIWLLGPPGVA
 Bac_rhodopsin:    VVLWLAYPVVWLLGPEGIG

AC  r18493xA;
    distance from previous block=(7,177)
DE  none BL  LLL motif=[6,0,17] motomat=[1,1,-10]
    width=12 seqs=8
 DIP:7371N  (  10) LALIILYLSIPL
 DIP:8128N  (  35) LSLRFLALIFDL
 DIP:4176N  ( 106) LVLTSLSLTLLL
 DIP:7280N  (  11) LSLFLPPVAVFL
 DIP:5331N  ( 178) LSFFVLCGLARL
 ...
 pdb 1h2s_B (  61)    VSAILGLII
 HAMP:                IALLLALLL
```

Table 3. Left block l18493xB aligning with the Bac_rhodopsin domain and right block r18493xA aligning with the HAMP domain. For brevity, only 5 sequences in each of the two blocks are shown. In line Bac_rhodopsin and line HAMP, each letter is the amino acid with the highest frequency in the corresponding column in the multiple alignment. Pdb 1h2s_A and pdb 1h2s_B are chain A and chain B in protein complex 1h2s, respectively.

8. Conclusion

We have proposed algorithms for finding the maximum vertex quasi-biclique problem. We illustrate the applications of the proposed algorithms for finding protein-protein binding sites. The general approach contains three steps: (1) find quasi-bicliques from PPI networks; (2) do multiple sequence alignment for each of the groups in the quasi-biclique and identify possible domains on the protein sequences. (3) use other methods, e.g., the one in (Guo & Wang, 2011), to further confirm the binding sites.

9. References

Yannakakis, M. (1981). Node deletion problems on bipartite graphs. *SIAM Journal on Computing*, Vol. 10, 1981, 310–327.

Thomas, A.; Cannings, R.; Monk, N. A. M. & Cannings, C. (2003). On the structure of protein-protein interaction networks. *Biochemical Society Transanctions*, Vol. 31, Dec 2003, 1491–1496.

Morrison, J. L.; Breitling, R.; Higham, D. J. & Gilbert, D. R. (2006). A lock-and-key model for protein-protein interactions. *Bioinformatics*, Vol. 22, No. 16, Aug 2006, 2012–2019.

Andreopoulos, B.; An, A.; Wang, X.; Faloutsos, M. & Schroeder, M. (2007). Clustering by common friends finds locally significant proteins mediating modules. *Bioinformatics*, Vol. 23, No. 9, 2007, 1124–1131.

Bu, D.; Zhao, Y.; Cai, L.; Xue, H.; Zhu, X.; Lu, H.; Zhang, J.; Sun, S.; Ling, L.; Zhang, N.; Li, G. & Chen, R. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, Vol. 31, No. 9, May 2003, 2443–2450.

Hishigaki, H.; Nakai, K.; Ono, T.; Tanigami, A. & Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, Vol. 18, No. 6, Apr 2001, 523–531.

Li, H.; Li, J.; Wong, L. (2006). Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, Vol. 22, No. 8, 2006, 989–996.

Garey, M. R. & Johnson, D.S. (1979). *Computers and Intractability, A guide to the theory of NP-completeness*. Freeman, San Francisco, 1979.

Peeters, R. (2003). The maximum edge biclique problem is NP-complete. *Discrete Applied Mathematics*, Vol. 131, No. 3, 2003, 651–654.

Karp, R. M. (1972). Reducibility among combinatorial problems. *Complexity of Computer Computations (R. E. Miller and J. W. Thatcher, eds.)*, 1972, 85–103.

Peleg, D.; Schechtman, G. & Wool, A. (1993). Approximating bounded 0-1 integer linear programs, *Proceedings of the 2nd Symposium on Theory of Computing and Systems.* IEEE Computer Society, 1993.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, Vol. 24, 1996, 3836–3845.

Attwood, T. K. & Beck, M. E. (1994). PRINTS-a protein motif fingerprint database. *Protein Engineering, Design and Selection*, Vol. 7, 1994, 841–848.

Finn, R. D.; Marshall, M.; & Bateman,A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, Vol. 21, No. 3, Feb 2005, 410–412.

Grahne, G. & Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. *Proceedings of the Workshop on Frequent Itemset Mining Implementataions (FIMI)*, 2003.

Sonnhammer, E. L.; Eddy, S. R. & Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function and Genetics*, Vol. 28, 1997, 405–420.

Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Birney, E.; Biswas, M.; Bucher, P.; Cerutti, L.; Corpet, F.; Croning, M. D.; Durbin, R.; Falquet, L.; Fleischmann, W.; Gouzy, J.; Hermjakob, H.; Hulo, N.; Jonassen, I.; Kahn, D.; Kanapin, A.; Karavidopoulou, Y.; Lopez, R.; Marx, B.; Mulder, N. J.; Oinn, T. M.; Pagni, M.; Servant, F.; Sigrist, C. J. & Zdobnov, E. M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, Vol. 29, No. 1, Jan 2001, 37–40.

Gordeliy, V. I.; Labahn, J.; Moukhametzianov, R.; Efremov, R.; Granzin, J.; Schlesinger, R.; Büldt, G.; Savopol, T.; Scheidig, A. J.; Klare, J. P. & Engelhard, M. (2002). Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. *Nature*, Vol. 419, No. 6906, Oct 2002, 484–487.

Bergo, V. B.; Spudich, E. N.; Rothschild, K. J. & Spudich, J. L. (2005). Photoactivation perturbs the membrane-embedded contacts between sensory rhodopsin II and its transducer. *Journal of Biological Chemistry*, Vol. 280, No. 31, Aug 2005, 28 365–28 369.

Inoue, K.; Sasaki, J.; Spudich, J. L. & Terazima, M. (2007). Laser-induced transient grating analysis of dynamics of interaction between sensory rhodopsin II D75N and the HtrII transducer. *Biophysical Journal*, Vol. 92, No. 6, Mar 2007, 2028–2040.

Sudo, Y.; Furutani, Y.; Spudich, J. L. & Kandori, H. (2007). Early photocycle structural changes in a bacteriorhodopsin mutant engineered to transmit photosensory signals. *Journal of Biological Chemistry*, Vol. 282, No. 21, May 2007, 15 550–15 558.

Martz, E. (2002). Protein Explorer: easy yet powerful macromolecular visualization. *Trends in Biochemical Sciences*, Vol. 27, No. 3, 2002, 107–109.

Sivaraman, J.; Li, Y.; Banks, J.; Cane, D. E.; Matte, A. & Cygler, M. (2003). Crystal structure of Escherichia coli PdxA, an enzyme involved in the pyridoxal phosphate biosynthesis pathway. *Journal of Biological Chemistry*, Vol. 278, No. 44, Oct 2003, 43 682–43 690.

Sakai, A.; Kita, M. & Tani, Y. (2004). Recent progress of vitamin B6 biosynthesis. *Journal of Nutritional Science and Vitaminology (Tokyo)*, Vol. 50, No. 2, Apr 2004, 69–77.

Fitzpatrick, T. B.; Amrhein, N.; Kappes, B.; Macheroux, P.; Tews, I. & Raschle, T. (2007). Two independent routes of de novo vitamin B6 biosynthesis: not that different after all. *Biochemistry Journal*, Vol. 407, No. 1, Oct 2007, 1–13.

Guo, F. & Wang, L. (2011). Computing the Protein Binding Sites. *ISBRA*, 2011, 25-36.

Liu, X.; Li, J. & Wang, L. (2010). Modeling Protein Interacting Groups by Quasi-Bicliques: Complexity, Algorithm, and Application. *IEEE/ACM Trans. Comput. Biology Bioinform.*, Vol. 7, No. 2, 2010, 354-364.

Wang, L. (2011). Near Optimal Solutions for Maximum Quasi-bicliques, *Journal of Combinatorial Optimization*, on-line available at DOI 10.1007/s10878-011-9392-4.

Li, M.; Ma, B. & Wang, L. (2002). On the closest string and substring problems. *Journal of the ACM*, Vol. 49, No. 2, 2010, 157-171.

Raghavan, P. (1988). Probabilistic construction of deterministic algorithms: Approximate packing integer programs. *JCSS*, Vol. 37, No. 2, 2010, 130-143.

6

Prediction of Combinatorial Protein-Protein Interaction from Expression Data Based on Conditional Probability

Takatoshi Fujiki, Etsuko Inoue,
Takuya Yoshihiro and Masaru Nakagawa
Wakayama University
Japan

1. Introduction

After the entire human DNA sequence was made public, many post-genome researchers began to investigate the systems of living creatures. Creatures consist of vast collections of proteins and their bodies are maintained by complex interactions among genes, proteins, and organic molecules. One major area of interest is how the characteristics of each creature are manifest and what kind of proteins, genes, and their interactions are related to them.

Much research to detect protein–protein interactions has been conducted. The most direct approach to tackle protein–protein interactions is to identify the evidence of the interactions through *in vitro* or *in vivo* experiments. Since several high-throughput experimental methods to detect physical interactions of proteins, such as yeast two-hybrid [1] and tandem affinity purification [2], have been developed, a significant number of protein interactions have been clarified that accelerated the exploration for protein functionality.

As vast amounts of genome sequences became available, computational approaches to infer protein–protein interactions became more focused. They typically assume some hypotheses of biological activity or property, and search biological databases with their own analytical methods for combinations of proteins to satisfy their hypotheses. Initially, many of these methods simply used gene or protein sequences, e.g., the method based on conservation of gene neighborhoods [3], the Rosetta Stone method [4][5], and the sequence-based co-evolution method [6]. Later, as various public databases became available, such as 3D-structures, domains, motifs, pathways, and phylogenetic profiles, various advanced methods to search for protein–protein interactions were developed. These methods and their results are available on the Web [7].

As one computational approach, gene or protein expression-based analysis is widely used to understand gene or protein interactions, which is the focus of this article. These methods were originally developed for microarray experiments that produced gene expression profiles, but they can apply to protein expression data as well. Because we can now obtain the expression profile of genes using high-throughput experiments such as microarray, protein chip, and 2D-electrophoresis, algorithms to derive interactions from expression data

are increasingly valuable. As a basic analysis, the correlation coefficient of expression levels between two proteins is often used to measure the interaction level of protein pairs. (Note that, in this article, we call this type of interaction the *sole effect*, which refers to the effect on a protein from another single protein.) However, since protein interactions have more complex structures, more sophisticated analyses such as Bayesian networks [8] have been used to understand *combinatorial effects* among proteins. A Bayesian network provides the optimal network computed from a set of expression data, which shows the landscape of interaction effects among proteins. Although this network does not infer direct physical interactions, it helps us gain a better understanding of protein functions. However, since the process of Bayesian network analysis considers the sole effects and the combinatorial effects together, it cannot recognize the combinatorial effects alone.

In this article, we treat interactions among three proteins. We derive the combinatorial effect level, which emerges only when the three proteins are together, besides the sole effects that emerge between two proteins. The combinatorial effect level is estimated in a statistical manner, which will lead to a better understanding of protein interactions and a guide to deeper investigations.

The remainder of this paper is organized as follows. In Section 2, we describe related work to understand the current state of the art in this research area. In Section 3, we describe the model of protein–protein interactions used in our method, and present the method to retrieve the combinatorial effect of three proteins. In Section 4, we evaluate our method by applying it to real protein expression data, and finally in Section 5 present the conclusions.

2. Related work

In this section, we give a short introduction of the major approaches used to predict protein–protein interactions.

Many computational methods to predict protein–protein interactions have been proposed. They utilize various kinds of public data such as genome sequences, amino-acid sequences, pathways, domains, 3D-structures, motifs, and phylogenetic profiles, to identify a property of protein pairs in order to predict protein–protein interactions. One typical genome-sequence-based technique is based on conservation of gene neighbourhood [3]. This technique assumes that genes with similar functions or genes that are in the same pathways are transcribed together as a single unit known as an operon. Thus, finding two proteins that are neighbours in several genomes infers that they interact or have similar functions. Another typical sequence-based technique is called the Rosetta Stone method [4][5]. This method is based on the fact that several pairs of proteins interacting with each other have their homologs in other single proteins, called Rosetta Stone proteins. The phylogenetic profile method [6] uses a series of gene sequences in evolution and detects the set of genes that are simultaneously present or absent in the sequences. Since proteins in interaction tend to disappear simultaneously, finding the set of such genes predicts that the corresponding proteins interact. In addition, the in silico two-hybrid system [9] provides a fully alignment-based protein–protein interaction prediction. This technique tries to detect physical interaction of proteins within their 3D structures by means of correlation of sequences of sites among target proteins. Recently, docking analysis using 3D structures of proteins has progressed rapidly. The main difficulty in docking analysis is that there are many potential

ways in which proteins can interact, and protein surfaces are flexible. Currently, one of the major approaches is a global search based on fast Fourier Transform [10]. Including the methods introduced in this brief discussion, there are a tremendous number of techniques to predict protein–protein interactions, and their algorithms and results are available in public databases. For more details, see [7][11].

Boolean networks [12] and Bayesian networks [8] are well known as computational methods to predict interactions from expression data. It is important to note that they treat gene interactions rather than protein interactions since most of them originally suppose microarray data as their source of analysis. However, they can also treat protein expression data.

A Boolean network [12] is a network that represents causal association and it is typically generated from a pattern of time-series expression data. In Boolean networks, a set of expression levels for a sample at time *t* is regarded as "state" at some time *t*, where each expression level is typically represented by "1 (expressed)" or "0 (not expressed)." To compute the network, the time-series state transition is analyzed to learn the functions to determine the state at time *t+1* from the current state at time *t*. As a result, an expression level of a protein at time *t+1* is determined depending on the expression level of several proteins at time *t*. This dependency indicates the protein–protein interaction, although it does not always indicate a direct interaction. There are several versions and extensions of Boolean networks. Akutsu et al. proposed a model and an algorithm of Boolean networks that is generated from non-time-series expression data [13]. Laubenbacher et al. proposed multistate Boolean networks [14]. However, these models cannot treat noise and, thus, often fail in computing networks. To overcome this problem, Shumulevich et al. proposed a model of probabilistic Boolean networks [15] that enables Boolean networks to apply to practical real expression data that includes noise.

A Bayesian network [8] is also a model of interactions often used in computational approaches that is typically built from expression data with discrete expression levels. Bayesian networks represent a joint distribution of random variables, and its direct edge between nodes represents causal association of those nodes. The learning process of a Bayesian network includes the optimization of network topology, where the evaluation of topologies is based on some information criterion, which is typically based on entropy. Note that it evaluates, for each node, the strength of the relationship between the node and its parents in the network, meaning that the sole effects and the combinatorial effects are evaluated together. Later, as an extension of the model, the Dynamic Bayesian network model was proposed [16], which handles time-series expression data. For details of this kind of network learning, there are several survey articles available, such as [17][18].

3. Method to retrieve combinatorial effects

3.1 Expression data used in our method

In this section, we explain the typical representation of protein expression data. Protein expression data represents the expression level of each protein *i* in sample *j*. Typically, the number of proteins in the data are several hundreds to thousands while the number of samples is usually several tens and at most hundreds.

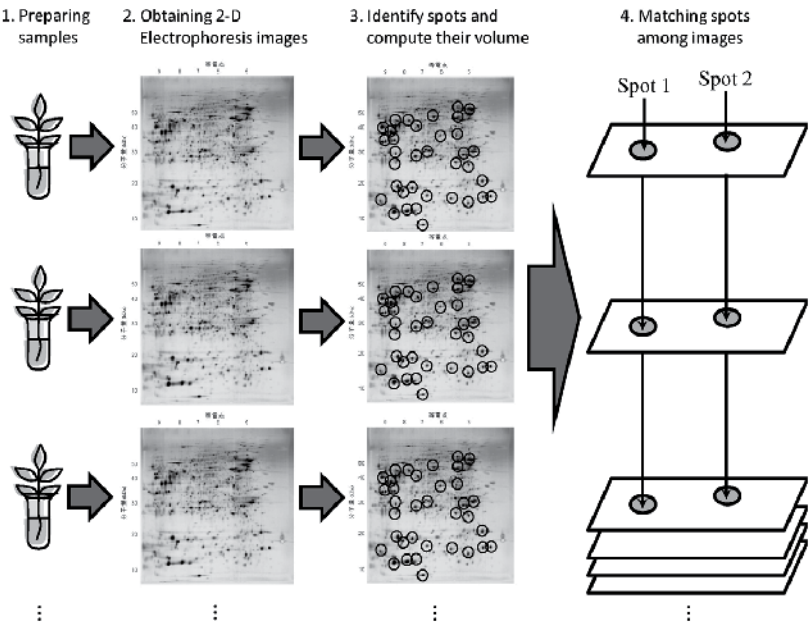


Fig. 1. The process of obtaining Proteome Expression Data.

Sample ID	Protain ID				
	A	B	C	D	…
1	0.000582	0.000107	0.000338	0.000451	…
2	0.000563		0.000475	0.000458	…
3	0.000495	0.000126	0.000433	0.000565	…
4	0.000553	0.000153	0.000382	0.000486	…
5	0.000536	0.000134	0.000536	0.000471	…
6	0.000601	0.000185	0.000457	0.000513	…
:	:	:	:	:	:

Fig. 2. The Data Format for Our Data Mining Process.

Protein expression data is obtained from several methods or devices such as protein arrays, 2D electrophoresis, and mass spectrometry. Among these, we now introduce a 2D electrophoresis-based method [19] as a typical way of generating protein expression data. The process of obtaining protein expression data is somewhat complicated compared to microarray data that measures gene expression levels (see Figure 1). First, we prepare target samples and obtain 2D electrophoresis images from each target sample through an experimental biological process. Second, we identify areas (in the rest of this article we call them *spots*) of separated proteins using image-processing software and measure the expression level of each spot. Third, we match the spots among different images such that the matched spots indicate the same protein. Finally, we normalize the values of expression levels using a normalization method as a preprocess to the data mining processes. As a result, we have a set of protein expression levels as shown in Figure 2, which shows the expression levels of each protein in each sample.

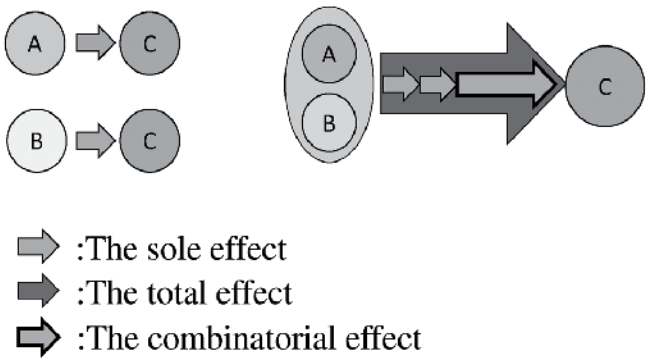


Fig. 3. The Interaction Model to Predict.

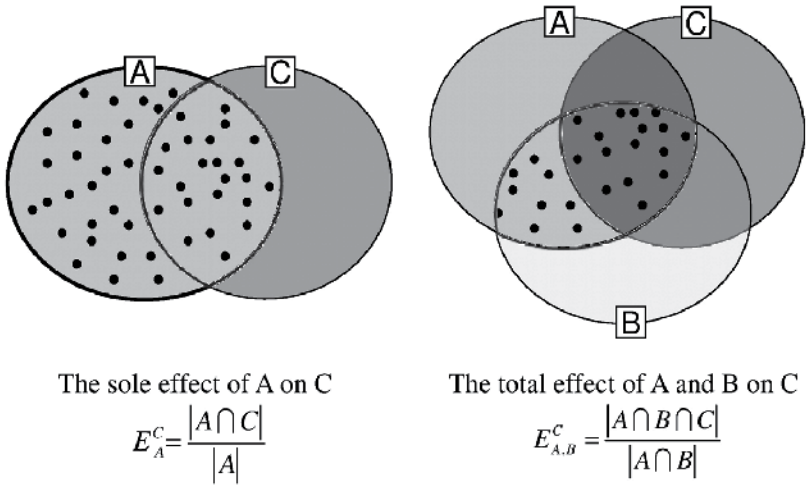


Fig. 4. How to Measure Sole and Total Effect Level of Protein A and B on C.

3.2 Combinatorial protein-protein interaction model

The protein–protein interaction model we try to predict in this paper is shown in Figure 3. Three proteins, A, B, and C, are related to this model, where A and B individually effect the expression level of C, but if both A and B are expressed together, they have a far larger effect on the expression level of C. We call the effect from A to C (resp. B to C) the *sole effect*, and we call the whole effect from A and B on C the *total effect*. Note that the total effect consists of two sole effects and the *combinatorial effect* appears only if both A and B express. What we want to retrieve from expression data is the combinatorial effect of A and B on C.

To measure the combinatorial effect, we first estimate the amount of total effect of A and B on C. Then from the estimated total effect level, we subtract the two sole effects, i.e., the effect of $A - C$ and $B - C$, to obtain the combinatorial effect level.

Note that the three proteins may interact directly or indirectly. We try to extract the three proteins that work in the same functional groups by identifying the behaviour of expression levels following our model of interaction.

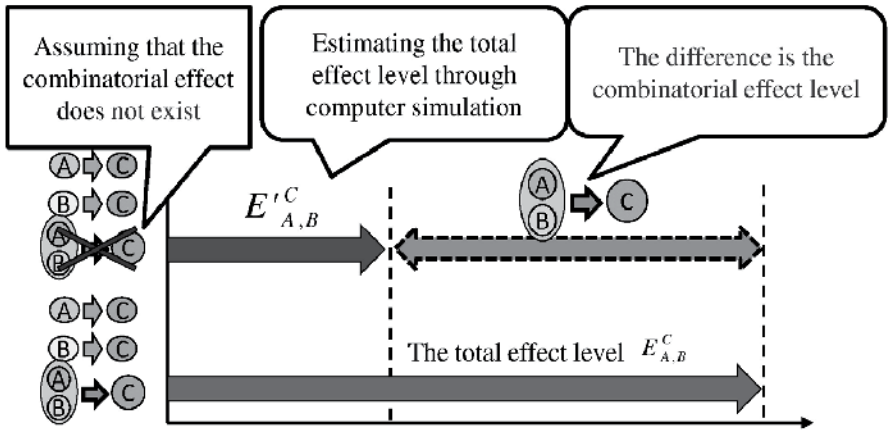


Fig. 5. Dividing Total Effect into Sole and Combinatorial Effect.

3.3 Estimating sole and total interaction levels based on conditional probability

We use conditional probability to retrieve this interaction from expression data. The probability of the sole interactions of $A-C$ and $B-C$ are measured by conditional probability, as shown in Figure 4. Namely, the sole interaction effect level of A on C is measured as the ratio of the number of samples in which the expression levels of both A and C are sufficiently high out of the number of samples in which the expression level of A is sufficiently high. The total interaction effect of A and B on C is also measured in a similar manner, i.e., the ratio of the number of samples in which the expression level of A, B, and C are all sufficiently high out of the number of samples in which the expression levels of both A and B are sufficiently high.

The definitions and formulation of our problems are as follows. We handle proteins i ($1 \le i \le I$) and samples j ($1 \le j \le J$), both of which are included in the input expression data. We also call the proteins A, B, C, ..., and so on. As a parameter, we define r ($0 < r < 1$) as the threshold of the ratio used to judge the expression, i.e., if the expression level of sample j for protein i is within the top r among all the expression levels of protein i, we call the protein i "expressed" in sample j. Let $|A|$ be the number of samples in which protein A is expressed, and similarly, let $|A \cap B|$ be the number of samples in which both protein A and B are expressed. Then, we define $E_A^C = \frac{|A \cap C|}{|A|}$ as the sole effect level of A on C. Similarly, the sole effect level of B on C is defined as $E_B^C = \frac{|B \cap C|}{|B|}$, and the total effect level of A and B on C is defined as $E_{A,B}^C = \frac{|A \cap B \cap C|}{|A \cap B|}$.

3.4 Retrieving combinatorial effect

What we want to estimate is the amount of the combinatorial interaction effect level, which can be estimated from the total interaction level (presented in the previous section) and the sole effect levels of $A-C$ and $B-C$ (see Figure 5). To estimate the combinatorial effect level for the combination of the three proteins A, B, and C, we split the total interaction effect into two parts, i.e., into two sole interaction effects and the combinatorial effect. Then, the

difference between them is regarded as the combinatorial effect level that we wish to compute. To obtain the combinatorial effect level, we compute the statistical distribution of the total effect levels $E'^{C}_{A,B} = \frac{|A \cap B \cap C|}{|A \cap B|}$, which are computed through the simulation executed under the assumption that no combinatorial effect exists over A, B, and C. From the distribution of $E'^{C}_{A,B} = \frac{|A \cap B \cap C|}{|A \cap B|}$ and the total effect score $E^{C}_{A,B} = \frac{|A \cap B \cap C|}{|A \cap B|}$, which is the total effect level presented in the previous subsection, we can estimate the combinatorial effect level.

The computer simulation to compute the distribution of $E'^{C}_{A,B} = \frac{|A \cap B \cap C|}{|A \cap B|}$ is performed as follows. For the corresponding value of α and β, which are the sole effect values for the combination $\mathrm{A} - \mathrm{C}$ and $\mathrm{B} - \mathrm{C}$, we first create distributions of A, B, and C randomly such that the sole effect levels of $\mathrm{A} - \mathrm{C}$ and $\mathrm{B} - \mathrm{C}$ are α and β, respectively. Since those distributions are created randomly, it is possible to assume that they do not include any combinatorial effect. Then we compute the total effect score of the combination A, B, and C. After a sufficient number of repetitions of this process, we obtain the distribution of $E'^{C}_{A,B}$ as the accumulation of the total effect scores. Note that we do not consider what kind of distribution A, B, and C follow in our method since we determine if the protein is expressed using the threshold *r* of the ranking in expression levels.

From this total effect distribution $E'^{C}_{A,B}$, we compute the combinatorial effect as a z-score in the distribution of $E'^{C}_{A,B}$. The z-score $z^{C}_{A,B}$ is defined as $z^{C}_{A,B} = \frac{(E^{C}_{A,B} - \mu)}{\sigma}$, where $E^{C}_{A,B}$ is the total effect level of A, B, and C obtained from the real data, and μ and σ are the average and the standard deviation of the distribution of $E'^{C}_{A,B}$ obtained from the computer simulation, respectively. Namely, the z-score is the difference between the average μ of the distribution of $E'^{C}_{A,B}$ and the real total effect level obtained from the real data, which is measured as the unit value σ. Intuitively, the z-score indicates the probability of the value $E^{C}_{A,B}$ assuming that the combinatorial effect does not exist, which implies the level of the combinatorial effect.

To compute the distribution of the total effect levels through the simulation, however, requires considerable computing time so it is desirable to precompute the distribution. Thus, we prepared a distribution table that shows the average and the standard deviation of the distribution for each value of α and β, as shown in Figure 6. Note that when we compute the distributions in Figure 6, we prepared the data of A, B, and C with 10,000 samples and we perform 5,000,000 trials for each pair of α and β. Because we computed the table for 20 values of α and β between 0 and 1, for obtaining the corresponding values of μ and σ we used the value in the table that is the closest to α and β of A, B, and C.

Now we summarize the proposed method. First, we enumerate every combination of the three proteins A, B, and C from the input data set. For each of the combinations, we compute the total effect level $E^{C}_{A,B}$ of A, B, and C. By referring to the precomputed distribution table, we find the distribution of $E'^{C}_{A,B}$ corresponding to the value α and β of A, B, and C. From the distribution of $E'^{C}_{A,B}$, and the total effect level $E^{C}_{A,B}$, we obtain the combinatorial effect level of A and B on C as the corresponding z-score. Finally, we create a ranking of all the combinations of the three proteins by ordering them by the z-score.

Average table

The sole effect of A on C \ The sole effect of B on C	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.05	0.0000	0.0064	0.0135	0.0212	0.0298	0.0393	0.0500	0.0620	0.0757	0.0913	0.1094	0.1305	0.1556	0.1858	0.2228	0.2694	0.3296	0.4107	0.5255	0.7005	1.0000
0.10	0.0000	0.0135	0.0280	0.0437	0.0609	0.0795	0.1000	0.1225	0.1473	0.1750	0.2059	0.2406	0.2800	0.3251	0.3770	0.4377	0.5093	0.5953	0.7002	0.8315	1.0000
0.15	0.0000	0.0212	0.0437	0.0678	0.0933	0.1207	0.1500	0.1815	0.2154	0.2520	0.2917	0.3348	0.3819	0.4334	0.4901	0.5528	0.6224	0.7002	0.7877	0.8868	1.0000
0.20	0.0000	0.0298	0.0609	0.0933	0.1273	0.1628	0.2000	0.2390	0.2800	0.3231	0.3685	0.4163	0.4668	0.5201	0.5766	0.6365	0.7001	0.7679	0.8401	0.9173	1.0000
0.25	0.0000	0.0393	0.0795	0.1207	0.1628	0.2059	0.2500	0.2953	0.3415	0.3889	0.4376	0.4874	0.5385	0.5910	0.6449	0.7001	0.7569	0.8152	0.8751	0.9367	1.0000
0.30	0.0000	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5001	0.5501	0.6001	0.6501	0.7001	0.7501	0.8001	0.8501	0.9001	0.9500	1.0000
0.35	0.0000	0.0620	0.1225	0.1815	0.2390	0.2952	0.3500	0.4035	0.4559	0.5069	0.5569	0.6057	0.6534	0.7001	0.7457	0.7904	0.8341	0.8769	0.9188	0.9598	1.0000
0.40	0.0000	0.0757	0.1473	0.2154	0.2800	0.3415	0.4000	0.4559	0.5091	0.5601	0.6088	0.6554	0.7001	0.7429	0.7841	0.8236	0.8616	0.8982	0.9334	0.9673	1.0000
0.45	0.0000	0.0913	0.1750	0.2520	0.3231	0.3889	0.4500	0.5069	0.5601	0.6097	0.6563	0.7001	0.7412	0.7801	0.8167	0.8514	0.8843	0.9154	0.9450	0.9732	1.0000
0.50	0.0000	0.1094	0.2059	0.2917	0.3685	0.4376	0.5001	0.5569	0.6088	0.6563	0.7000	0.7405	0.7779	0.8126	0.8449	0.8750	0.9033	0.9297	0.9546	0.9780	1.0000
0.55	0.0000	0.1305	0.2406	0.3348	0.4163	0.4874	0.5501	0.6057	0.6554	0.7001	0.7405	0.7770	0.8106	0.8412	0.8694	0.8954	0.9194	0.9417	0.9625	0.9819	1.0000
0.60	0.0000	0.1556	0.2800	0.3819	0.4668	0.5385	0.6001	0.6534	0.7001	0.7412	0.7779	0.8106	0.8399	0.8667	0.8909	0.9131	0.9334	0.9520	0.9692	0.9852	1.0000
0.65	0.0000	0.1858	0.3251	0.4334	0.5201	0.5910	0.6501	0.7001	0.7429	0.7801	0.8126	0.8412	0.8667	0.8894	0.9100	0.9286	0.9455	0.9609	0.9750	0.9880	1.0000
0.70	0.0000	0.2228	0.3770	0.4901	0.5766	0.6449	0.7001	0.7457	0.7841	0.8167	0.8449	0.8694	0.8909	0.9100	0.9269	0.9423	0.9561	0.9686	0.9800	0.9904	1.0000
0.75	0.0000	0.2694	0.4377	0.5528	0.6365	0.7001	0.7501	0.7904	0.8236	0.8514	0.8750	0.8954	0.9131	0.9286	0.9423	0.9543	0.9655	0.9759	0.9854	0.9925	1.0000
0.80	0.0000	0.3296	0.5093	0.6224	0.7001	0.7569	0.8001	0.8341	0.8616	0.8843	0.9033	0.9194	0.9334	0.9455	0.9561	0.9655	0.9738	0.9815	0.9882	0.9944	1.0000
0.85	0.0000	0.4107	0.5953	0.7002	0.7679	0.8152	0.8501	0.8769	0.8982	0.9154	0.9297	0.9417	0.9520	0.9609	0.9686	0.9754	0.9815	0.9868	0.9917	0.9960	1.0000
0.90	0.0000	0.5255	0.7003	0.7877	0.8401	0.8751	0.9001	0.9188	0.9334	0.9450	0.9546	0.9625	0.9692	0.9750	0.9800	0.9844	0.9882	0.9917	0.9947	0.9975	1.0000
0.95	0.0000	0.7005	0.8315	0.8868	0.9173	0.9367	0.9500	0.9598	0.9673	0.9732	0.9780	0.9819	0.9852	0.9880	0.9904	0.9925	0.9944	0.9960	0.9975	0.9988	1.0000
1.00	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Standard deviation table

The sole effect of A on C \ The sole effect of B on C	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.05	0.0000	0.0024	0.0032	0.0040	0.0046	0.0053	0.0059	0.0065	0.0070	0.0076	0.0082	0.0088	0.0095	0.0103	0.0111	0.0122	0.0137	0.0156	0.0183	0.0209	0.0000
0.10	0.0000	0.0032	0.0048	0.0053	0.0064	0.0071	0.0078	0.0084	0.0090	0.0096	0.0101	0.0107	0.0112	0.0118	0.0125	0.0133	0.0141	0.0150	0.0155	0.0144	0.0000
0.15	0.0000	0.0040	0.0055	0.0069	0.0076	0.0084	0.0090	0.0096	0.0101	0.0105	0.0109	0.0114	0.0117	0.0121	0.0125	0.0129	0.0132	0.0132	0.0127	0.0108	0.0000
0.20	0.0000	0.0046	0.0064	0.0076	0.0087	0.0092	0.0098	0.0103	0.0107	0.0110	0.0113	0.0115	0.0117	0.0119	0.0120	0.0121	0.0120	0.0116	0.0108	0.0085	0.0000
0.25	0.0000	0.0053	0.0071	0.0084	0.0092	0.0099	0.0103	0.0107	0.0110	0.0112	0.0113	0.0114	0.0115	0.0115	0.0114	0.0112	0.0109	0.0102	0.0091	0.0070	0.0000
0.30	0.0000	0.0059	0.0078	0.0090	0.0098	0.0103	0.0107	0.0109	0.0111	0.0112	0.0112	0.0112	0.0111	0.0109	0.0107	0.0103	0.0098	0.0090	0.0078	0.0059	0.0000
0.35	0.0000	0.0065	0.0084	0.0096	0.0103	0.0107	0.0109	0.0111	0.0111	0.0111	0.0110	0.0108	0.0106	0.0104	0.0100	0.0095	0.0089	0.0080	0.0068	0.0050	0.0000
0.40	0.0000	0.0070	0.0090	0.0101	0.0107	0.0110	0.0111	0.0111	0.0113	0.0109	0.0107	0.0104	0.0101	0.0098	0.0093	0.0087	0.0080	0.0071	0.0060	0.0044	0.0000
0.45	0.0000	0.0076	0.0096	0.0105	0.0110	0.0112	0.0112	0.0111	0.0109	0.0112	0.0103	0.0100	0.0096	0.0092	0.0086	0.0080	0.0073	0.0064	0.0053	0.0038	0.0000
0.50	0.0000	0.0082	0.0101	0.0109	0.0113	0.0115	0.0112	0.0110	0.0107	0.0103	0.0109	0.0095	0.0091	0.0085	0.0080	0.0073	0.0066	0.0057	0.0047	0.0033	0.0000
0.55	0.0000	0.0088	0.0107	0.0114	0.0115	0.0114	0.0112	0.0108	0.0104	0.0100	0.0095	0.0104	0.0085	0.0079	0.0073	0.0067	0.0059	0.0051	0.0042	0.0029	0.0000
0.60	0.0000	0.0095	0.0112	0.0117	0.0117	0.0115	0.0111	0.0106	0.0101	0.0096	0.0091	0.0085	0.0096	0.0073	0.0067	0.0061	0.0054	0.0046	0.0037	0.0026	0.0000
0.65	0.0000	0.0103	0.0118	0.0121	0.0119	0.0115	0.0109	0.0104	0.0098	0.0092	0.0085	0.0079	0.0073	0.0085	0.0061	0.0055	0.0048	0.0041	0.0033	0.0023	0.0000
0.70	0.0000	0.0111	0.0125	0.0125	0.0120	0.0114	0.0107	0.0100	0.0093	0.0086	0.0080	0.0073	0.0067	0.0061	0.0075	0.0049	0.0043	0.0036	0.0029	0.0020	0.0000
0.75	0.0000	0.0122	0.0133	0.0129	0.0121	0.0112	0.0103	0.0095	0.0087	0.0080	0.0073	0.0067	0.0061	0.0055	0.0049	0.0062	0.0038	0.0032	0.0025	0.0017	0.0000
0.80	0.0000	0.0137	0.0141	0.0132	0.0120	0.0109	0.0098	0.0089	0.0080	0.0073	0.0066	0.0059	0.0054	0.0048	0.0043	0.0038	0.0050	0.0027	0.0021	0.0015	0.0000
0.85	0.0000	0.0156	0.0150	0.0132	0.0116	0.0102	0.0090	0.0080	0.0071	0.0064	0.0057	0.0051	0.0046	0.0041	0.0036	0.0032	0.0027	0.0039	0.0018	0.0012	0.0000
0.90	0.0000	0.0183	0.0155	0.0127	0.0108	0.0091	0.0078	0.0068	0.0060	0.0052	0.0047	0.0042	0.0037	0.0033	0.0029	0.0025	0.0021	0.0018	0.0025	0.0010	0.0000
0.95	0.0000	0.0209	0.0144	0.0108	0.0085	0.0070	0.0059	0.0050	0.0044	0.0038	0.0033	0.0029	0.0026	0.0023	0.0020	0.0017	0.0015	0.0012	0.0010	0.0013	0.0000
1.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Fig. 6. The Distribution Table of $E'^{C}_{A,B}$ Created through Simulation.

4. Evaluation

4.1 Property of expression data used in our method

In this section, we explain the preprocess applied to the expression data, and also describe the basic property of the data. The expression data used in this experiment originated from the sample of fat near the kidney of black cattle. We performed 2D electrophoresis on each sample and measured the volume of each separated spot that corresponds to each protein. For details of the protocol of the experiment, see [19].

We preprocessed the expression data to improve the reliability of the expression data. Our preprocess consists of the following three steps. First, we removed from the data the

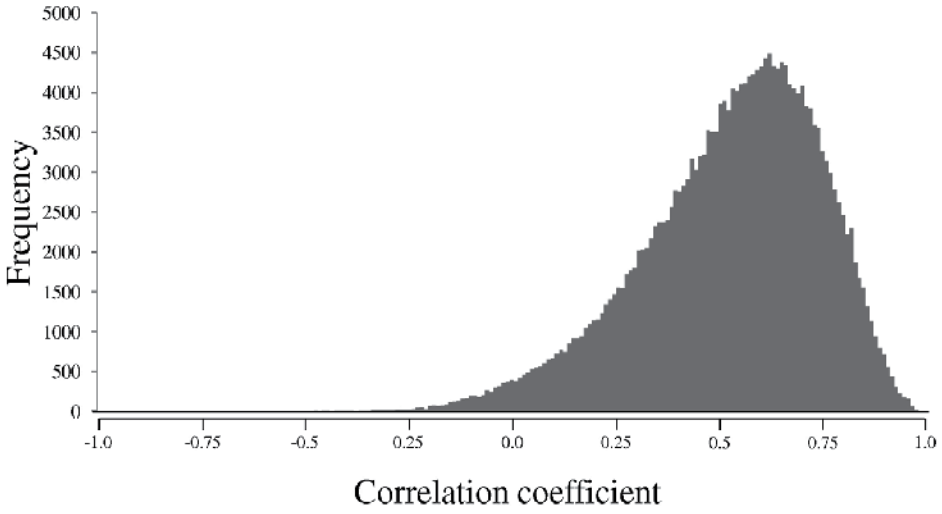


Fig. 7. The histogram of correlation coefficient between proteins.

samples and the proteins that included more than 10% of null expression levels. This was done because samples or proteins with so many null values significantly reduce the reliability of the expression data. Next, we normalized the expression data with the global scaling method [20], where for every sample a scale factor is applied such that the total sum of the protein expression levels in the sample is 1. Finally, we removed the samples with high repetition error. Note that, in fact, in this data set, we performed 2D electrophoresis twice for each sample to confirm the accuracy of each electrophoresis experiment. To maintain the reliability of the data, we removed the sample in which more than 30% of the spots have a high repetation error or null value. Specifically, we consider a spot to have high repetation error if the larger expression level is larger than 1.3 times the value of the smaller expression level. Otherwise, the average of the two expression levels is used for each sample-protein pair. As a result, the expression data used for our evaluation consist of 124 samples and 670 proteins.

In order to indicate a characteristic of this data, we investigated the correlation between proteins. See Figure 7 for the results of calculating correlation coefficients for all pairs of the proteins. Note that the number of pairs is ${}_{670}C_2$ in total. Figure 7 is the histogram where the horizontal axis shows the correlation coefficient separated into classes with 0.05 intervals and the vertical axis shows the frequency of each class. From this result, we can see that most of the correlation coefficients take positive values, and many of them take relatively large values.

4.2 Evaluation experiment of retrieving combinatorial effect

4.2.1 Methods

We performed the experiment to evaluate the performance of the proposed method by applying it to the expression data described in Section 4.1. As a parameter of the experiment, we used the values of 50% and 30% as the threshold r to define the phenomenon that a protein is expressed.

To maintain statistical reliability, we excluded from the analysis the combinations of three proteins where the number of samples was insufficient. Namely, we ignored the

combinations of the three proteins if $|A \cap B|$, which is the denominator in the total effect level $E^C_{A,B}$, was less than 35 in case of r is 50%, and less than 20 in case r is 30%. Similarly, we also removed the combinations if $|A \cap B \cap C|$ was less than 18 in case of r is 50%, and less than 10 in case r is 30%. Furthermore, for the computation, we only used the samples in which all the expression levels of the three proteins are not null.

4.2.2 Results

In this section, we describe the results of the evaluation experiments. Figure 8 shows the histogram of the case of r = 50%, where the horizontal axis indicates the z-scores separated into classes with 0.5 intervals, and the vertical axis indicates the number of combinations in each class. Figure 9 shows the ranking of the top 30 combinations of proteins in terms of z-score. This table includes the columns of the spot numbers of proteins A, B, C, z-score of the combinations, E^C_A and E^C_B (the sole effect levels), $E^C_{A,B}$ (the total effect level), $|A \cap B|$ and $|A \cap B \cap C|$ (the number of samples contained in each phenomenon).

Under the significance level of 1%, we extracted 462,706 combinations in which a strong combinatorial effect is inferred. Here, we caluculate the corresponding p-value to the significance level of 1% using the formula of the Bonferroni correction presented in [21], i.e., p-value $= 1 - e^{\frac{log(1-\gamma)}{n}}$, where n is the number of combinations of three proteins and γ is the significance level. This suggests that if p-value $= 1 - e^{\frac{log(1-0.01)}{149,708,820}} = 6.713 \times 10^{-11}$ or less, the combinatorial effect exists. When the p-value is 6.713×10^{-11}, then the corresponding z-score is 6.423. This is computed as the point in the normal distribution where the probability that the value will become more than the point is p-value $= 6.713 \times 10^{-11}$. Figure 8 shows only the part where the z-score is larger than 6.423. Note that the probability of a z-score larger than 6.423 is only 6.713×10^{-11} if we assume that there is no combinatorial effect. This and the results of Figure 8 imply that our expression data includes many combinations in which the combinatorial effect exists.

Figure 9 shows that most of the sole effects of the shown combinations occur between 0.4 and 0.45, and the total effects occur between 0.45 and 0.55. Moreover, in most of the combinations, $|A \cap B|$ takes values close to 35, which is the threshold value to judge statistical reliability. This implies that combinations of lower $|A \cap B|$ tend to have larger z-scores. Although it is not shown in Figure 9, the combinations of lower ranks have larger values of $|A \cap B|$.

Figures 10 and 11 show the results with r = 30%. Compared to Figure 8, z-scores tend to have lower values. In addition, the number of combinations with z-scores larger than 6.423 decreases to 167,320. Here, 6.423 is the corresponding p-value with the significance level of 1%. In Figure 11, all of the total effects take a value of 1.0 and all of $|A \cap B|$ take a value of 20, which is the threshold value to judge statistical reliability. Furthermore, about 97.8% of the total effects take 1.0 in the retrieved 167,320 combinations. This means that in most of retrieved combinations, protein C is expressed in all the samples in which both proteins A and B are expressed. This appears to be an unusual tendency. Since in the case of 30% the number of samples in the phenomenon "express" is smaller than in the case of 50%, it is possible that the number of samples is not sufficient to ensure a reliable statistical analysis. One of our future projects will be to clarify why this result appears in the case of r = 30%.

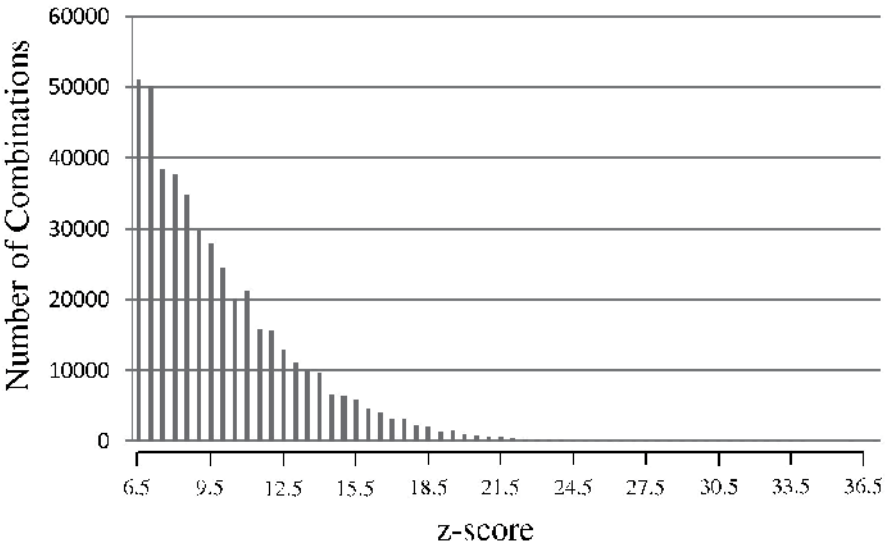


Fig. 8. The histogram of z-score (r=50%).

rank	A (spot No.)	B (spot No.)	C (spot No.)	z-score	E_A^C	E_B^C	$E_{A,B}^C$	|A∩B|	|A∩B∩C|
1	5052	6080	5895	37.3456	0.4583	0.3958	0.5429	35	19
2	3554	5639	5895	36.2082	0.4600	0.4000	0.5429	35	19
3	2742	3554	5895	34.4911	0.4490	0.4490	0.5714	35	20
4	4015	5735	3100	33.7957	0.4348	0.4348	0.5405	37	20
5	5812	5866	1767	33.7458	0.4468	0.4255	0.5429	35	19
6	4798	6080	4849	33.5581	0.4000	0.4000	0.4737	38	18
7	5052	5731	5895	33.4141	0.4468	0.3830	0.5000	36	18
8	5739	6043	4838	33.2666	0.4043	0.4255	0.5000	38	19
9	5812	5866	5895	32.7405	0.4490	0.4286	0.5429	35	19
10	5052	5730	5895	32.6462	0.4375	0.3958	0.5000	36	18
11	3861	6111	5649	32.6423	0.3958	0.3958	0.4615	39	18
12	2318	5940	1765	32.5554	0.4130	0.4348	0.5135	37	19
13	926	5739	5895	32.3921	0.4800	0.4000	0.5429	35	19
14	168	6162	5695	31.9159	0.4667	0.4444	0.5714	35	20
15	5738	6043	3657	31.8151	0.4222	0.4444	0.5278	36	19
16	5639	6242	5895	31.3446	0.3600	0.4400	0.4615	39	18
17	5612	5732	5895	31.3436	0.4375	0.4167	0.5135	37	19
18	6043	6080	4849	31.2987	0.4348	0.3913	0.4865	37	18
19	4015	5735	4838	31.2948	0.4130	0.4565	0.5278	36	19
20	5735	6043	4838	31.2367	0.4348	0.4348	0.5278	36	19
21	4201	5808	3646	31.1739	0.4468	0.4681	0.5714	35	20
22	5726	6242	5895	30.9849	0.4082	0.4490	0.5143	35	18
23	5940	6080	1767	30.7235	0.4255	0.4468	0.5278	36	19
24	2318	4134	5895	30.6533	0.4082	0.5102	0.5714	35	20
25	5734	5866	3467	30.5461	0.4255	0.4043	0.4865	37	18
26	3880	6162	1763	30.5325	0.4444	0.4444	0.5429	35	19
27	5620	5639	5895	30.4974	0.4800	0.3800	0.5135	37	19
28	3554	5621	5895	30.4920	0.4490	0.4694	0.5714	35	20
29	4015	5849	3100	30.4629	0.4222	0.4222	0.5000	36	18
30	5622	5731	5895	30.3865	0.4800	0.4200	0.5526	38	21

Fig. 9. The Top 30 Combinations in z-score (r=50%).

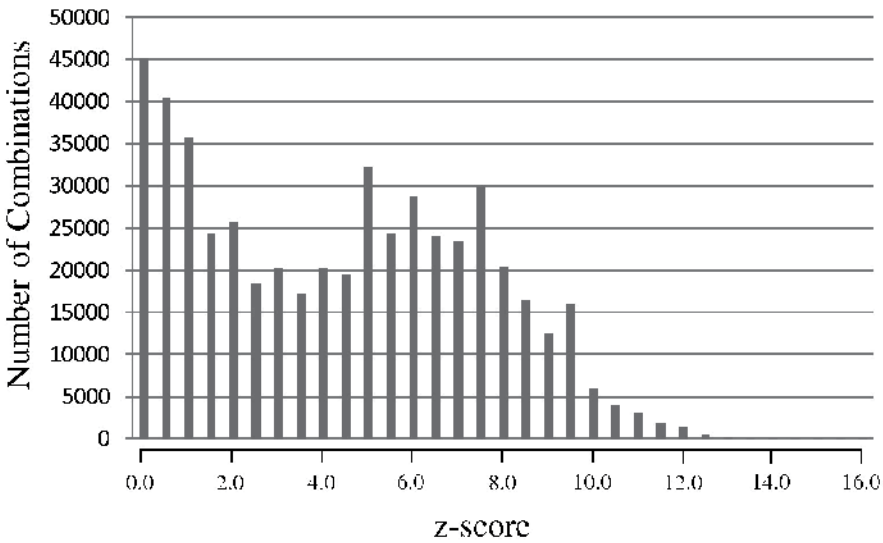


Fig. 10. The histogram of z-score (r=30%).

rank	A (spot No.)	B (spot No.)	C (spot No.)	z-score	E_A^C	E_B^C	$E_{A,B}^C$	|A∩B|	|A∩B∩C|
1	932	4257	4284	16.1614	0.6061	0.6970	1.0000	20	20
2	932	4284	4257	16.1614	0.6061	0.6970	1.0000	20	20
3	934	5140	828	15.9453	0.6176	0.6765	1.0000	20	20
4	932	6240	4134	15.5417	0.7059	0.6176	1.0000	20	20
5	2319	4056	6039	15.5417	0.6176	0.7059	1.0000	20	20
6	934	4284	4257	15.0960	0.6364	0.6970	1.0000	20	20
7	975	4134	4284	15.0960	0.6364	0.6970	1.0000	20	20
8	3998	4795	5045	15.0960	0.6970	0.6364	1.0000	20	20
9	4479	5724	4009	15.0960	0.6970	0.6364	1.0000	20	20
10	5045	5715	5194	15.0960	0.6970	0.6364	1.0000	20	20
11	5573	5954	5218	15.0960	0.6970	0.6364	1.0000	20	20
12	5615	6240	5965	15.0960	0.6970	0.6364	1.0000	20	20
13	2318	4013	5943	15.0958	0.6061	0.7273	1.0000	20	20
14	2318	5943	4013	15.0958	0.6061	0.7273	1.0000	20	20
15	932	4755	5954	14.9425	0.6286	0.7143	1.0000	20	20
16	3972	4755	4476	14.8927	0.7188	0.6250	1.0000	20	20
17	4476	4755	3972	14.8927	0.7188	0.6250	1.0000	20	20
18	4134	6240	5724	14.8927	0.7188	0.6250	1.0000	20	20
19	5724	6240	4134	14.8927	0.7188	0.6250	1.0000	20	20
20	5731	6065	6158	14.8927	0.6250	0.7188	1.0000	20	20
21	5731	6158	6065	14.8927	0.6250	0.7188	1.0000	20	20
22	5733	6065	6158	14.8927	0.6250	0.7188	1.0000	20	20
23	5733	6158	6065	14.8927	0.6250	0.7188	1.0000	20	20
24	934	6240	4134	14.6466	0.7059	0.6471	1.0000	20	20
25	2319	6158	5207	14.6466	0.6471	0.7059	1.0000	20	20
26	5622	5639	5955	14.6466	0.7059	0.6471	1.0000	20	20
27	1762	6034	5965	14.5887	0.7097	0.6452	1.0000	20	20
28	5965	6034	1762	14.5887	0.7097	0.6452	1.0000	20	20
29	1764	3626	5965	14.5887	0.7097	0.6452	1.0000	20	20
30	3626	5965	1764	14.5887	0.6452	0.7097	1.0000	20	20

Fig. 11. The Top 30 Combinations in z-score (r=30%).

4.3 Evaluation experiment of exchangeable proteins

4.3.1 Procedure to exchange proteins

In this section, for the combinations that have high z-scores, we investigate the z-scores when we exchange protein A with protein D in the case where D has a high correlation coefficient with A. Figure 9 shows that many high z-score combinations include C as the common protein, although A and B are also found as common proteins. Since our method defines the samples with the top r expression levels as expressed, having similar z-scores is intuitively inferred if we exchange A with D when D has a high correlation coefficient with A. We believe this is because there are many pairs of proteins in our data set that have a high correlation coefficient allowing us to retrieve so many combinations with a high combinatorial effect. In order to confirm this, we performed an experiment where we exchanged proteins.

The experiment is as follows. First, we create the list of proteins for D that have correlation coefficients against A that are larger than a certain threshold value. Next, we exchange A with D, and calculate the z-score $z^C_{D,B}$ for all combinations of proteins D, B, and C.

4.3.2 Result of exchanging protein

Figure 12 shows the value of the z-scores $z^C_{D,B}$ when A and D are exchanged in the highest z-score combination of A, B, and C in the case r = 50%, where A is exchanged with D if D has the correlation coefficient with A larger than 0.8. This table includes the columns of the spot numbers of proteins A, B, C, protein D exchanged with A, correl(A,D) (the correlation coefficient of A and D), E^C_D (the sole effect level when A and D are exchanged), E^C_B (the sole effect level of before exchanging), $E^C_{D,B}$ (the total effect level), $|D \cap B|$ and $|D \cap B \cap C|$ (the number of samples contained in each phenomenon). In addition, this table is sorted in descending order of z-score.

Figure 12 shows that the lowest z-score as a result of exchanging is 5.503. Note that there are only three combinations that have a z-score less than 6.423, by which the combinatorial effect is inferred under the significance level of 1%. This means that the z-score tends to be high when two proteins with a strong correlation are exchanged. Accordingly, one of the reasons that so many combinations that have a combinatorial effect are retrieved in our data seems to be that our data includes so many pairs of proteins in which the correlation coefficient is high.

5. Conclusion

In this paper, we proposed a method to retrieve the combinatorial protein–protein (or gene-gene) interactions from expression data using statistics of conditional probability. We suppose a model of protein–protein interactions in which the expression level of C takes a large value only if proteins A and B are expressed together. This is the first study to estimate the combinatorial effect level apart from the sole effect. In this study we described our method to treat protein interactions, but note that our method is also applicable to gene expression data generated from microarray experiments.

We evaluated our method using real expression data obtained from a 2D electrophoresis-based experiment. We performed two evaluation experiments with two different parameters, i.e., r = 50% and r = 30%. As a result, the real expression data used in our experiment

A (spot No.)	B (spot No.)	C (spot No.)	D (spot No.)	correl(A,D)	z_score	E_D^C	E_B^C	$E_{D,B}^C$	|D∩B|	|D∩B∩C|
5052	6080	5895	5142	0.8222	30.7904	0.5306	0.4286	0.6129	31	19
			6019	0.9351	27.2615	0.4898	0.3878	0.5143	35	18
			4275	0.8750	26.9008	0.5600	0.4000	0.5938	32	19
			2312	0.8205	26.3425	0.5000	0.4565	0.5882	34	20
			6043	0.8442	26.2674	0.4600	0.4200	0.5128	39	20
			926	0.8302	26.1577	0.4902	0.4314	0.5526	38	21
			4001	0.8393	25.2836	0.5600	0.4200	0.6061	33	20
			4269	0.8817	25.0023	0.5882	0.4118	0.6250	32	20
			5706	0.9268	24.8728	0.5319	0.3617	0.5161	31	16
			5281	0.8255	24.7993	0.5000	0.3913	0.5152	33	17
			4225	0.8406	24.7963	0.5417	0.3542	0.5172	29	15
			5298	0.8360	24.7883	0.4783	0.4130	0.5161	31	16
			4243	0.8686	24.6493	0.5102	0.3673	0.5000	34	17
			5612	0.8929	24.4295	0.4706	0.4314	0.5250	40	21
			4256	0.8447	24.2406	0.6939	0.4082	0.7308	26	19
			6020	0.9019	24.0511	0.5000	0.3800	0.5000	34	17
			5703	0.9148	24.0087	0.4800	0.4400	0.5405	37	20
			5961	0.8195	23.9841	0.5490	0.4314	0.6000	35	21
			2595	0.8112	23.9176	0.5400	0.4000	0.5588	34	19
			⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
			4257	0.8637	14.3240	0.6800	0.4200	0.6774	31	21
			914	0.8112	14.2823	0.5686	0.4314	0.5714	35	20
			4185	0.8303	14.0312	0.6531	0.4286	0.6552	29	19
			4710	0.8103	13.5711	0.5200	0.4400	0.5278	36	19
			5978	0.8501	13.5634	0.7000	0.4400	0.7143	28	20
			5207	0.8104	13.4411	0.5294	0.4314	0.5278	36	19
			2589	0.8287	13.1487	0.5294	0.4314	0.5263	38	20
			921	0.8219	12.5048	0.6000	0.4000	0.5625	32	18
			6057	0.8128	12.4964	0.5652	0.4348	0.5625	32	18
			6012	0.8380	12.4960	0.5625	0.4375	0.5625	32	18
			6181	0.8033	11.6221	0.5714	0.4490	0.5789	38	22
			5060	0.8653	11.1105	0.5800	0.4200	0.5556	36	20
			942	0.8278	10.6944	0.7000	0.4400	0.7000	30	21
			5193	0.8156	8.1077	0.5800	0.4200	0.5405	37	20
			6276	0.8043	8.0482	0.6739	0.4348	0.6538	26	17
			5968	0.8026	7.7789	0.6939	0.4490	0.6875	32	22
			5615	0.8195	6.9890	0.6000	0.4400	0.5758	33	19
			975	0.8314	6.2982	0.7200	0.4400	0.7000	30	21
			4261	0.8536	5.7678	0.6600	0.4200	0.6129	31	19
			978	0.8142	5.5027	0.7000	0.4200	0.6552	29	19

Fig. 12. The ranking of z-score about exchangeable proteins (r=50%).

included a considerable number of combinations in which combinatorial effect is inferred. However, the results are quite different between the two parameters of r that we used in our expeirment. This may be because the number of samples is not sufficient for statistical analysis, and we hope to clarify the validity of our method in detail in our future work. Further, we confirmed that we can exchange protein of A with D when D has strong correlation with A, and we found that the combinatorial effect is still strong even when A is exchanged with D.

In the future, we would like to perform more experiments to further validate our proposed method. In addition, we would like to develop an algorithm for the analytical computation

of the statistical distribution under the assumption of no combinatorial effect, i.e., we would like to compute the distribution shown in Figure 6 without simulation. If such fast computation is possible, it enables us to easily vary the threshold *r*, and it also enables us to compute a more accurate analysis. Finally, we also would like to find the known interactions in our results verify the value of this data-mining method.

6. Acknowledgment

This work was partly supported by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry.

7. References

[1] Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions, *Nature*, Vol. 340, pp. 245-246.

[2] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration, *Nature Biotechnology*, Vol. 17, pp. 1030 - 1032.

[3] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling, *Proc. Natl Acad Sci U S A*, Vol. 96, No. 6, pp. 2896-2901.

[4] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, CA. (1999). Protein interaction maps for complete genomes based on gene fusion events, *Nature*, Vol. 402, No. 6757, pp. 86-90.

[5] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O.& Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences, *Science*, Vol. 285, No. 5428, pp. 751-753.

[6] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl Acad Sci U S A*, Vol. 96, No. 8, pp. 4285-4288.

[7] Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. (2009). A Survey of Available Tools and Web Servers for Analysis of Protein-Protein Interactions and Interfaces, *Briefings in Bioinformatics*, Vol. 10, No.3, pp.217-232.

[8] Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data, *Journal of Computational Biology*, Vol. 7, No. 3/4, pp. 601-620.

[9] Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function and Genetics*, Vol. 47, No. 2, pp. 219-227.

[10] Comeau, S.R., Gatchell, D.W., Vajda, S. & Camacho, C.J. (2004). ClusPro: A Fully Automated Algorithm for Protein-protein Docking, *Nucleic Acids Research*, Vol. 32(Web server issue), pp. W96-99.

[11] Jothi, R. & Przytycka, T.M. (2008). Computational approaches to predict protein-protein and domain-domain interactions, In: *Bioinformatics Algorithms: Techniques and Applications*, Mondoiu, I.I. and Zelikovsky, A. of Editors, pp. 465-492, Wiley Press, ISBN 978-047-0097-73-1.

[12] Liang, S., Fuhrman, S. & Somogyi, R. (1998). REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures, *Proc. Pacific Symposium on Biocomputing '98*, pp. 18-29.

[13] Akutsu, T., Kuhara, S., Maruyama, O. & Miyano, S. (1998). A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions, *Genome Informatics*, Vol. 9, pp. 151-160.

[14] Laubenbacher, R. & Stigler, B. (2004). A Computational Algebra Approach to the Reverse Engineering of Gene Regulatory Network, *Journal of Theoretical Biology*, Vol. 229, No. 4, pp. 523-537.

[15] Shmulevich, I., Dougherty, E.R., Kim, S. & Zhang, W. (2002). Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks, *Bioinformatics*, Vol. 18, No. 2, pp. 261-274

[16] Husmeier, D. (2003). Sensitivity & Specificity of Inferring Genetic Regulatory Interactions From Microarray Experiments with Dynamic Bayesian Networks, *Bioinformatics*, Vol. 19, No. 17, pp. 2271-2282.

[17] Huang, Y., Tienda-Luna, I.M. & Wang, Y. (2009). A Survey of Statistical Models for Reverse Engineering Gene Regulatory Networks, *IEEE Signal Process Mag*, Vol. 26, No. 1, pp. 76-97.

[18] Sima, C., Hua, J. & Jung, S. (2009). Inference of Gene Regulatory Networks Using Time-Series Data: A Survey, *Current Genomics*, Vol. 10, No. 6, pp. 416-429.

[19] Nagai, K., Yoshihiro, T., Inoue, E., Ikegami, H., Sono, Y., Kawaji, H., Kobayashi, N., Matsuhashi, T., Ohtani, T., Morimoto, K., Nakagawa, M., Iritani, A. & Matsumoto, K. (2008). Developing an Integrated Database System for the Large-scale Proteomic Analysis of Japanese Black Cattle, *Animal Science Journal*, Vol. 79, No. 4. (in Japanese)

[20] Lu, C. (2004). Improving the Scaling Normalization for High-density Oligonucleotide GeneChip Expression microarrays, *BMC Bioinformatics*, Vol. 5, pp. 103.

[21] Spelman, R.J., Coppieters, W., Karim, L., van-Arendonk, J.A.M., & Bovenhuis, H. (1996). Quantitative Trait Loci Analysis for Five Milk Production Traits on Chromosome Six in the Dutch Holstein-Friesian Population, *Genetics,* Vol. 144, No. 4, pp. 1799-1808.

7

Inferring Protein-Protein Interactions (PPIs) Based on Computational Methods

Shuichi Hirose
Nagase & Co. Ltd. Research & Development Center, 2-2-3 Murotani, Nishi-ku, Kobe, Hyogo,
Computational Biology Research Center, Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo,
Japan

1. Introduction

Proteins are involved in many essential cellular processes, such as metabolism and signalling. They function by interacting with other molecules within the cell. Thus, protein interaction is one of the important keys to understand protein functions. As a consequence of the development of high-throughput experimental methods for detecting protein interactions, large volumes of data are now available. Although the data are valuable, there are limitations to their application. Therefore, computational methods are helpful tools for predicting protein interactions. With the increase in genome sequence data, the importance of computational methods in this field is growing more and more.

Another important factor to understand protein function is flexibility, because a protein molecule is not a rigid body. Flexible regions are often necessary for proteins to perform their functions, e.g. by enabling their flexible conformations to interact with other molecules and proteins. Therefore, it is important to understand the relationship between protein flexibility and protein interactions. In accordance with the increasing numbers of available protein structures, several databases that deal with protein flexibility have been built. Computational methods for analyzing protein motion are also being developed, for applications to PPI (protein-protein interaction) data.

The aim of this chapter is to provide a review on PPI prediction by computational techniques. In the first half of this chapter, the concepts and applications of several methods for inferring PPIs are introduced. They use genomic information based on evolutionary events. In the second half, the databases and prediction methods that deal with protein flexibility are introduced, and the possibility of inferring PPIs from protein flexibility will be discussed.

2. Computational methods to infer PPIs

The prediction of PPIs can be regarded as a binary classification problem, whereby the aim is to identify pairs of proteins as either interacting or non-interacting. PPIs can be divided into three types (Brown *et al.*, 2010). The first is direct protein interactions, which involve

direct physical contacts between proteins. The second is indirect functional association. In this case, the interacting protein pair does not have a direct physical contact, but it indirectly interacts, such as in the formation of a complex. The third is a member of biological pathway. In this case, the protein pairs do not form a complex, but their interactions occur in a logical order (for instance, in a signalling pathway).

Proteins exert their biological functions by participating in a PPI network. Protein interactions, as well as their biochemical functions, work as a type of selective pressure during evolution. Therefore, they influence the genome structure associated with the protein interactions. Conversely, analyzing the changes in the patterns on the genome makes it possible to infer PPIs. Several evolutionary events function as factors that impact the evolution of PPIs, including horizontal gene transfer, operon structure, co-evolution, co-expression, and lineage-specific gene loss. Several methods that infer PPIs using these various types of evolutionary information have been developed (Valencia and Pazos, 2002; Skrabanek *et al*., 2008). In this section, the principles and applications of PPI prediction methods are introduced. The different PPI prediction methods (Shoemaker and Panchenko, 2007) are listed in Table 1. First, the genomic inference methods (Rosetta stone, conservation of gene neighborhood, and phylogenetic profile), which predict functional association using genomic context, are presented. Next, the methods (mirror tree, *in silico* two-hybrid system) based on co-evolution are introduced, which are applicable to domain interactions as well as protein interactions. Finally, the sequence signature and machine learning based-methods are presented.

Method	Interaction Type	Interaction
Rosetta stone	Indirect functional association	Protein
Conservation of gene neighborhood	Indirect functional association	Protein
Phylogenetic profiles	Indirect functional association	Protein/domain
Mirror tree	Indirect functional association	Protein/domain
In silico two-hybrid system	Direct physical interaction	Protein/domain
Sequence signature	Direct physical interaction	Protein/domain
Supervised classification	Direct physical interaction	Protein/domain

Table 1. Summary of PPI prediction methods.

The second column represents the interaction type predicted by the method. The third column shows whether the method is designed to predict protein or domain interaction.

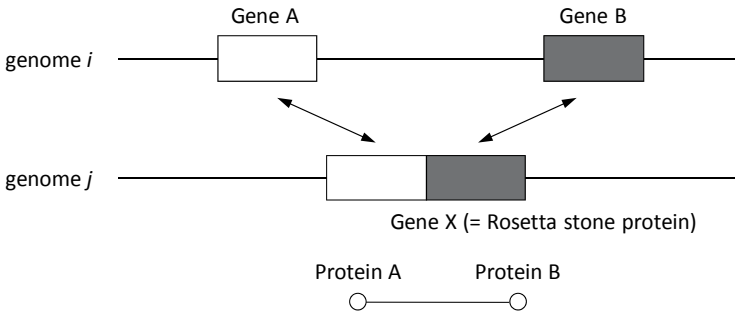
2.1 Rosetta stone

The Rosetta stone approach infers PPIs by comparing different genomes. It is often observed that two proteins that interact in genome *i* have a homologous protein that is fused into one protein in genome *j* (Fig. 1). These two gene products are functionally related in many cases (Enright *et al*., 1999). The fused protein is called a 'Rosetta stone protein', since it serves as a key for unlocking the functional relationship between two genes that are encoded independently in the genome. The Rosetta stone approach estimates functionally related protein pairs based on such a concept. The benefit of this approach is applicable to all

genomes, including those of Eukaryote. Not surprisingly, the inference of a protein interaction is restricted to the case where the gene fusion can be detected.

Hence, the approach searches for the proteins that are conserved between different organisms. The following two points must be considered, in order to obtain higher prediction accuracy. First, the proteins that interact with many other proteins, such as the HRG domain, and the CBS domain, which binds to DNA, should be removed. Second, the analysis is focused on the case where the pairs of genes that are fused together are orthologous. As an extension of the Rosetta stone approach, it can predict a functionally related gene cluster by combining several results. Four proteins, A, B, C, and D, are considered to be functionally related if the Rosetta stone proteins of the A-B, B-C, and C-D pairs are found.

Applying the Rosetta stone approach to many genomes revealed 6,809 potentially interacting pairs in *Escherichia coli* and 45,502 pairs in yeast (Marcotte *et al.*, 1999). The two proteins in each pair have significant sequence similarity to a single fused protein in another genome. Some proteins interact with several other proteins, and these connections apparently represent functional interactions, such as complexes or pathways.



Gene X is the Rosetta stone protein, indicating that protein A and protein B are functionally related.

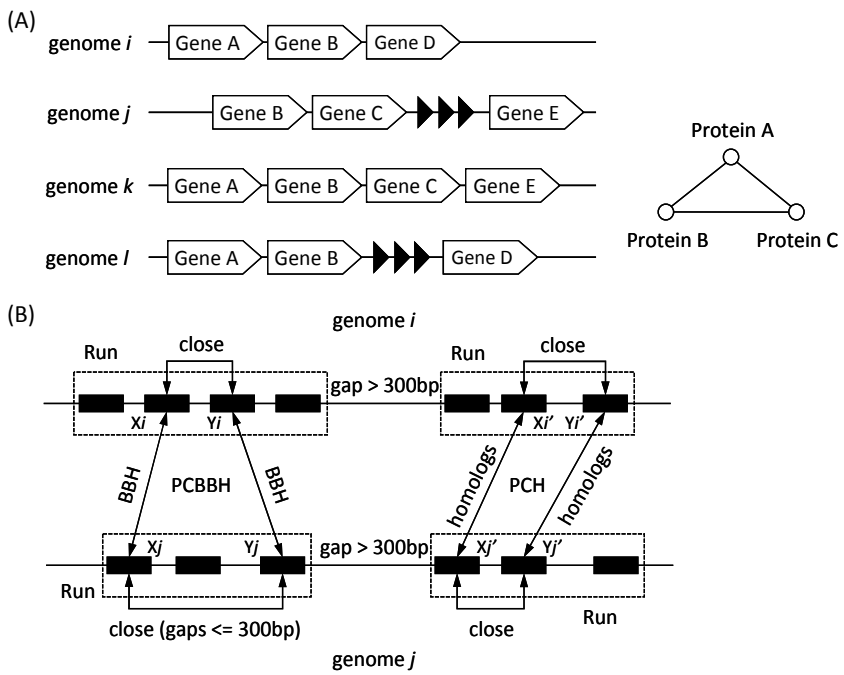
Fig. 1. The concept of the Rosetta stone approach.

2.2 Conservation of gene neighborhood

The genome comparison among Bacteria or Archaea indicated that the gene order and the operon structure are not conserved on the genome. This is because they have changed with evolutionary events, such as recombination, gene disruption, gene formation, and horizontal gene transfer. These phenomena suggest that the gene order is basically not subjected to selective pressure. However, the gene order or gene clusters on a genome are conserved if the gene products physically interact with each other, such as by complex formation, or if the proteins are transcribed as a single unit (Dandekar *et al.*, 1998). Briefly, it is often observed that the genes encoding proteins that either form a complex via physical interaction or work together in the same pathway are encoded in the same operon in different genomes. Thus, the gene order is conserved among different genomes, although the operon structure is fundamentally unstable during evolution (Fig. 2(A)). The conservation of the gene neighborhood approach infers proteins that are involved in the same biological process, using genome information. Many of the functionally related genes predicted by this approach encode proteins that either interact with each other directly, participate in the formation of the same complex, or work in the same metabolic pathway.

The conserved clusters of genes in an operon are detected by various concepts, such as Run or BBH (bidirectional best hit) (Overbeek *et al*., 1999) (Fig. 2(B)). A set of genes is called a "run" if they all occur on the same strand and the gaps between adjacent genes are 300 bases or less. Any pair of genes occurring within a single run is called "close". If gene X*i* in genome *i* is closest to X*j* in genome *j* and X*j* is closest to X*i*, then X*i* and X*j* are called BBH. Genes (X*i*, Y*i*) from genome *i* and genes (X*j*, Y*j)* from genome *j* form a PCBBH (pair of close bidirectional best hit) if two pairs of BBHs are considered. The conservation of gene neighborhood approach uses such virtual operons and orthologs to infer PPIs. That is, two orthologous groups are considered to have a connection if they co-occur in the same potential operon two or more times. The advantage of this approach is that the conservation of gene order or gene co-occurrence in the Run is stricter than the Rosetta stone and phylogenetic profile approaches, and it can cover a wider range of genes. However, the application of this approach is limited to Bacteria or Archaea that have operon structures.

Snel *et al*. reported 3,033 orthologous groups with 8,178 pairwise significant associations, by comparing 38 genomes (Snel *et al*., 2002). Among them, 88% of the 516 small, disjointed clusters, containing 2.7 orthologous groups on average, have a more homogeneous functional composition, in terms of the COG functional category. They are regarded as functional modules.



Different boxes signify different genes. The triangles represent genes that lack a conserved gene order. Protein A, protein B, and protein C, which line up in the same order among different organisms, are considered to be functionally associated.

Fig. 2. Illustration of (A) the concept of the conservation of gene neighborhood approach and (B) the definitions of BBH, PCBBH, and PCH (pairs of close homologs).

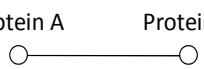
2.3 Phylogenetic profile

This approach is based on a concept derived from the lineage-specific gene loss. The genes encoding proteins that interact with each other co-occur in different genomes. If one gene is absent in a genome, and then the other gene that interacts with it also is lost. On the basis of this hypothesis, the phylogenetic profile approach infers PPIs from genome comparisons. The phylogenetic profile approach is based on the co-occurrence of gene pairs, while the conservation of gene neighborhood approach is based on the gene order or co-occurrence of genes. The advantage is that it is applicable to Eukaryote, since it is not necessary to consider operon structure. In addition, this approach is different from the prediction method based on the operon, in that the rate of predicted genes that belong to the same biological process is higher. The approach has two disadvantages. The first point is that the analysis targets are limited to the organisms with completely sequenced genomes, because whether a certain gene is actually encoded in the genome must be known. The second is that this approach is not applicable for the proteins encoded in all organisms that are analysis subjects.

The functional relationship between two genes is detected by comparing their phylogenetic profiles (Fig. 3) (Pellegrini *et al*., 1999). A phylogenetic profile is constructed for each protein, as a vector of N elements, where N is the number of genomes. Each position of the profile represents whether the protein that is homologous to the target protein is absent (signified by 0) or present (1) in each genome. Consequently, the phylogenetic distribution is shown by a long binary number along with each genome. A functionally related protein pair is detected by searching for the same phylogenetic distribution patterns. This method is applicable to domains as well as proteins (Pagel *et al*., 2004).

Pellegrini *et al*. applied a phylogenetic profile approach to the *Escherichia coli* genome and 16 other fully sequenced genomes, in order to predict the functions of uncharacterized proteins. When the function of a protein is assumed to be the same as that of its neighbors in the phylogenetic-profile space, 18% of the neighbor keywords overlapped the known keywords of the query protein. This indicates that the phylogenetic profile approach has the ability to assign functions to uncharacterized proteins.

	genome *i*	genome *j*	genome *k*	genome *l*	genome *m*
gene A	1	0	1	0	1
gene B	1	1	0	1	1
gene C	1	0	1	0	1
gene D	1	0	1	1	1
gene E	0	1	1	0	1



Phylogenetic profile (10101)

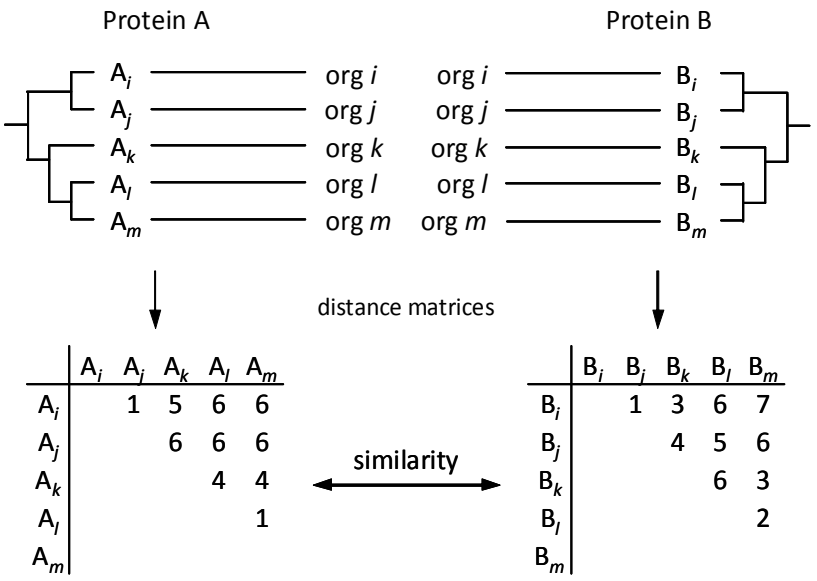
Protein A and protein C are considered to interact with each other, since they have the same profile (10101).

Fig. 3. An example of the phylogenetic profile approach.

2.4 Mirror tree

Pairs of physically contacting proteins co-evolve, such as insulin and its receptors (Fryxell, 1996). Co-evolution refers to the phenomenon in which the evolution found in one protein has a considerable effect on the evolution of its partner protein, in order to maintain the protein interaction. Therefore, the amino acid substitutions are expected to occur at the same time in the interacting proteins. As a result, the two phylogenetic trees drawn for the interacting proteins show a greater degree of similarity than those drawn for proteins without interactions (Goh *et al.*, 2000). The mirror tree approach infers two protein/domain interaction pairs, using the similarity between the phylogenetic trees as an indicator. The advantage of this approach is that it can be applied to an organism whose genome has not been completely sequenced. Conversely, the approach is not applicable to a gene that shows a species-specific loss. In addition, the applications of this approach are limited to the cases where high-quality and complete multiple sequence alignments, including sequences from the common organisms, can be obtained.

The similarity between two proteins/domains can be quantified as follows (Fig. 4) (Pazos and Valencia, 2001). First, for two proteins or domains, the multiple sequence alignments are built using orthologous proteins that are collected from *N* organisms. Next, the distant matrices are constructed from the genetic distances among all sequences, based on the multiple sequence alignment. The correlation coefficient between the two distance matrices is calculated. The value can be considered as an indicator that shows the intensity of co-evolution. Hence, if the value is close to one, it is judged that the two phylogenetic trees , and the two proteins are considered to interact. The mirror tree approach does not depend on the method used to construct the phylogenetic tree, since it does not compare them directly.



The trees have the same number of leaves and the same organisms in the leaves.

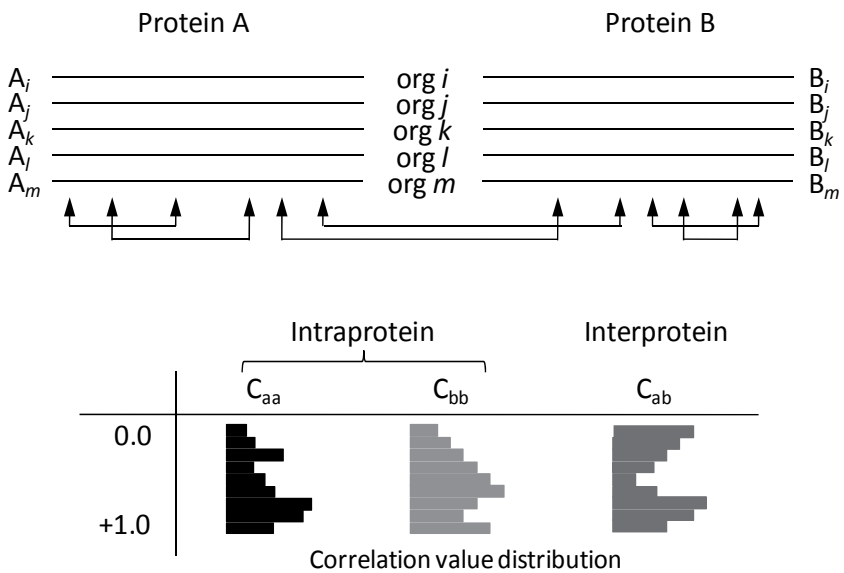
Fig. 4. The flow of the mirror tree approach.

The mirror tree method was applied to six protein families of ligand-receptor pairs, to predict the interaction partners (Goh and Cohen, 2002). Consequently, 79% of all known binding partners on average were detected. In addition, potentially new binding partner in the syntaxin/Unc-18 protein and TGF-β/TGF-β receptor families were found among previously characterized proteins.

2.5 *In silico* two-hybrid system

The *in silico* two-hybrid system infers physical contact sites by computing the correlation coefficient of amino acid variation between two sites, using the multiple sequence alignments for protein pairs (Göbel *et al*., 1994). That is, in the residue pairs that are in physical contact or related functionally, the amino acids tend to change at the same time. This type of correlated mutation is called co-variation. The similarity of the variation patterns is thought to be related to compensatory mutation. The *in cilico* two-hybrid system infers PPIs by expanding this concept. This system can detect an interaction accompanied by physical contact, and estimates the protein binding sites as well as interacting protein pairs. Meanwhile, the main limitation of this system is the requirement of high quality alignments that include a wide range of common organisms encoding the two proteins, in the same manner as the mirror tree approach.

The *in silico* two-hybrid system quantifies the degree of co-variation between pairs of residues (Fig. 5) (Pazos and Valencia, 2002). First, a multiple sequence alignments are built



On the top, the alignments are built for two different proteins (protein A and protein B), including the corresponding sequences from different organisms (org *i*, *j*, *k*…). On the bottom, the distributions of the correlated coefficients for the pair of residues internal to the two proteins (C_{aa} and C_{bb}) and for the pair of residues from each of the two proteins (C_{ab}) are represented.

Fig. 5. A schematic representation of the *in silico* two-hybrid system.

using orthologs derived from the common organism for two proteins. Next, the correlation coefficients between all combinations of sites in a protein are computed, and the frequency distribution of the values that are computed between two sites is investigated. Similarly, the correlation coefficients between all combinations of two sites from different proteins are calculated, and the frequency distribution is computed. Finally, the interaction index score is computed by using the three frequency distributions of correlation coefficients. If the value is close to one, then the two proteins are considered to interact.

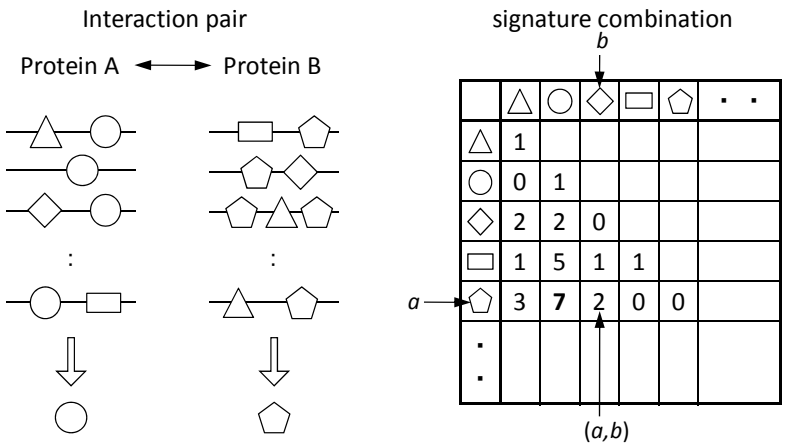
Pazos *et al.* applied this system to four test sets: 1) 14 two-domain proteins with a tight intradomain interaction, from the PDB, 2) 53 proteins including 31 known interactions, 3) 195 pairs with 15 possible interactions, derived from 749 predicted interactions, and 4) 321 pairs, 17 of which are known to interact, from the SPIN database. As a result, it discriminated between true and false interactions in a significant number of cases.

2.6 Sequence signature

The sequence signature approach, which predicts interacting proteins based on domain information, has developed separately from the methods using genome comparison or protein sequence analysis. This approach utilizes sequence and/or structure motifs in order to discriminate interacting proteins. In this approach, the characteristic pairs of sequence signatures are prepared from a database including experimentally determined interacting proteins, where one protein contains one sequence-signature and its interacting partner contains the other sequence-signature. The pairs that occur with high frequency are termed "correlated sequence-signatures", and they can be used for the prediction of putative interacting partners. The prediction result provides the pairs of protein/domain groups that include the correlated sequence-signature, while the other methods described above predict one-on-one protein pairs. Combining this approach with other techniques can yield higher performance.

In this approach, the sequence-signature of the signature combinations must be constructed to identify the correlated sequence-signatures (Fig. 6) (Sprinzak and Margalit, 2001). First, the experimentally determined interacting protein pairs are collected. Then, the sequence-signatures, defined by a motif database such as InterPro, are identified for each sequence. Each entry (a,b) in the table shows the number of protein pairs, composed of one protein containing signature a and its partner containing signature b. Next, the occurrence frequencies of the sequence-signature are converted into the log-odds. The sequence-signature with a positive log-odds value is considered to be observed more frequently in the interactive pairs. Therefore, they are regarding as having a correlated sequence-signature. Finally, this approach searches for the protein or domain pairs that contain the correlated sequence-signature.

An example of applying the Myb domain and the Bromodomain that are correlated sequence-signature to the yeast *S. cerevisiae* is shown (Sprinzak and Margalit, 2001). There are 19 and 10 protein sequences containing the Myb domain and the Bromodomain, respectively. Therefore, in this case, 190 protein interaction pairs are predicted, out of which five interactions were already known.



In the left panel, each row contains the sequences of the pair of protein A and protein B. Each sequence has a sequence-signature, illustrated by shapes. In the right panel, a contingency table of the signature combination is described, where each entry (*a*,*b*) in the table shows the number of protein pairs. For example, the sequence-signature pair represented by a square and a pentagon appears in two pairs of interacting proteins. The most abundant pair of sequence-signatures is indicated by bold type.

Fig. 6. A scheme for detecting correlated sequence-signatures in interacting proteins.

2.7 Supervised classification

The PPI prediction can be defined as a binary classification problem. Therefore, a statistical model or machine learning method can be applied to the problem of determining whether a pair of proteins is interacting or non-interacting. The K-Nearest Neighbor (KNN) (Qi *et al*., 2006), Naïve Bayesian (NB) (Jansen *et al*., 2003; Lu *et al*., 2005), support vector machines (SVM) (Lo *et al*., 2005), Artificial Neural Networks (ANN) (Ma *et al*., 2007), and Random Forest (RF); (Chen and Liu, 2005; Qi *et al*., 2005) methods were previously applied to this problem. The advantage of these methods is to use data that integrated different datasets. Datasets that do not directly measure PPI, such as sequence and structure information, can be used to infer PPIs. Conversely, the weak point is that the predictive performance varies widely, depending on the quality of the dataset and the selection of statistical methods.

In a statistical model, protein pairs are expressed by N dimensional vectors, where N is the number of features. For example, gene co-expression, GO biological process similarity, MIPS functional similarity, and essentiality are used as features in Jansen's work (Jansen *et al*., 2003). In addition, sequence information, such as homology and domain data, is used. Two points must be considered when the prediction model is built. The first is that it is necessary to pay attention to the quality of the experimental data used for training and evaluating statistical model, since the performance of the prediction model strongly depends on them. A high-throughput experimental method, such as Yeast Two-Hybrid (Y2H), Mass Spectrometry and Tandem Affinity Purification (MS TAP), and gene co-expression, can detect proteomic-wide PPIs, yielding vast amounts of protein interaction data within the cell. However, these data are often noisy, incomplete, and low-reproducible, since they contain contradictory values. The second is that the selection of an appropriate classification technique is an important task.

The statistical model was developed to infer PPIs in the human and yeast genomes (Lee *et al*., 2004; Rhodes *et al*., 2005). Qi *et al*. applied six different classifiers (RF, KNN, NB, Decision Tree, Logistic Regression, and SVM) to predict PPIs, and among them, the RF classifier exhibited the highest performance (Qi *et al*., 2006). In addition, gene expression is the most important feature for prediction.

3. Computational methods to infer protein flexibility

A protein molecule is not a rigid body. The scale of protein motions is very broad: motions range from local fluctuations, such as those seen in loop regions, to global ones involving changes in the relative positions of rigid domains. Protein motion is often necessary for proteins to perform their specific biological functions. For example, a protein possesses certain conformations in order to interact with its partner protein in many cases. Therefore, structural flexibility is an important feature to consider for understanding protein functions.

Experimental methods that analyze protein dynamics have been developed. Nuclear magnetic resonance (NMR) is a powerful experimental technique (Williams, 1989). NOEs and relaxation experiments provide information related to picosecond-microsecond-scale motions of the backbone atoms (Chill *et al*., 2004; Gitti *et al*., 2005). Also, model-free analysis enables quantitative determination of the fluctuations and slow conformational changes of the backbone amide vectors (Lipari and Szabo, 1982a; Lipari and Szabo, 1982b). Although NMR provides a detailed view of protein dynamics, it is time-consuming and suffers from size limitations.

In contrast, computational methods are useful to calculate the dynamics of proteins for which structures are available. They are divided into two types of method. One method compares the structures of a protein crystallized under different conditions or different conformers obtained by NMR. The structural differences indicate flexible regions (Shatsky *et al*., 2002; Ye and Godzik, 2004). Another computational method is to simulate protein dynamics by methods such as Normal Model Analysis (NMA) and Molecular Dynamics (MD). With the increasing number of available protein structures and the development of high-performance computers, databases that treat protein dynamics have been developed (Table 2). Some databases are introduced below.

ProMode

ProMode is a database including NMA results from analyses performed with a full-atom model for many proteins. It displays realistic three-dimensional motions at an atomic level, using a free plug-in, Charm. In addition, the dynamic domains and their mutual screw motions defined from NMA results are displayed.

MolMovDB

The database of macromolecular movements (MolMovDB) is a collection of quantitative data for flexibility and a number of graphical representations. The motions are generated from alignments of pairs of structures from the Protein Data Bank (PDB). The motions are divided into various classes (e.g. 'hinged domain' or 'allosteric'), according to the type of conformational change.

DynDom database

DynDom, a domain motion analysis program, analyzes the conformational change in terms of dynamic domains, interdomain screw axes, and interdomain bending regions, by comparing two structures when at least two X-ray conformers are available. The DynDom database displays details on the conformational changes obtained from the DynDom analysis results.

iGNM

The database contains visual and quantitative information on the collective modes predicted by the Gaussian Network Model (GNM) for the structure in the PDB. The output includes the equilibrium fluctuations of residues and comparisons with X-ray crystallographic B-factors, the sizes of residue motions in different collective modes, the cross-correlations between the residue fluctuations or domain motions, and other useful information.

Database name	HTTP address	Description	Reference
ProMode	http://cube.socs.waseda.ac.jp/pages/jsp/index.jsp	Large-scale collection of animations of the normal mode vibrating proteins with the full-atom models.	Wako *et al*., 2004
MolMovDB	http://www.molmovdb.org/	Visualization and classification of molecular motions according to their size and mechanism.	Echols *et al*., 2003
DynDom database	http://fizz.cmp.uea.ac.uk/dyndom/	Collection of domains, hinge axes and hinge bending residues in proteins.	Lee *et al*., 2003
*i*GNM	http://ignm.ccbb.pitt.edu/	Static and animated images for describing the conformational mobility of proteins by computing the GNM dynamics.	Yang *et al*., 2005

Table 2. List of databases that deal with conformational changes.

4. PPI prediction from protein flexibility

A Structural flexibility is an important characteristic of protein that is frequently related to their functions, as reviewed in section 3. Flexible regions are often necessary for proteins to bind a ligand or another protein. When we focus on the motion of a protein backbone segment, the movement can be classified conceptually into two forms: internal motion and external motion (Nishikawa and Go, 1987). An internal motion is the deformation of the segment itself, while an external motion involves only rotational and translational motions, as a rigid body. The segment fluctuates as a rigid body by changes in the dihedral angles of the flanking residues. For this reason, internal and external motions are considered to be fundamentally different.

This section introduces a means for the calculation of internal and external motions in a protein, by the constriction of statistical models, called "FlexRetriver", and its application to PPI data (Hirose *et al*., 2010).

4.1 Development of a method for predicting internal and external motions

This subsection introduces the RF-based method for predicting the internal and external motions defined by the NMA from the sequence information.

4.1.1 Calculation of internal and external motions

Using FEDER/2 (Wako *et al.*, 2004), the NMA was performed for the energy-minimized conformation, with the PDB data as the starting conformation. In the NMA, the mean-square displacement of atom a, $<D_a^2>$, in the thermal fluctuations is given as the sum of contributions from individual modes

$$<D_a^2>=\sum_{k=1}^{N} D_{ak}^2 ,$$

where D_{ak} is the displacement vector of atom a in the k-th normal mode, and N is the number of dihedral angles used as independent variables, i.e., the number of normal modes.

In this study, two conformations for a nine-residue segment in each normal mode are considered. The displacement vector of atom a by this purely translational and rotational motion is designated as D_{ak}^e, and the residual one is designated as D_{ak}^i. Then, D_{ak} is decomposed as

$$D_{ak} = D_{ak}^e + D_{ak}^i .$$

The superscripts e and i respectively stand for external and internal. The mean square deviation of atom a is given as

$$<D_a^2>=\sum_k \left|D_{ak}^e\right|^2 +\sum_k \left|D_{ak}^i\right|^2 +\sum_k 2D_{ak}^e \cdot D_{ak}^i$$

$$=<\left|D_a^e\right|^2>+<\left|D_a^i\right|^2>+2<\left|D_a^e \cdot D_a^i\right|> .$$

The third term on the right-hand side of this equation is usually much smaller than the first two terms. Therefore, the mean-square deviation of atom a is decomposed approximately into external (first term) and internal (second term) ones. In this case, we are interested in the main-chain fluctuation; for simplicity, only the C_α atom in this decomposition is considered. This means that we selected data for the C_α atoms from the results obtained using NMA with a full-atom model.

4.1.2 Dataset

The dataset was created by selecting protein chains from ProMode, as follows. Proteins with a root mean square deviation (RMSD) of more than 2Å between the energy-minimized structure and the PDB structure were excluded. Protein chains with redundant SCOP IDs were excluded, multi-domain proteins defined by SCOP were then removed. Next, some proteins were discarded so that the maximum pairwise sequence identity was limited to 25%. The resulting dataset comprised 481 chains (87,236 residues).

We calculated the internal and external motions using NMA with a full-atom model for all proteins in the dataset. Raw NMA values were normalized to correct for the variability among proteins in the dataset.

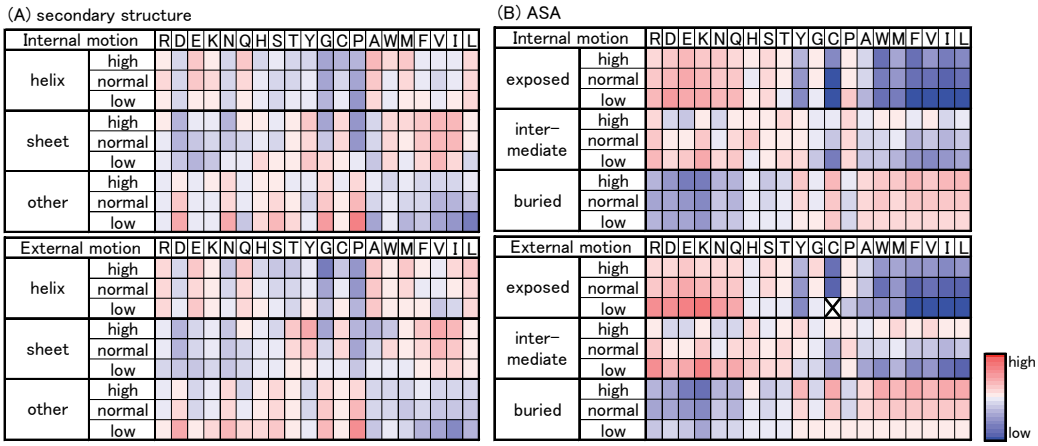
4.1.3 Structure-specific protein mobility propensity

The protein mobility propensities of amino acids are associated with their secondary structures and accessible surface areas (ASAs). The protein mobility propensity was divided into three types of protein mobility: the high and low groups comprised amino acids with normalized NMA scores higher than 1 and lower than -1, respectively, while the normal group comprised amino acids with normalized NMA scores between 1 and -1. The structure-specific protein mobility propensity (*SpecProg(n,s,g)*) was calculated as

$$SpecProp(n,s,g) = log_2\, freq(n,s,g)/freq(n,s)\ ,$$

where *freq(n,s,g)* and *freq(n,s)* respectively represent the relative frequencies of amino acid *n* in the *g* protein mobility group of the *s* state dataset and in the *s* dataset. The *s* state indicates a secondary structure or ASA.

The results of the structure-specific protein mobility propensity are shown in Fig. 7. For most amino acids, the protein mobility propensity pattern (in the high, normal, and low groups in the same type of secondary structure) depended on the type of secondary structure (Fig. 7(A)). For example, for both motions, the high mobility propensity of proline (Pro) was low in alpha helices and beta sheets, but high in other structures. This might be because Pro is a secondary structure breaker and its amide nitrogen cannot form a hydrogen bond. On the other hand, the low mobility propensity of hydrophobic amino acids tended to be high in



The upper and lower tables represent the results of internal and external motions, respectively. The terms high, normal, and low stand for the protein mobility of the high, normal and low states, respectively. The protein mobility propensity is colored with a gradient from negative (blue) to positive (red). The other in secondary structure is a region without helix and sheet. The cross mark signifies that no data exist.

Fig. 7. The protein mobility propensity associated with (A) secondary structures and (B) ASA.

alpha helices and beta sheets, but low in other structures. The distribution of high mobility propensity in other structures is similar to that of the propensity in hinge regions (Flores *et al*., 2007). Similarly, the protein mobility propensity pattern changes, depending on the ASA (Fig. 7(B)). The high mobility propensity became higher with increasing ASA, as seen for hydrophilic amino acids. In contrast, high mobility propensity of hydrophobic amino acids was lower with increasing ASA. The external motion might be more strongly influenced by the ASA, as compared to the internal motion. Altogether, these results strongly suggest that the secondary structure and the ASA influence the degrees of the internal and external motions.

4.1.4 Construction of a prediction method

A method for predicting internal and external motions was built by applying RF, which is a type of supervised classification algorithm. The sequence in the sliding window (with sizes of 11 and 17 residues for internal and external motions, respectively) was encoded by using paired amino acid information, corresponding to the variable. The variables were obtained by adding two features, which are derived from the amino acid pairs of the central amino acid with the other amino acids in the window. In total, 18 features were defined, and they were divided into four groups, designated as physicochemical, mobility, secondary structure (predicted by psipred (Jones, 1999) or PHD (Rost, 1996)), and ASA (predicted by sable (Adamczak *et al*., 2005) or RVPnet (Ahmad *et al*., 2003)). The profile-based predictors (psipred and sable) have higher prediction accuracy than the amino-acid propensity-based predictors (PHD and RVPnet). The value of a feature of an amino acid was set to one if the amino acid satisfied a feature's definition, and to zero otherwise.

The RF algorithm was used to build a prediction model for classifying amino acids into the three classes: flexible, intermediate, and rigid. Three RF prediction models were trained for the three categories of window location: the center of a secondary structure (CS), the remote area from a secondary structure (RS), and the periphery of a secondary structure (PS). The RF prediction models classified the windows into the three classes, and their prediction results were attributed to the central residue in the window. The results of the classification obtained from the RF were then converted into a score.

4.1.5 Prediction performance

The prediction results were assessed on a residue basis, by which the predicted score in the sequence was compared to the normalized NMA score. The prediction performance was evaluated by using three criteria: the mean absolute error (MAE), correlated coefficient (CC), and Receiver Operating Characteristic (ROC) curves. The MAE was defined as the absolute difference between two values. The MAE value approaches 0 as the prediction improves. The CC was also computed between two values. The CC ranges from -1 to 1, and a large, positive value represents a better prediction. The ROC curve was obtained by plotting the false positive rate against the true positive rate. A larger area under the ROC curve (AUC) indicates a more robust algorithm.

The prediction performance of FlexRetriever was compared with those of three published methods (PROFbval (Schlessinger *et al*., 2006), POODLE-S (Shimizu *et al*., 2007), FlexPred

(Kuznetsov and McDuffie, 2008)), and the naïve model. The naïve model is based on the simple idea that protein motion tends to be large in a coil or loop region and small in a secondary structure. The FlexRetriver, which implemented psipred and sable, yielded the lowest MAE and the highest CC among all prediction methods for both motions (Table 3). However, it is noteworthy that the distribution of CC varied widely. Additionally, in AUC, FlexRetriever exhibited the best performance among all methods.

(A) Internal motion			
Method	MAE	CC	AUC
FlexRetriever (PHD & RVPnet)	0.621	0.482	0.765
FlexRetriever (psipred & sable)	<u>0.605</u>	<u>0.525</u>	<u>0.786</u>
Naïve model (PHD)	0.988	0.248	0.653
Naïve model (psipred)	0.952	0.293	0.672
PROFbval	0.743	0.367	0.693
POODLE-S	-	-	0.730
FlexPred	-	-	0.741
(B) External motion			
Method	MAE	CC	AUC
FlexRetriever (PHD & RVPnet)	0.571	0.541	0.777
FlexRetriever (psipred & sable)	<u>0.542</u>	<u>0.597</u>	<u>0.806</u>
Naïve model (PHD)	0.970	0.262	0.661
Naïve model (psipred)	0.929	0.320	0.681
PROFbval	0.608	0.547	0.784
POODLE-S	-	-	0.783
FlexPred	-	-	0.777

The CC and MAE were estimated by performing a five-fold cross validation test. The highest scores in each criterion are underlined. PHD and psipred in parentheses signify the secondary structure predictor. Similarly, RVPnet and sable represent the ASA predictor. "-": scores could not be calculated.

Table 3. Comparison of prediction performance.

4.2 Applying FlexRetriever to PPI data

In this study, we utilized the set of 20 proteins that undergo large conformational changes upon association (> 2Å C_α RMSD) created by Dobbins *et al.*, with which they demonstrated the relationship between normal mode fluctuations and conformational change (Dobbins *et al.*, 2008). They regarded protein motions as being associated with their functions, because they are observed along with the PPI. We compared the internal motion with the observed conformational change region, because it was defined as the deformation of a segment itself. To begin with, we present three typical results, in which the observed conformational change regions are located in a binding site, a hinge region, and other regions. We will then discuss the overall results.

i. Ecotin

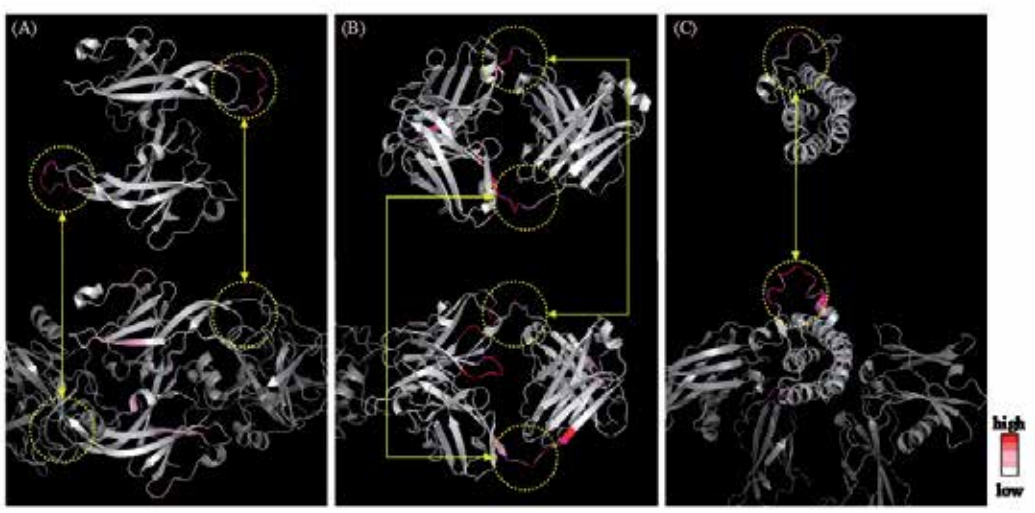
Ecotin, a homodimeric protein, is an inhibitor of a group of homologous serine proteases, such as trypsin, chymotrypsin, and elastase. One dimeric inhibitor binds to a protease molecule. From a comparison of two structures determined with different crystalline environments, an inherent flexible loop was identified in the binding site with trypsin. It was necessary for its inhibitory function (Shin *et al*., 1996). FlexRetriever predicted high internal motion for the corresponding loop (Fig. 8(A)).

ii. Fab fragment

The fragment antigen binding (Fab fragment) region is the site where an antibody binds to antigens. It is a heterodimer of the heavy and light chains in each of the two composed domains. The hinge region between the two domains changed its conformation when Fab bound to hemagglutinin derived from a flu virus (Fleury *et al*., 1998). FlexRetriever predicted high internal motion at the hinge region in each chain (Fig. 8(B)).

iii. Erythropoietin

Erythropoietin (EPO) is a hormone produced primarily in the kidneys. It has a four-helical bundle topology with two long loops, and binds to the extracellular domain of the EPO receptor. The CD loop, which is located in a region remote from the binding site, changed its conformation (Cheetham *et al*., 1998). FlexRetriever predicted high internal motion for the corresponding loop (Fig. 8(C)).



The observed degrees of conformational change and the predicted scores for internal motion are mapped, respectively, with a gradient from zero (white) to a high score (dark red) onto their structures in the upper and lower sections. The regions enclosed with a yellow dotted line are the regions with observed conformational changes. The free-state and complex-state structures are displayed, respectively, in the upper and lower sections.

Fig. 8. Example of the relationship between the predicted internal motions and the observed conformational changes of (A) ectin, (B) Fab fragment, and (C) erythropoietin.

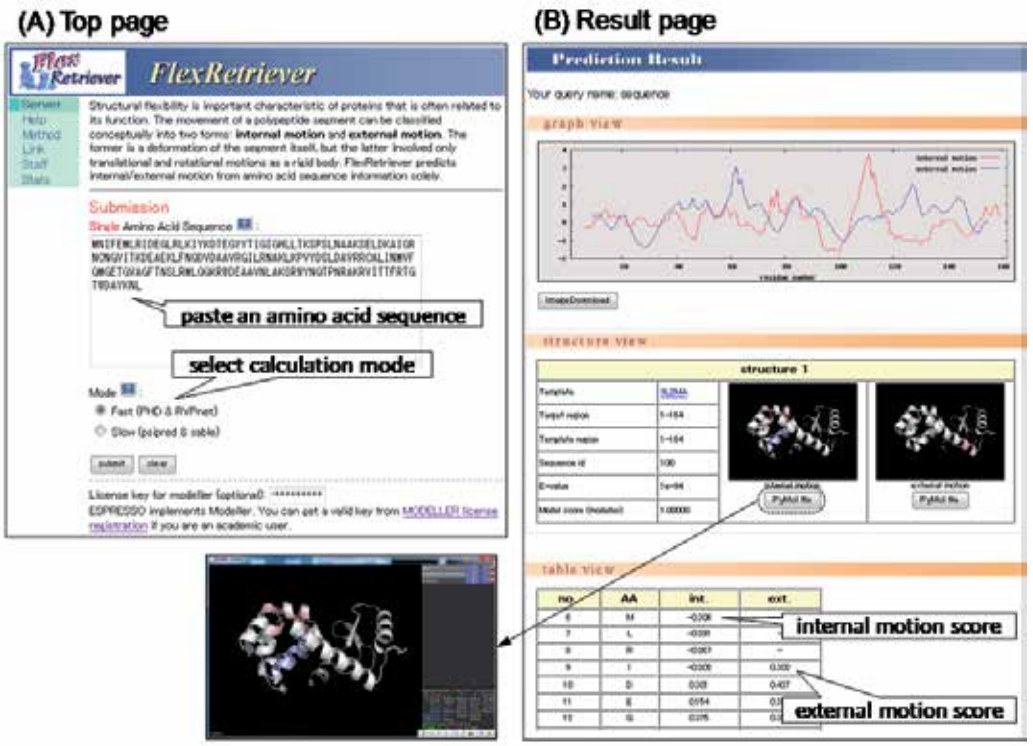
Overall results

When FlexRetriever was applied to a set of 20 proteins, three or more consecutive residues with high internal scores were regarded as candidates for the regions undergoing conformational changes. From a comparison between the observed conformational change regions with the predicted high internal motion regions, at least one overlap was found in 85% of the proteins studied. If the analysis object was limited to the 16 proteins that interact with only one partner, then the overlap was observed in 15 proteins (94% of the proteins studied). These observations suggest that FlexRetriever is a sensitive method for the detection of protein motions related to PPIs, including binding sites.

4.3 Web server

The presented method is implemented in the FlexRetriever server, which has been designed with a user-friendly interface to provide easily interpretable prediction results.

The server accepts the submission of a single amino acid sequence with less than 1,000 amino acids in the FASTA format (Fig. 9(A)). The user is asked to choose a calculation mode.



On the result page, a graph is displayed on the top, and two structures on which the scores of the internal and external motions are mapped are shown in the middle. They can be downloaded as a PyMol file. The table with the raw scores is displayed below the structures.

Fig. 9. Images of FlexRetriever's (A) top page and (B) result page.

The calculation time of the fast mode, which uses PHD for secondary structure prediction and RVPnet for ASA prediction, is shorter than that of the slow mode, but its performance is poorer.

The results page is divided into three sections (Fig. 9(B)). The first section (graph view) provides the graph which contains the prediction results of both motions. The second section (structure view) presents the degrees of internal and external motions on the three-dimensional structure. The third section (table view) lists the amino acids and the raw scores of their internal and external motions.

FlexRetriever is available at http://mbs.cbrc.jp/FlexRetriever and is free.

5. Conclusion

This chapter provides an overview of the computational methods to infer pairs of interacting proteins and to study the relevance of protein flexibility. Genomic information and experimental data are now readily available, and thus computational methods will become more important tools in the field of analyzing or inferring PPIs. In addition, as a novel attempt to predict PPIs, we have presented an efficient algorithm for predicting flexible regions in proteins, and shown its application to PPIs. The tool is expected to be useful for inferring motions associated with PPIs.

6. References

Adamczak, R.; Porollo, A. & Meller, J. (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins*, Vol.59, No.3, (May 15), pp. 467-475, ISSN 0887-3585

Ahmad, S.; Gromiha, M.M & Sarai, A. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, Vol.50, No.4, (Mar 1), pp. 629-635, ISSN 0887-3585

Brown, F.; Zheng, H.; Wang, H.Y. & Azuaje, F. (2010) From experimental approaches to computational techniques: A review on the prediction of protein-protein interaction. *Advances in Artificial Intelligence*, Vol.2010, ISSN 16877470

Chill, J.H.; Quadt, S.R. & Anglister, J. (2004) NMR backbone dynamics of the human type I interferon binding subunit, a representative cytokine receptor. *Biochemistry*, Vol.43, No.31, (Aug 10), pp. 10127-10137, ISSN 0006-2960

Cheetham, J.C.; Smith, D.M.; Aoki, K.H.; Stevenson, J.L.; Hoeffel, T.J.; Syed, R.S.; Egrid, J. & Harvey, T.S. (1998) NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Biol*, Vol.5, No.10, (Oct 1998), pp. 861-868, ISSN 1072-8368

Chen, X.W. & Liu, M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, Vol.21, No.24, (Dec 15), pp. 4394-4400, ISSN 0973-2063

Dandekar, T.; Snel, B.; Huynen, M. & Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, Vol.23, No.9, (Sep 1998), pp. 324-328, ISSN 0968-0004

Dobbins, S.E.; Lesk, V.I. & Sternberg, M.J. (2008) Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A*, Vol.105, No.30, (July 29), pp. 10390-10395, ISSN 0027-8424

Echols, N.; Milburn, D. & Gerstein, M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res*, Vol.31, No.1, (Jan 1), pp. 478-482, ISSN 0305-1048

Enright, A.J.; Iliopoulos, I.; Kyrpides N.C. & Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, Vol.402, No.6757, (Nov 4), pp. 86-90, ISSN 0028-0836

Fleury, D.; Wharton, S.A.; Skehel, J.J.; Knossow, M. & Bizebard, T. (1998) Antigen distortion allows influenza virus to escape neutralization. *Nat Struct Biol*, Vol.5, No.2, (Feb 1998), pp. 119-123, ISSN 1072-8368

Flores, S.C.; Lu, L.J.; Yang, J.; Carriero, N. & Gerstein, M.B. (2007) Hinge Atlas: relating protein sequence to site of structural flexibility. *BMC Bioinformatics*, Vol.8, No.167, (May 22), ISSN 1471-2105

Fryxell, K.J. (1996) The coevolution of gene family trees. *Trends Genet*, Vol.12, No.9, (Sep 1996), pp. 364-369, ISSN 0168-9525

Gitti, R.K.; Wright, N.T.; Margolis, J.W.; Varney, K.M.; Weber, D.J. & Margolis, F.L. (2005) Backbone dynamics of the olfactory marker protein as studied by 15N NMR relaxation measurements. *Biochemistry*, Vol.44, No.28, (Jul 19), pp. 9673-9679, ISSN 0006-2960

Göbel, U.; Sander, C.; Schneider, R. & Valencis, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, Vol.18, No.4, (Apr 1994), pp. 309-317, ISSN 0887-3585

Goh, C.S.; Bogan, A.A.; Joachimiak, M.; Walther, D. & Cohen, F.E. (2000) Co-evolution of proteins with their interaction partners. *J Mol Biol*, Vol.299, No.2, (Jun 2), pp. 283-293, ISSN 0022-2836

Goh, C.S. & Cohen, F.E. (2002) Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol*, Vol.324, No.1, (Nov 15), pp. 177-192, ISSN 0022-2836

Hirose, S.; Yokota, K.; Kuroda, Y.; Wako, H.; Endo, S.; Kanai, S. & Noguchi, T. (2010) Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC Struct Biol*, Vol.10, No.20, (Jul 13), ISSN 1472-6807

Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N.J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J.F. & Gerstein, M. (2003) A Baysian networks approach for predicting protein-protein interactions from genomic data. *Science*, Vol. 302, No.5644, (Oct 17), pp. 449-453, ISSN 0036-8075

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, Vol.292, No.2, (Sep 17), pp. 195-202, ISSN 0022-2836

Kuznetsov, I.B. & McDuffie M. (2008) FlexPred: a web-server for predicting residue positions involved in conformational switches in proteins. *Bioinformation*, Vol.3, No.3, (Nov 5), pp. 134-136, ISSN 0973-2063

Lee, I.; Date, S.V.; Adai, A.T. & Marcotte E.M. (2004) A probabilistic functional network of yeast genes. *Science*, Vol.306, No.5701, (Nov 26), pp. 1555-1558, ISSN 0036-8075

Lee, R.A.; Razaz, M. & Hayward, S. (2003) The DynDom database of protein domain motions. *Bioinformatics*, Vol.19, No.10, (Jul 1), pp.1290-1291, ISSN 1460-2059

Lipari, G. & Szabo, A. (1982a) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J Am Chem Soc*, Vol.104, No.17, (August 1982), pp. 4546-4559, ISSN 0002-7863

Lipari, G. & Szabo, A. (1982b) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J Am Chem Soc*, Vol.104, No.17, (August 1982), pp. 4559-4570, ISSN 0002-7863

Lo, S.L.; Cai, C.Z.; Chen, Y.Z. & Chung, M.C. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, Vol.5, No.4, (Mar 2005), pp. 876-884, ISSN 1615-9861

Lu, L.J.; Xia, Y.; Paccanaro, A.; Yu, H. & Gerstein, H. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, Vol.15, No.7, (Jul 2005), pp. 945-953, ISSN 1088-9051

Ma, Z.; Zhou, C.; Lu, L.; Lu, L.; Ma, Y.; Sun, P. & Cui, Y. Predicting protein-protein interactions based on BP neural network, *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pp. 3-7, ISBN 978-1-4244-1604-2, Fremont, CA, Nov 2-4, 2007

Marcotte, E.M.; Pellegrini, M.; Ng, H.L.; Rice D.W.; Yeates, T.O. & Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequence. *Science*, Vol.285, No.5428, (Jul 30), pp. 751-753, ISSN 0036-8075

Nishikawa, T. & Go, N. (1987) Normal modes of vibration in bovine pancreatic trypsin inhibitor and its mechanical property. *Proteins*, Vol.2, No.4, (1987), pp. 308-329, ISSN 0887-3585

Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D. & Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, Vol.96, No.6, (Mar 16), pp. 2896-2901, ISSN 0027-8424

Pagel, P.; Wong, P. & Frishman, D. (2004) A domain interaction map based on phylogenetic profiling. *J Mol Biol*, Vol.344, No.5, (Dec 10), pp. 1331-1346, ISSN 0022-2836

Pazos, F. & Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, Vol.14, No.9, (Sep 14), pp. 609-614, ISSN 0269-2139

Pazos, F. & Valencia, A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, Vol.47, No.2, (May 1), pp. 219-227, ISSN 0887-3585

Pellegrini, M.; Marcotte, E.M.; Thompson, M.J.; Eisenberg, D. & Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, Vol.96, No.8, (Apr 13), pp. 4285-4288, ISSN 0027-8424

Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*, Vol.266, (1996), pp. 525-539, ISSN 0076-6879

Qi, Y.; Klein-Seetharaman, J & Bar-Joseph, Z. (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, pp. 531-542, ISSN 1793-5091

Qi, Y.; Bar-Joseph, Z. & Klein-Seetharaman, J. (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, Vol.63, No.3, (May 15), pp. 490-500, ISSN 0887-3585

Rhodes, D.R.; Tomlins, S.A.; Varambally, S.; Mahavisno, V.; Barrette, T.; Kalyana-Sundaram, S.; Ghosh, D.; Pandey, A. & Chinnalyan, A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, Vol.23, No.8, (Aug 2005), pp. 951-959, ISSN 1087-0156

Schlessinger, A.; Yachday, G. & Rost, B. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, Vol.22, No.7, (Apr 1), pp. 891-893, ISSN 1460-2059

Shatsky, M.; Nussinov, R. & Wolfson, H.J. (2002) Flexible protein alignment and hinge detection. *Proteins*, Vol.48, No.2, (Aug 1), pp. 242-256, ISSN 0887-3585

Shimizu, K.; Hirose, S. & Noguchi, T. (2007) POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, Vol.23, No.17, (Sep 1), pp. 2337-2338, ISSN 1460-2059

Shin, D.H.; Song, H.K.; Seong, I.S.; Lee, C.S.; Chung C.H & Suh, S.W. (1996) Crystal structure analyses of uncomplexed ecotin in two crystal forms: implications for its function and stability. *Protein Sci*, Vol.5, No.11, (Nov 1996), pp. 2236-2247, ISSN 0961-8368

Shoemaker, B.A. & Panchenko, A.R. (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, Vol.3, No.4, (Apr 27), pp. e43, ISSN 1553-734X

Skrabanek, L.; Saini, H.K.; Bader, G.D. & Enright, A.J. (2008) Computational prediction of protein-protein interactions. *Mol Biotechnol*, Vol.38, No.1, (Jan 2008), pp. 1-17, ISSN 1073-6085

Snel, B.; Bork, P. & Huynen, M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci U S A*, Vol.99, No.9, (Apr 30), pp. 5890-5895, ISSN 0027-8424

Sprinzak, E. & Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, Vol.311, No.4, (Aug 24), pp.681-692, ISSN 0022-2836

Valencia, A. & Pozos, F. (2002) Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, Vol.12, No.3, (Jun 2002), pp. 368-373, ISSN 0959-440X

Wako, H.; Kato, M. & Endo, S. (2004) ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, Vol.20, No.13, (Sep 1), pp. 2035-2043, ISSN 1460-2059

Williams, R.J. (1989) NMR studies of mobility within protein structure. *Eur J Biochem*, Vol.183, No.3, (Aug 15), pp. 479-497, ISSN 0014-2956

Yang, L.W.; Liu, X.; Jursa, C.J.; Holliman, M.; Rader, A.J.; Karimi, H.A. & Bahar, I. (2005) *i*GMN: a database of protein functional motions based on Gaussian Network Model. *Bioinformatics*, Vol.21, No.13, (Jul 1), pp. 2978-2987, ISSN 1460-2059

Ye, Y. & Godzik, A. (2004) Database searching by flexible protein structure alignment. *Protein Sci*, Vol.13, No.7, (Jul 2004), pp. 1841-1850, ISSN 0961-8368

8

Slow Protein Conformational Change, Allostery and Network Dynamics

Fan Bai[1], Zhanghan Wu[2], Jianshi Jin[1],
Phillip Hochendoner[3] and Jianhua Xing[2]
[1]*Biodynamic Optical Imaging Centre, Peking University, People's Republic of China, Beijing*
[2]*Department of Biological Sciences, Virginia Tech,*
[3]*Department of Physics, Virginia Tech,*
[1]*China*
[2,3]*USA*

1. Introduction

Macromolecules such as proteins contain a large number of atoms, which lead to complex dynamic behaviors not usually seen in simpler molecular systems with only a few to tens of atoms. Characterizing the biochemical and biophysical properties of macromolecules, including their interactions with other molecules, has been a central research theme for many decades. The field is especially accelerated by recent advances in experimental techniques, such as nuclear magnetic resonance (NMR) and single-molecule measurements, and computational powers that has been facilitated to simulate molecular dynamics at large scales.

Chemical kinetics has been well developed for simple molecular systems, and most of the small molecular reactions can be described accurately by kinetic equations. However, it's hard to describe a macromolecular system using simple mathematical equations, because reactions at the macromolecular level usually involve complicated processes and dynamic behaviors. Even so, biochemists have done many efforts to find a way to describe the biological systems. Many equations and models have been published by using approximate treatments or hypothesis.

If biochemists were asked what is the most important mathematical equation they know, most likely the answer you will hear is the Michaelis-Menten equation. Michaelis–Menten equation is one of the simplest and best-known equations describing enzyme kinetics (Menten and Michaelis, 1913). It is named after American biochemist Leonor Michaelis and Canadian physician Maud Menten. For a typical enzymatic reaction one often finds that the following scheme works reasonably well,

$$S + E \underset{\alpha_{-1}}{\overset{\alpha}{\rightleftarrows}} \mathrm{ES} \xrightarrow{k} E + P \tag{1}$$

with S, E, ES, P representing the substrate, the free enzyme, the enzyme-substrate complex, and the product. Then one has the rate of product formation: (after certain assumptions, such as the enzyme concentration being much less than the substrate concentration)

$$\frac{d[P]}{dt} = \frac{k[E]_t[S]}{(\alpha_{-1} + k)/\alpha + [S]} \tag{2}$$

In this model, the rate of product formation increases along with the substrate concentration [S] with the characteristic hyperbolic relationship, asymptotically approaching its maximum rate $V_{max} = k[E]_t$, ($[E]_t$ is the total enzyme concentration) attained when all enzymes are bound to substrates .We can use K_m to represent $(a_{-1}+k)/a$, named Michaelis constant. It is the substrate concentration at which the reaction rate is at half the maximum rate, and is a measure of the substrate's affinity for the enzyme. A small K_m indicates high affinity, meaning that the rate approaches V_{max} more quickly.

The Michaelis–Menten equation was first proposed for investigating the kinetics of an enzymatic (invertase) reaction mechanism in 1913 (Menten and Michaelis, 1913). Later, it has been widely used in a variety of biochemical transitions other than enzyme-substrate interaction, which includes antigen-antibody binding, DNA-DNA hybridization and protein-protein interaction. There is no exaggeration to say that the Michaelis–Menten model has greatly pushed forward our understanding of enzymatic reactions.

However, biochemists also found that many enzymes show kinetics are more complicated than the Michaelis-Menten kinetics. Frieden coined the name "hysteretic enzyme" referring to "those enzymes which respond slowly (in terms of some kinetic characteristic) to a rapid change in ligand, either substrate or modifier, concentration" (Frieden, 1970). Since then a sizable literature exists on the enzyme behavior. The list of hysteretic enzymes cover proteins working in many organisms from bacteria to mammalians (Frieden, 1979), with one of the latest examples related to the protein secreted by bacteria *Staphylococcus aureus* to induce host blood coagulation (Kroh et al., 2009). The kinetics, especially the enzymatic activity of a hysteretic enzyme, cannot adapt to new environmental conditions quickly. The delay time can be surprisingly long. For example, upon changing the solution's pH value, it takes more than two hours for alkaline phosphatase to relax to the enzymatic activity corresponding to the new pH value (Behzadi et al., 1999). The mnemonic behavior is another key example of slow conformational dynamic disorder advocated by Richard and his colleagues (Cornish-Bowden and Cardenas, 1987; Frieden, 1970; Frieden, 1979; Ricard and Cornish-Bowden, 1987). It refers to the phenomenon that "the free enzyme alone which undergoes the 'slow' transition…upon the desorption of the last product from the active site, the enzyme retains for a while the conformation stabilized by that product before relapsing to another conformation" (Ricard and Cornish-Bowden, 1987). Their observation revealed that Mnemonic enzymes show non-Michaelis-Menten (NMM) behaviors. The concepts of mnemonic and hysteretic enzymes emphasize the steady-state kinetics and the transient kinetics leading to the steady state, respectively. However, the conformational change in a protein is the rate limiting step in both enzymatic reactions which are slower than the actual chemical reaction step (chemical bond breaking and forming). To this end, a unified model exists (Ainslie et al., 1972).

A deeper understanding on the origin of the mnemonic and hysteretic behaviors comes from biophysical studies. A related phenomenon called dynamic disorder has been discussed extensively in the physical chemistry and biophysics communities. Dynamic disorder refers to the phenomena that the 'rate constant' of a process is actually a random function of time, and is affected by some slow protein conformational motions (Frauenfelder

et al., 1999; Zwanzig, 1990). A molecule fluctuates constantly at finite temperature. The Reaction Coordinate (RC) is an important concept in chemical rate theories (Hanggi et al., 1990). The RC is a special coordinate in the configurational space (expanded by the spatial coordinates of all the atoms in the system), which leads the system from the reactant configuration to the product configuration. A fundamental assumption in most rate theories (such as the transition state theory) states that the dynamics along the RC is much slower than fluctuations along all other coordinates. Consequently, for any given RC position, one may assume other degrees of freedom approaches approximately equilibrium. This is the so-called adiabatic approximation. Deviation from this assumption is treated as secondary correction (Grote and Hynes, 1980). Chemical rate theories based on this assumption are remarkably successful in explaining the dynamics involving small molecules. The dynamics of a system can be well characterized by a rate constant. However, the situation is much more complicated in macromolecules like proteins, RNAs, and DNAs. Macromolecules have a large number of atoms and possible conformations. The conformational fluctuation time scales of macromolecules span from tens of femtoseconds to hundreds of seconds (McCammon and Harvey, 1987). Consequently, conformational fluctuations can be comparable or even slower than the process involving chemical bond breaking and formation. The adiabatic approximation seriously breaks down at this regime. If one focuses on the dynamics of processes involving chemical reactions, the canonical concept of "rate constant" no longer holds. Since the pioneering work of Frauenfelder and coworkers on ligand binding to myoglobin (Austin et al., 1975), extensive experimental and theoretical studies have been performed on this subject (see for example ref. (Zwanzig, 1990) for further references). Additionally, the conformational fluctuation of a macromolecule is an individual behaviour, many dynamic processes were hidden under the ensemble measurements. Fortunately, recent advances in room-temperature single-molecule fluorescence techniques gave us an opportunity to investigate the conformational dynamics on the single-molecule level. Hence, the dynamic disorders in an individual macromolecule has been demonstrated directly through single molecule enzymology measurements recently (English et al., 2006; Min et al., 2005b; Xie and Lu, 1999). For example, Xie and coworkers showed that both enzymes' conformation and catalytic activity fluctuate over time, especially the turnover time distribution of one β-galactosidase molecule spans several orders of magnitude (10^{-3} s to 10 s). Their results revealed that although a fluctuating enzyme still exhibits MM steady-state kinetics in a large region of time scales, the apparent Michaelis and catalytic rate constants do have different microscopic interpretations. It is also shown that at certain conditions dynamic disorder results in Non-Michaelis-Menten kinetics (Min et al., 2006). Single molecule measurements on several enzymes suggested that the existence of dynamic disorder in biomolecules is a rule rather than exception (Min et al., 2005a). So if problems arise, when there are only a few copies of a particular enzyme in a living cell, do these fluctuations result in a noticeable physiological effect?

Therefore, an important question we need to ask is: What is the biological consequence of dynamic disorder? Frieden insightfully noticed that "*it is of interest that the majority of enzymes exhibiting this type of (hysteretic) behavior can be classed as regulatory enzymes*" (Frieden, 1979). A series of important questions emerge naturally: Is the existence of complex enzymatic kinetic behaviors an evolutional byproduct or selected trait? Is there any biological function for it? How can such diverse and complex enzymatic kinetic behaviors affect our understanding of regulatory protein interaction networks?

In recent years, studying interactions of molecules in a cell from a systems perspective has been gaining popularity. Researchers in this newly formed field "systems biology" emphasize that to characterize a complex system, it is insufficient to take the reductionist's view. Combining several reactions together, one can form reaction networks with emerging dynamic behaviors such as switches, oscillators, etc, and ultimately the life form (Alon, 2007; Kholodenko, 2006; Tyson et al., 2001). In the new era of systems biology, a modeler may deal with hundreds to thousands ordinary differential rate equations describing various biological processes. The hope is that by knowing the network topology and associated rate constants (which requires daunting experimental efforts), one can reveal the secret of life and even synthesize life.

On modeling such regulatory protein interaction networks, it is common practice to assume that each enzymatic reaction can be described by a simple rate process, especially by the Michaelis-Menten kinetics. In our opinion, most contemporary researches on biological network dynamics emphasize the effect of network topology without giving sufficient consideration of the biochemical/biophysical properties of each composing macromolecule. One of the reasons that account for the current state of affair is due to a lack of experimental data and theoretical understanding in the "intermediate regime" between single-molecule studies of individual enzymes (relatively simple) and cellular dynamics (too complex). Recent advances in single-molecule techniques give us hope to study larger systems. One of its unique advantages is the ability to study macromolecular dynamics under room temperature and nonequilibrium state, which well mimics physiological conditions of a living cell. Using these single-molecule experimental results to build the cellular dynamics model will be a promising and significative research field.

In this chapter, we will present a unified mathematical formalism describing both conformational change and chemical reactions. Then we will discuss some implications of slow conformational changes in protein allostery and network dynamics.

2. Coarse grained mathematical description of conformational changes

Substrate binding often induces considerable changes of the protein conformation, especially in the binding pocket. This is the so-called induced-fit model. To explicitly take into account the induced conformational change, one can generalize the scheme given in Equation 1 to what shown in Fig. 1A. The substrate and protein form a loosely bound complex first. Their mutual interactions drive further conformational change of the binding pocket to form a tight bound complex, where atoms are properly aligned for chemical bond breaking and forming to take place. Next the binding pocket opens to release the product and is ready for another cycle. Mathematically one can write a set of ordinary differential based rate equations to describe the dynamics, or perform stochastic simulations of the process.

For a more complete description of the continuous nature of conformational changes, one can reduce the conformational complexity of the system to a few well defined degrees of freedom with slow dynamics (Xing, 2007). For example, let's denote x to represent the conformational coordinate of the enzyme from open to close of the binding pocket, and $U(x)$ the potential of mean force along x. In general $U(x)$ is affected by substrate binding. Therefore, in a minimal model the chemical state of the binding pocket (the catalytic site)

can be: Emp (empty), Rec (reactant bound), or Prod (product bound). As shown in Figure 1B, each state is described by a potential curve $U_i(x)$ along the conformational coordinate, and localized transitions can happen between two potentials. For an enzymatic cycle, a reactant molecule first binds onto the catalytic site (Emp→Rec), then forms a more compact complex, next the chemical reaction happens (Rec→Prod), and finally the catalytic site is open and the product is released (Prod→Emp). Notice that binding molecules may shift both the curve shape and minimum position, and some conformational motion is necessary during the cycle. The harmonic shape of the curves shown in Figure 1B is only illustrative. A more complete description is to use the two (or higher) dimensional potential surfaces plotted in Figure 1C. The plot should be only viewed as illustrative. Within an enzymatic cycle, the system zigzags through the potential surface, with motions along both the conformational and reaction coordinates coupled. Figure 1D gives projection of the potential surface along the reaction coordinate at two conformational coordinate values. The curves have the characteristic double well shape. For barrier crossing processes, a system spends most of the time at potential wells, and the actual barrier-crossing time is transient and fast. Therefore, one can reduce the two-dimensional surface (Figure 1C) to one-dimensional projections along the conformational coordinate (Figure 1B), and approximate transitions along the reaction coordinate by rate processes among the one-dimensional potential curves.

With the above introduction of potential curves, we can now formulate the governing dynamic equations by a set of over-damped Langevin equations coupled to Markov chemical transitions (Xing, 2007; Zwanzig, 2001),

$$\zeta_i \frac{dx(t)}{dt} = -\frac{dU_i(x)}{dx} + f_i(t)\,, \tag{3}$$

where x and U_i as defined above, ζ_i is the drag coefficient along the molecular conformational coordinate, and f is the random fluctuation force with the property $\langle f(t)f(t')\rangle = 2\,k_B T\zeta\delta(t-t')$, with k_B the Boltamann's constant, T the temperature. Chemical transitions accompany motions along the conformational coordinate with x-dependent transition rates. In general the dynamics may be non-Markovian and contain a memory effect (Zwanzig, 2001). Min et al. observed a power law memory kernel for single protein conformational fluctuations (Min et al., 2005b). Xing and Kim showed that the observation can be well reproduced using a coarse-grained protein fluctuation model, with both of two adjustable parameters agree with other independent studies (Xing and Kim, 2006). However here we will assume Markovian dynamics for simplicity. The Langevin dynamics described by Equation 3 can be equally described by a set of coupled Fokker-Planck equations,

$$\frac{\partial}{\partial t}\rho_i(x) = -\frac{D_i}{k_B T}\cdot\frac{\partial}{\partial x}\left(-\frac{\partial U_i(x)}{\partial x}\rho_i(x)\right) + D_i\frac{\partial^2\rho_i(x)}{\partial x^2} + \sum_{j\neq i}\left(K_{ij}(x)\rho_j(x) - K_{ji}(x)\rho_i(x)\right) \tag{4}$$

Where $D_i = k_B T/\zeta_i$ is the diffusion constant, K_{ij} is the transition matrix element, and $\rho_i(x)$ is the probability density to find the system at position x and state i.

The formalism given by Equations 3 and 4 is widely used to model systems such as electron transfer reactions, protein motors (Bustamante et al., 2001; Julicher et al., 1997; Wang and Oster, 1998; Xing et al., 2006; Xing et al., 2005), as well as enzymatic reactions here (Gopich and Szabo, 2006; Min et al., 2008; Qian et al., 2009; Xing, 2007).

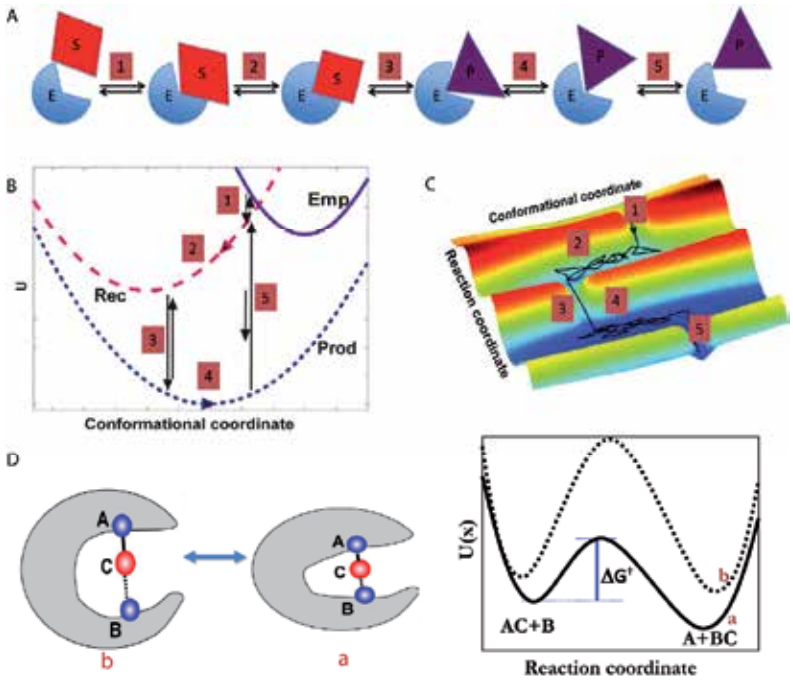


Fig. 1. Descriptions of coupling between chemical reactions and conformational changes.
(A) A discrete enzymatic cycle model with conformational changes. (B) A minimal continuous model representing three potentials of mean force along a conformational coordinate.
(C) A continuous model with explicit reaction and conformational coordinates.
(D) Two protein conformations and the corresponding potentials of mean force along the reaction coordinate.

The continuous form of Equation 4 can also be discretized to a form more familiar to biochemists[1],

$$\begin{array}{ccccccc} E_1 + S & \underset{\alpha_{-1}}{\overset{\alpha_1}{\rightleftharpoons}} & E_1S & \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} & E_1P & \underset{\beta_{-1}}{\overset{\beta_1}{\rightleftharpoons}} & E_1 + P \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ E_2 + S & \underset{\alpha_{-2}}{\overset{\alpha_2}{\rightleftharpoons}} & E_2S & \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} & E_2P & \underset{\beta_{-2}}{\overset{\beta_2}{\rightleftharpoons}} & E_2 + P \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ \cdots & & & & & & \\ \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\ E_N + S & \underset{\alpha_{-N}}{\overset{\alpha_N}{\rightleftharpoons}} & E_NS & \underset{k_{-N}}{\overset{k_N}{\rightleftharpoons}} & E_NP & \underset{\beta_{-N}}{\overset{\beta_N}{\rightleftharpoons}} & E_N + P \end{array} \qquad (5)$$

[1] A mathematical procedure for the discretization is given in Xing, J., Wang, H.-Y., and Oster, G. (2005). From continuum Fokker-Planck models to discrete kinetic models. Biophys J *89*, 1551-1563.

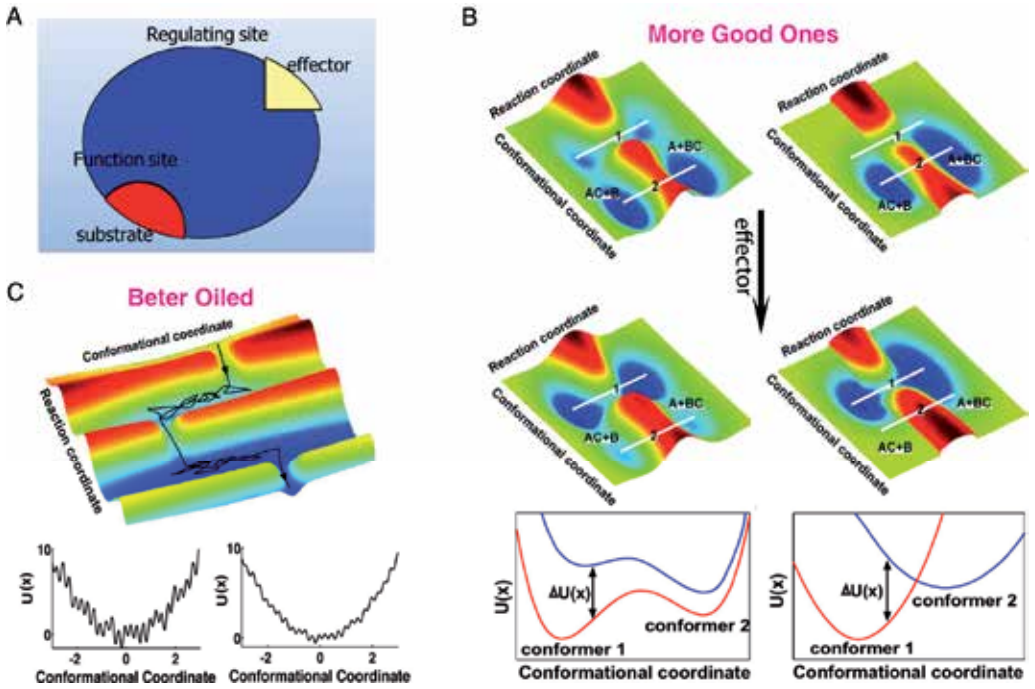


Fig. 2. Different models for allostery. (A) Schematic illustration of allosteric regulation. (B) Schematic potentials of mean force illustrating the MWC (left) and the KNF (right) models. (C) A nonequilibrium dynamic model.

Equations 3-5 describe richer physics than the simple induced fit model does. The conformational changes include contributions from binding induction as well as enzyme spontaneous fluctuations. There may be a number of parallel pathways for an enzymatic reaction corresponding to different protein conformations. An optimal conformation for one step of the reaction may not be the optimal conformation for another step. If an enzyme can transit among these conformations faster than a chemical transition event (including substrate/product binding and release), then the system can mainly follow the tortuous optimal pathway involving different conformations shown in Figure 1B and C. If the conformational change is comparable or slower than chemical events, multiple pathways may contribute significantly to the dynamics, and one observes time varying enzyme activity at the single molecule level, which leads to the phenomenon "dynamic disorder". One origin of the slow dynamics of intramolecular dynamics comes from diffusion along rugged potential surfaces with numerous potential barriers (Frauenfelder et al., 1991). Zwanzig shows that the effectic diffusion constant is greatly reduced along a rugged potential (Zwanzig, 1988). For example, for a rugged potential with a gaussian distributed barrier height, and root-mean-square ε, the so-called roughness parameter, the effective diffusion constant is scaled as,

$$D_{effective} = D\exp\left[-(\varepsilon / k_B T)^2\right] \tag{6}$$

which can be greatly reduced from the bare value of D.

3. Thermodynamic versus dynamic models for allostery

A cell needs to adjust its metabolic, transcriptional, and translational activities to respond to changes in the external and internal environment. Allostery and covalent modification are two fundamental mechanisms for regulating protein activities (Alberts et al., 2002). Allostery refers to the phenomenon that binding of an effector molecule to a protein's allosteric site affects the protein activity at its active site, which is usually physically distinct from where the effector binds. The discovery of allosteric regulations was in the 1950s, followed by a general description of allostery in the early 1960s, has been regarded as revolutionary at that time (Alberts et al., 2002). Not surprisingly, to understand the mechanism of allosteric regulation is an important topic in structural biology. Below we will focus on allosteric enzymes. For simplicity, we will restrict our discussions to positive allosteric effect, i.e., effector binding increases enzymatic activity. The discussions can be easily generalized to negative allosteric effects.

3.1 Conventional models of allostery

There are two popular models proposed to explain the allosteric effects. The concerted MWC model by Monod, Wyman, and Changeux, assumes that an allosteric protein can exist in two (or more) conformations with different reactivity, and effector binding modifies the thermal equilibrium distribution of the conformers (Monod et al., 1965). Recent population shift models re-emphasize the idea of preexisting populations (Goodey and Benkovic, 2008; Kern and Zuiderweg, 2003; Pan et al., 2000; Volkman et al., 2001). The sequential model described by Koshland, Nemethy, and Filmer is based on the induced-fit mechanism, and assumes that effector binding results in (slight) structural change at another site and affects the substrate affinity (Koshland et al., 1966). While different in details, both of the above models assume that the allosteric mechanism is through modification of the equilibrium conformation distribution of the allosteric protein by effector binding. For later discussions, we denote the mechanisms as "thermodynamic regulation".

The mechanisms of thermodynamic regulation impose strong requirements on the mechanical properties of an allosteric protein. The distance between the two binding sites of an allosteric protein can be far. For example, the bacterial chemotaxis receptor has the two reaction regions separated as far as 15 nm (Kim et al., 2002). In this case, signal propagation requires a network of mechanical strain relaying residues with mechanical properties distinguishing them well from the surroundings to minimize thermal dissipation – Notice that distortion of a soft donut at one side has negligible effect on another side of the donut. Mechanical stresses due to effector molecule binding irradiate from the binding site, propagate through the relaying network, and con-focus on the reaction region at the other side of the protein (Amaro et al., 2009; Amaro et al., 2007; Balabin et al., 2009; Cecchini et al., 2008; Cui and Karplus, 2008; Horovitz and Willison, 2005; Ranson et al., 2006). However, it is challenging to transmit the mechanical energy faithfully against thermal dissipation over a long distance. A possible solution is the attraction shift model proposed by Yu and Koshland(Yu and Koshland, 2001).

From a chemical physics perspective, current existing models on allosteric effects differ in some details of the potential shapes. The MWC and the recent population-shift model emphasizes that there are pre-existing populations for all the possible forms, as exemplified by the double well shaped potentials and the two corresponding conformers in the left panel of

Figure 2C. Effector binding only shifts their relative populations. The KNF model emphasizes that without the effector the protein exists mainly in one form (conformer 2 in the right panel of Figure 2C). Effector binding shifts the protein to another form (conformer 1) with different reactivity. The functions $U(x)$ are potentials of mean force, which suggests that the effect of effector binding can be enthalpic or entropic (Cooper and Dryden, 1984). Therefore in some sense there is no fundamental difference between the KNF and MWC models. They differ only in the extent of each conformer being populated, which is related to the free energy difference between conformers ΔU (in Figure 2C) through the Boltzman factor.

3.2 Possibly neglected dynamic aspect of allostery

The above allosteric models focus on the conformational changes decoupled from those changes associated with an enzymatic cycle. Consequently, the distribution along the conformational coordinate can be described as thermodynamic equilibrium. However, as discussed in section 2, an enzymatic cycle usually inevitably involves enzyme conformational changes, so the distribution of the latter is in general driven out of equilibrium due to coupling to the nonequilibrium chemical reactions. In many cases, as Frieden wrote, "conformational changes after substrate addition but preceding the chemical transformation, or after the chemical transformation but preceding product release may be rate-limiting" (Frieden, 1979). Recent NMR studies further demonstrate conformational changes as rate-limiting steps (Boehr et al., 2006; Cole and Loria, 2002). Based on these experimental observations, Xing proposed that the conformational change dynamics within an enzymatic cycle can be subject to allosteric modulation (Xing, 2007).

Enzyme conformational changes can be thermally activated barrier crossing events, and effectors function by modifying the height of the dominant barrier. Alternatively, effectors may accelerate conformational changes through decreasing the potential roughness (see Figure 2D). Intuitively, for the latter mechanism effectors transform rusty engines (enzymes) into better-oiled ones.

Figure 3 schematically summarizes possible effector binding induced changes of the potentials of mean force along a conformational coordinate, which then affects the

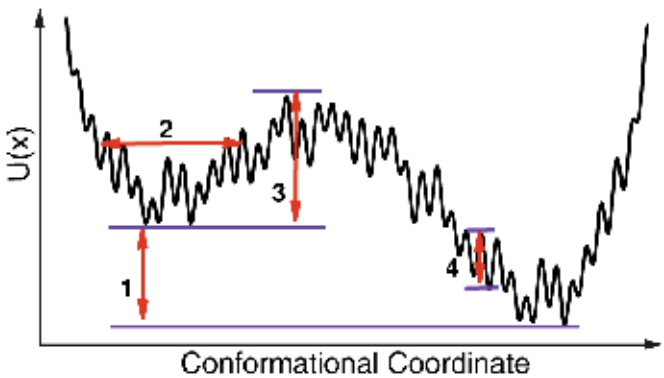


Fig. 3. Summary of effects of effector binding on the potential of mean force: (1) relative free energy difference of the two conformers; (2) Width of the potential well; (3) Barrier height; (4) Potential roughness.

enzymatic reaction dynamics. The changes can be the relative height of the potential wells representing different conformers (labelled 1 in Figure 3, enthalpic), the widths of potential wells (labelled 2, entropic), the barrier height (labelled 3) and the potential roughness (labelled 4) (dynamic). For a given enzyme subject to a given effector regulation, one or more effects may play the dominant role.

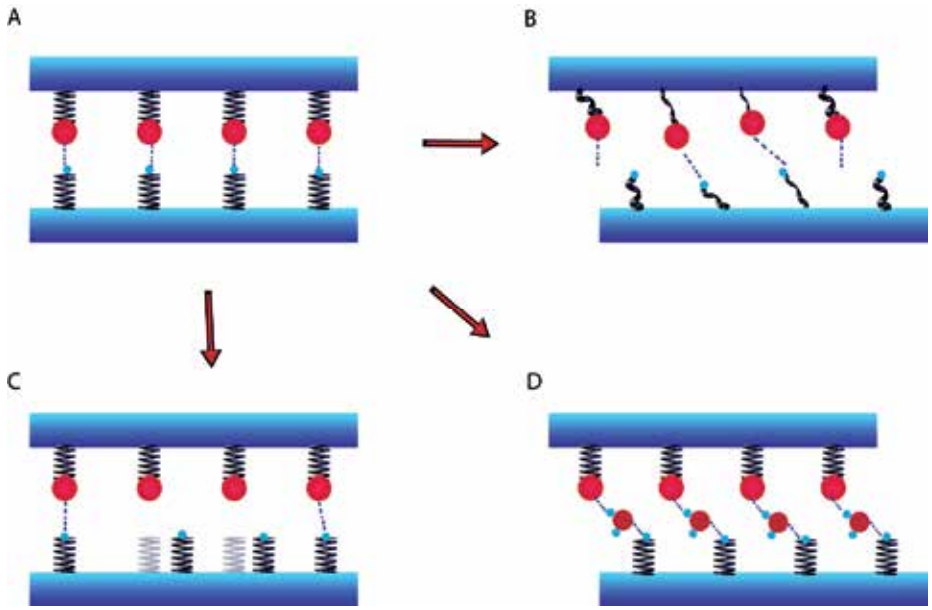


Fig. 4. Possible scenarios of modifying potential roughness. Relative motion between two protein surfaces (A) can be modulated through changing the linkage stiffness (B) or the arrangement of surface residues (C), or solvant accessibility (D).

Further experimental and theoretical studies are necessary to reveal the detailed molecular mechanisms for the proposed potential roughness regulation. Figure 4 gives some possible scenarios. Suppose during the process of conformational change, two protein surfaces need to move along each other, with numerous residues dangling on the surfaces forming and breaking noncovalent interaction pairs, e.g., hydrogen bonds. If these residues are rigidly connected to the protein body, one can treat the process as two rigid bodies moving relative to each other. At a given instance moving of the two surfaces requires breaking of all the previously formed interaction pairs (see Figure 4A). The repetitive breaking and forming interaction pairs result in rugged potentials along the moving coordinate. Effector binding may increase the elasticity of the residue linkages or the protein body. Then the two surfaces can move with some of the existing interaction pairs being stretched but not necessarily broken (see Figure 4B). Formation of new interaction pairs may energetically facilitate eventual broken of these bonds. This increased elasticity effectively smoothen the potential of mean force. Similarly, effector binding induced displacement of some residues may also reduce the average number of interaction pairs formed at a given relative position of the two surfaces. Effector binding may also increase solvent (water) molecule accessibility to the protein interface. Water molecules are effective on bridging interactions between displaced residues, and thus stabilizing the intermediate configurations (see Figure 4D).

3.3 Allosteric regulation of bacterial flagellar motor switching

Here we specifically discuss allosteric regulation in the bacterial flagellar motor system. Although the flagellar motor switching process does not involve enzymatic cycles directly, the process shares some features common to what we discussed in section 3.2.

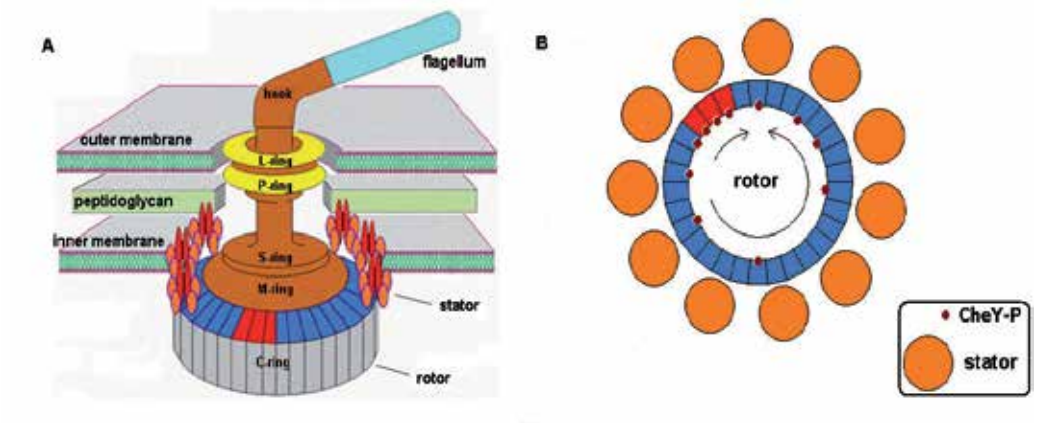


Fig. 5. Cartoon illustrations of the BFM torque generation/switching structure and the concept of conformational spread on the rotor ring (A) Schematic plot of the main structural components of the BFM. In this figure some rotor units (red) are in CW state against majority of the rotor units (blue) driving the motor rotating along CCW direction. (B) Top-view of the rotor ring complex with putative binding positions of the CheY-P molecules.

The bacterial flagellar motor (BFM) is a molecular device most bacteria use to rotate their flagella when swimming in aqueous environment. Using the transmembrane electrochemical proton (or sodium) motive force as the power source, the bacterial flagellar motor can rotate at an impressive high speed of a few hundred Hz and consequently, free-swimming bacteria can propel their cell body at a speed of 15-100 $\mu m/s$, or up to 100 cell body lengths per second (Berg, 2003, 2004; Sowa and Berry, 2008). Figure 5A shows a schematic cartoon plot of the major components of the *E. coli* BFM derived from previous research of electron microscopy, sequencing and mutational studies. These structural components can be categorized into two groups according to their function: the rotor and the stators. In the center of the motor, a long extracellular flagellum (about 5 or 10 times the length of the cell body) is connected to the basal body of the motor through a flexible hook domain. The basal body consists of a few protein rings, functioning as the rotor of the machine, and spans across the outer membrane, peptidoglycan and inner membrane into the cytoplasm of the cell (Berg, 2004). Around the periphery of the rotor, a circular array of 8-11 stator complexes are located. Each stator complex functions independently as a torque generation unit. When ions (proton or sodium) flow from periplasm to cytoplasm through an ion channel on the stator complex, conformational changes are triggered by ion binding on/off events, and therefore deliver torque to the rotor at the interface between the cytoplasmic domain of the stator complex and C-terminal domain of one of the 26 copies of FliG monomers on the rotor (Sowa et al., 2005). A series of mathematical models haven been proposed to explain the working mechanism of the BFM (Bai et al., 2009; Meacci and Tu, 2009; Mora et al., 2009; Xing et al., 2006).

The bacterial flagellar motor is not only important for the propulsion of the cell, but also crucial for bacterial chemotaxis. In the *E. coli* chemotaxis system, chemical gradients (attractant or repellent) are sensed through multiple transmembrane methyl-accepting chemotaxis proteins (MCPs) (Berg, 2004). When extracellular chemotactic attractants (or repellents) bind to MCPs, conformational changes through the membrane inhibit (or trigger) the autophosphorylation in the histidine kinase, CheA. CheA in turn transfers phosphoryl groups to conserved aspartate residues in the response regulators CheY. The phosphorylated form of CheY, CheY-P, diffuses away across the cytoplasm of the cell and binds to the bottom of the FliM/FliN complex of the flagellar motor. When attractant gradient is sensed, CheY-P concentration is low in the cytoplasm and therefore less CheY-P molecules bind to the flagellar motor, which favours counter-clockwise (CCW) rotation of the motor. When most of the motors on the membrane spin CCW, flagellar filaments form a bundle and propel the cell steadily forward. When repellent gradient is sensed, CheY-P concentration is raised and more CheY-P binds to the flagellar motor, which leads to clockwise (CW) rotation of the motor. When a few motors (can be as few as one) spin CW, flagellar filaments fly apart and the cell tumbles. The bacterial flagellar motor (BFM) switches stochastically between CCW and CW states and therefore the cell repeats a 'run'-'tumble'-'run' pattern. This enables a chemotactic navigation in a low Reynolds number environment (reviewed in Berg, book E. coli in motion). The ratio of the rotation direction CCW/CW is tuned by the concentration of the signalling protein, CheY-P.

The problem of BFM switching response to cytoplasmic CheY-P concentration is essentially a protein allosteric regulation. When the effector (CheY-P) binds to the bottom of each rotor unit (a protein complex formed by roughly 1:1:1 of FliG, FliM, FliN protein), it makes CW rotation more favourable (Figure 5B). However, a careful examination of the BFM switching shows that the allosteric regulation here has distinct features: 1) in previous *in vivo* experiment (Cluzel et al., 2000), Cluzel et al. monitored in real time the relationship between BFM switching bias and CheY-P concentration in the cell and found that the response curve is ultrasensitive with a Hill coefficient of ~ 10. Later FRET experiment further showed that binding of CheY-P to FliM is much less cooperative than motor switching response (Sourjik and Berg, 2002). The molecular mechanism of this high cooperativity in BFM switching response remains unknown. 2) the BFM rotor has a ring structure, which is a large multisubunit protein complexes formed by 26 identical rotor units. For such a large multisubunit protein complex, an absolute coupling between subunits as the MWC model requires seems very unlikely. 3) the BFM rotates in full speed stably in CCW or CW directions, and transitions between these two states are brief and fast. This indicates that the 26 rotor units on the basal body of the BFM are in a coherent conformation for most of the time and switching of the whole ring can finish within a very short time period. The above facts also put the KNF model in doubt. As in the KNF model, coupling between effector binding and conformation is absolute: When an effector binds a rotor unit, that rotor unit switches direction.

Therefore a new type of model is in needed to explain the molecular mechanism of the BFM switching. Duke et al. constructed a mathematical model of the general allosteric scheme based on the idea proposed by Eigen (Eigen, 1968) in which both types of coupling are probabilistic (Duke et al., 2001; Duke and Bray, 1999). This model encompasses the classical mechanisms at its limits and introduces the mechanism of conformational spread, with

domains of a particular conformational state growing or shrinking faster than ligand binding. Particular regions in the parameter space of the conformational spread model reproduce the classical WMC and KNF model (Duke et al., 2001).

Here we introduce the conformational spread model modified for studying the BFM switching mechanism. In this model, first we assumed that each rotor unit can take two conformations: CCW and CW. The rotor unit in CCW state generates torque along CCW direction when interacting with a stator unit; the rotor unit in CW state generates torque along CW direction when interacting with a stator unit. Each rotor unit undergoes rapid flipping between these two conformations and may also bind a single CheY-P molecule. On the free energy diagram, we further assumed that for each rotor unit the CCW state is energetically favoured by Ea while the binding of CheY-P stabilizes the CW state. As shown in Figure 6A, the free energy of the CW state (red) changes from $+E_A$ to $-E_A$ relative to the CCW state (blue), when a rotor unit binds CheY-P.

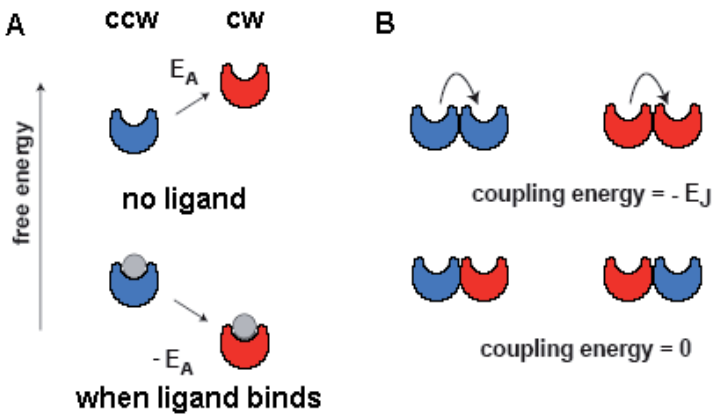


Fig. 6. Energy states of a rotor unit in the BFM switch complex. (A) The free energy of the CW state (red) changes from $+E_A$ to $-E_A$ relative to the CCW state (blue), when a rotor unit binds CheY-P. (B) The rotor unit is stabilized by E_J if the adjacent neighbor is in the same conformation.

In order to reproduce the ultrasensitivity of the BFM switching, a coupling energy E_J between adjacent neighbors in the ring is introduced. The free energy of a rotor unit is further stabilized by a coupling energy E_J when each neighboring rotor unit is in the same conformational state (Figure 6B), an idea inspired by the classical Ising phase transition theory from condensed matter physics.

In this conformational spread model, the rotor ring shows distinct features upon increasing of E_J. Below a critical coupling energy, the ring exhibits a random pattern of states as the rotor units flip independently of each other. Above the critical coupling energy, switch-like behaviour emerges: the ring spends the majority of time in a coherent configuration, either all in CCW or CW states, with abrupt stochastic switching between these two states. Unlike the MWC model, the conformational spread model allows the

existence of an intermediate (or mixed) configuration of the rotor units on the ring; and unlike the KNF model, the conformational spread model also allows rotor units stay in its original conformation without being switched by effector binding events. By implementing parallel Monte Carlo processes, one can simulate BFM switching response to CheY-P concentration. In each iteration, each rotor unit on the ring is visited and polled to determine whether to stay in the old state or jump to a new state according to the free energy difference between the two states as a function of 1) free energy of the rotor unit itself 2) binding condition of the regulator molecule CheY-P 3) energy coupling of adjacent neighboring subunits.

The conformational spread model has successfully reproduced previous experimental observations that 1) the BFM switching bias responses ultrasensitively to changes in CheY-P concentration 2) the motor rotates stably in CCW and CW states with occasional fast transitions from one coherent state to the other. The model also made several new predictions: 3) creation of domains of the opposite conformation is frequent due to fast flipping of single rotor unit, but most of them shrink and disappear, failing to occupy the whole ring. However, some big fluctuations can still produce obvious slowdowns and pausing of the motor. Therefore, speed traces of the BFM should have frequent transient speed slowdowns and pauses. 4) the switch interval (the time that the motor spends in the CCW or CW state) follows a single exponential distribution. 5) the switch time, the time that the motor takes to complete a switch, is non-instantaneous. It can be modeled as a biased random walk along the ring. The characteristic switching time depends on the size of the ring and flipping rate of each rotor unit in a complicated manner. Due to the stochastic nature of this conformational spread, we expect to see a wide distribution of switching times.

With the cutting-edge single molecule detection technique, the above predictions of the conformational spread model has recently been confirmed (Bai et al., 2010). Instead of instant transition , switches between CCW and CW rotor states were found to follow a broad distribution, with switching time ranging from less than 2 milliseconds to several hundred milliseconds, and transient intermediate states containing a mixture of CW/CCW rotor units have been observed. The conformational spread model has provided a molecular mechanism for the BFM switching, and more importantly, it sheds light on allosteric regulation in large protein complexes. In addition to the canonical MWC and KNF models, the conformational spread model provides a new comprehensive approach to allostery, and is consistent with the discussion in section 3.2 that both kinetic and thermodynamic aspects should be considered.

4. Coupling between slow conformational change and network dynamics

A biological network usually functions in a noisy ever-changing environment. Therefore, the network should be: 1) robust — functioning normally despite environmental noises; 2) adaptive — the tendency to function optimally by adjusting to the environmental changes; 3) sensitive — sharp response to the regulating signals. It is not-fully understood how a biological network can achieve these requirements simultaneously. Contemporary researches emphasize that the dynamic properties of a network is closely related to its topology.

Many *in vivo* biological processes involve only a small number of substrate molecules. When this number is in the range of hundreds or even smaller, stochastic effect becomes predominant. Chemical reactions take place in a stochastic rather than deterministic way. Therefore one should track the discrete numbers of individual species explicitly in the rate equation formalism. So far, many studies have shown that one might make erroneous conclusions without considering the stochastic effect (Samoilov et al., 2005; Wylie et al., 2007). Noise propagation through a network is currently an important research topic (Levine et al., 2007; Paulsson et al., 2000; Pedraza and van Oudenaarden, 2005; Rao et al., 2002; Rosenfeld et al., 2005; Samoilov et al., 2005; Shibata and Fujimoto, 2005; Suel et al., 2007; Swain et al., 2002). One usually assumes that the stochastic effect mainly arises from small number of identical molecules, and rate constants are still assumed well defined.

With the existence of dynamic disorder, the activity of a single enzyme (and so of a small number of enzymes) is a varying quantity. This adds another noise source with unique (multi-time scale, non-white noise) properties (Min and Xie, 2006; Xing and Kim, 2006). For bulk concentrations, fluctuations due to dynamic disorder are suppressed by averaging over a large number of molecules. However, existence of NMM kinetics can still manifest itself in a network. If there are only a small number of protein molecules, as in many *in vivo* processes, dynamic disorder will greatly affect the network dynamics. The conventionally considered stochastic effect is mainly due to number variations of identical molecules. Here a new source of stochastic effect arises from small numbers of molecules with the same chemical structure but different conformations. Dynamic disorder induced stochastic effect has some unique properties, which require special theoretical treatment, and may result in novel dynamic behaviors. First, direct fluctuation of the rate constants over several orders of magnitude may have dramatic effects on the network dynamics. Second, the associated time scales have broad range. The Gaussian white noise approximation is widely used in stochastic modeling of network dynamics with the assumption that some processes are much faster than others (Gillespie, 2000). Existence of broad time scale distribution makes the situation more complicated. Furthermore, a biological system may actively utilize this new source of noise. Noises from different sources may not necessarily add up. Instead they may cancel each other and result in smaller overall fluctuations (Paulsson et al., 2000; Samoilov et al., 2005). We expect that the existence of dynamic disorder not only further complicates the situation, but may also provide additional degrees of freedom for regulation since the rates can be continuously tuned. Especially we expect that existence of dynamic disorder may require dramatic modification on our understanding of signal transduction networks. Many of these processes involve a small number of molecules, and are featured by short reaction time scales (within minutes), high sensitivity and specificity (responding to specific molecules only).

Wu et al. examined the coupling between enzyme conformational fluctuations and a phosphorylation-dephosphorylation cycle (PdPC) (Wu et al., 2009). The PdPC is a common protein interaction network structure found in biological systems. In a PdPC, the substrate can be in phosphorylated and dephosphorylated forms with distinct chemical properties. The conversions are catalyzed by a kinase (E1 in Figure 7A) and a phosphatase (E2 in Figure 7A) at the expense of ATP hydrolysis. Under the condition that the enzymes are saturated by the substrates, the system shows ultrasensitivity (Goldbeter and Koshland, 1981). As shown in Figure 7B, the fraction of the phosphorylated substrate form, f(W-P), is close to zero if the ratio between E1 and E2 enzymatic activities $\theta < 1$, but close to 1 if $\theta > 1$. Now

consider a system with a finite size, e.g. 50 E1 molecules, 50 E2 molecules, and a total of 1500 substrate molecules as used to generate results in Figure 7C & D. Enzyme activities fluctuate due to conformational fluctuations. For simplicity let us assume that E1 can stochastically convert between an active and a less active forms. While the average value of $\theta = 1.1$, it fluctuates within the range [0.7, 1.5], depending on the number of E1 molecules in the active form. For convenience of discussion, let us also define the response time of the PdPC, τ, as the time it takes for the fraction of W-P reaching half given at time 0 the system jumps from $\theta < 1$ to $\theta > 1$ due to enzyme conformational fluctuations. The response time clearly related to the enzymatic turnover rate. As the trajectories in Figure7C show, for slow θ fluctuation $\Delta\theta$ is amplified to ΔW-P due to the ultrasensitivity of the PdPC and the much larger number of substrate molecules compared to enzymes. However, with θ fluctuation much faster than τ, the PdPC only responses to the average value of θ . Therefore depending on the relative time scale between θ fluctuation and the response time of the PdPC to θ change, fluctuations of θ can be either amplified or suppressed.

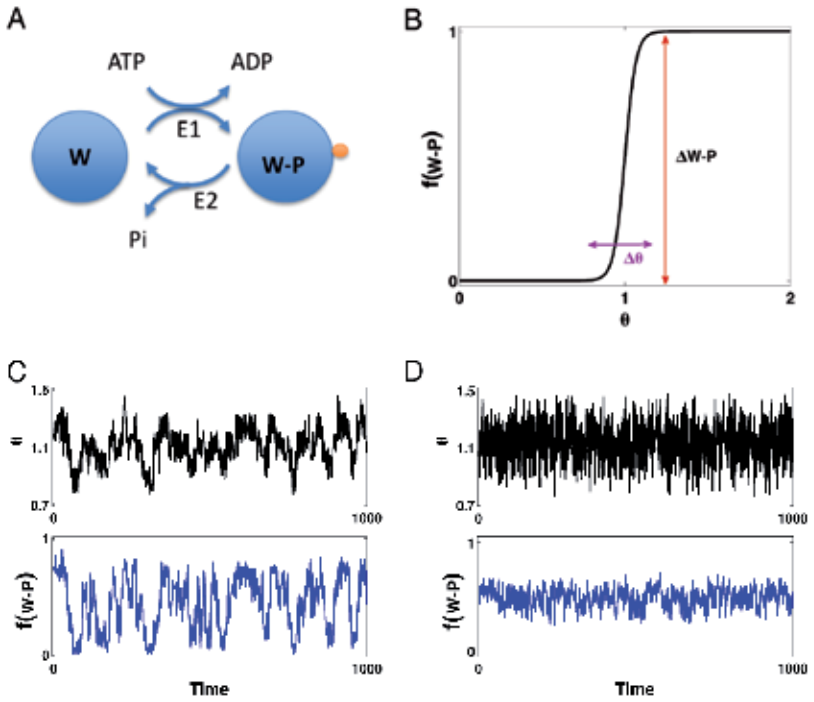


Fig. 7. Coupling between enzyme conformational fluctuations and a phosphorylation-dephosphorylation cycle. (A) A phosphorylation-dephosphorylation cycle (PdPC). (B) Ultrasensitivity of a PdPC. (C) Trajectories of enzyme activity due to slow conformational fluctuations and the corresponding substrate fluctuation. (D) Similar to C but with fast conformational fluctuations.

5. Conclusion

Slow conformational motions in macromolecules play crucial roles in their unique function in enzymatic reactions as well as biological networks. We suggest that these motions are of

great functional importance, which can only be fully appreciated in the context of regulatory networks. Collaborative researches from molecular and cellular level studies are urgently needed for this largely unexplored area.

6. Acknowledgment

ZW and JX were supported by an NSF grant (EF-1038636) and a grant from the William and Mary Jeffress Memorial Trust.

7. References

Ainslie, G.R., Jr., Shill, J.P., and Neet, K.E. (1972). Transients and Cooperativity. A slow transition model for relating transients and cooperative kinetics of enzymes. J Biol Chem 247, 7088-7096.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell, 4d edn (New York, Garland).

Alon, U. (2007). An introduction to systems biology: Design principles of biological circuits, 1 edn (Chapman and Hall/CRC).

Amaro, R.E., Cheng, X., Ivanov, I., Xu, D., and McCammon, J.A. (2009). Characterizing Loop Dynamics and Ligand Recognition in Human- and Avian-Type Influenza Neuraminidases via Generalized Born Molecular Dynamics and End-Point Free Energy Calculations. J Am Chem Soc 131, 4702-4709.

Amaro, R.E., Sethi, A., Myers, R.S., Davisson, V.J., and Luthey-Schulten, Z.A. (2007). A Network of Conserved Interactions Regulates the Allosteric Signal in a Glutamine Amidotransferase,Ä†. Biochemistry 46, 2156-2173.

Austin, R.H., Beeson, K.W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I.C. (1975). Dynamics of Ligand-Binding to Myoglobin. Biochemistry 14, 5355-5373.

Bai, F., Branch, R.W., Nicolau, Dan V.,, Jr., Pilizota, T., Steel, B.C., Maini, P.K., Berry, R.M. (2010). Conformational Spread as a Mechanism for Cooperativity in the Bacterial Flagellar Switch. Science 327, 685-689.

Bai, F., Lo, C.-J., Berry, R.M., and Xing, J. (2009). Model Studies of the Dynamics of Bacterial Flagellar Motors. Biophys J 96, 3154-3167.

Balabin, I.A., Yang, W., and Beratan, D.N. (2009). Coarse-grained modeling of allosteric regulation in protein receptors. Proceedings of the National Academy of Sciences 106, 14253-14258.

Behzadi, A., Hatleskog, R., and Ruoff, P. (1999). Hysteretic enzyme adaptation to environmental pH: change in storage pH of alkaline phosphatase leads to a pH-optimum in the opposite direction to the applied change. Biophys Chem 77, 99-109.

Berg, H.C. (2003). The Rotary Motor of Bacterial Flagella. Annu Rev Biochem 72, 19-54.

Berg, H.C. (2004). E. Coli in Motion (New York, Springer-Verlag Press).

Boehr, D.D., McElheny, D., Dyson, H.J., and Wright, P.E. (2006). The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. Science 313, 1638-1642.

Bustamante, C., Keller, D., and Oster, G. (2001). The physics of molecular motors. Acc Chem Res 34, 412-420.

Cecchini, M., Houdusse, A., and Karplus, M. (2008). Allosteric Communication in Myosin V: From Small Conformational Changes to Large Directed Movements. PLoS Comput Biol 4, e1000129.

Cluzel, P., Surette, M., and Leibler, S. (2000). An Ultrasensitive Bacterial Motor Revealed by Monitoring Signaling Proteins in Single Cells. Science 287, 1652-1655.

Cole, R., and Loria, J.P. (2002). Evidence for Flexibility in the Function of Ribonuclease A,Ä†. Biochemistry 41, 6072-6081.

Cooper, A., and Dryden, D.T.F. (1984). Allostery without Conformational Change - a Plausible Model. Eur Biophys J Biophys Lett 11, 103-109.

Cornish-Bowden, A., and Cardenas, M.L. (1987). Co-operativity in monomeric enzymes. J Theor Biol 124, 1-23.

Cui, Q., and Karplus, M. (2008). Allostery and cooperativity revisited. Protein Science 17, 1295-1307.

Duke, T., Novere, N.L., and Bray, D. (2001). Conformational spread in a ring of proteins: A stochastic approach to allostery. J Mol Biol 308.

Duke, T.A.J., and Bray, D. (1999). Heightened sensitivity of a lattice of membrane receptors. Proc Nat Acad Sci USA 96, 10104-10108.

Eigen, M. (1968). Kinetics of reaction control and information transfer in enzymes and nucleic acids. In Nobel Symp 5 on Fast Reactions and Primary Processes, Chem Kinetics, S. Claesson, ed. (Stockholm, Almquist & Wiksell).

English, B.P., Min, W., van Oijen, A.M., Lee, K.T., Luo, G.B., Sun, H.Y., Cherayil, B.J., Kou, S.C., and Xie, X.S. (2006). Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nat Chem Biol 2, 87-94.

Frauenfelder, H., Sligar, S.G., and Wolynes, P.G. (1991). The Energy Landscapes and Motions of Proteins. Science 254, 1598-1603.

Frauenfelder, H., Wolynes, P.G., and Austin, R.H. (1999). Biological physics. Rev Mod Phys 71, S419-S430.

Frieden, C. (1970). Kinetic Aspects of Regulation of Metabolic Processes. The hysteretic enzyme concept. J Biol Chem 245, 5788-5799.

Frieden, C. (1979). Slow Transitions and Hysteretic Behavior in Enzymes. Ann Rev Biochem 48, 471-489.

Gillespie, D.T. (2000). The chemical Langevin Equation. J Chem Phys 113, 297-306.

Goldbeter, A., and Koshland, D.E. (1981). An Amplified Sensitivity Arising from Covalent Modification in Biological-Systems. Proc Natl Acad Sci USA 78, 6840-6844.

Goodey, N.M., and Benkovic, S.J. (2008). Allosteric regulation and catalysis emerge via a common route. Nat Chem Biol 4, 474-482.

Gopich, I.V., and Szabo, A. (2006). Theory of the statistics of kinetic transitions with application to single-molecule enzyme catalysis. J Chem Phys 124, 154712.

Grote, R.F., and Hynes, J.T. (1980). The stable states picture of chemical-reactions 2: Rate constants for condensed and gas-phase reaction models. J Chem Phys 73, 2715-2732.

Hanggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate theory: 50 years after Kramers. Rev Mod Phys 62, 254-341.

Horovitz, A., and Willison, K.R. (2005). Allosteric regulation of chaperonins. Current Opinion in Structural Biology 15, 646-651.

Julicher, F., Ajdari, A., and Prost, J. (1997). Modeling molecular motors. Rev Mod Phys 69, 1269-1281.

Kern, D., and Zuiderweg, E.R.P. (2003). The role of dynamics in allosteric regulation. Curr Opin Struc Biol 13, 748-757.

Kholodenko, B.N. (2006). Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7, 165-176.

Kim, S.H., Wang, W.R., and Kim, K.K. (2002). Dynamic and clustering model of bacterial chemotaxis receptors: Structural basis for signaling and high sensitivity. Proc Natl Acad Sci USA 99, 11611-11615.

Koshland, D.E., Nemethy, G., and Filmer, D. (1966). Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. Biochemistry 5, 365-385.

Kroh, H.K., Panizzi, P., and Bock, P.E. (2009). Von Willebrand factor-binding protein is a hysteretic conformational activator of prothrombin. Proc Natl Acad Sci U S A 106, 7786-7791.

Levine, J., Kueh, H.Y., and Mirny, L. (2007). Intrinsic fluctuations, robustness, and tunability in signaling cycles. Biophys J 92, 4473-4481.

McCammon, J.A., and Harvey, S.C. (1987). Dynamics of Proteins and Nucleic Acids (New York, Cambridge Univ Press).

Meacci, G., and Tu, Y. (2009). Dynamics of the bacterial flagellar motor with multiple stators. Proc Natl Acad Sci USA 106, 3746-3751.

Menten, L., and Michaelis, M.I. (1913). Die Kinetik der Invertinwirkung. Biochem Z 49, 333-369.

Min, W., English, B.P., Luo, G.B., Cherayil, B.J., Kou, S.C., and Xie, X.S. (2005a). Fluctuating enzymes: Lessons from single-molecule studies. Acc Chem Res 38, 923-931.

Min, W., Gopich, I.V., English, B.P., Kou, S.C., Xie, X.S., and Szabo, A. (2006). When Does the Michaelis-Menten Equation Hold for Fluctuating Enzymes? J Phys Chem B 110, 20093-20097.

Min, W., Luo, G.B., Cherayil, B.J., Kou, S.C., and Xie, X.S. (2005b). Observation of a power-law memory kernel for fluctuations within a single protein molecule. Phys Rev Lett 94, 198302.

Min, W., and Xie, X.S. (2006). Kramers model with a power-law friction kernel: Dispersed kinetics and dynamic disorder of biochemical reactions. Phys Rev E 73, -.

Min, W., Xie, X.S., and Bagchi, B. (2008). Two-Dimensional Reaction Free Energy Surfaces of Catalytic Reaction: Effects of Protein Conformational Dynamics on Enzyme Catalysis. J Phys Chem B 112, 454-466.

Monod, J., Wyman, J., and Changeux, J.P. (1965). On Nature of Allosteric Transitions - a Plausible Model. J Mol Biol 12, 88-118.

Mora, T., Yu, H., and Wingreen, N.S. (2009). Modeling Torque Versus Speed, Shot Noise, and Rotational Diffusion of the Bacterial Flagellar Motor. Phys Rev Lett 103, 248102.

Pan, H., Lee, J.C., and Hilser, V.J. (2000). Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. Proceedings of the National Academy of Sciences 97, 12020-12025.

Paulsson, J., Berg, O.G., and Ehrenberg, M. (2000). Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. Proc Natl Acad Sci USA 97, 7148-7153.

Pedraza, J.M., and van Oudenaarden, A. (2005). Noise Propagation in Gene Networks. Science 307, 1965-1969.

Qian, H., Shi, P.-Z., and Xing, J. (2009). Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. Phys Chem Chem Phys 11, 4861-4870.

Ranson, N.A., Clare, D.K., Farr, G.W., Houldershaw, D., Horwich, A.L., and Saibil, H.R. (2006). Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes. Nat Struct Mol Biol 13, 147-152.

Rao, C.V., Wolf, D.M., and Arkin, A.P. (2002). Control, exploitation and tolerance of intracellular noise. Nature 420, 231-237.

Ricard, J., and Cornish-Bowden, A. (1987). Co-operative and allosteric enzymes: 20 years on. Eur J Biochem 166, 255-272.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene Regulation at the Single-Cell Level. Science 307, 1962-1965.

Samoilov, M., Plyasunov, S., and Arkin, A.P. (2005). Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci USA 102, 2310-2315.

Shibata, T., and Fujimoto, K. (2005). Noisy signal amplification in ultrasensitive signal transduction. Proc Natl Acad Sci USA 102, 331-336.

Sourjik, V., and Berg, H.C. (2002). Binding of the Escherichia coli response regulator CheY to its target measured in vivo by fluorescence resonance energy transfer. Proc Natl Acad Sci USA 99, 12669-12674.

Sowa, Y., and Berry, R.M. (2008). Bacterial flagellar motor. Quart Rev Biophys 41, 103-132.

Sowa, Y., Rowe, A.D., Leake, M.C., Yakushi, T., Homma, M., Ishijima, A., and Berry, R.M. (2005). Direct observation of steps in rotation of the bacterial flagellar motor. Nature 437, 916-919.

Suel, G.M., Kulkarni, R.P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M.B. (2007). Tunability and noise dependence in differentiation dynamics. Science 315, 1716-1719.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA 99, 12795-12800.

Tyson, J.J., Chen, K., and Novak, B. (2001). Network dynamics and cell physiology. Nat Rev Mol Cell Biol 2, 908-916.

Volkman, B.F., Lipson, D., Wemmer, D.E., and Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. Science 291, 2429-2433.

Wang, H., and Oster, G. (1998). Energy transduction in the F1 motor of ATP synthase. Nature 396, 279-282.

Wu, Z., Elgart, V., Qian, H., and Xing, J. (2009). Amplification and Detection of Single-Molecule Conformational Fluctuation through a Protein Interaction Network with Bimodal Distributions. J Phys Chem B 113, 12375-12381.

Wylie, D.C., Das, J., and Chakraborty, A.K. (2007). Sensitivity of T cells to antigen and antagonism emerges from differential regulation of the same molecular signaling module. Proc Natl Acad Sci USA, 0611482104.

Xie, X.S., and Lu, H.P. (1999). Single-molecule enzymology. J Biol Chem 274, 15967-15970.

Xing, J. (2007). Nonequilibrium dynamic mechanism for allosteric effect. Phys Rev Lett 99, 168103.

Xing, J., Bai, F., Berry, R., and Oster, G. (2006). Torque-speed relationship for the bacterial flagellar motor. Proc Natl Acad Sci USA 103, 1260-1265.

Xing, J., and Kim, K.S. (2006). Protein fluctuations and breakdown of time-scale separation in rate theories. Phys Rev E 74, 061911.

Xing, J., Wang, H.-Y., and Oster, G. (2005). From continuum Fokker-Planck models to discrete kinetic models. Biophys J 89, 1551-1563.

Yu, E.W., and Koshland, D.E. (2001). Propagating conformational changes over long (and short) distances in proteins. Proc Natl Acad Sci USA 98, 9517-9520.

Zwanzig, R. (1988). Diffusion in a Rough Potential. Proc Natl Acad Sci USA 85, 2029-2030.

Zwanzig, R. (1990). Rate-Processes with Dynamic Disorder. Acc Chem Res 23, 148-152.

Zwanzig, R. (2001). Nonequilibrium statistical mechanics (Oxford, Oxford University Press).

9

Prediction of Protein Interaction Sites Using Mimotope Analysis

Jian Huang, Beibei Ru and Ping Dai
School of Life Science and Technology
University of Electronic Science and Technology of China
China

1. Introduction

Biological functions depend on all kinds of interaction networks; life is a miracle of all types of molecular interactions. Among them, proteins interacting with proteins, nucleic acids and small compounds play a central role (Barabasi & Oltvai, 2004; Przulj, 2011; Vidal, et al., 2011). To guide protein engineering studies for better enzymes, antibodies and drugs, structural and functional characterization of protein interaction sites at the residue or atom level is of great help. Experimental approaches such as X-ray diffraction of protein complex can define structural binding sites at the atomic level (Bickerton, et al., 2011; Higurashi, et al., 2009); mutagenesis and binding test are capable of identifying functional binding sites at the residue or group level (Moreira, et al., 2007; Peng, et al., 2011). However, these means are costly, time-consuming and sometimes technically difficult or even impossible. Moreover, they are not always applicable on a large scale. As a result, computer tools for the prediction of protein interaction sites have been increasingly popular for complementing experimental techniques (Fernández-Recio, 2011; Wass, et al., 2011).

The existing methods for the prediction of protein interaction sites can be grouped into three categories based on the main input data used. The first category consists of methods using protein sequence as the only input (Ofran & Rost, 2007; Res, et al., 2005). Methods in the second category such as molecular docking and simulation solely use structure data as input (Kozakov, et al., 2010; Mashiach, et al., 2010). Methods of the third category make use of a mimotope motif or a set of mimotope sequences together with protein sequence or structure as input (Huang, et al., 2011).

In this chapter, we review methods of the third category, focusing on their current statuses, discussing challenges and providing suggestions to advance this field.

2. Mapping protein-protein interaction sites using mimotope analysis

Mimotopes are peptides mimicking protein interaction sites; they are initially acquired from chemical synthesis (Geysen, et al., 1986). High-throughput obtainment of mimotopes has achieved since phage display and other surface display technologies became available (Smith, 1985; Smith & Petrenko, 1997). Taking phage display as an example, random DNA

sequences can be inserted into genes coding for coat proteins of bacteriophage to make combinatorial libraries. As shown in Figure 1, the combinatorial library can be incubated and selected with an immobilized protein, termed as the target. The natural partner of the target is called as the template. Phages without affinity to the target are washed away with buffer. Then, bound phages are eluted with the target, the template or stronger buffer only. The bound phages are further amplified by infecting bacteria to form a secondary library, which is then used for the next round of incubating, washing, eluting and amplifying. After several rounds of such processes which are well known as biopanning, phage clones are picked randomly from the isolation of bound phages and sequenced. The affinities of these phage clones or corresponding peptides to the target are measured by surface plasmon resonance, enzyme-linked immunosorbent assay or other binding assays. The foreign inserts which enable corresponding phage clones to bind the target competitively with a template are considered as mimotopes.

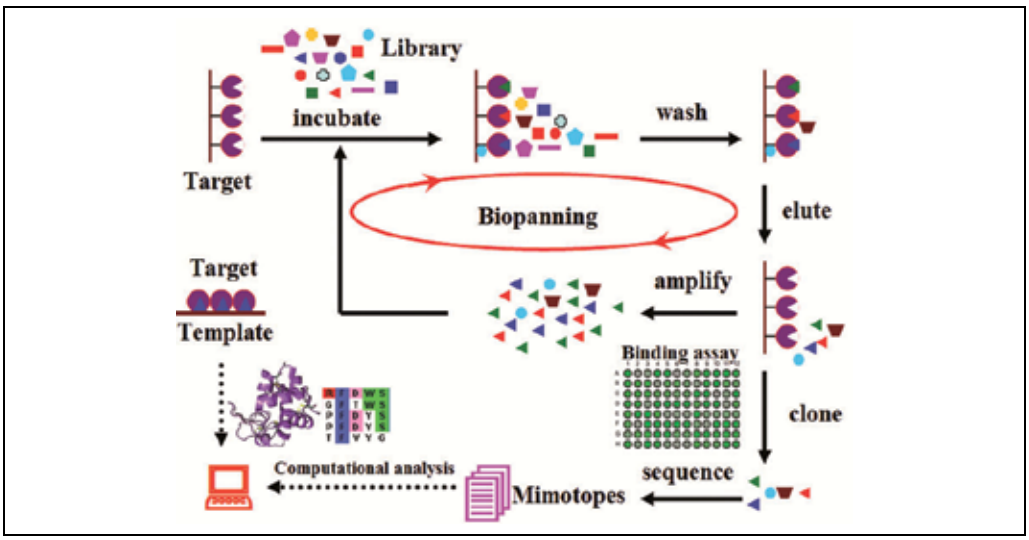


Fig. 1. Schematic view of in vitro phage display and mimotope analysis.

As described above, a set of mimotopes can be readily obtained via phage display. They are capable of binding to the target and blocking the interaction between the target and the template. Therefore, it implies that the information of protein interaction sites is encoded in mimotopes and can be predicted by decoding mimotopes. It is only natural to suppose that the mimotopes are similar to the binding site on the template at the sequential or structural level. Indeed, all approaches to prediction of protein interaction sites based on mimotopes depend on either the sequence or the structure of the template. Thus, the existing methods can be divided into the following two groups:

2.1 Methods based on template sequence

Various methods based on template sequence are summarized in Figure 2. In brief, a set of mimotopes are aligned with the corresponding template to find out the similar region in sequence, which is thought to be at least a part of the target-binding site on the template

protein. Sometimes, sequences of paralogs or orthologs of the template are also aligned to help the identification of the protein interaction site. In some studies, consensus sequences or motifs are derived from the blocks of mimotope alignments. Then, consensus sequences are aligned to the template sequence; motifs are scanned along the template sequence. And the template segments similar to consensus sequences or matching the motifs are considered to be a part of the protein interaction sites. If the template itself is not determined, local alignment search with each mimotope or the consensus sequence against the protein database would help to predict reasonable candidates of template and its binding sites. The template and corresponding interaction sites can also be predicted through pattern search with mimotope motifs against the protein database.

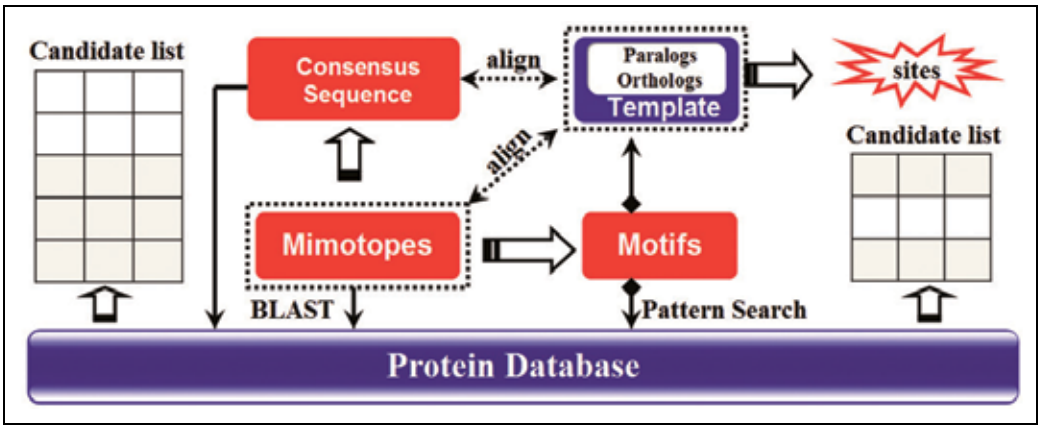


Fig. 2. Flow chart of methods based on template sequence.

As shown in Figure 2, methods based on template sequence involve several steps and tasks such as aligning sequence, inferring consensus sequence or motifs, searching local alignment or motif against the protein database. Among them, sequence alignment is undoubtedly the most important one. Methods based on template sequence can be fulfilled with visual inspection, general-purpose programs and tools specially designed for mimotope analysis.

2.1.1 Manual sequence analysis with visual inspection

Some mimotopes are very similar or even identical to some part of the template sequence every now and then, indicating the segment involving in binding the target protein. In this situation, the protein binding site can be easily depicted through aligning mimotope and template sequence manually by visual inspection. A 6mer random library was screened with the monoclonal antibody GDO5 raised against the Hantaan virus glycoprotein G2. After three rounds of panning, the mimotope obtained had the sequence LEYPWH, which was very similar to the template sequence 94YEYPWH99, implying the site where GDO5 bound (Fack, et al., 1997). The Ph.D.-7 random phage library was panned using the anti-SEB monoclonal antibody ab53981 (Urushibata, et al., 2010). Among the mimotopes obtained, SPDELHK was almost identical to 8PDELHK13 of the staphylococcal enterotoxin B. The ab53981 binding site was thus located. Four anti-HBsAg

monoclonal antibodies, namely H5, H35, H53 and H166 were characterized using phage display (Chen, et al., 1996). Manually aligning 16 H166-binding mimotopes with the HBsAg sequences from subtype adw2, ayw2 and ayw3, Chen et al found that most mimotopes have the CRTC or CKTC subsequences by visual inspection, which were identical to the segments from 121 to 124 of the HBsAg. The epitope recognized by H166 was thus indicated. Nonetheless, sequence alignment programs are necessary when there are a lot of sequences to be aligned or the similarity between the mimotope and template sequence is not obvious.

2.1.2 General-purpose sequence analysis tools

General-purpose tools for sequence alignment, local alignment and pattern search have been widely used in the prediction of protein interaction sites based on mimotope and template sequences. As we described above, Chen et al identified the H166-binding site by visual inspection. However, the software GENEWORK was used in the left three cases (Chen, et al., 1996). Significant matches were found by manual analysis on the dot-matrix diagrams produced by GENEWORK. For example, ARARCEHRSGLSL as one part of an H35-selected mimotope was aligned to 166ASARFSWLSL175 of the HBsAg sequence from subtype ayw3, locating the H35-binding site (Chen, et al., 1996). MDM2-binding peptides were obtained using mRNA display (Shiheido, et al., 2011); the peptides were aligned using the ClustalW program (Larkin, et al., 2007). Compared with the sequence of P53, a similar segment 17-28 was found be the MDM2-P53 interaction site (Shiheido, et al., 2011). A monoclonal antibody against the West Nile virus capsid protein was generated and designated as 6D3 (E. C. Sun, et al., 2011). A 12mer peptide library was screened with 6D3 to produce a set of mimotopes. Alignment revealed a consensus segment KKPGGPG, which was same to the subsequence 3-9 of West Nile virus capsid protein. A monoclonal antibody 2A10G6 was raised against the heat-inactivated dengue virus and used to screen the Ph.D.-12 random phage library. Alignment of mimotopes revealed a consensus FFDRTWP, which corresponded well with 98DRGW101 located at the tip of the fusion loop of E protein of dengue virus (Deng, et al., 2011). In the studies of Sun and Deng, MegAlign software within the Lasergene suite was used to align the orthologs of the template. In the study of Urushibata, the ClustalW program was used to align the paralogs of the template (Urushibata, et al., 2010). These studies showed that orthologs or paralogs were helpful for locating the binding sites.

Unlike the interaction between an antigen and corresponding antibody, a protein may have quite a few partners sometimes or its natural partner may be unknown. In these situations, the sequence alignment between mimotope and template cannot be done directly. However, a local alignment search against the protein sequence database helps to identify candidate templates and binding sites. The AC3 protein of geminiviruses was characterized using phage display (Pasumarthy, et al., 2011). Each AC3-specific peptide sequence obtained was then searched for local alignment against the Arabidopsis non-redundant protein database at NCBI through BLASTP program adjusted for short sequence (Mount, 2007). Proteins from a few metabolic pathways were identified as putative AC3-interacting proteins. For example, YALKHLPESTIP was very similar with 704YALKHIRES712 of the Hua Enhancer 1 (HEN1). Thus HEN1 might interact with the AC3 protein around 704YALKHIRES712

(Pasumarthy, et al., 2011). LipL32, the major outer membrane protein of pathogenic Leptospira, was panned against the Ph.D.-7 phage library. For each mimotope obtained, a BLASTP search against the protein database was performed. Quite a few proteins expressed on the surface of target cells of pathogenic Leptospira were suggested to interact with LipL32. For example, the mimotope HLPPNHT is similar with the sequence PLPPEHT of Collagen XX, indicating LipL32 might bind to Collagen XX at that site (Chaemchuen, et al., 2011). The strategy of mimotope blast against protein databases has also been used to deduce small molecule binding sites and drug targets (Chen, et al., 2006; Takakusagi, et al., 2010; Takami, et al., 2011), infer proteins involving in cell interactions (Kanki, et al., 2011; Zhao, et al., 2010).

Besides tools for local alignment search, pattern search against the protein database has also been used to find possible templates and their target-binding sites. SdrC is important in the interactions between Staphylococcus aureus and its host. However, the host ligand interacting with SdrC was not previously identified. The Ph.D.-12 phage library was screened with SdrC and eight phage clones displayed significantly higher affinity to SdrC. These clones were sequenced, and an alignment revealed the consensus sequence HHIHHFH. It was then used for a pattern search against the human protein database allowing for zero, one and two residue mismatches. The results showed that human neurexin 1β, 2β, 3β and a T-type voltage-dependent calcium channel might be the host ligands interacting with SdrC. Among them, the subsequence 10-16 of human neurexin 1β was identical to the consensus sequence, which implied SdrC might bind to human neurexin 1β at the site 10HHIHHFH16 (Perosa, et al., 2010). Autoantibodies against centromere associated protein A (CENP-A) were purified from sera of eight systemic sclerosis patients with the immunodominant epitope of CENP-A (Ap17–30). These antibodies were used to screen a phage library. The binding phage clones were sequenced, and the inserted peptides were aligned with MULTALIN (Corpet, 1988) to derive antigenic motifs. Human proteins containing such motifs were searched in the SwissProt Protein Sequence Database using the ScanProsite tool. Taking the PTPxxGPxxR motif as an example, 20PTPTPGPSRR29 of human CENP-A was certainly found. However, 53PTPAPGPGRR62 of human Forkhead box protein E3 (FOXE3) also matched the motif, indicating those autoantibodies could interact with FOXE3 at the site around 53-62. Indeed, the peptide 53-62 of FOXE3 was confirmed to behave similarly in binding and inhibition assays with anti-Ap17–30 IgG (Barbu, et al., 2010).

2.1.3 Specially designed sequence analysis tools

Even very recently, general-purpose tools for sequence alignment, local alignment and pattern search remain popular in the study of mapping protein interaction sites based on mimotopes. One reason for this is that these tools are freely, stably and conveniently available. However, these general-purpose tools have their limits. For example, most of them are not good at aligning a very short sequence (mimotope) to quite a long sequence (template). Furthermore, they are less efficient to deduce conformational binding sites, which are made of segments far away in primary sequence but close on the surface of template structure. Specially designed tools are thus needed for sequence analyses of mimotopes and templates.

To delineate conformational binding sites on protein, the program FINDMAP was proposed (Mumey, et al., 2003). FINDMAP allowed any permutations (e.g. inversion) of the mimotope sequence to align its template sequence. Furthermore, gaps even large gaps were permitted in both mimotope and template sequences. Such alignment was proven to be NP-complete and a branch-and-bound algorithm was used to solve the problem in practice. As FINDMAP could deal with only one mimotope each time, an improved version called EPIMAP was introduced later. It was capable of aligning each mimotope to the template, producing a set of top-scoring alignments, selecting the most mutually compatible alignments and filtering out spurious alignments (Mumey, et al., 2006). MimAlign was a meta-method. It combined results from four multiple sequence alignments of the template and its mimotopes (Moreau, et al., 2006). In the RELIC suite, there were quite a few tools specially designed for analysis on mimotopes (Mandava, et al., 2004). For example, MOTIF1 and MOTIF2 were designed to identify weak sequence motifs within short peptide sequence; MATCH, FASTAcon and FASTAskan were designed for optimal sequence alignments between mimotopes and its template. Although the RELIC suite focused on the interaction between small molecule and protein, its sequence tools were often used in the analysis of protein-protein interaction sites. For instance, MMACHC-binding peptides were aligned to MMADHC with tools in RELIC and five MMACHC-binding sites on the protein MMADHC were predicted (Plesa, et al., 2011). Mouse monoclonal antibodies against the predominant VSGs LiTat 1.3 and LiTat 1.5 of T.b. gambiense were used to screen PhD.-12 and Ph.D.-C7C phage libraries. Epitopes were identified by sequence alignment performed manually and with RELIC suite (Van Nieuwenhove, et al., 2011). For example, ALLPFKDHLPYP selected with the monoclonal antibody H12H3 against VSG LiTat 1.5 was aligned to 269AQAVYKDHDPDQ280 of VSG LiTat 1.5. The following experiment did show that the binding of H12H3 to synthetic ALLPFKDHLPYP was inhibited by human African trypanosomiasis sera. Regretfully, all the special tools described here are now hard or impossible to access.

2.1.4 Methods based on template sequence: challenges and suggestions

Methods based on template sequence are of their advantages. For example, they can be used in any condition because no structural information is required during prediction. Even if the template sequence is not given, local alignment or pattern search against protein databases may fulfil the task of inferring possible templates and protein interaction sites. However, to evaluate the results of sequence alignment, local alignment search and pattern search is still a great challenge.

Two formulae have been proposed to compute the frequency of finding similar sequences in two random sequences with different lengths (Chen, et al., 1996). One formula is for a single sequence match; another is for nearby matches within a pair of two sequences. This was a good attempt to evaluate if a continuous or discontinuous match was significant or just by chance. Chen et al assumed that 20 different residues were with equal probability at each position of the two sequences. However, it is not true in real case. To be more reasonable, we suggest using the residue frequency of the corresponding phage library for mimotopes and the actual frequency for template with long sequence. For short or unknown template, use the amino acids frequency of SwissProt.

Although the BLAST program has its statistical means to evaluate a match, they are not fit for short peptides such as mimotopes. In the study of Pasumarthy et al, a lot of matches were found. Among them, the mimotope FPKAFHHHKIY was found to be similar with 317HKIY310 of the Retinoblastoma like protein (pRBR) with an E-value of 1250 and pRBR was known to be an AC3-interacting protein. The mimotope DAMIMKKHWHRF was found to be similar with 164MIMK167 of the Geminivirus Rep interacting kinase 1 (GRIK1) with an E-value of 517 and GRIK1 did interact with the AC1 protein. Thus, they used the E-value 1250 as the threshold to filter the blast results. The candidate list was further shortened with one of the following conditions: (1) at least two hits from the same or different peptides; (2) with E-value less than 517 (Pasumarthy, et al., 2011). In another study, only a tri-peptide or longer sequence match was considered (Kanki, et al., 2011). It seems that the evaluation of sequence matches found by sequence alignment, local alignment search and pattern search are rather arbitrary. As the standard is different case by case, the results from these tools are more like a kind of indication rather than a formal prediction. The results can be confirmed only when more background information is available. Results from sequence alignment, local alignment search and pattern search are same in nature: similarity matches between mimotope and a protein sequence. Thus, a general statistics model or method that evaluates the similarity match reasonably is urgently needed.

As described previously, methods based on template sequence have succeeded in many cases. However, it is more frequent that mimotopes show little similarities to the template, especially when the interaction sites are conformational. Thus methods based on template sequence often fail too. TSOL18 is a host-protective oncosphere antigen of Taenia solium, which is a cestode parasite causing cysticercosis in humans and pigs. The Ph.D.-12 phage library was screened with the anti-TSOL18 monoclonal antibody 17E1. The mimotopes were aligned to the TSOL18 protein sequence using ClustalW software. No significant match was found (Guo, et al., 2010). Intact oocytes surrounded by canine zona pellucida proteins were used to identify peptide sequences from phage display libraries that could recognize and bind to zona pellucida proteins (Samoylova, et al., 2010). The selection of a 12mer library resulted in identification of four sequences with the common NNXXPIL motif discovered by the MOTIF2 program in the RELIC suite. Among them, NNQSPILKLSIH was synthesized and immunized in dogs. The anti-NNQSPILKLSIH antibodies did bind to the acrosomal region of the canine sperm cell. However, BLAST search did not result in identification of homologies to known sperm proteins or other mammalian proteins. Thus, to predict protein interaction sites that are discontinuous using only sequences of mimotope and template is a great challenge. Though the FINDMAP program is a good attempt on this, it is still far from satisfactory. As the entry number of the PDB database increases exponentially, more and more protein structures become available to be used in the prediction of protein interaction sites based on mimotope analysis (Rose, et al., 2011).

2.2 Methods based on template structure

When sequence similarities are not found, it is very likely that mimotopes resemble a special region on the surface of template rather than a linear segment of template sequence. The

prediction of protein interaction sites based on mimotope sequences and corresponding template structure is actually to identify and evaluate surface regions on the template that are similar to mimotopes.

2.2.1 Algorithms, programs and web servers

In 1995, Pizzi et al described the first method that predicted discontinuous antibody binding site based on mimotopes and the antigen structure (Pizzi, et al., 1995). Since then, quite a few algorithms, programs and web servers have been published by different teams around the world. All these methods can be divided into four groups. The first one is the motif-based group, which align a motif or consensus sequence to template structure. This group includes 3DEX (Schreiber, et al., 2005), MIMOX (Huang, et al., 2006) and the MimCons section of MIMOP program (Moreau, et al., 2006). The second group includes Mapitope (Bublil, et al., 2006; Bublil, et al., 2007; Enshell-Seijffers, et al., 2003; Tarnovitski, et al., 2006) and its derivatives (Denisov, et al., 2009; Denisova, et al., 2008; Denisova, et al., 2009; Denisova, et al., 2010). It can be called the pairs-based group because amino acid pairs on the template surface are considered to be simulated by amino acid pairs in the mimotope sequence. The third one is the patch-based group, which evaluates similarities between surface patches on template and mimotopes. SiteLight (Halperin, et al., 2003) and EpiSearch (Negi & Braun, 2009) belong to this group. The fourth is the graph-based group, which aligns a set of query peptides to a graph representing the template surface. Pepsurf (Mayrose, Shlomi, et al., 2007) and Pep-3D-Search (Huang, et al., 2008) belong to this group. To improve the performances of existing programs, hybrid methods such as MimoPro (Chen, et al., 2011) and meta-servers such as Pepitope (Mayrose, Penn, et al., 2007) were also proposed.

As tools mentioned above have been reviewed in detail recently (Huang, et al., 2011), here we only introduce LocaPep, a tool proposed very recently (Pacios, et al., 2011). For each mimotope, this program firstly scans the template surface to select seeds. Then it searches residues adjacent to each seed to form a cluster. For each residue in a cluster, its total score is the weighted sum of the area, exposure, contacts and distance score. At last, the final consensus cluster is calculated to form the binding site predicted. LocaPep is written with Fortran90 independent of any specific library and runs in command line mode. Its source code, manual and binaries are available at http://atenea.montes.upm.es.

2.2.2 Benchmarking tools of the trade

As described above, quite a few methods based on template structure are available for the phage display community to predict protein interaction sites. All these methods have succeeded in some case studies. These test cases were either compiled from published papers or from special databases such as the ASPD database (Valuev, et al., 2002) and the MimoDB database (Huang, et al., 2012; Ru, et al., 2010). However, no systematic evaluations were done when these methods were published. This is due to a relative lack of the type of data where the target-template complex is solved and the relevant mimotope data is available simultaneously.

As the protein structure and mimotope data increase rapidly (Huang, et al., 2012; Rose, et al., 2011), now it becomes possible to make benchmarks for the trade to evaluate its tools at a

larger scale. Sun et al compiled a benchmark from the PDB database (Sun, et al., 2011) and the MimoDB database. It included 47 test cases in which 18 cases were with structures of the antigen-antibody complexes and 29 cases had structures of other protein-protein complexes. They further kept only one test case for each complex with the same template, which made a representative dataset with 30 test cases. Five popular tools, i.e. Mapitope, PepSurf, Pepitope, EpiSearch and Pep-3D-Search, were evaluated with the benchmark and the representative dataset. The results showed that performances of these tools were better than random predictions. However, their overall performances were still not satisfactory. Most tools were good at some cases but failed with other cases.

Our group has also compiled a benchmark called MimoBench (Huang, et al., 2012). It can be freely accessed from http://immunet.cn/mimodb/mimobench.php. Currently, MimoBench has 23, 23 and 27 sets of data for antibody–antigen complex, receptor–ligand complex and other protein-protein complex respectively. Using this benchmark, we have performed a preliminary evaluation on Mapitope, Episearch and MimoPro by their default parameters. Our results showed that performances of these tools were poor in many cases. However, they made quite accurate predictions in some cases. Taking the AUC value 0.8 as a cutoff, the three benchmarked tools succeeded in overlapping but different cases, which suggested that these tools complemented each other. Thus, it is recommended to use several tools together in the prediction of protein–protein interaction sites based on mimotopes.

2.2.3 Methods based on template structure: Challenges and suggestions

Methods based on template structure are capable of predicting the conformational sites of protein-protein interactions. However, the existing tools are not robust enough. Sun et al reported that many test cases in their benchmark dataset could not be applied to the five tools they evaluated due to software limitations (Sun, et al., 2011). We met the same problem when we compared Mapitope, Episearch and MimoPro using MimoBench. For example, four test cases were excluded from benchmarking because these tools did not work on the template with two or more chains. Another 10 cases were dropped because MimoPro returned no results for unknown reason (Huang, et al., 2012). Hence, tools in the future should be more robust. Furthermore, they should also be more convenient to access. It is hoped that web sites of these tools are stable and easy to access. No login is required. Thus, they can be utilized more conveniently whether they are standalone tools or web servers.

As described in the previous section, performances of the existing tools based on template structure are poor in many cases. To improve their performances is one of the greatest challenges in this field. We have suggested that the poor performance might partly due to information loss and noise inclusion during the experimental and computational process (Huang, et al., 2009). Considering the two points in mind, the accuracy of deciphering protein interaction sites using mimotopes might be improved. We will discuss on this issue in the following section.

2.3 Data cleaning tools

Due to the limitation of experiments, the biopanning results are noisy. They are usually a mixture of mimotopes (desired signal) and target-unrelated peptides (unwanted noise).

Target-unrelated peptides (TUPs) can be divided into two categories. One is called selection-related TUP. They appear in the biopanning results because they are selected by contaminants or other components of the screening system rather than the target (Menendez & Scott, 2005; Vodnik, et al., 2011). Propagation-related TUP makes another category (Brammer, et al., 2008; Derda, et al., 2011; Thomas, et al., 2010). They sneak into the output of biopanning because they have a higher infection rate or faster secretion rate. Phages with growth advantage can be not only noise but also decrease the library diversity and lead to a loss of useful mimotopes. Simulations and experiments showed that subtle differences in growth rate yielded drastic differences in clone abundances after rounds of amplifications (Derda, et al., 2011). Thus, propagation-related TUP may even dominate the biopanning results. As TUPs are peptides unrelated to the target, they undoubtedly interfere with the prediction of protein interaction sites based on mimotopes if a TUP is taken as a mimotope. Changing experimental conditions and improving experimental methods can decrease TUPs. For example, increasing the stringency of panning may reduce TUPs; subtractive procedures may decrease selection-related TUPs; amplification in isolated compartment can mitigate the growth advantage of propagation-related TUPs (Derda, et al., 2010). However, TUPs cannot be eradicated experimentally. To exclude TUPs from the biopanning with computational tools has become an alternative and more convenient choice.

2.3.1 Data cleaning tool based on information theory

Based on the information theory, the program INFO in the RELIC suite (Mandava, et al., 2004) calculates information content for each peptide of the panning result. Two input files are required. The first one is a text file with a minimum of 50 peptide sequences from clones randomly selected from a naive library. The second file is the query of users, one or more peptide sequences selected from that same library. INFO first uses AAFREQ to calculate the amino acid frequency distributions at each position of the inserted peptide sequences from the parent library. The probability of random occurrence of any peptide can be calculated by multiplying the probability of each amino acid occurring at each position. The natural logarithm of the probability of a peptide multiplied by -1 is defined as its information content. Obviously, if the query peptide has very high information content, it is less possible to appear in the panning result. If it does occur in the result, it is more likely to be the mimotope selected by specific binding to the target. On the contrary, when a peptide with very low information content is observed in the result, it is less confident of taking it as a mimotope because it may be a propagation-related TUP. The INFO program was also integrated in other tools in the RELIC suite such as MATCH, HETEROalign and FASTAskan. However, all these tools have regretfully been inaccessible for about one year, which makes the RELIC suite now a real relic.

2.3.2 Data cleaning tool based on TUP motif

We have developed a free web server called SAROTUP, which is short for scanner and reporter of target-unrelated peptides (Huang, et al., 2010). It can be used to scan and exclude possible target-unrelated peptides from biopanning result. SAROTUP is based on known TUP motifs and sequences. In the current version, a set of 26 TUP motifs and 27 known TUP

sequences are collected from literature and compiled into the program. Among them, nine sequences are known or highly suspected to be propagation-related TUPs; the left 42 motifs or sequences are for selection-related TUPs, including 14 for albumin binders, six for unrelated antibody binders, five for immunoglobulin Fc region binders, five for streptavidin binders, five for plastic binders, four for bivalent metal ion binders and one for biotin binders, protein A binders and lipid A binders respectively. We had tested SAROTUP before the MimoDB database was constructed. The results showed that: (1) TUPs were often seen and taken as mimotopes; (2) epitope prediction based on mimotopes was greatly interfered if TUPs were used in the analysis; (3) SAROTUP improved performances of epitope mapping based on mimotopes through cleaning the input data; (4) SAROTUP also helped to explain experiment results. However, as a tool based on pattern search, SAROTUP cannot deal with TUPs without known motifs.

2.3.3 Data cleaning tool based on database search

The problem mentioned above was partly solved when the MimoDB database became available (Huang, et al., 2012; Ru, et al., 2010). With a lot of biopanning results and relevant background information collected, this database can be used as a virtual and comprehensive control for experimental biologists. In the MimoDB database version 2.0, a batched peptide search tool can be used for a set of peptides to search against the database. If a peptide has been reported by different groups with different targets, it may be a TUP rather than a mimotope. This is because the chance of obtaining an identical peptide from a library having millions or billions of different peptides with a completely different target is extremely small. If this happens, the peptide obtained may be due to some common factors in the biopanning systems rather than by the target. The MimoBlast tool of the MimoDB database can further find out those peptides not identical but highly similar to the query peptides. Such peptides may also be TUPs. New TUP motifs can be derived from analyzing the result of MimoBlast. With these tools, we studied the peptides in the MimoDB database and claimed confidently that GETRAPL, SILPYPY, LLADTTHHRPWT, TMGFTAPRFPHY, SAHGTSTGVPWP and HLPTSSLFDTTH are TUPs which were not reported before (Huang, et al., 2012).

2.3.4 Data cleaning tools: Challenges and suggestions

Although the data cleaning tools described in this section complement each other, none of them are real classifiers but rather reminders. Without a solid statistical estimation, they can only tell users that a peptide in the result may be a TUP rather than a mimotope. However, as the entries in the MimoDB database increases rapidly, it is now practical to construct various TUP predictors based on machine learning methods. Secondly, the data cleaning procedure was ignored by most existing tools for the prediction of protein interaction sites based on mimotopes. This situation should be changed in the future.

3. Conclusion

Identification of the protein interaction site is very important for basic and applied research. Computational analysis on mimotopes obtained from phage display or other

surface display experiments is a relatively cheap, convenient and efficient strategy to locate a protein interaction site at the segment or residue level. Although used mostly in epitope prediction, this strategy can also be used to other types of protein interaction sites. Insights can be gained by methods based on template sequence, which find sequence similarities between mimotopes and template through sequence alignment, local alignment search and pattern search. Conformational sites can also be mapped by methods based on template structure. However, performances of all existing methods are not satisfactory enough. This is at least partly due to TUPs that crept into the biopanning result. Several tools are available to detect TUPs based on information theory, known TUP motifs or special database. With the rapid accumulation of experimental data and improvement of methods, an evidence-based virtual phage display platform is expected to be established and the performance of predicting protein interaction sites based on mimotopes will substantially be increased.

4. Acknowledgment

This work was supported by the National Natural Science Foundation of China (grant 61071177) and the Scientific Research Foundation of UESTC for Youth (grant JX0769).

5. References

Barabasi, A. L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nat Rev Genet, Vol. 5, No. 2, (Feb), pp. 101-113, ISSN 1471-0056

Barbu, E. M.; Ganesh, V. K.; Gurusiddappa, S.; Mackenzie, R. C.; Foster, T. J.; Sudhof, T. C. & Hook, M. (2010). beta-Neurexin is a ligand for the Staphylococcus aureus MSCRAMM SdrC. PLoS Pathog, Vol. 6, No. 1, (Jan), p. e1000726, ISSN 1553-7374

Bickerton, G. R.; Higueruelo, A. P. & Blundell, T. L. (2011). Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. BMC Bioinformatics, Vol. 12, p. 313, ISSN 1471-2105

Brammer, L. A.; Bolduc, B.; Kass, J. L.; Felice, K. M.; Noren, C. J. & Hall, M. F. (2008). A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site. Anal Biochem, Vol. 373, No. 1, (Feb 1), pp. 88-98, ISSN 0003-2697

Bublil, E. M.; Yeger-Azuz, S. & Gershoni, J. M. (2006). Computational prediction of the cross-reactive neutralizing epitope corresponding to the monoclonal antibody b12 specific for HIV-1 gp120. FASEB J, Vol. 20, No. 11, (Sep), pp. 1762-1774, ISSN 1530-6860

Bublil, E. M.; Freund, N. T.; Mayrose, I.; Penn, O.; Roitburd-Berman, A.; Rubinstein, N. D.; Pupko, T. & Gershoni, J. M. (2007). Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. Proteins, Vol. 68, No. 1, (Jul 1), pp. 294-304, ISSN 1097-0134

Chaemchuen, S.; Rungpragayphan, S.; Poovorawan, Y. & Patarakul, K. (2011). Identification of candidate host proteins that interact with LipL32, the major outer membrane protein of pathogenic Leptospira, by random phage display peptide library. Vet Microbiol, Vol. 153, No. 1-2, (Nov 21), pp. 178-185, ISSN 1873-2542

Chen, T.; Cui, J.; Liang, Y.; Xin, X.; Owen Young, D.; Chen, C. & Shen, P. (2006). Identification of human liver mitochondrial aldehyde dehydrogenase as a potential target for microcystin-LR. Toxicology, Vol. 220, No. 1, (Mar 1), pp. 71-80, ISSN 0300-483X

Chen, W. H.; Sun, P. P.; Lu, Y.; Guo, W. W.; Huang, Y. X. & Ma, Z. Q. (2011). MimoPro: a more efficient Web-based tool for epitope prediction using phage display libraries. BMC Bioinformatics, Vol. 12, p. 199, ISSN 1471-2105

Chen, Y. C.; Delbrook, K.; Dealwis, C.; Mimms, L.; Mushahwar, I. K. & Mandecki, W. (1996). Discontinuous epitopes of hepatitis B surface antigen derived from a filamentous phage peptide library. Proc Natl Acad Sci U S A, Vol. 93, No. 5, (Mar 5), pp. 1997-2001, ISSN 0027-8424

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res, Vol. 16, No. 22, (Nov 25), pp. 10881-10890, ISSN 0305-1048

Deng, Y. Q.; Dai, J. X.; Ji, G. H.; Jiang, T.; Wang, H. J.; Yang, H. O.; Tan, W. L.; Liu, R.; Yu, M.; Ge, B. X.; Zhu, Q. Y.; Qin, E. D.; Guo, Y. J. & Qin, C. F. (2011). A broadly flavivirus cross-neutralizing monoclonal antibody that recognizes a novel epitope within the fusion loop of E protein. PLoS One, Vol. 6, No. 1, p. e16059, ISSN 1932-6203

Denisov, D. A.; Denisova, G. F.; Lelic, A.; Loeb, M. B. & Bramson, J. L. (2009). Deciphering epitope specificities within polyserum using affinity selection of random peptides and a novel algorithm based on pattern recognition theory. Mol Immunol, Vol. 46, No. 3, (Jan), pp. 429-436, ISSN 0161-5890

Denisova, G.; Denisov, D.; Evelegh, C.; Weissgram, M.; Beck, J.; Foley, S. R. & Bramson, J. L. (2009). Characterizing complex polysera produced by antigen-specific immunization through the use of affinity-selected mimotopes. PLoS One, Vol. 4, No. 4, p. e5309, ISSN 1932-6203

Denisova, G. F.; Denisov, D. A.; Yeung, J.; Loeb, M. B.; Diamond, M. S. & Bramson, J. L. (2008). A novel computer algorithm improves antibody epitope prediction using affinity-selected mimotopes: a case study using monoclonal antibodies against the West Nile virus E protein. Mol Immunol, Vol. 46, No. 1, (Nov), pp. 125-134, ISSN 0161-5890

Denisova, G. F.; Denisov, D. A. & Bramson, J. L. (2010). Applying bioinformatics for antibody epitope prediction using affinity-selected mimotopes - relevance for vaccine design. Immunome Res, Vol. 6 Suppl 2, p. S6, ISSN 1745-7580

Derda, R.; Tang, S. K. & Whitesides, G. M. (2010). Uniform amplification of phage with different growth characteristics in individual compartments consisting of monodisperse droplets. Angew Chem Int Ed Engl, Vol. 49, No. 31, (Jul 19), pp. 5301-5304, ISSN 1521-3773

Derda, R.; Tang, S. K.; Li, S. C.; Ng, S.; Matochko, W. & Jafari, M. R. (2011). Diversity of phage-displayed libraries of peptides during panning and amplification. Molecules, Vol. 16, No. 2, pp. 1776-1803, ISSN 1420-3049

Enshell-Seijffers, D.; Denisov, D.; Groisman, B.; Smelyanski, L.; Meyuhas, R.; Gross, G.; Denisova, G. & Gershoni, J. M. (2003). The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1. J Mol Biol, Vol. 334, No. 1, (Nov 14), pp. 87-101, ISSN 0022-2836

Fack, F.; Hugle-Dorr, B.; Song, D.; Queitsch, I.; Petersen, G. & Bautz, E. K. (1997). Epitope mapping by phage display: random versus gene-fragment libraries. J Immunol Methods, Vol. 206, No. 1-2, (Aug 7), pp. 43-52, ISSN 0022-1759

Fernández-Recio, J. (2011). Prediction of protein binding sites and hot spots. WIREs Computational Molecular Science, Vol. 1, No. 5, pp. 680-698, ISSN 1759-0884

Geysen, H. M.; Rodda, S. J. & Mason, T. J. (1986). A priori delineation of a peptide which mimics a discontinuous antigenic determinant. Mol Immunol, Vol. 23, No. 7, (Jul), pp. 709-715, ISSN 0161-5890

Guo, A.; Cai, X.; Jia, W.; Liu, B.; Zhang, S.; Wang, P.; Yan, H. & Luo, X. (2010). Mapping of Taenia solium TSOL18 antigenic epitopes by phage display library. Parasitol Res, Vol. 106, No. 5, (Apr), pp. 1151-1157, ISSN 1432-1955

Halperin, I.; Wolfson, H. & Nussinov, R. (2003). SiteLight: binding-site prediction using phage display libraries. Protein Sci, Vol. 12, No. 7, (Jul), pp. 1344-1359, ISSN 0961-8368

Higurashi, M.; Ishida, T. & Kinoshita, K. (2009). PiSite: a database of protein interaction sites using multiple binding states in the PDB. Nucleic Acids Res, Vol. 37, No. Database issue, (Jan), pp. D360-364, ISSN 1362-4962

Huang, J.; Gutteridge, A.; Honda, W. & Kanehisa, M. (2006). MIMOX: a web tool for phage display based epitope mapping. BMC Bioinformatics, Vol. 7, p. 451, ISSN 1471-2105

Huang, J.; Xia, M.; Lin, H. & Guo, F.B. Information loss and noise inclusion risk in mimotope based epitope mapping, The 3rd International Conference on Bioinformatics and Biomedical Engineering(iCBBE2009), pp. 1-3, Beijing, 2009

Huang, J.; Ru, B.; Li, S.; Lin, H. & Guo, F. B. (2010). SAROTUP: scanner and reporter of target-unrelated peptides. J Biomed Biotechnol, Vol. 2010, p. 101932, ISSN 1110-7251

Huang, J.; Ru, B. & Dai, P. (2011). Bioinformatics resources and tools for phage display. Molecules, Vol. 16, No. 1, pp. 694-709, ISSN 1420-3049

Huang, J.; Ru, B.; Zhu, P.; Nie, F.; Yang, J.; Wang, X.; Dai, P.; Lin, H.; Guo, F. B. & Rao, N. (2012). MimoDB 2.0: a mimotope database and beyond. Nucleic Acids Res, Vol. 40, No. D1, pp. D271-D277, ISSN 1362-4962

Huang, Y. X.; Bao, Y. L.; Guo, S. Y.; Wang, Y.; Zhou, C. G. & Li, Y. X. (2008). Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis. BMC Bioinformatics, Vol. 9, p. 538, ISSN 1471-2105

Kanki, S.; Jaalouk, D. E.; Lee, S.; Yu, A. Y.; Gannon, J. & Lee, R. T. (2011). Identification of targeting peptides for ischemic myocardium by in vivo phage display. J Mol Cell Cardiol, Vol. 50, No. 5, (May), pp. 841-848, ISSN 1095-8584

Kozakov, D.; Hall, D. R.; Beglov, D.; Brenke, R.; Comeau, S. R.; Shen, Y.; Li, K.; Zheng, J.; Vakili, P.; Paschalidis, I. & Vajda, S. (2010). Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. Proteins, Vol. 78, No. 15, (Nov 15), pp. 3124-3130, ISSN 1097-0134

Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J. &

Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics, Vol. 23, No. 21, (Nov 1), pp. 2947-2948, ISSN 1367-4811

Mandava, S.; Makowski, L.; Devarapalli, S.; Uzubell, J. & Rodi, D. J. (2004). RELIC--a bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction sites. Proteomics, Vol. 4, No. 5, (May), pp. 1439-1460, ISSN 1615-9853

Mashiach, E.; Schneidman-Duhovny, D.; Peri, A.; Shavit, Y.; Nussinov, R. & Wolfson, H. J. (2010). An integrated suite of fast docking algorithms. Proteins, Vol. 78, No. 15, (Nov 15), pp. 3197-3204, ISSN 1097-0134

Mayrose, I.; Penn, O.; Erez, E.; Rubinstein, N. D.; Shlomi, T.; Freund, N. T.; Bublil, E. M.; Ruppin, E.; Sharan, R.; Gershoni, J. M.; Martz, E. & Pupko, T. (2007). Pepitope: epitope mapping from affinity-selected peptides. Bioinformatics, Vol. 23, No. 23, (Dec 1), pp. 3244-3246, ISSN 1367-4803

Mayrose, I.; Shlomi, T.; Rubinstein, N. D.; Gershoni, J. M.; Ruppin, E.; Sharan, R. & Pupko, T. (2007). Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. Nucleic Acids Res, Vol. 35, No. 1, pp. 69-78, ISSN 0305-1048

Menendez, A. & Scott, J. K. (2005). The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. Anal Biochem, Vol. 336, No. 2, (Jan 15), pp. 145-157, ISSN 0003-2697

Moreau, V.; Granier, C.; Villard, S.; Laune, D. & Molina, F. (2006). Discontinuous epitope prediction based on mimotope analysis. Bioinformatics, Vol. 22, No. 9, (May 1), pp. 1088-1095, ISSN 1367-4803

Moreira, I. S.; Fernandes, P. A. & Ramos, M. J. (2007). Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins, Vol. 68, No. 4, (Sep 1), pp. 803-812, ISSN 1097-0134

Mount, D. W. (2007). Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc, Vol. 2007, p. pdb top17, ISSN 1559-6095

Mumey, B.; Ohler, N.; Angel, T.; Jesaitis, A. & Dratz, E. (2006). Filtering Epitope Alignments to Improve Protein Surface Prediction, In: Frontiers of High Performance Computing and Networking – ISPA 2006 Workshops, G. Min, B. Di Martino, L. Yang, M. Guo and G. Ruenger, (Eds.), 648-657, Springer,

Mumey, B. M.; Bailey, B. W.; Kirkpatrick, B.; Jesaitis, A. J.; Angel, T. & Dratz, E. A. (2003). A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins. J Comput Biol, Vol. 10, No. 3-4, pp. 555-567, ISSN 1066-5277

Negi, S. S. & Braun, W. (2009). Automated detection of conformational epitopes using phage display Peptide sequences. Bioinform Biol Insights, Vol. 3, pp. 71-81, ISSN 1177-9322

Ofran, Y. & Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics, Vol. 23, No. 2, (Jan 15), pp. e13-16, ISSN 1367-4811

Pacios, L. F.; Tordesillas, L.; Palacin, A.; Sanchez-Monge, R.; Salcedo, G. & Diaz-Perales, A. (2011). LocaPep: localization of epitopes on protein surfaces using peptides from phage display libraries. J Chem Inf Model, Vol. 51, No. 6, (Jun 27), pp. 1465-1473, ISSN 1549-960X

Pasumarthy, K. K.; Mukherjee, S. K. & Choudhury, N. R. (2011). The presence of tomato leaf curl Kerala virus AC3 protein enhances viral DNA replication and modulates virus induced gene-silencing mechanism in tomato plants. Virol J, Vol. 8, p. 178, ISSN 1743-422X

Peng, L.; Oganesyan, V.; Damschroder, M. M.; Wu, H. & Dall'acqua, W. F. (2011). Structural and Functional Characterization of an Agonistic Anti-Human EphA2 Monoclonal Antibody. J Mol Biol, Vol. 413, No. 2, (Oct 21), pp. 390-405, ISSN 1089-8638

Perosa, F.; Vicenti, C.; Racanelli, V.; Leone, P.; Valentini, G. & Dammacco, F. (2010). The immunodominant epitope of centromere-associated protein A displays homology with the transcription factor forkhead box E3 (FOXE3). Clin Immunol, Vol. 137, No. 1, (Oct), pp. 60-73, ISSN 1521-7035

Pizzi, E.; Cortese, R. & Tramontano, A. (1995). Mapping epitopes on protein surfaces. Biopolymers, Vol. 36, No. 5, (Nov), pp. 675-680, ISSN 0006-3525

Plesa, M.; Kim, J.; Paquette, S. G.; Gagnon, H.; Ng-Thow-Hing, C.; Gibbs, B. F.; Hancock, M. A.; Rosenblatt, D. S. & Coulton, J. W. (2011). Interaction between MMACHC and MMADHC, two human proteins participating in intracellular vitamin B metabolism. Mol Genet Metab, Vol. 102, No. 2, (Feb), pp. 139-148, ISSN 1096-7206

Przulj, N. (2011). Protein-protein interactions: making sense of networks via graph-theoretic modeling. Bioessays, Vol. 33, No. 2, (Feb), pp. 115-123, ISSN 1521-1878

Res, I.; Mihalek, I. & Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics, Vol. 21, No. 10, (May 15), pp. 2496-2501, ISSN 1367-4803

Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M. & Bourne, P. E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res, Vol. 39, No. Database issue, (Jan), pp. D392-401, ISSN 1362-4962

Ru, B.; Huang, J.; Dai, P.; Li, S.; Xia, Z.; Ding, H.; Lin, H.; Guo, F. & Wang, X. (2010). MimoDB: a New Repository for Mimotope Data Derived from Phage Display Technology. Molecules, Vol. 15, No. 11, pp. 8279-8288, ISSN 1420-3049

Samoylova, T. I.; Cox, N. R.; Cochran, A. M.; Samoylov, A. M.; Griffin, B. & Baker, H. J. (2010). ZP-binding peptides identified via phage display stimulate production of sperm antibodies in dogs. Anim Reprod Sci, Vol. 120, No. 1-4, (Jul), pp. 151-157, ISSN 1873-2232

Schreiber, A.; Humbert, M.; Benz, A. & Dietrich, U. (2005). 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. J Comput Chem, Vol. 26, No. 9, (Jul 15), pp. 879-887, ISSN 0192-8651

Shiheido, H.; Takashima, H.; Doi, N. & Yanagawa, H. (2011). mRNA display selection of an optimized MDM2-binding peptide that potently inhibits MDM2-p53 interaction. PLoS One, Vol. 6, No. 3, p. e17898, ISSN 1932-6203

Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science, Vol. 228, No. 4705, (Jun 14), pp. 1315-1317, ISSN 0036-8075

Smith, G. P. & Petrenko, V. A. (1997). Phage Display. Chem Rev, Vol. 97, No. 2, (Apr 1), pp. 391-410, ISSN 1520-6890

Sun, E. C.; Zhao, J.; Yang, T.; Liu, N. H.; Geng, H. W.; Qin, Y. L.; Wang, L. F.; Bu, Z. G.; Yang, Y. H.; Lunt, R. A. & Wu, D. L. (2011). Identification of a conserved JEV serocomplex B-cell epitope by screening a phage-display peptide library with a mAb generated against West Nile virus capsid protein. Virol J, Vol. 8, p. 100, ISSN 1743-422X

Sun, P.; Chen, W.; Huang, Y.; Wang, H.; Ma, Z. & Lv, Y. (2011). Epitope prediction based on random peptide library screening: benchmark dataset and prediction tools evaluation. Molecules, Vol. 16, No. 6, pp. 4971-4993, ISSN 1420-3049

Takakusagi, Y.; Takakusagi, K.; Sugawara, F. & Sakaguchi, K. (2010). Use of phage display technology for the determination of the targets for small-molecule therapeutics. Expert Opinion on Drug Discovery, Vol. 5, No. 4, pp. 361-389, ISSN 1746-0441

Takami, M.; Takakusagi, Y.; Kuramochi, K.; Tsukuda, S.; Aoki, S.; Morohashi, K.; Ohta, K.; Kobayashi, S.; Sakaguchi, K. & Sugawara, F. (2011). A screening of a library of T7 phage-displayed peptide identifies E2F-4 as an etoposide-binding protein. Molecules, Vol. 16, No. 5, pp. 4278-4294, ISSN 1420-3049

Tarnovitski, N.; Matthews, L. J.; Sui, J.; Gershoni, J. M. & Marasco, W. A. (2006). Mapping a neutralizing epitope on the SARS coronavirus spike protein: computational prediction based on affinity-selected peptides. J Mol Biol, Vol. 359, No. 1, (May 26), pp. 190-201, ISSN 0022-2836

Thomas, W. D.; Golomb, M. & Smith, G. P. (2010). Corruption of phage display libraries by target-unrelated clones: diagnosis and countermeasures. Anal Biochem, Vol. 407, No. 2, (Dec 15), pp. 237-240, ISSN 1096-0309

Urushibata, Y.; Itoh, K.; Ohshima, M. & Seto, Y. (2010). Generation of Fab fragment-like molecular recognition proteins against staphylococcal enterotoxin B by phage display technology. Clin Vaccine Immunol, Vol. 17, No. 11, (Nov), pp. 1708-1717, ISSN 1556-679X

Valuev, V. P.; Afonnikov, D. A.; Ponomarenko, M. P.; Milanesi, L. & Kolchanov, N. A. (2002). ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved in vitro. Nucleic Acids Res, Vol. 30, No. 1, (Jan 1), pp. 200-202, ISSN 1362-4962

Van Nieuwenhove, L. C.; Roge, S.; Balharbi, F.; Dieltjens, T.; Laurent, T.; Guisez, Y.; Buscher, P. & Lejon, V. (2011). Identification of peptide mimotopes of Trypanosoma brucei gambiense variant surface glycoproteins. PLoS Negl Trop Dis, Vol. 5, No. 6, (Jun), p. e1189, ISSN 1935-2735

Vidal, M.; Cusick, M. E. & Barabasi, A. L. (2011). Interactome networks and human disease. Cell, Vol. 144, No. 6, (Mar 18), pp. 986-998, ISSN 1097-4172

Vodnik, M.; Zager, U.; Strukelj, B. & Lunder, M. (2011). Phage display: selecting straws instead of a needle from a haystack. Molecules, Vol. 16, No. 1, pp. 790-817, ISSN 1420-3049

Wass, M. N.; David, A. & Sternberg, M. J. (2011). Challenges for the prediction of macromolecular interactions. Curr Opin Struct Biol, Vol. 21, No. 3, (Jun), pp. 382-390, ISSN 1879-033X

Zhao, S.; Zhao, W. & Ma, L. (2010). Novel peptide ligands that bind specifically to mouse embryonic stem cells. Peptides, Vol. 31, No. 11, (Nov), pp. 2027-2034, ISSN 1873-5169

10

Structural Bioinformatics of Proteins: Predicting the Tertiary and Quaternary Structure of Proteins from Sequence

J. Planas-Iglesias, J. Bonet,
M.A. Marín-López, E. Feliu, A. Gursoy and B. Oliva
Structural Bioinformatics Lab. Universitat Pompeu Fabra, Catalunya
Spain

1. Introduction

Although the regulatory role of non-coding nucleic acids is currently being unraveled, the role of proteins is still a major issue as they mediate most biological functions. Thus, understanding how proteins fulfill their intricate functions is one of the most relevant current challenges in biology. It is well known that a protein's function is determined by its three-dimensional (3D) structure known as tertiary structure, which in turn is mainly dictated by its sequence (Thornton and cols. reviewed this issue in detail in (Watson et al., 2005)). Despite the exponential increase of available sequences and 3D structures, the number of sequences highly exceeds that of 3D structures. This difference in numbers is proportional to the disparity of the costs for experimentally obtaining either the sequence or the structure of a protein. Therefore, covering the gap between sequence and structure becomes a compelling requirement to achieve a molecular understanding of the protein function. Theoretical methods can help to bridge this gap by inferring the 3D structure from the sequence. These methods are classified into three different groups: comparative modeling, fold recognition and new fold or *ab initio* methods.

Besides the tertiary structure of a protein, other contextual factors may modulate its function. Among these, the ability of the proteins to interact with others and the particular partners with which they form complexes are one of the most important. This is because proteins rarely act alone; they rather constitute a mingled network of physical interactions, some times to form large macro-complexes and sometimes to produce transient interactions. In this context, understanding the function of a protein implies to recognize its partners and to grasp how they associate, even at the atomic level. The structure of these complexes is known as quaternary structure. To this end, computational techniques have been developed to dock one protein onto another (Janin, 2010; Vajda and Kozakov, 2009), and can help to infer 3D structure of a protein from the knowledge of the protein interactions (Fornes et al., 2009) and vice-versa (Stein et al., 2005). Furthermore, the combined use of data from multiple resources allows us to obtain an accurate model of large molecular complexes such as nucleopore (Alber et al., 2007).

There are two strategies for modeling the interaction between two proteins from sequence data. The first one is to model the unbound interactors and to dock them into the final

complex (i.e. solving first the tertiary structure of the proteins and afterwards the quaternary). The second is to model the interacting pair or complex from the scratch, using as template the structural knowledge of an available homologous interacting pair (interolog, (Matthews et al., 2001)). When the template is not available the strategy can only play with the docking of the unbound partners. Figure 1 summarizes these possibilities. Here, we will cover these strategies and methods to infer and assess the 3D structure of binary protein interactions, and we will review the existing techniques to model large cellular macro-complexes.

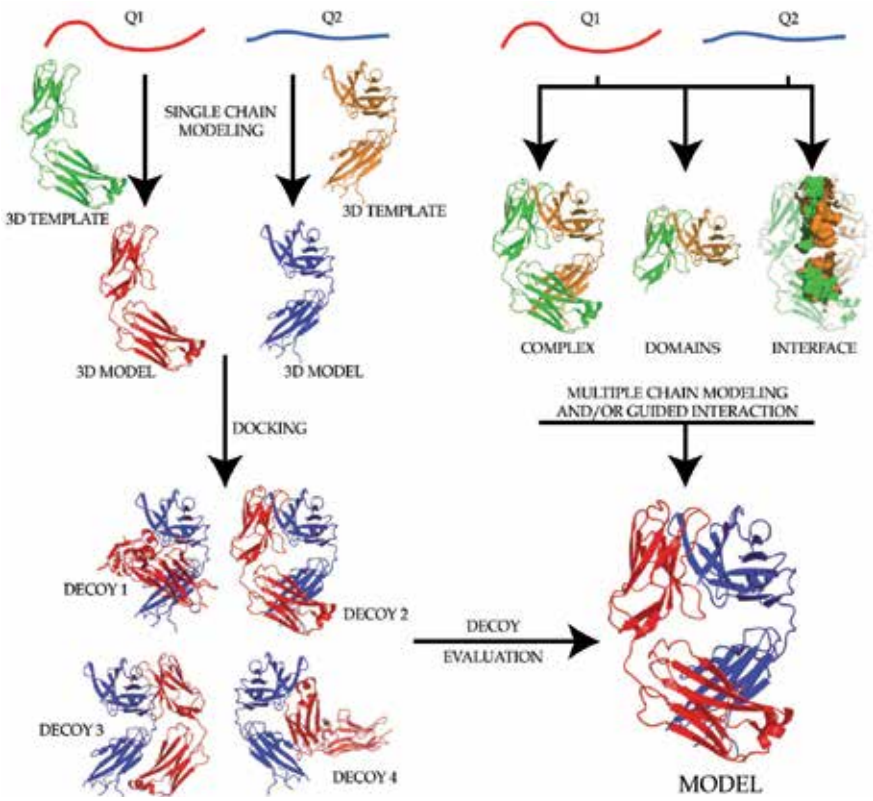


Fig. 1. Different strategies for modeling a protein interaction: The 3D structure of a binary protein interaction can be inferred by modeling individual interacting partners apart and subsequently docking them (left side) or modeling the interaction with one template, taking advantage of the available information of homologous complexes. Templates can be obtained from structural resources of information containing the full complex, a partial complex (in general formed by interacting domains) or with only the interacting interface (right side).

2. Modeling the tertiary structure of proteins

In order to obtain a complete model of a protein interaction, the interacting partners can be modeled separately and then docked into a functional complex. The first step of this approach is to obtain the 3D structure of each of the interacting partners. Comparative modeling, fold recognition, and *ab initio* (for new folds) are computational methods that

may overcome the lack of experimental structural data. Models obtained using these approaches may be further assessed, in order to ensure that the inferred 3D structure contains no errors (see section 4). In case of persistence, such errors would damper the deduction of further biological conclusions such as the mode the modeled protein can interact with others. Figure 2 summarizes the different steps and strategies that can be exploited to achieve these objectives.

2.1 Homology modeling

Homology or comparative modeling techniques are those devoted to the prediction and construction of the 3D conformation of proteins. These methods are based on the assumption that structural features in proteins are more conserved than its sequences. Thus, two proteins with enough sequence similarity will fold in a similar way and share the same conformation in space. The process through which a tertiary structure is assigned to a given sequence is carried out in three steps, namely: template identification, template alignment, and model building. Finally, the produced model should be assessed (see section 4).

Template identification is the key step in the molecular modeling process. Templates are defined as the set of known structures used to build the tertiary structure of the query (target or problem protein). Known 3D data of proteins is stored in the Protein Data Bank (PDB) (Berman et al., 2000). Thus, the identification of the template refers to the process of identifying the structure of the PDB whose sequence is the closest homolog of the target. Such sequence homology search can be performed using sequence alignment tools like BLAST and PSI-BLAST (Altschul et al., 1997), or Hidden Markov Model (HMM) profile methods like HMMER (Eddy, 1998). While BLAST will reveal if there is any relatively close homolog to our query, PSI-BLAST and HMMER will also reveal the possibility of remote homologues. The homology threshold that can be used to define whether or not a template assignation is correct may be fuzzy. Those templates assigned with low percentage of identity, low homology, or in short parts of the sequence fall into what is known as the twilight zone (Rost, 1999). Some rules have been described to shed some light into that twilight region in order to better describe the viability of a template for a given query (Fornes et al., 2009).

Provided that a good template has been selected, the sequence alignment between the query and the template can be directly extracted or easily inferred (in case of the HMM) from the template search. Depending on specific requirements, the alignments can be redone with other sequence alignment methods such as CLUSTALW (Chenna et al., 2003) or T-COFFE (Notredame et al., 2000). Additionally, some methods optimize the sequence alignment through a genetic algorithm protocol that iterates the alignment, model building and model evaluation in order to obtain the best possible alignment (Fernandez-Fuentes et al., 2007) .

Model building is the process by which the three-dimensional data of the template(s) is applied on the query sequence. MODELLER is one of the most used and comprehensive pieces of modeling software (Sali et al., 1995). Provided the sequence alignment the modeling process is practically automatic. As many other modeling tools, it is based on satisfying a set of spatial constraints. Specifically it satisfies three spatial constraints being (1) homology-derived constraints, (2) stereochemical constraints such as bond angles, and (3) statistical preferences for dihedral angles and non-bonded interatomic distances.

Optionally, manually curated restraints from secondary structure packing to site-directed mutagenesis can be added to the modeling process.

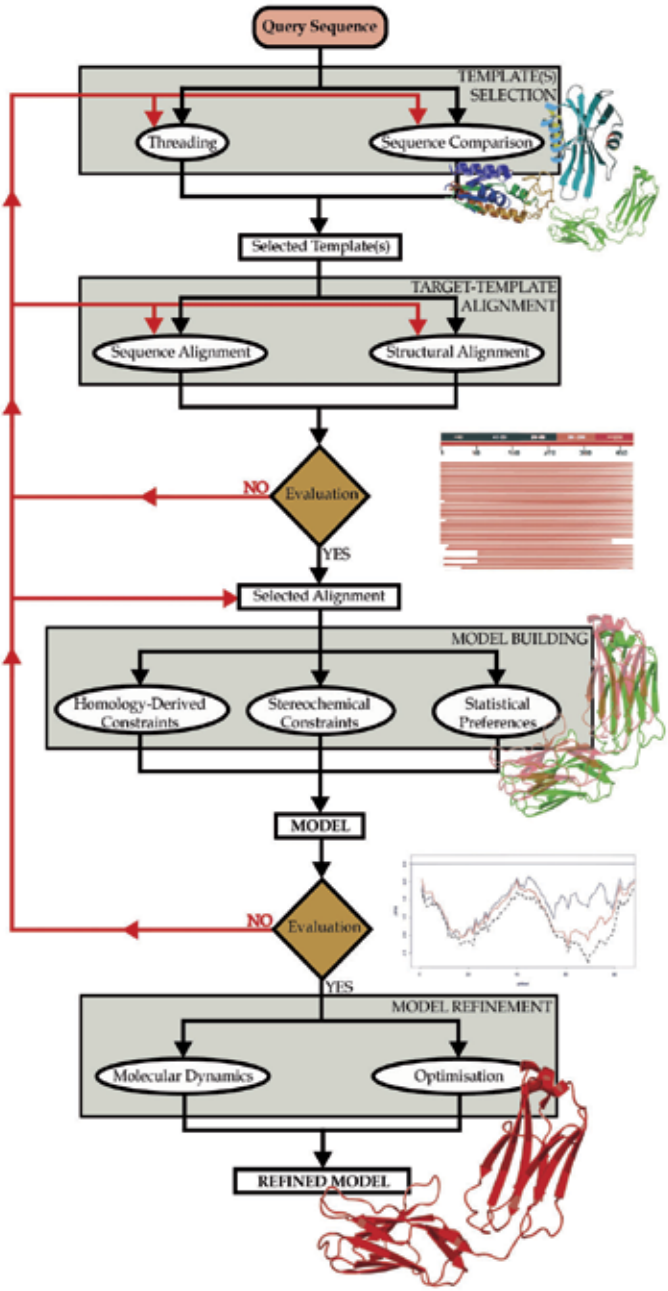


Fig. 2. Flowchart for single protein modeling: Scheme of the methods used for modeling, comprising template(s) selection, template-target alignment, model building, model evaluation, and model refinement steps.

2.2 Threading

In the same way as homology modeling, threading is a method to determine the tertiary structure of a protein based on the fairly small number of different folds contained in nature. The main difference resides in the fact that threading does not use specific protein 3D structures as templates but uses statistical knowledge extracted from all structures in PDB. Thus, threading is especially useful when no suitable template for the protein can be found. Basically, the prediction is made by aligning each amino acid of a given query sequence to a position in a set of structural templates. Once the optimal template is selected through this method the structural model is built according to the alignment between the query and the template. The threading process can be divided in four different steps: the template database construction, the scoring function, the threading alignment and the threading prediction. Among the most used programs on threading and fold prediction are GeneThreader (Jones, 1999) and Phyre (Kelley and Sternberg, 2009).

Gathering representative structures for the different folds avoiding redundancy creates the template database. This means extracting all the structures from PDB and picking one representative of each known fold, eliminating redundancy and sequence homology. The homology filtering is key to ensure that the predictions over the query sequence are not going to be biased because of the database composition.

Designing an optimal function to score the suitability of the templates for modeling the query protein is determinant. The scoring function should be based on the known relationships between structure and sequence. A good scoring function should contain as much information as possible such as pairwise potential, secondary structure compatibilities, environment fitness potential, and gap penalties. The accuracy of the alignment and the prediction will be directly related to the quality of the scoring function. During the threading alignment the query sequence is going to be tested against each given possible template. This part of the process is, by far, the most computationally costly as it takes into account pairwise contact potentials (see section 4.2) and cannot be substituted by the classic dynamic programming algorithm for sequence alignment. Finally the threading prediction uses the scoring function and all the provided alignments to select the better template and build the protein model by placing the backbone atoms of the query according to the position of their aligned counterparts in the template.

2.3 *Ab Initio* methods

Ab initio or *de novo* protein structure prediction tries to predict the tertiary structure of proteins directly from its sequence properties. The idea is that the structure of proteins can be determined without any explicit templates by means of applying the general principles that govern protein folding and the statistical tendencies of conformational features gathered from structural knowledge. Those predictions involve sampling the conformational space, which means that a large set of decoys (structural candidates) is likely to be generated. Scoring functions, either physics-based or knowledge based, are then used to select those decoys that can be identified as more native-like conformations. Optionally, high-resolution refinement is used to optimize those native-like structures. Few programs have been successful in this task, and among them the most flourishing are ROSETTA (Leaver-Fay et al., 2011) and TASSER (Chen and Skolnick, 2008).

3. Modeling the quaternary structure of proteins

As the structural data on protein complexes keep increasing steadily, using known protein complex structures has become an important approach for modeling protein interactions. This increase of structural knowledge of protein complexes (even if it is only partial) opens a new window of possibilities to infer the quaternary structure of proteins. However, for a large quantity of protein complexes this knowledge is still limited, and alternative techniques are required to infer their 3D structure. Docking methods surmount this lack of data providing predictions of the quaternary structure of the complex based on the physical, chemical, and biological known properties of protein complexes. New approaches have introduced the possibility to integrate different sources of experimental information, such as high-resolution electron-microscopy, SAXS, NMR, yeast-two-hybrid, and affinity purifications to extract restraints that can be applied to model the quaternary structure of macro-complexes (Alber et al., 2007).

3.1 Comparative modeling of protein binary complexes

Provided that a homolog structure of an interaction is known, homology modeling can be used to model a protein-protein interaction of interest. Two different approaches can be taken: (1) direct interaction modeling or (2) protein modeling and reorientation (see Figure 1).

When directly modeling a protein interaction, it has to be taken into account that both query proteins need to have a couple of acceptable templates that share the same crystal structure. If that is the case, MODELLER is able to directly model the protein-protein interaction. However, in not only each separate structure needs to be evaluated but also the interface created between them (see section 4.2).

An alternative is to model each protein separately and afterwards use a known interaction as a guide to reposition each structure in the way the interaction is supposed to be taking place. To do so it is required to perform a structural alignment between the model and the template for the interaction. That can be done with strictly devoted tools such as STAMP (Russell and Barton, 1992) or through a variety of protein structure graphical interfaces such as PYMOL (http://www.pymol.org). This approach should be selected if the resolution of the structure of the templates in the interaction is largely worse than the unbound templates. However, it has to be taken into account the need to introduce some structural flexibility produced to construct the interaction. Considering the principal motions and intrinsic fluctuations to accommodate the unbound structures (Dobbins et al., 2008) may help to this purpose. The final structure needs to be refined, and additional restraints are applied to keep the partners on the orientation defined by the template of the binary complex.

3.2 Modeling of protein binary complexes from partial structural information

The sequence and structural homology methods described in the previous section require global similarity (sequence or structure). However, recent research shows that the binding sites of proteins are somewhat more distinguishable from the rest of the protein surface. The binding site of two interacting proteins is called a protein-protein interface. If the structure of a protein complex is available, determining the interface is fairly simple. The interface can

be found either by finding contacting residues (distance based) or by calculating the accessible surface area of the residues. Since proteins interact through interfaces, physico-chemical properties of interfaces are important to study protein interactions. Statistical studies of known protein complexes have revealed general characteristics of interfaces. Interfaces in general have electrostatic and shape complementary. Compared to the rest of the protein surface, interfaces are found to be slightly more conserved (Caffrey et al., 2004). Depending on the interaction type, properties of interfaces display variation. Homo-oligomeric complexes have more hydrophobic and larger interfaces than the hetero complexes. Homo-oligomers are usually permanent and their interfaces resemble interior of globular proteins. Transient interactions, on the other hand, are mediated by smaller interfaces (less than 1500 $Å^2$) and have more polar and charged amino acids than the interfaces of permanent interactions (Nooren and Thornton, 2003). The small surface-area of transient interfaces are partly due to requirement of individual partners of the interaction to fold independently and to be soluble. The secondary structure content of interfaces shows differences between permanent and transient interfaces as well. For example, turns are observed more frequently in non-obligatory interfaces since flexibility is required to repeatedly associate/disassociate (De et al., 2005). Even within an interface, the properties and organization of residues are not uniform. The interface area may be dissected into regions where a set of buried residues forming a core region is surrounded by a rim of residues that are partially solvent accessible. The composition of residues are distinct between these two regions (Guharoy and Chakrabarti, 2005). Alanine scanning mutagenesis of interface residues has also revealed that some residues contribute more to the binding energy (Clackson and Wells, 1995). These areas, called hot spots, are particularly enriched in Trp, Tyr, and Arg residues and are structurally conserved, which can be used to differentiate binding sites from the rest of the surface (Ma et al., 2003).

All these characteristics can be used to identify binding sites of proteins either from sequence (Ofran and Rost, 2007) or from unbound structures (Neuvirth et al., 2004), and potentially for modeling protein interactions. Therefore, a systematic collection and categorization of protein interfaces play important role. Several databases of interfaces have been compiled along this direction, including PiBASE (Davis and Sali, 2005), SCOWLP (Teyra et al., 2006), SCOPPI (Winter et al., 2006) and PRINT (Tuncbag et al., 2008). These databases, in general, present interfaces extracted form known protein complexes together with features of the interfaces such as change in accessible surface area, conservation, or residue composition (reviewed in (Tuncbag et al., 2009)). PRINT database presents all interfaces from PDB (as of 2006) clustered by structural similarity where each cluster represents a different interface architecture. Some interface architectures are observed to be more favorable and reused frequently. These interface architectures are found to be similar to domain folds, consistent with earlier studies indicating that the folding and binding are similar processes (Tsai et al., 1997).

The analysis of protein interactions and interfaces has suggested that the number of possible interfaces is much smaller than the possible number of protein interactions (Aloy and Russell, 2004; Tuncbag et al., 2008). In addition, interfaces are observed to be reused in different protein interactions that are not globally similar (the same interface used by proteins with different fold architectures) (Keskin et al., 2004). This information can be used to overcome the global similarity of requirement of the homology based modeling methods.

That is, modeling protein interactions using only the similarity of protein interfaces. PRISM (Aytuna et al., 2005) is one of the first approaches that has used interface similarity along this direction. It was originally developed to predict protein interactions between proteins (target set) from a set of known protein interfaces (template set). If the two complementary sides of a template interface are found to be structurally similar to two target proteins (one side on one protein, the other on another protein), then the proteins are predicted to be interacting and modeled using the binding site dictated by the template interface. A schematic description of the method is illustrated in Figure 3. Putative interactions are then re-ordered by flexible refinement. PRISM protocol is a collection of scripts that performs a) preparing a set of template interfaces from known complexes, b) preparing surfaces of target proteins that interactions among them to be predicted, c) structural alignment of templates to targets, d) scoring with flexible refinement. The method can be used to model a protein interaction by selecting the two potentially interacting proteins as targets, and using all non-redundant interfaces as the template set. Although the method is limited by the availability and coverage of known protein-protein interfaces from PDB, the continuous growth of the PDB database will increase the applicability of the method. In fact, a recent study on the structural coverage of known protein interfaces already points out that the coverage is close to complete (Gao and Skolnick, 2010).

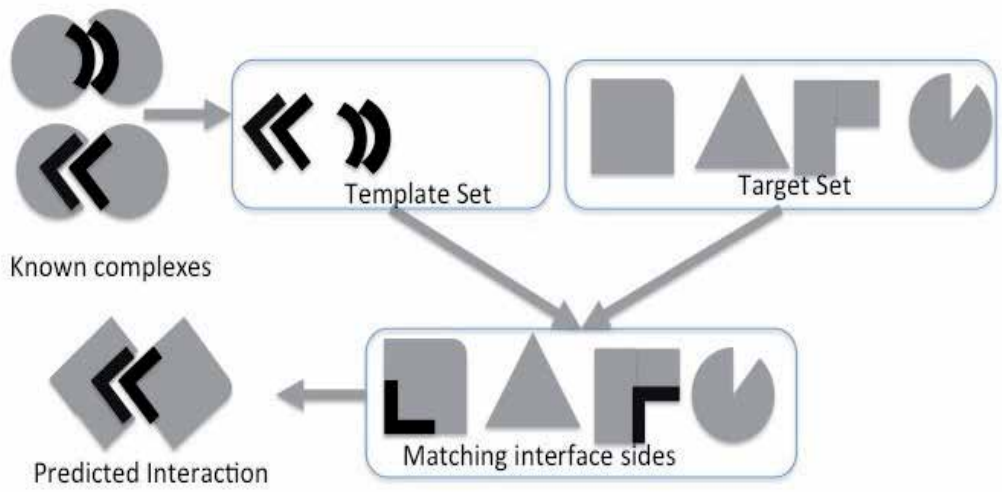


Fig. 3. Schematic representation of PRISM: Two target proteins are predicted to interact if the two complementary sides of a template interface are found to be structurally similar to them (a different side on each protein).

3.3 Protein-protein docking

In contrast to previously described methods (which are based on the structural knowledge of the interaction), protein docking is one of the computational techniques for elucidating the structures of binary bio-molecules (e.g. two proteins) when experimental data regarding the structure of the complex is lacking but the structures of the interacting proteins are known. Docking techniques sample the orientation of two unbound protein

structures to produce several predictions about their interaction, followed by a scoring step to rank the predictions. These methods were introduced in 1978 (Wodak and Janin, 1978). Since then, docking algorithms have substantially improved, with a breakthrough in algorithm speed given by the introduction of the Fast Fourier Transform (FFT) (Katchalski-Katzir et al., 1992) (e.g. FTDock (Gabb et al., 1997), ZDock (Mintseris et al., 2007), PIPER (Kozakov et al., 2006)), and by some other very successful geometry-based methods (e.g. FRODOCK (Garzon et al., 2009), Hex (Ritchie and Kemp, 2000), MolFit (Katchalski-Katzir et al., 1992)). A docking procedure usually involves several steps (Vajda and Kozakov, 2009). First, a rigid-docking search is performed by treating the two proteins as rigid bodies. One of the proteins, called the receptor, is kept fixed while the other protein, the ligand, is rotated and translated around the first. Next, further refinement of some structures takes place, allowing changes in conformation of the two unbound structures upon binding (Dobbins et al., 2008; Shen et al., 2008); this step may or may not be supported by experimental evidence.

3.3.1 The ranking problem

Docking methods yield a large number of output conformations (ranging from 10000 to more than 50000), which include a large number of false positives. Thus, a crucial point after a rigid-docking search is the discrimination of near-native structures for further consideration and refinement. The number of selected conformations typically spans from 10 to 2000. There are two non-excluding strategies to perform such selection. The first strategy is to re-rank the docked conformations with a scoring function, which is supposed to rank near-native structures at the top (i.e. describe the molecular environment of the molecular interaction). Scoring functions are usually built upon different properties of protein-protein interactions observed in known binary complexes. These properties include physical and chemical characteristics of the binding site, at the level of residue or atomic contacts (Z-rank (Pierce and Weng, 2007), Fold X (Guerois et al., 2002)). Among these scoring functions, statistical potential is a term that refers to a knowledge-based scoring function that depends on specific properties of known protein-protein interactions stored in some database. Initially, statistical potentials were derived in order to distinguish a correct protein fold (i.e. near-native) of a model from a plethora of generated solutions (see section 4.2). In contrast to atomistic-detailed scoring functions, statistical potentials represent a much faster approach to solve this problem. It has been recently shown that the performance of split statistical potentials to rank docking poses (see following sections) may surpass that of scoring functions encoding atomistic energy terms or other statistical potentials (Feliu et al., 2011).

The atomistic scoring potentials of Z-rank and FoldX split the score into a linear combination of energetic terms and further obtained the best parameterization. In FoldX (1) the energy terms were split in the van der Waals (Gvdw), electrostatic (Gel), solvation (Gsolv) and hydrogen bonding (GHbond) contributions, and the entropy was also included. Some of these terms were split with different weights (i.e. the solvation of hydrophobic residues had a different weight than the solvation of polar residues, and the entropy of the main-chain had different weight than the entropy of side-chains). The parameters optimizing the final score were obtained using single-point mutations of nine different proteins and the corresponding free energies obtained with their 3D conformations.

$$\begin{aligned}
\Delta G &= \alpha_{vdw}\Delta G_{vdw} + \Delta G_{solv} + \Delta G_{Hbond} + \Delta G_{el} + T\Delta S \\
\Delta G_{solv} &= \alpha_{sh}\Delta G^{solv}_{hydrophobic} + \alpha_{sp}\Delta G^{solv}_{polar} \\
\Delta G_{Hbond} &= \Delta G_{water-bridge} + \left(G^{prot}_{Hbond} - G^{water}_{Hbond}\right) \\
\Delta S &= \alpha_{mc}\Delta S_{main-chain} + \alpha_{sc}\Delta S_{side-chain}
\end{aligned} \tag{1}$$

In Z-rank the energies were also split in van der Waals, electrostatic and solvation terms, but the weights of van der Waals and electrostatic interactions were different for attractive (a) and repulsive (r) interactions, and also different for short-range (<5Å) and long-range (>5 Å) interactions (sr and lr, for short and long ranges, respectively):

$$\begin{aligned}
score &= E_{vdw} + E_{el} + E_{solv} \\
E_{vdw} &= \alpha^{lrr}_{vdw}E^{lrr}_{vdw} + \alpha^{lra}_{vdw}E^{lra}_{vdw} + \alpha^{srr}_{vdw}E^{srr}_{vdw} + \alpha^{sra}_{vdw}E^{sra}_{vdw} \\
E_{el} &= \alpha^{lrr}_{el}E^{lrr}_{el} + \alpha^{lra}_{el}E^{lra}_{el} + \alpha^{srr}_{el}E^{srr}_{el} + \alpha^{sra}_{el}E^{sra}_{el}
\end{aligned} \tag{2}$$

The second strategy follows the rationale that near-native structures will show a broader and deeper well in the energy landscape compared to non–near-native structures. This assumption is the basis of clustering a collection of output conformations (around 1000–2000) as a function of the number of similar structures. Clustering is performed using as the similarity measure either the Cα binding site root mean square deviation (named I-RMSD) (Comeau et al., 2004) or the ligand Cα RMSD (Ritchie and Kemp, 2000). Selection based on the clustering methodology has proved to be better for determining near-native conformations than selection based solely on scoring functions (Ritchie and Kemp, 2000; Vajda and Kozakov, 2009). Consequently, the clustering method has become popular, mainly in combination with a re-ranking given by a scoring function that guides the selection of structures to cluster (Comeau et al., 2004; Shen et al., 2008).

3.3.2 Knowledge based potentials

In knowledge-based potentials, also named statistical potentials, the interaction between two residues is scored by the potential of mean force (PMF) obtained from the probability of finding a pair of residues at a given distance (Sippl, 1990). Let k_B denote the Boltzmann constant and let T be the standard temperature (300K). If A and B are the two interacting chains and a,b are two residues in chains A and B (respectively) at distance d, the potential of mean force is given by:

$$PMF(a,b,d) = PMF_{std}(d) - k_B T \log\left(\frac{P(a,b \mid d)}{P(a)P(b)}\right) \tag{3}$$

where $PFM_{std}(d)=k_BTlog(P(d))$; P(a), P(b) are the individual probabilities of residues a, b; P(a,b|d) is the conditional probability of residues a,b at distance smaller or equal to d and P(d) the probability of any pairs of residues at distance smaller or equal to d. All probabilities correspond to the observed frequencies of the events in the reference database (i.e. 3DID (Stein et al., 2005))

The score of the interaction is then defined as the sum over all interacting pairs of the pair residue scores. Formally, if $a_1,...a_s$ is the residue sequence of chain A, $b_1,...,b_r$ is the residue

sequence of chain B, Γ is the set of pair position indices (i,j) of interacting residues a_i,b_j at distance d_{ij}, then the statistical potential E_{pair} is:

$$E_{pair} = \sum_{(i,j)\in\Gamma} PMF(a_i, b_j, d_{ij}) \tag{4}$$

As energy can usually be split in independent terms from which different forces are derived, the statistical potential can also be split in terms that would describe the different parts of the interaction as particular forces. Particularly, considering a residue *condition* θ as the triplet formed by (secondary structure, polarity, degree of exposure), then the PMF in (3) can be decomposed using:

$$\begin{aligned}
PMF_{pair}(a,b) &= -k_B T \log\left(\frac{P(a,b \mid d_{ab})}{P(a)P(b)P(d_{ab})}\right) \\
PMF_{local}(a,b) &= k_B T \log\left(\frac{P(a \mid \theta_a)P(\theta_a)}{P(a)}\right) + k_B T \log\left(\frac{P(b \mid \theta_b)P(\theta_b)}{P(b)}\right) \\
PMF_{3D}(a,b) &= k_B T \log(P(d_{ab})) \\
PMF_{3DC}(a,b) &= k_B T \log\left(\frac{P(\theta_a,\theta_b \mid d_{ab})}{P(\theta_a,\theta_b)}\right) \\
PMF_{S3DC}(a,b) &= -k_B T \log\left(\frac{P(a,b \mid d_{ab},\theta_a,\theta_b)P(\theta_a,\theta_b)}{P(a,b \mid \theta_a,\theta_b)P(\theta_a,\theta_b \mid d_{ab})}\right)
\end{aligned} \tag{5}$$

Finally, the split statistical potentials E_{pair}, E_{local}, E_{3D}, E_{3DC}, and E_{S3DC} can be obtained by applying the formula (4) to the decomposed PMFs (5), with corresponding subindexes between E_ and PMF_. It was shown (Aloy and Oliva, 2009) that Epair admits a decomposition of the form:

$$E_{pair} = E_{S3DC} - E_{3DC} + E_{3D} - E_{Local} + E_{cmp} \tag{6}$$

where E_{cmp} is a residual energy term depending only on the conditions of the interacting residues and accounts for the reference state (first term in PMF equations). This equation was initially derived for the scoring of protein folds, but it remains valid when applied to the residues in the interface between two interacting proteins (Feliu et al., 2011).

Note that the statistical potential E_{S3DC} is a refinement of the residue-pair statistical potential E_{pair}, in the sense that it takes into account not only the residues that interact but also the condition in which each of them sits. On the contrary, the statistical potential E_{3DC} depends on the occurrence of interacting conditions, disregarding the specific interacting residues. The score E_{local} reflects the probability of placing a residue on a specific condition. Moreover, it splits into two terms, each of them depending only on the probability of placing a certain residue in some condition for each chain separately. The energy term E_{3D} concerns only the distance at which pairs of residues interact, and increases together with the number of interacting residue-pairs, thus being proportional to the number of residues implied in the interface.

3.3.3 Using split statistical potentials to rank docking poses

To test the scoring functions, the benchmark decoy dataset of Weng and cols. (Hwang et al., 2008) is widely used as gold standard. This dataset is based on a set of non-redundant real interactions for which both the complex 3D structure and the individual chain structures are available. It consists of a collection of binary complexes (124) with known structure (named targets) and a set of decoys for each of them (named target set). The 54,000 decoys generated using the rigid-body docking algorithm ZDock3.0 (Mintseris et al., 2007) from the individual chain structures were considered. The set of binary-complex conformations of a rigid-body prediction are classified according to the expected difficulties to obtain a near-native solution of the target. They deal with three types named: easy, medium and difficult cases. In total, the dataset consists of 124 cases, 88 of which are straight forward for rigid-body docking, 19 are medium and 17 are difficult cases for which further conformational changes are required upon binding. Only 97 of them (88 rigid-body and 9 medium) fit into the common near-native decoy criterion of structures differing from the native one at most 2.5Å (computed in terms of I-RMSD from the native structure). For difficult cases it is not possible to have near-native poses because of the deformation suffered by one or two of the protein partners. Thus, a different definition of a successful prediction is required in these cases. A selected pose was considered good if its I-RMSD differs less than 0.5Å from the lowest I-RMSD among all the decoys in the target set. This measure enables to determine if the scoring function top-ranks the best available decoys of the set. Figure 4 shows the success curves for the split potentials, revealing the relative importance of E_{local} and E_{3D} in the composition of the residue-pair statistical potential E_{pair}.

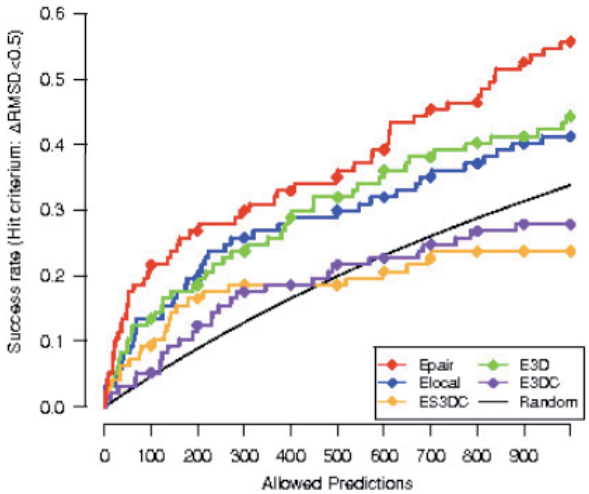


Fig. 4. Success curves for the split potentials: Success curves on the whole benchmark dataset are plotted for the five statistical potentials E_{pair} (red), E_{S3DC} (orange), E_{local} (blue), E_{3D} (light green) and E_{3DC} (purple), plus the success curve expected by random (black).

Based on the observation that E_{pair} and E_{S3DC} provided a fairly amount of non-overlapping hits, a new ranking strategy was defined: "MixRank". This strategy consists of first considering the lists of decoys ranked by different scoring functions separately, and then alternatively selecting one decoy from each list. Then, in order to avoid repetitions, we

apply a removal of redundant predictions (Feliu and Oliva, 2010). That is, we do not include decoys that are less than 5Å of I-RMSD from an already selected decoy. This way of removal of redundancies was analyzed (Feliu and Oliva, 2010) and was proved to provide better selection of near-native decoys. This ranking strategy proved to be able to compete with other ranking strategies based on atomistic-detailed scoring functions if large conformational changes of the interacting partners are required for the interaction. These are the cases typically included in the medium and difficult categories of the benchmark data set. This is shown in Figure 5, where E_{pair} and MixRank surpass ranking system based either on a reference statistical potential (RPScore (Moont et al., 1999)) or on an atomistic-detailed scoring function (ZRank) when predicting near-native poses within the medium and difficult categories of the benchmark data set.

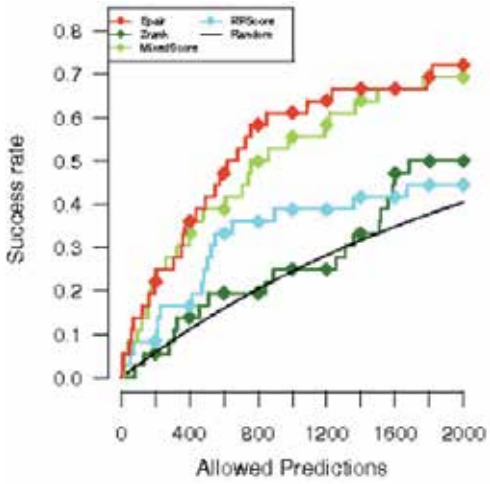


Fig. 5. Different ranking approaches compared for difficult cases of the benchmark data set: Success curves are plotted after removal of redundant solutions for the MixRank strategy (light green), Epair (red), Zrank (dark green) and RPScore (blue) scoring functions, and also compared with the success curve expected by random (black), only with the medium and difficult cases of the benchmark dataset.

4. Errors in models

The quality of the obtained model establishes the limits of the biological information that can be safely extracted from it. Although all structural models may enclose errors, these become less of a problem if correctly detected and assessed: once an error is identified, it is possible to discriminate whether it affects key structural or functional regions. Therefore an essential step in any structural modeling process is the detection of the wrongly modeled regions.

4.1 Sources of errors

In comparative modeling (homology modeling and threading), wrongly modeled regions are expected to be more frequent as the sequence identity between the query protein and the template decreases. Errors can be expected to occur at any step of the process, thus, they can

be catalogued according to the step in which they can be found and, therefore, the step in which they can be corrected or compensated. Docking techniques may incorporate similar errors during the step of molecular refinement.

Wrong template selection is the most costly error that can be found in a modeling process. Being a key step in the process, the selection of a wrong template cannot be overcome at any other part of the process and will inevitable yield to a wrong model. Correcting such error implies going back to the beginning of the modeling process and start all over again. The selection of a wrong template usually derives from the lack of a sequence homologous enough to sequence the query protein. A lot of effort is being put into trying to describe the optimal thresholds of identity and similarity to decide whether or not a sequence can be chosen as template.

Misalignment errors tend to appear under a 40% sequence identity. Their abundance rapidly increases below 30% of identity, as the occurrence of local regions with very low sequence identity makes wrong alignments more feasible. These errors are specially focused on gap misplacements in the alignment, and are one of the major sources of problems in homology modeling. As with the detection of the correct template, the sequence-template alignment is key, and correcting it requires redoing the alignment.

Structural distortions can be found both in well-aligned and in unaligned regions. Those in aligned regions appear when the sequence identity is too low in a local region and the sequence does, in fact, acquire a different secondary structure than that of the template. This problem can be overcome by using several templates in low identity regions in order to explore the possibilities. The regions that, even with multiple templates, are not aligned to any template have to be predicted by energy-based methods of database searching. The alignments at the sequence boundaries and 3D boundaries of such regions will determine the accuracy of the prediction.

Finally, side chain packing needs to be optimized especially as sequence identity decreases. Such optimization can be a major issue, specifically when it involves residues implicated in the protein's function and mostly in the interface of interacting proteins.

4.2 Detecting the errors

Automated methods for detecting errors in 3D models rely on the knowledge of previously solved structures in the PDB. This knowledge has lead to identify stereochemical and energy-related restrictions in the final 3D conformation of a protein. Considering stereochemical restrictions, perhaps the most obvious is that two amino-acids cannot clash (i.e. they cannot occupy the same spatial region). In addition, not all possible relative orientations of two correlative amino-acids in the protein sequence are allowed. These orientations are defined by the Φ and Ψ angles of the amino-acidic bond and the applicable restrictions are summarized in the Ramachandran diagram (Ramachandran et al., 1963), which represents the allowed conformations as a function of the Φ and Ψ angles. PROCHECK program (Laskowski et al., 1996) assess the overall quality of a protein model based on these parameters.

Besides stereochemistry, there are other protein spatial features in the proteins that could be used as indicators of errors in the models: packing, creation of a hydrophobic core,

residue and atomic solvent accessibilities, spatial distribution of charged groups, distribution of atom-atom distances, and main-chain hydrogen bonding structures (Sali, 1995). These are key features to understand the mechanisms by which a protein finds its native state. This mechanism is known as the folding pathway and the possibilities space for the folding of a protein is vast (Levinthal, 1968). Solving this problem requires an accurate potential describing the interactions among different amino-acid residues (Dinner et al., 2000). However, the use of such atomistic-detailed potentials (Brooks et al., 2009) is quasi-prohibitive and it does not ensure the native and biologically active conformation.

An alternative approach to the full atomistic description is to construct a coarse grained potential. The aim of such potential would be to approximate the function: a) whose global minimum corresponds to the native structure (Sippl, 1990), and b) capable to drive the structure from incorrect folding states toward native-like conformations (i.e. the having a correlation with native structure similarity (Keasar and Levitt, 2003)) describing a funnel-like energy surface. This scoring function, termed knowledge-based or statistical potential, works as a coarse-grained descriptor of the environment of the protein, and can be used to assess the quality of a protein 3D model. Based on this approach PROSAII (Sippl, 1993) is probably the most widely used program to assess the quality of a protein 3D model. Similarly, specific potentials have been derived for the interaction between macromolecules in order to assess protein-protein interactions (e.g., M-TASSER (Chen and Skolnick, 2008), MULTIPROSPECTOR (Lu et al., 2002) or InterPreTS (Aloy and Russell, 2003). Nevertheless, a funneling theory such as the Levinthal paradox in protein folding is still under development and some explanations are recently found (Wass et al., 2011).

5. Integrative modeling

The previous detailed methods could be useful in small complexes, where the docking of few subunits can solve the quaternary structure. However, the assembly of large macromolecular complexes such as the nucleopore complex, which contains more than 450 proteins, is unaffordable. In these cases, the presence of such amount of subunits forces the necessity to find methods that could manage the assembly problem in terms of costs and time.

During the last years, the integration of the maximum amount of structural information available about the structurally unknown macromolecular complex has become the state of the art solution to this problem. The main idea of this methodology is to use particular characteristics of the complex that can be synergistically combined in order to restrict the possible solutions to only those consistent with these features.

Electron microscopy has been established as a crucial technique for studying the structure of macromolecular assemblies (Alber et al., 2007). The resolution is insufficient to construct an atomic model but reveals insights into the shape and size of the whole complex. Thus, fitting atomic-resolution structures into the electron density maps is a suitable method for determining not only large macromolecular assemblies but also small ones.

Several methods have been developed for simultaneously fitting the individual protein subunits into the density map of their assembly. MultiFit (Lasker et al., 2010b) solve the position and orientation of each component within a protein structure using a function score that maximizes the quality of fit in the electron density map, the protrusion from the density map envelope, and the complementary shape between subunits. An optimizer algorithm DOMINO (Discrete Optimization of Multiple Interacting Objects) (Lasker et al., 2009) searches like a puzzle the positions of the subunits within a discrete sampling space. Each subunit is placed in a particular position inside the density map, conditioning the position of the rest of the subunits. This algorithm is used to efficiently find the global minimum in an affordable way.

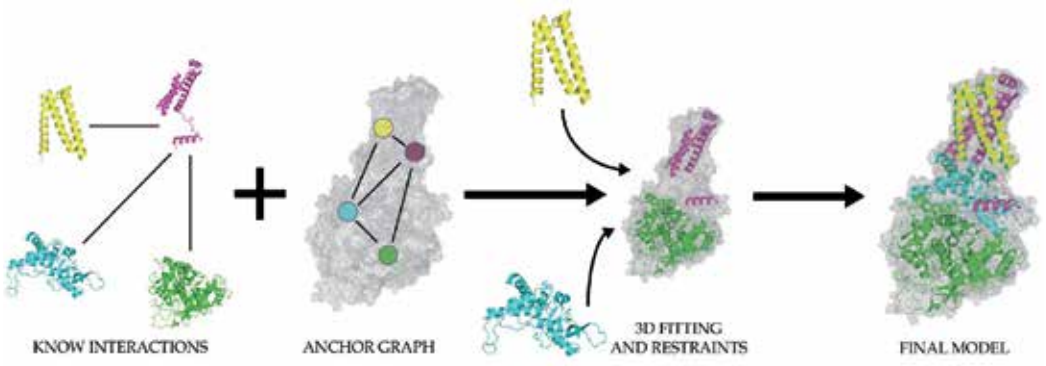


Fig. 6. Schematic representation of integrative modeling.

Often, the electron density map or the high-resolution structure of the subunits is not available. In these cases, it is not possible to apply the fitting procedure mentioned above. However, the integrative approach is not restricted to this data. There are different techniques that provide different types of information that can be used to understand particular features of the assembled complex. Table 1 highlights a list of proteomics, biophysical and computational methods used to obtain this valuable data.

In this way, Sali and collaborators developed an integrative modeling platform (IMP) (Lasker et al., 2010a) that collect this information and consider them simultaneously to generate models consistent with the data. This platform was used to describe the nuclear pore complex (Alber et al., 2007) and the structure of chromatin at megabase scale (Bau et al., 2010). Moreover, this platform can be used to solve any kind of 3D structure when enough data is provided.

IMP performs its function in an iterative series of four different steps. Below, a brief description of each step gave us an insight into how this heterogeneous data can be combined to deliver such large complex models.

Type of Structural Information	Techniques
Composition	Mass spectrometry and quantitative immunoblotting
Interactions	Genetic interactions and bioinformatics predictions
Connectivity	Affinity purification and surface plasmon resonance (SPR)
Interaction partners	Yeast to hybrid, protein microarrays, protein-fragment complementation assay (PCA) and calorimetry
Interaction distances	Fluorescence resonance energy transfer (FRET), bioluminescence resonance energy transfer (BRET) and cross-linking
Complex shape	X-ray scattering (SAXS) Cryo-electron microscopy, Cryo-electron tomography and Negative stain electron microscopy
Protein positions	High resolution electron microscopy, gold-labelling, green fluorescence protein (GFP) labelling and Docking
Residue positions	Crosslinking, hydrogen/deuterium exchange, Limited Proteolysis and Footprinting
Atomic positions	X-rays crystallography and nuclear magnetic resonance (NMR)

Table 1. Proteomics, biophysical and computational methods used to obtain information for modelling macromolecular complexes.

5.1 Data gathering

The collection of structural information is the first requirement needed to start the assembly process. The techniques listed in table 1 are appropriate generators of this data. In addition, a large amount of biological information is available through databases. Table 2 lists some databases with structural relevant information.

5.2 System representation and data translation into spatial restraints

One of the most characteristic features of the integrative modeling process is the ability to use structures that are not solved in high-resolution. In those cases, it is necessary to find an appropriate representation of the system. For example, on one hand, an atomic-resolution structure can be represented with particles corresponding to atoms and, on the other hand, in a low-resolution structure a single particle can represent a sphere corresponding to a group of atoms, residues or domains. Consequently, the resolution of the final complex is dictated by the resolution of the available data.

The raw data gathered in the first step must to be translated into spatial restraints, which specify values for the encoded data in order to decide if the model satisfies or not the experimental information about it. A restraint is a scoring function that reaches its minimum if the feature is consistent with the experimental data. A 0 indicates a model that is perfectly consistent with the restraint, whereas the result of the function is higher when the restraint is violated. In Table 3, most common types of restrains are reviewed.

Database	Description
PDB	PDB (Protein Data Bank) is the worldwide repository of information of 3D biological molecules structures
ModBase	ModBase is a relational database of protein structure models calculated by comparative homology modelling of known structures
SCOWLP	SCOWLP (Structural Characterization Of Water, Ligands and Proteins) is a relational database for detailed structural analysis of PDB protein interfaces at atomic level. The SCOWLP includes proteins, ligands and water as descriptors of interfaces
3DID	3did (3D interacting domains) is a collection of domain-domain interactions extracted from atomic-resolution structures. Each domain is associated to a Pfam domain and the database is GO term functional annotated
EMdataBank	EM Data Bank is a database of cryo-electron microscopy maps, models and associated metadata
BioGRID	BioGRID (Biological General for Interaction Datasets) is a database that archives genetic and proteomic interactions curated from high-throughput datasets and individual studies
PRISM	PRISM (Protein Interactions by Structural Matching) is a web-served compilation of protein-protein interaction interfaces
SCOPPI	SCOPPI (Structural Classification of Protein-Protein Interactions) is a database of all domain-domain interactions and their interfaces derived from PDB structure files and SCOP domain definitions

Table 2. Databases of structural information suitable for the integrative modelling process.

Type of restraint	Description of the restraint
Distance restraints	Restraints the distance between two particles
Connectivity restraints	Restraints all proteins in a set to interact or not.
Quality of fit restraint	Restraints the overlapping of the particles in an electron density map
Excluded volume	Restraint steric clashes
Geometric complementary	Maintains the geometric complementary between two particles interfaces
Statistical potential restraint	Restraint depending on the frequencies of contacts in previous solved complexes
Angle restraint	Restraint the angle between three particles
Protein localization restraint	Restraints a particle in a specific position
Complex diameter restraint	Restraints the maximum distance between the two most distance particles
Symmetry restraint	Maintains the same configuration of equivalent particles across multiple symmetry units
Radial distribution function restraint	Restraints the correlation between experimentally measured and computed radial distribution functions

Table 3. Most common types of spatial restrains obtained from structural data.

5.3 Calculation of an ensemble consistent with the restraints

At this point, the different restraints are combined into a final scoring function, which is commonly the sum of the singular scoring functions corresponding to each restraint. Then, the configuration of the constituent protein beads is determined by optimizing this scoring function.

The optimization process consists of searching through the configuration space the positions and orientations of the structural subunits that minimizes this function. It starts from random positions and iteratively moves them to minimize the violation of the restraints. In essence, a kind of 'force' pulls the proteins together to the native complex configuration. For this task, it is possible to use methods that explore the scoring function landscape in an efficient manner, such as conjugate gradient, molecular dynamics with simulated annealing or personalized optimizers, such as DOMINO (Lasker et al., 2009).

5.4 Analysis of the ensemble

Assuming a unique native state of the complex, the optimization process it is supposed to give a single model that satisfies all restraints. However, if the data used to encode the restraints is insufficient, more than one solution might be obtained. This problem could be solved introducing new restrains and running the process again. Conversely, in case of incorrect restraints, it is possible that no solution is obtained because there is not a model that satisfies all the restraints. In conclusion, the integrative method is a very powerful tool but it is clearly conditioned by the quality of the gathered information. Finally, the structure of the complex needs to be evaluated using similar approaches as in modelling, but adding the quality of the accomplishment of the restraints applied to construct the macro-complex.

6. Conclusions

Protein sequences are totally valueless if meaningful information about their biological function is not reported. In the past 30 years clear relationships between proteins sequence, structure, and function have been proven. Thus, the knowledge of a protein's 3D structure is normally required to completely understand its function. Since many proteins act in association with others, the knowledge of the structure of the complex formed by this association (named quaternary structure) is crucial to understand how proteins perform their functions. In this chapter we have attempted to establish the capabilities and limitations of currently available computational methods for predicting the tertiary and quaternary structure of proteins. Different strategies can be followed depending on the data available, and this review hopefully could serve as a practical guide for modelling the tertiary structures of proteins and its association into complexes.

When known structures of homologous proteins are available, these can be used as a template to model the structure of a target protein or a protein complex in a process termed comparative modelling. Being the current knowledge on the structure of protein complexes much more limited than that of single proteins, several databases of protein-protein interfaces such as PRISM (Aytuna et al., 2005) have been developed to overcome this problem. In comparative modelling, the percentage of sequence identity between the problem proteins and the templates is crucial. Below a certain threshold of sequence identity

(~30%) comparative modelling becomes a difficult task even for experts. In any case, models must be critically evaluated to be sure that they are correct, devoting most efforts to the region involved in the function (usually implying its interaction with other proteins or compounds).

On the lack of experimental data of the structure of a complex of proteins, protein docking is one of the computational techniques for elucidating the structures of binary interactions. We have shown that the use of split knowledge-based statistical potentials to score and rank the different docking solutions can be as accurate as atomistic-detailed potentials (Feliu et al., 2011) in any type of docking. Furthermore, these statistical poteintals surpass atomistic detailed scores when the complex requires large conformational changes of the interacting partners upon the interaction and we apply a rigid docking protocol.

Finally, we reviewed how different sources of experimental data are synergistically used to model large macromolecular complexes by the Integrative Modelling Platform (Lasker et al., 2010a). This approach has successfully been used to elucidate the structure of the nucleopore complex or the structure of chromatin at megabase scale.

7. References

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., *et al.* (2007). Determining the architectures of macromolecular assemblies. *Nature 450*, 683-694.

Aloy, P., and Oliva, B. (2009). Splitting statistical potentials into meaningful scoring functions: testing the prediction of near-native structures from decoy conformations. *BMC Struct Biol 9*, 71.

Aloy, P., and Russell, R.B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics 19*, 161-162.

Aloy, P., and Russell, R.B. (2004). Ten thousand interactions for the molecular biologist. *Nat Biotechnol 22*, 1317-1321.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*, 3389-3402.

Aytuna, A.S., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics 21*, 2850-2855.

Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2010). The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol 18*, 107-114.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res 28*, 235-242.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., *et al.* (2009). CHARMM: the biomolecular simulation program. *J Comput Chem 30*, 1545-1614.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci 13*, 190-202.

Chen, H., and Skolnick, J. (2008). M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J 94*, 918-928.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res 31*, 3497-3500.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science 267*, 383-386.

Comeau, S.R., Gatchell, D.W., Vajda, S., and Camacho, C.J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics 20*, 45-50.

Davis, F.P., and Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics 21*, 1901-1907.

De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol 5*, 15.

Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M., and Karplus, M. (2000). Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem Sci 25*, 331-339.

Dobbins, S.E., Lesk, V.I., and Sternberg, M.J. (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A 105*, 10390-10395.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics 14*, 755-763.

Feliu, E., Aloy, P., and Oliva, B. (2011). On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. *Protein Sci*.

Feliu, E., and Oliva, B. (2010). How different from random are docking predictions when ranked by scoring functions? *Proteins 78*, 3376-3385.

Fernandez-Fuentes, N., Rai, B.K., Madrid-Aliste, C.J., Fajardo, J.E., and Fiser, A. (2007). Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics 23*, 2558-2565.

Fornes, O., Aragues, R., Espadaler, J., Marti-Renom, M.A., Sali, A., and Oliva, B. (2009). ModLink+: improving fold recognition by using protein-protein interactions. *Bioinformatics 25*, 1506-1512.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol 272*, 106-120.

Gao, M., and Skolnick, J. (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc Natl Acad Sci U S A 107*, 22517-22522.

Garzon, J.I., Lopez-Blanco, J.R., Pons, C., Kovacs, J., Abagyan, R., Fernandez-Recio, J., and Chacon, P. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics 25*, 2544-2551.

Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol 320*, 369-387.

Guharoy, M., and Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A 102*, 15447-15452.

Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins 73*, 705-709.

Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst 6*, 2351-2362.

Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol 287*, 797-815.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A 89*, 2195-2199.

Keasar, C., and Levitt, M. (2003). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J Mol Biol 329*, 159-174.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc 4*, 363-371.

Keskin, O., Tsai, C.J., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci 13*, 1043-1055.

Kozakov, D., Brenke, R., Comeau, S.R., and Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins 65*, 392-406.

Lasker, K., Phillips, J.L., Russel, D., Velazquez-Muriel, J., Schneidman-Duhovny, D., Tjioe, E., Webb, B., Schlessinger, A., and Sali, A. (2010a). Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol Cell Proteomics 9*, 1689-1702.

Lasker, K., Sali, A., and Wolfson, H.J. (2010b). Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins 78*, 3205-3211.

Lasker, K., Topf, M., Sali, A., and Wolfson, H.J. (2009). Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J Mol Biol 388*, 180-194.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR 8*, 477-486.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., *et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol 487*, 545-574.

Levinthal, C. (1968). Are there pathways for protein folding? *J Chem Phys*, 44-45.

Lu, L., Lu, H., and Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins 49*, 350-364.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A 100*, 5772-5777.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res 11*, 2120-2126.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. (2007). Integrating statistical pair potentials into protein complex prediction. *Proteins 69*, 511-520.

Moont, G., Gabb, H.A., and Sternberg, M.J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins 35*, 364-373.

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol 338*, 181-199.

Nooren, I.M., and Thornton, J.M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol 325*, 991-1018.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol 302*, 205-217.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics 23*, e13-16.

Pierce, B., and Weng, Z. (2007). ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins 67*, 1078-1086.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol 7*, 95-99.

Ritchie, D.W., and Kemp, G.J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*, 178-194.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng 12*, 85-94.

Russell, R.B., and Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins 14*, 309-323.

Sali, A. (1995). Modeling mutations and homologous proteins. *Curr Opin Biotechnol 6*, 437-451.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins 23*, 318-326.

Shen, Y., Paschalidis, I., Vakili, P., and Vajda, S. (2008). Protein docking by the underestimation of free energy funnels in the space of encounter complexes. *PLoS Comput Biol 4*, e1000191.

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol 213*, 859-883.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins 17*, 355-362.

Stein, A., Russell, R.B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res 33*, D413-417.

Teyra, J., Doms, A., Schroeder, M., and Pisabarro, M.T. (2006). SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics 7*, 104.

Tsai, C.J., Xu, D., and Nussinov, R. (1997). Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci 6*, 1793-1805.

Tuncbag, N., Gursoy, A., Guney, E., Nussinov, R., and Keskin, O. (2008). Architectures and functional coverage of protein-protein interfaces. *J Mol Biol 381*, 785-802.

Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform 10*, 217-232.

Vajda, S., and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol 19*, 164-170.

Wass, M.N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology 7*, 469.

Watson, J.D., Laskowski, R.A., and Thornton, J.M. (2005). Predicting protein function from sequence and structural data. *Curr Opin Struct Biol 15*, 275-284.

Winter, C., Henschel, A., Kim, W.K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res 34*, D310-314.

Wodak, S.J., and Janin, J. (1978). Computer analysis of protein-protein interaction. *J Mol Biol 124*, 323-342.

11

Computational Approaches to Predict Protein Interaction

Darby Tien-Hao Chang
National Cheng Kung University
Taiwan

1. Introduction

Recently there have been large advances in high-throughput experimental approaches to identifying protein interactions. However, these experimental verified interactions still account for a small proportion of the complete interaction network. For example, based on current understanding (Stumpf, Thorne et al. 2008), less than 10% of interactions of the human protein interaction network (PIN) are identified and collected in the Human Protein Reference Database (HPRD) (Peri, Navarro et al. 2003; Stumpf, Thorne et al. 2008). The low coverage can be complemented by the computational approaches methods to predict protein interaction. This chapter describes approaches based on different biological observations and/or different computational techniques. Another focus of this chapter is to highlight the importance of creating a benchmark - especially negative samples since there are very limited techniques developed to confirm that two proteins do not interact (Doerr 2010; Smialowski, Pagel et al. 2010) - in evaluating computational approaches.

Computational methods can be roughly divided into two categories. Methods in the first category utilize the observation that functionally related proteins have patterns of co-occurrence, such as co-evolution or co-expression; while methods in the second category compile proteins into features potentially related to protein interaction, such as protein surface area, and resort to machine learning (ML) techniques for prediction. Different co-occurrence-based methods are distinct in where, namely which biological properties, the co-occurrence is observed and in the implementation details to record the co-occurrence. For example, Salgado et al. suggested that some related genes are close in genome to make the transcription more efficient (Salgado, Moreno-Hagelsieb et al. 2000). Methods based on this observation utilize co-localization in genome and could, for example, use the distance between two genes to record the co-occurrence. Section 2 will introduce seven categories of co-occurrence as follows.

1. **Genomic location**—some genes producing proteins that will interact are close in genome to facilitate transcription;
2. **Cellular compartment**—interacting proteins should appear in the same area in a cell to interact, another co-localization pattern which consider the cellular location;

3. **Phylogenetic tree**—if one protein was mutated in evolution, its cooperating protein should have a corresponding mutation to keep their interaction/function and thus the species survival, i.e. cooperating proteins should have similar phylogenetic trees;
4. **Existence in close species**—if two proteins co-work for a function to a species, then the species will have both of them, otherwise the species will have none of them, i.e. some related proteins are present in/absent from species together;
5. **Interacting domains**—interacting proteins usually have complementary parts of a interacting domain pair;
6. **Literature**—related proteins, since there must be some papers describing their relations, are prone to be mentioned together in literature, as opposed to other proteins existing only in articles describing their individual characteristics;
7. **Gene fusion**—some interacting proteins whose homologues form a fused protein chain, a special biological phenomenon named Rosetta Stone protein.

Different ML-based (or feature-based) methods, however, may share partial features to previous studies but develop new features at the same time. This led to more complicated relations than that among co-occurrence-based methods. For example, Shen et al. (Shen, Zhang et al. 2007) proposed to use a composition of short sequences as protein features and a following work by Chang et al. (Chang, Syu et al. 2010) combined these features with protein surface information. In addition to the overlap of features among different ML-based methods, they may use identical or different ML techniques. Using the two ML-based methods as an example, Shen et al. chose the widely used support vector machine (SVM) (Vapnik and Vapnik 1998), while Chang et al. used a relaxed variable kernel density estimator (RVKDE) (Oyang, Hwang et al. 2005) developed by their group. Thus to keep the description structure compact, we will focus on the features in section 3. We will provide only a minimum introduction to several well-known ML techniques in section 4 since they are beyond the scope of this chapter. Knowing the concepts of these ML techniques may help to understand the design of different ML-based PPI predictors and to select appropriate features. This chapter roughly divides features into four categories.

1. **Sequence information**—many studies extracted features only from protein sequences. Methods using only such features are very challenging but provide much applicability. Some derived features such as protein polarity (by summing the polarity index of its amino acids) are also included in this category.
2. **Evolution information**—features involving alignment with other sequences fall into this category. Methods using such features usually require a collection of protein sequences of many species.
3. **Structure information**—methods of this category can perform geometry and even energy analyses. Many useful features such as protein surface, secondary structure and binding affinity can be derived. These methods are usually time-consuming, where researchers will expect to obtain extremely accurate predictions.
4. **Auxiliary information**—some studies used auxiliary information such as function annotation. These studies usually used such features to analyze rather than to predict protein interactions, since some features were manually curated. It was hard to perform a fair comparison with other methods not using such features.

After the features used in recent ML-based methods there is an introduction to three well-known ML techniques.

1. **Decision tree**—a time-honored tool, which is less accurate than modern ML tools but preferred by many biologists because its learning model is more interpretable to human;
2. **SVM**—a state-of-the-art tool that overwhelmingly prevails in the field of computational biology because of its accuracy;
3. **RVKDE**—another modern ML tool that solves the most critical problem of SVM, unacceptable execution time on large data, by slightly sacrificing accuracy.

This chapter ends up with the important issue of computational approaches - evaluation. Computational approaches of identifying protein interactions have a fateful difference to experimental approaches. That is, their results are considered as "prediction" rather than the answer. So it is an inevitable step for the studies of computational methods that they must test their algorithms and report the prediction accuracy compared to a benchmark with the answers already known.

As a summary, this chapter will first introduce the concept of co-occurrence pattern and the implementation details of some co-occurrence-based methods. For the ML-based methods, this chapter focuses on the features and a little on the ML techniques. Finally, three contradictions are used to describe to readers the importance of evaluating these computational methods and explain how to interpret the accuracy they see in literature.

2. Co-occurrence-based approaches

This section introduces seven concepts of co-occurrence patterns that have been adopted to predict protein interactions. An identical concept, based on the available materials, may have different implementation details. In this section, the concept of each co-occurrence pattern is first introduced followed by the implementation details of several methods as examples of that co-occurrence pattern.

2.1 Genomic location

The advance of sequencing leads to the opportunity not only of identifying the genomic locations of genes, but also of analyzing genomic context to predict interactions between genes (Huynen and Snel 2000). The genomic location, also known as genomic context, co-occurrence pattern relies on the fact that operons and some adjacent genes are likely to encode functionally related proteins (Rogozin, Makarova et al. 2002). Huynen and Snel proposed a method to assess the probability that two genes occur as neighbours in a genome only by chance (Huynen and Snel 2000). They randomized the genes in each genome over the loci in that genome. The expected number of the co-occurrences of two genes, namely those that occur as neighbours, in the randomized genomes was less than one. A functional interaction between genes was inferred if the observed number of co-occurrences is significantly higher than the expectation. Rogozin et al. proposed a procedure to compare the orders of orthologous genes (Rogozin, Makarova et al. 2002). They clustered genes into orthologous groups which were then projected onto genomes to identify the neighborhood genes. The results show that the gene neighbours have good functional coherence.

2.2 Cellular compartment

Proteins that occur in different cellular compartments are, in principle, considered not to interact since they do not have the chance to meet. However, some *in vitro* experiments such as tandem affinity purification-mass spectroscopy (TAP-MS) method (Krogan, Cagney et al. 2006), might report such interactions of two proteins in different cellular compartments. It is difficult to determine if these *in vitro* interactions are correct. Thus, this co-occurrence pattern is usually used to increase the prediction reliability of another method or to generate a reliable benchmark rather than an individual interaction predictor. For example, the Eukaryotic Linear Motif (ELM) server used cellular compartment information as a filter to double-check gene function (Davey, Van Roey et al. 2011). The gene function, represented by its Gene Ontology (GO) terms (Ashburner, Ball et al. 2000), was required to be consistent with its cellular compartment. Guo et al. used cellular compartment information to build the negative data of protein interaction (Guo, Yu et al. 2008). They assumed that proteins that occur in different cellular compartments do not interact. They grouped proteins into eight subsets based on the eight main types of cellular compartment—cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, golgi apparatus, peroxisome, vacuole and cytoplasm and nucleus. The negative samples of non-interacting pairs were generated by pairing proteins from different subsets.

2.3 Phylogenetic tree

The phylogenetic tree was proposed to reflect the evolution information. Thus, the similarity between phylogenetic tress provides a good measure of gene co-evolution. Interacting proteins usually co-evolve since mutations in one protein led to the loss of function or a compensation mutation of the other protein to preserve the interaction (Walhout, Sordella et al. 2000). Jothi et al. proposed the MORPH, an algorithm to search the best superimposition between evolutionary trees based on the tree *automorphism* group in 2005 (Jothi, Kann et al. 2005). The search was done by Monte Carlo algorithm that probes the search space of all possible superimpositions, which is computationally intensive. In graph theory, two trees are isomorphic if there is a one-to-one mapping between their vertices (genes) and edges (interactions). Jothi et al. extended this definition to automorphic whereby a tree is isomorphic to itself. The search space was largely reduced to the automorphism group of a phylogenetic tree. The same group proposed another method to assess the degree of co-evolution of domain pairs in interacting proteins in 2006 (Jothi, Cherukuri et al. 2006). Multiple sequence alignments of two proteins/domains to a reference set of genomes were used to construct phylogenetic trees and similarity matrices. The degree of co-evolution of two domains was then estimated by the correlation coefficient of the two corresponding similarity matrices.

2.4 Existence in close species

The co-occurrence pattern of the existence in close species, known as phylogenetic profile, is based on the fact that functionally related proteins usually co-evolve and have homologues in the close genomes (Snitkin, Gustafson et al. 2006). A phylogenetic profile of a gene is a vector, representing the presence or absence of homologues to that gene across a collection

of reference organisms. There are two major components in a phylogenetic profile-based method: i) how to construct a phylogenetic profile of a given gene and ii) how to determine the similarity of two phylogenetic profiles. First, the presence or absence of homologues can be determined by sequence alignment scores, such as a BLAST (Altschul, Madden et al. 1997) E-value, with a threshold of presence (Sun, Xu et al. 2005). Such binary vectors were improved as real valued vectors of normalized alignment scores without arbitrarily determining a score threshold (Enault, Suhre et al. 2003). Second, any similarity or distance function between two vectors can be used to define the similarity of two phylogenetic profile vectors. Enault et al. have examined two Euclidean-like distance funtions and another two correlation coefficient variants (Enault, Suhre et al. 2003). They concluded that inner product, shown as follows, is a good indicator in predicting *Escherichia coli* protein interactions.

$$Sim(i,j) = \frac{\sum_{k=1}^{n} R_{ik} \times R_{jk}}{\left[\left(\sum_{k=1}^{n} R_{ij}^2\right) \times \left(\sum_{k=1}^{n} R_{jk}^2\right)\right]^{1/2}}$$

2.5 Interacting domains

Proteins usually depend on a short sequence of residues to perform interactions with other molecules. The functional short sequences between two interacting proteins form the contact interfaces, also known as interaction sites (Sheu, Lancia et al. 2005). These interaction sites are usually represented by domains/motifs. Li et al. proposed a method to detect interaction sites, which required only protein sequences (Li, Li et al. 2006). They developed an efficient itemset mining algorithm that can identify the most conserved motifs within two interacting protein groups. Here interacting protein groups indicate two groups, *A* and *B*, of proteins where all proteins in group *A* interact with all proteins in group *B*, denoted an *all-versus-all* interaction network. The conserved motifs within group *A* were considered, in principle, related to the conserved motifs within group *B*. The identified interacting motif pairs can be then used to predict novel interacting proteins. Tan et al. proposed a method, D-STAR to find correlated motifs that were overrepresented in interacting protein pairs (Tan, Hugo et al. 2006). The basic idea of D-STAR is to check all possible (*l*, *d*)-motif pairs, where (*l*, *d*) indicate an alignment of length *l* with at most *d* mismatches. Tan et al. speeded up the brute force procedure by transforming the problem into a clique-finding problem (Pevzner and Sze 2000).

2.6 Literature

Owing to the advance of Internet technologies, the scale of public accessible biomedical literature has increased astonishingly in the last decade. Text mining tools are critical to maximize the usage of such a large-scale knowledge base. Extracting protein interactions from literature is generally categorized as *relationship* mining, which aims to detect co-occurrences of a pair of entities of specific types (such as gene, protein, drug or disease) to a pre-specified relationship (such as interact, regulate, activate or inhibit) in the same article (Cohen and Hersh 2005). Albert et al. proposed a method to retrieve abstracts reporting

nuclear receptors (NRs) (Albert, Gaudan et al. 2003). The retrieved data were reviewed manually. Albert et al. generated a dictionary focusing on NRs, cofactors and other NR-binding proteins of human, mouse and rat. The extraction process as follows was performed on MEDLINE abstracts: i) identify abstracts with at least one NR in the generated dictionary, ii) tag entities (proteins) and relationships (interactions) according to the generated dictionary and iii) extract sentences contains two tagged proteins and a tagged interaction. In the current genomic era, the text-minded information is widely applied in database annotation. Many popular protein interaction databases such the Database of Interacting Proteins (DIP) database (Salwinski, Miller et al. 2004) and the Search Tool for the Retrieval of Interacting Genes (STRING) database (Szklarczyk, Franceschini et al. 2011) included automatically extracted literature information as an additional line of evidence.

2.7 Gene fusion

Gene fusion is a special genomic organization whereby some interacting proteins have orthologues in the close genomes fused as a single protein (Enright, Iliopoulos et al. 1999). The fused protein is usually called a Rosetta Stone protein, thus this method is sometime called the Rosetta Stone method. This genomic organization of gene fusion is formed for efficiently transcribing related genes together, thus it is preserved evolutionarily. Marcotte et al. applied the gene fusion method on *Escherichia coli* (Marcotte, Pellegrini et al. 1999). They identified 6,809 protein pairs of which both protein sequences were significantly similar to the same protein sequence of at least a genome. More than half of these 6,809 protein pairs have been shown to be related. This method, unlike previous co-occurrence patterns, is a very specific genomic organization rather than a concept of co-occurrence. Thus, there is very limited space for the algorithm development and implementation details. For any new genome, researchers can always search for Rosetta Stone proteins first. But other methods are required since many interacting proteins are not Rosetta Stone proteins. For example, in the DIP database that deposits experimentally confirmed protein interactions, only 6.4% interacting protein pairs formed Rosetta Stone proteins (Shoemaker and Panchenko 2007).

3. Machine learning-based approaches

This chapter roughly divides features into four categories: sequential, evolutionary, structural and other. Note that the power of ML tools allows researchers to submit any features, with or without obvious biological glues to protein interaction, into a magical black box and wait for the prediction without knowing how the prediction was made. For example, amino acid composition (20 features) and number of search results in PubMed can be used as features. Namely, each co-occurrence pattern can be used as a feature—the only thing to do is designing a rule to record the pattern with one or more real numbers. So in this chapter we only demonstrate several features that have been shown to help the prediction accuracy in published articles, but cannot list all features in a category.

3.1 Sequence information

One of the most widely used data to encode proteins is their primary sequence. Methods that only rely on protein sequences have a great advantage of the wide applicability. Because such methods do not rely on other information, they are sometime called *de novo* (*ab*

initio) predictors of protein interaction. Yu et al. proposed a method that encoded protein sequences as feature vectors by considering the amino acid triads observed in it (Yu, Chou et al. 2010). An amino acid triad regards three continuous residues as a unit. However, considering all 20^3 amino acid triads requires an 8000-dimensional feature vector to represent a protein, which is too large for contemporary machine learning tools. Thus, the 20 amino acid types were clustered into seven groups based on their dipole strength and side chain volumes to reduce the dimensions of the feature vector (Shen, Zhang et al. 2007). The frequencies of the $7^3 = 343$ triads can be used to encode a protein sequence. However, such a frequency is highly correlated to the distribution of amino acids. To overcome this problem, Yu et al. proposed a significance calculation by answering the question: how rare is the number of observed occurrences considering the amino acid composition of the protein? The significances of all triads were used to encode protein sequences.

Methods based on sequence motifs/domains also fall into this category since sequence motifs are mined from protein sequences. One may notice that the co-occurrence-based methods mentioned in subsection 2.5 used similar features. In this regard, the co-occurrence-based methods use domain as features with a straightforward rule: if two proteins have interacting domains, then they are predicted as interacting. On the other hand, ML-based methods use domain as features but resort to ML tools for the final decision/prediction. Depending on the ML tools used, the decision rules could be very complicated models of, for example, non-linear equations or a combination of multiple individual components (it could be a 'sum' of multiple functions). Dijk et al. proposed a ML-based method to select relevant motifs from a set of pre-mined motifs (Van Dijk, Ter Braak et al. 2008). They first invoked the D-STAR (Tan, Hugo et al. 2006) algorithm to identify correlated motifs that overrepresented in interacting protein pairs. The vector of the presence or absence of the identified motif pairs were used to encode proteins.

3.2 Evolution information

Methods that require not only the sequences of the query protein pairs but also a collection of supporting sequences fall into this category. The supporting collection is usually from other species for calculating the conservation score. Position-specific scoring matrix (PSSM) is a widely used scheme to encode a protein sequence while considering its orthologues. For a protein sequence, PSSM describes the likelihood of a particular residue substitution at a specific position based on evolutionary information (Altschul, Madden et al. 1997). It is outputted by BLAST when aligning the query protein seuqence to a seqeucne database, e.g. the non-redundant (NR) database from National Center for Biotechnology Information (NCBI). The likelihood values are scaled to [0,1] using the following logistic function:

$$x' = \frac{1}{1 - \exp(-x)},$$

where x is the raw value in PSSM profile and x' is the value corresponding to x after scaling. Each position of a protein sequence is represented by a 21-dimensional vector where 20 elements take the likelihood values of 20 amino acid types from the scaled PSSM profile and the last element is a terminal flag. Finally, the feature vector of a residue comprises a window of positions. Chang et al. proposed a method based on the assumption that protein interactions are more related to amino acids at the surface than those at the core (Chang, Syu

et al.). They first used PSSM to encode protein sequences for surface prediction and then used the surface sequence for interaction prediction.

Espadaler et al. proposed a method that made use of conservation of protein pairs (Espadaler, Romero-Isart et al. 2005). They first collected 855 protein complexes with known three-dimensional structure with <80% sequence identity. The 855 complexes were further classified into 16 groups. In a protein complex, the distance between a residue pair from two proteins was defined as the distance of the nearest heavy atoms of the two residues. Via setting a cut-off of the contact distance, one can identify the interface of two proteins in these complexes. These identified interfaces were actually unordered sequence fragments, among which Espadaler et al. defined more than five contiguous residues a *patch*. The conservation of the patches obtained by multiple sequence alignment was considered to select the final patches. These conserved structural patch pairs can be used to predict novel protein interactions. Notice that this method proposed by Espadaler et al. also used the structure information which will be introduced in the next subsection. This also reveals that with the ML tools, combining multiple resources becomes relatively easy since it is no longer dependent on a single co-occurrence pattern.

3.3 Structure information

The most critical problem of sequence-based methods is the reliability. Conversely, researchers usually resort to structure-based methods for verification since the results delivered by structure-based methods can be visualized. Aloy and Russell proposed a method to detect interactions based on protein tertiary structures. They used empirical potentials to compute the fitness between two protein structures. Thus, success of such a method is highly dependent on the performance of the underlying potential function. The adopted potential function did not rely on model proteins, which enlarges its applicability. Aloy and Russell defined interacting residues as those having at least one i) hydrogen bonds (N—O distances ≤ 3.5 Å), salt bridges (N—O distances ≤ 5.5 Å), or van de Waals interactions (C—C distances ≤ 5 Å). Buried side-chains were excluded by filtering out residues with relative accessibility ≥10%. The identified interacting residues were used to train the empirical potentials based on a molar-fraction random state model as follows:

$$S_{ab} = \log_{10}\left(\frac{O_{ab}}{E_{ab}}\right) E_{ab} = N \frac{n_a}{\sum_{a=1}^{20} n_a} \frac{n_b}{\sum_{b=1}^{20} n_b},$$

where a and b are amino acid types, O_{ab} and E_{ab} are the number of observed/expected contacts, N is the number of analyzed residue pairs and n_a and n_b are number of residues of the corresponding types. The method of Aloy and Russell provided ranks of analyzed protein pairs so that researchers can pick the most promising prediction for further biological experiments.

3.4 Auxiliary information

Important data that is not mentioned above is microarray data, which has been broadly utilized in various biomedical problems. The Gene Expression Omnibus (GEO) database (Barrett, Troup et al. 2006) of NCBI holds more than 20 thousands microarray experiments.

A problem of microarray data is that they are usually full of noises. Soong et al. used principal component analysis (PCA) to reduce such noises (Soong, Wrzeszczynski et al. 2008). PCA is a statistical technique used to find hidden factors from observed factors, expression values in this case. Lee and Batzoglou have shown that proteins with extreme principal components are prone to participate in relevant biological processes (Lee and Batzoglou 2003). The transformation of expression values to principal components can be represented as follows:

$$PX = Y,$$

where P is a $l \times m$ transformation matrix obtained by PCA, X is a $m \times n$ matrix of the raw expression values from m microarrays and n samples while Y is a $l \times n$ matrix containing every sample's l principal components. The final feature vector of two proteins a and b was the concatenation of a's principal components, b's principal components and the Pearson correlation of both.

This section ends with a method based on literature data, which has been discussed in subsection 2.6. Demonstrating literature data in a ML-based method is to reinforce the impression that in principle any data can be used as features with appropriate encoding schemes. Thus, one can consider combining any of the features discussed in section 2 with ML-based tools. Donaldson et al. proposed an extraction procedure for identifying protein interactions in literature (Donaldson, Martin et al. 2003). They first used a parser to collect synonyms for proteins and their encoding loci. The collected protein names were then used to search the title and abstract of articles in the PubMed literature database. An article was encoded by terms it contained. The weight of each term was the *tf-idf* score (term frequency-inverse document frequency), where term frequency is the number of occurrences of the term in the document and inverse document frequency is the inverse of the number of documents having the term. Here a term was a word or two adjacent words (usually called 2-gram) that appear in at least three documents.

4. Machine learning techniques

After encoding proteins into feature vectors, the next step is to choose a ML tool to generate a model describing these feature vectors. The generated model can be used to predict novel protein interactions. Most ML tools provide a user-friendly interface, where all that researchers need to do is encode their data. The remaining task is very trivial: i) a command to train the model and ii) a command to predict with the trained model. In this regard, researchers who want to adopt ML-based methods can focus on features without caring about the ML algorithms. This section briefly lists three ML algorithms that have been used in recent studies of protein interaction, which can be considered as a basic introduction for researchers who have no idea how to choose an appropriate ML tool.

4.1 Decision tree

Decision trees are usually constructed recursively (Witten, Frank et al. 2011). The first step is to select a feature to split samples (branch the decision tree) based on the selected feature. This step divides the original dataset into several disjointed subsets, each of them can be considered as another dataset. Thus, the same procedure can be applied recursively to each

subset and the further sub-subsets. Such a recursive fashion stops at several conditions of, for example, all samples in a branch belonging to the same class or all features have been examined. The above descriptions, however, missed an important detail in decision trees: how to select a feature to branch. A trivial strategy is to select the feature that can result in the purest subsets, namely most samples in the same subset belong to the same class. Thus, a measure of set purity is required.

So far, there have been many purity measurements proposed. This subsection introduces the most fundamental one, entropy, as follows. Indeed, larger entropy indicates the less purity. Thus, negative entropy, in definition, is a measurement of purity. Many mature decision tree algorithms use variants of entropy.

$$\text{entropy}(p_1, p_2, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n ,$$

where p_i is the fraction of class-i samples in the subset and n is the number of total classes. For example, suppose that a dataset has nine positive samples and five negative samples. Before any branching, the entropy of the original dataset is –(9/14)log(9/14)-(5/14)log(5/14) = 0.940, where p_1 is 9/14, p_2 is 5/14 and n is 2 (positive and negative). If after a branch, the 14 samples are split into three nodes that contain 2-3, 4-0, 3-2 positive-negative samples, respectively. The entropies of the three nodes are –(2/5)log(2/5)-(3/5)log(3/5) = 0.971, –(4/4)log(4/4)-0 = 0 and –(3/5)log(3/5)-(2/5)log(2/5) = 0.971, respectively. The total entropy of the branched tree became (5/14)×0.971+(4/14)×0+(5/14)×0.971 = 0.693, a weighted sum of the three entropies corresponding to the subset size. It is observed that the entropy decreased from 0.940 to 0.693, revealing that this operation of branch did increase the purity of the dataset. For a purity measurement, the following three conditions must be satisfied:

1. when a subset is pure (all samples belong to the same class), the measurement is zero;
2. when all possible classes appear equally, the measurement is maximized;
3. the measurements must be the same without depending on the order of branches.

The third condition requires that if a dataset is first split into two nodes of *a*-*b* and *c*-*d*, positive-negative samples and then the second node is further split into two more sub-nodes of *e*-*f* and *g*-*h*, the entropy should be the same as split into three nodes of *a*-*b*, *e*-*f* and *g*-*h* in a single branch while using another feature. Entropy is the only one function that fits all these conditions (Witten, Frank et al. 2011). This explains the high popularity of entropy and its variants in decision trees.

4.2 Support Vector Machine (SVM)

Currently, SVM is the state-of-the-art ML tool. It prevails in biomedical data because of its high accuracy. SVM first transforms the original data to a higher dimensional space with a non-linear transformation and then finds the maximum margin hyperplane to separate samples of different classes in the transformed space (Witten, Frank et al. 2011). This strategy has two advantages: i) it can generate non-linear model and ii) it prevent overfitting as the decision boundary is still linear in the transformed space. Overfitting is a critical issue in ML. It indicates that the constructed model overfit the training dataset, so that which cannot be used to predict novel data. This problem becomes more serious when using more complicated model. However, some complex data do need complicated

models to describe. Thus most advanced ML algorithms still favor complicated models and then try to solve the overfitting issue. In this regard, SVM finds an excellent balance, which can generate very complicated models depending on the adopted transformation while choosing a very simple decision, a hyperplane, which equals to a one stage decision tree of two branches.

Mathematically, SVM uses support vectors to model the transformation and hyperplane. That is the reason for the name. Transforming the original data from the sample space with a non-linear function to a new space means that a linear model (a straight line in a two dimensional space, a plane in a three dimensional space and a hyperplane in a higher dimensional space) in the new space becomes non-linear in the original sample space. For example, for a two dimensional sample $x = (a, b)$, a non-linear transformation to a three dimensional space could be $x' = (a^2, ab, b^2)$. If any ML tool finds a decision boundary in the new space, it does not look like a straight line in the original space. Notice that, in principle, any tool, such as a decision tree, could be used to make the decision in the transformed space. SVM advances in already developing a robust mathematical system with efficient optimization algorithms to find good hyperplanes.

4.3 Relaxed Variable Kernel Density Estimation (RVKDE)

The biggest drawback of SVM is the computational cost. Yu et al. reported that using SVM to perform a complete interaction analysis on human genome may take years (Yu, Chou et al. 2010). In this regard, efficient ML algorithms with acceptable accuracy are reasonable alternatives to SVM. The relaxed variable kernel density estimation (RVKDE) algorithm (Oyang, Hwang et al. 2005) has been practically used in recent interaction studies (Chang, Syu et al. 2010; Yu, Chou et al. 2010). The time complexity of RVKDE is an order faster than SVM. Furthermore, unlike other fast ML algorithms, such as decision trees, the descriptive capability of the constructed model of RVKDE is comparable to SVM.

The kernel of RVKDE is an approximate probability density function. Let $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ be a set of samples randomly and independently taken from the distribution governed by f_x in a m-dimensional vector space. RVKDE estimates the value of f_x at point $\mathbf{v}$ as follows:

$$\hat{f}(\mathbf{v}) = \frac{1}{n}\sum_{\mathbf{s}_i}\left(\frac{1}{\sqrt{2\pi}\cdot\sigma_i}\right)^m \exp\left(-\frac{\|\mathbf{v}-\mathbf{s}_i\|^2}{2\sigma_i^2}\right), \text{ where}$$

1. $\sigma_i = \beta\dfrac{R(\mathbf{s}_i)\sqrt{\pi}}{\sqrt[m]{(k+1)\Gamma(\frac{m}{2}+1)}}$;
2. $R(\mathbf{s}_i)$ is the maximum distance between $\mathbf{s}_i$ and its ks-th nearest training sample;
3. $\Gamma(\cdot)$ is the Gamma function (Artin 1964);
4. β and ks are parameters to be set either through cross-validation or by the user.

For prediction, a kernel density estimators is constructed to approximate the distribution of each class. Then, a query sample located at $\mathbf{v}$ is predicted to the class that gives the maximum value among the likelihood functions defined as follows:

$$L_j(\mathbf{v}) = \frac{|S_j| \cdot \hat{f}_j(\mathbf{v})}{\sum_h |S_h| \cdot \hat{f}_h(\mathbf{v})},$$

where $|S_j|$ is the number of class-*j* training samples and $\hat{f}_j(\cdot)$ is the kernel density estimator corresponding to class-*j* training samples.

RVKDE belongs to the radial basis function network (RBFN), a special type of neural networks with several distinctive features (Mitchell 1997; Kecman 2001). The decision function of two-class RVKDE can be simplified as follows:

$$f_{\text{RVKDE}}(\mathbf{v}) = \sum_{\mathbf{s}_i} y_i \cdot \frac{1}{\sigma_i} \cdot \exp\left(-\frac{\|\mathbf{v} - \mathbf{s}_i\|^2}{2\sigma_i^2} \right),$$

where $\mathbf{v}$ is a testing sample, y_i is the class value as either +1 (positive) or -1 (negative) of a training sample $\mathbf{s}_i$, and σ_i is the local density of the proximity of $\mathbf{s}_i$, estimated by the kernel density estimation algorithm. The testing sample $\mathbf{v}$ is classified as positive if $f_{\text{RVKDE}}(\mathbf{v}) \geq 0$, and as negative otherwise. Interestingly, the decision function of RVKDE is very similar to that of SVM using the radial basis function (RBF) kernel:

$$f_{\text{SVM}}(\mathbf{v}) = \sum_{\mathbf{s}_i} y_i \cdot \alpha_i \cdot \exp\left(-\gamma \|\mathbf{v} - \mathbf{s}_i\|^2 \right),$$

where a_i (corresponds to the inverse of σ_i in f_{RVKDE}) and γ (corresponds to $1/2\sigma_i^2$ in f_{RVKDE}) are user-specified parameters. Thus, the mathematical models of RVKDE and SVM are analogous. The main difference between RVKDE and SVM is the criteria used to determine σ_i and a_i.

5. Evaluation

A paradoxical situation is that a benchmark requires negative samples - proteins known not to interact. A benchmark that contains only interacting protein pairs is useless, since a trivial predictor predicting any protein pairs as interacting can achieve a perfect accuracy. However, there are very limited techniques developed to confirm that two proteins do not interact. Recently, several studies have addressed this problem in evaluating computational methods of identifying protein interactions (Yu, Chou et al. 2010; Yu, Guo et al. 2010). This issue is still in a chaos stage and there is no perfect solution that fit everyone's requirements. Instead, this chapter demonstrates this issue via three major contradictions in this area.

1. **Sampled vs. entire data (also efficiency issue)**—most ML-based methods adopted SVM and have to reduce the data size because of its high time complexity. However, sampled data must lose some information and may bias the evaluation. This contradiction is especially important when comparing co-occurrence- and ML-based methods, where the former usually can be applied on entire data. Using more computing power or switching more efficient ML tool is a compromising solution.

2. **Balanced vs. unbalanced**—once sampled data is adopted, (most studies of ML-based methods adopted using sampled data even without carefully considering the previous contradiction), how to sample is another serious problem. Random sampling can preserve the data distribution (ratio of positive and negative samples) but loss too many positive samples. However, balanced sampling, which forces the inclusion of all positive samples and thus change the data distribution, has also been shown bias the evaluation accuracy (Yu, Chou et al. 2010).
3. **Distinct vs. similar**—one philosophy of creating negative data is to choose the samples which can never be positive. For example, proteins appear in different cellular compartments are possible negative samples. An opposite philosophy is that if a method can discriminate between the negative samples that are very similar to the positive ones, then this method can discriminate those dissimilar ones. The first philosophy prevents collecting negative samples that are actually positive but somehow makes the problem easier while the second philosophy has opposite advantage and disadvantage.

6. Conclusions

In this chapter, various computational methods of protein interaction are reviewed. These methods used various data sources, including localization data, structural data, expression data and/or interactions from orthologs. As a result, all of them are limited to the experimental technologies that generate such data and the incompleteness of verified data. Based on current understanding, the size of protein interaction network (PIN) of human comprises ~650,000 interactions (Stumpf, Thorne et al. 2008). However, the Human Protein Reference Database (HPRD) deposits less than 3% of them (Peri, Navarro et al. 2003; Mishra, Suresh et al. 2006). Even under such a challenging circumstance, computational methods have shown to achieve satisfying performance. This encourages more effort in developing computational methods of protein interaction to complement experimental technologies.

7. References

Albert, S., S. Gaudan, et al. (2003). "Computer-assisted generation of a protein-interaction database for nuclear receptors." *Molecular Endocrinology* 17(8): 1555-1567.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Research 25*(17): 3389.

Artin, E. (1964). *The Gamma Function*. New York, Holt, Rinehart and Winston.

Ashburner, M., C. A. Ball, et al. (2000). "Gene Ontology: tool for the unification of biology." *Nature genetics 25*(1): 25.

Barrett, T., D. B. Troup, et al. (2006). "NCBI GEO: mining tens of millions of expression profiles—database and tools update." *Nucleic Acids Research 35*(suppl 1): D760.

Chang, D., Y. T. Syu, et al. (2010). "Predicting the protein-protein interactions using primary structures with predicted protein surface." *BMC Bioinformatics 11*(Suppl 1): S3.

Chang, D. T. H., Y. T. Syu, et al. "Predicting the protein-protein interactions using primary structures with predicted protein surface." *BMC bioinformatics 11*.

Cohen, A. M. and W. R. Hersh (2005). "A survey of current work in biomedical text mining." *Briefings in bioinformatics 6*(1): 57.

Davey, N. E., K. Van Roey, et al. (2011). "Attributes of short linear motifs." *Mol. BioSyst.*

Doerr, A. (2010). "The importance of being negative." *Nature Methods 7*(1): 10-11.

Donaldson, I., J. Martin, et al. (2003). "PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine." *BMC bioinformatics 4*(1): 11.

Enault, F., K. Suhre, et al. (2003). "Annotation of bacterial genomes using improved phylogenomic profiles." *Bioinformatics 19*(Suppl 1): i105.

Enault, F., K. Suhre, et al. (2003). "Annotation of bacterial genomes using improved phylogenomic profiles." *Bioinformatics 19*(suppl 1): i105.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." *Nature 402*(6757): 86-90.

Espadaler, J., O. Romero-Isart, et al. (2005). "Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships." *Bioinformatics 21*(16): 3360.

Guo, Y., L. Yu, et al. (2008). "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences." *Nucleic Acids Research 36*(9): 3025.

Huynen, M. A. and B. Snel (2000). "Gene and context: integrative approaches to genome analysis." *Advances in Protein Chemistry 54*: 345-379.

Jothi, R., P. F. Cherukuri, et al. (2006). "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions." *Journal of molecular biology 362*(4): 861-875.

Jothi, R., M. G. Kann, et al. (2005). "Predicting protein–protein interaction by searching evolutionary tree automorphism space." *Bioinformatics 21*(suppl 1): i241.

Kecman, V. (2001). *Learning and soft computing : support vector machines, neural networks, and fuzzy logic models*. Cambridge, Mass., MIT Press.

Krogan, N. J., G. Cagney, et al. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." *Nature 440*(7084): 637-643.

Lee, S. I. and S. Batzoglou (2003). "Application of independent component analysis to microarrays." *Genome Biology 4*(11): R76.

Li, H., J. Li, et al. (2006). "Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale." *Bioinformatics 22*(8): 989.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." *Science 285*(5428): 751.

Mishra, G. R., M. Suresh, et al. (2006). "Human protein reference database - 2006 update." *Nucleic Acids Research 34*: D411-D414.

Mitchell, T. M. (1997). *Machine learning*. New York, McGraw-Hill.

Oyang, Y. J., S. C. Hwang, et al. (2005). "Data classification with radial basis function networks based on a novel kernel density estimation algorithm." *IEEE Transactions on Neural Networks 16*(1): 225-236.

Peri, S., J. D. Navarro, et al. (2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans." *Genome research 13*(10): 2363.

Peri, S., J. D. Navarro, et al. (2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans." *Genome Research 13*(10): 2363-2371.

Pevzner, P. A. and S. H. Sze (2000). *Combinatorial approaches to finding subtle signals in DNA sequences*, Citeseer.

Rogozin, I. B., K. S. Makarova, et al. (2002). "Connected gene neighborhoods in prokaryotic genomes." *Nucleic Acids Research 30*(10): 2212.

Salgado, H., G. Moreno-Hagelsieb, et al. (2000). "Operons in Escherichia coli: genomic analyses and predictions." *Proceedings of the National Academy of Sciences 97*(12): 6652.

Salwinski, L., C. S. Miller, et al. (2004). "The database of interacting proteins: 2004 update." *Nucleic Acids Research 32*(suppl 1): D449.

Shen, J., J. Zhang, et al. (2007). "Predicting protein–protein interactions based only on sequences information." *Proceedings of the National Academy of Sciences 104*(11): 4337.

Sheu, S. H., D. R. Lancia, et al. (2005). "PRECISE: a database of predicted and consensus interaction sites in enzymes." *Nucleic Acids Research 33*(suppl 1): D206.

Shoemaker, B. A. and A. R. Panchenko (2007). "Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS computational biology 3*(4): e43.

Smialowski, P., P. Pagel, et al. (2010). "The Negatome database: a reference set of non-interacting protein pairs." *Nucleic acids research 38*(suppl 1): D540.

Snitkin, E., A. Gustafson, et al. (2006). "Comparative assessment of performance and genome dependence among phylogenetic profiling methods." *BMC bioinformatics 7*(1): 420.

Soong, T., K. O. Wrzeszczynski, et al. (2008). "Physical protein–protein interactions predicted from microarrays." *Bioinformatics* 24(22): 2608-2614.

Stumpf, M. P. H., T. Thorne, et al. (2008). "Estimating the size of the human interactome." *Proceedings of the National Academy of Sciences* 105(19): 6959.Sun, J., J. Xu, et al. (2005). "Refined phylogenetic profiles method for predicting protein–protein interactions." *Bioinformatics 21*(16): 3409.

Szklarczyk, D., A. Franceschini, et al. (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." *Nucleic Acids Research 39*(suppl 1): D561.

Tan, S. H., W. Hugo, et al. (2006). "A correlated motif approach for finding short linear motifs from protein interaction networks." *BMC bioinformatics 7*(1): 502.

Van Dijk, A., C. Ter Braak, et al. (2008). "Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control." *Bioinformatics 24*(1): 26.

Vapnik, V. and V. Vapnik (1998). *Statistical learning theory*, Wiley New York.

Walhout, A. J. M., R. Sordella, et al. (2000). "Protein interaction mapping in C. elegans using proteins involved in vulval development." *Science 287*(5450): 116.

Witten, I. H., E. Frank, et al. (2011). *Data mining : practical machine learning tools and techniques*. Burlington, MA, Morgan Kaufmann.

Yu, C. Y., L. C. Chou, et al. (2010). "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins." *BMC Bioinformatics 11*(1): 167.

Yu, J., M. Guo, et al. (2010). "Simple sequence-based kernels do not predict protein-protein interactions." *Bioinformatics 26*(20): 2610.

12

G-Protein Coupled Receptors: Experimental and Computational Approaches

Amirhossein Sakhteman, Hamid Nadri and Alireza Moradi
Faculty of Pharmacy, Shahid Sadoughi University of Medical Sciences, Yazd Iran

1. Introduction

Guanine Nucleotide binding protein coupled receptors (GPCRs) are among the most important targets in the treatment of cancer, endocrine, neural and many other types of disorders. (Katritch, V. & Abagyan, R. 2011) It is believed that activation of some GPCRs is involved in conditions such as immunosupression and response to ischemia of the brain and heart. Therefore, antagonists and agonists of GPCRs are potential therapeutic agents in treatment of inflammatory and ischemic diseases. (Moro, S., Spalluto, G., & Jacobson, K. A. 2005)

The superfamily of GPCRs consists of about 800 receptors which can be divided into different families regarding the similarities in the protein sequence. (Marshall, F. H. & Foord, S. M. 2010)

1. Family A, including rhodopsin and adrenoreceptor
2. Family B, Secretin vasointestinal peptide (VIP), the members of this family bind to hormones and neuropeptides
3. Family C, which include at least eight subtypes of glutamate receptors, the major excitatory receptor in the CNS.
4. Family D, the fungal pheromone p family
5. Family E, the fungal pheromone A family
6. Family F, CAMP receptors of *Dictyostelium discoideum*

The family A receptors is the best studied family of GPCRs in terms of functional and structural viewpoints and is therefore the most important target of GPCRs in drug discovery. (Moro, S. et al. 2005) It was reported that about 30% of the market prescription drugs act on these targets. (Marshall, F. H. et al. 2010)

2. Features and functions of GPCRs

A common feature in class A GPCRs is a core consisting of seven transmembrane domains (TM) connected by three intracellular loops (IL1, IL2 and IL3) and three extracellular loops (EL1, EL2, EL3).(Fig 1) Another feature observed in this class is the two cysteine residues, one in TM3 and the other in EL2. These two cysteines form a disulfide bridge which is responsible for the packing and stabilization of a restricted number of conformations for the seven TM domains (Fig 1).

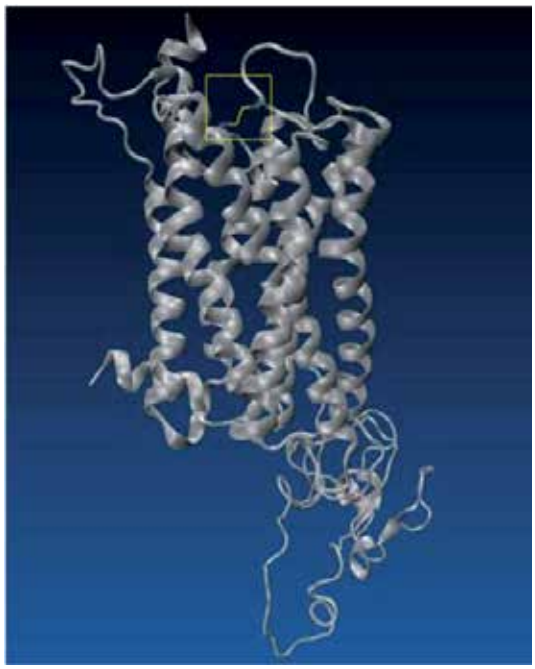


Fig. 1. The seven transmembrane structure of GPCRs. A disulfide bridge between TM3 and EL2 is conserved in most class A GPCRs.

GPCRs differ in the length and function of their N-terminal extracellular domains, C-terminal domain and intracellular loops. (Moro, S. et al. 2005) As an instance, glycoprotein hormone receptor (GPHR) family tether large amino terminal extracellular extensions which are responsible for the recognition and binding of dimeric agonists. Some studies in the area of GPCRs focused on interaction of ligands with GPCRs. There is strong evidence that in case of small molecules like the biogenic amines, the agonists of GPCRs interact directly with specific residues of TM helices of the receptor. On the other hand, for neuropeptides and small protein agonists like neurokinin the interactions involve both exoloops and amino portion of the receptor in association with the residues in TM helices. (Gilbert Vassart & Sabine Costagliola 2003)

2.1 Receptor activation in GPCRs

Many physiological procedures in the body are controlled by the GTPase including signal transduction, control of cellular growth, vesicle and protein transport and cytoskeletal assembly.(Smith, B., Hill, C., Godfrey, E. L., Rand, D., van den Berg, H., Thornton, S. et al. 2009). (Kobilka, B. K. 2007) Activation of a GPCR leads to nucleotide exchange on the Gα subunit and cause dissociation of the heterodimer and effector activation. GPCRs are mostly activated by diverse set of signals including small molecules, peptides and light. (Schneider, M., Wolf, S., Schlitter, J., & Gerwert, K. 2011) Members of GPCRs transduce signals by activation of at least one member of homologous hetrotrimeric G proteins. For example in FSH (Follicle Stimulating Hormone), receptor is activated by adrenaline which binds to the TM regions.

It should be noticed that in GPCRs, the ligand binds from the extracellular side and blocks the receptor. The activation or reduction in the basal activity of the heterotrimeric G-protein complex is dependent to the nature of the ligand such as agonists, antagonists and reverse agonists. The activation of G protein is in such a way that an exchange of guanosine diphosphate takes place in α subunit of G-protein. This exchange causes a conformational change in α subunit and leads to dissociation of α subunit from βγ subunits. (Fig 2) The two subunits introduce transduction systems in different ways. (Jaakola, V. P. & Ijzerman, A. P. 2010)

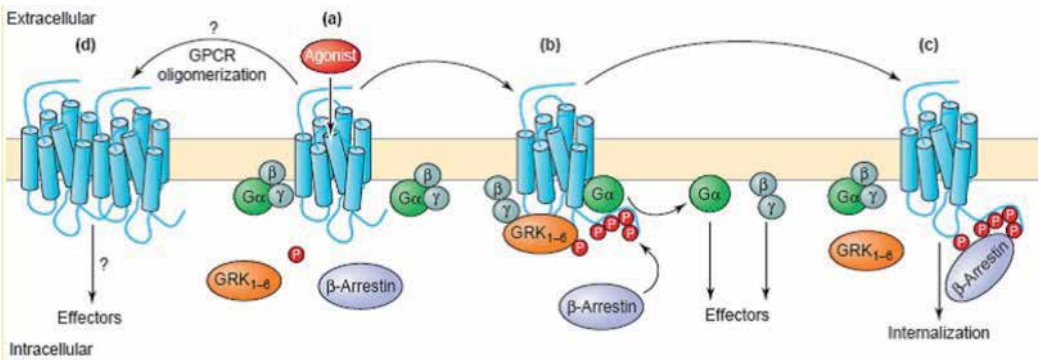


Fig. 2. a) Activation of GPCRs upon agonist binding b) conformational changes in hetrotrimeric G c) dissociation in G protein subunits proteins. d) A common feature observed in most GPCRs is the formation of dimers.

Some evidence verified that GPCRs exist as ensembles of conformations and two factors including binding of agonist and intracellular signaling proteins stabilize the active site of the receptor and accounts for the basal activity of GPCRs in the absence of agonists. The crystal structures of β2AR bound to an agonist and G-protein shows a large conformational change in the intracellular region of the receptor. It is likely that upon agonist binding reshuffling of short range intracellular contact cause large scale domain motions in the receptor. (Vaidehi, N. & Bhattacharya, S. 2011)

With the exception of rhodopsin, most family A GPCRs have considerable basal activity and both modeling and crystallographic data suggest that agonist dependent activation can vary between GPCRs. (Ahuja, S. & Smith, S. O. 2009;Deupi, X. & Kobilka, B. 2007;Katritch, V. et al. 2011) In contrast to rhodopsin where more information is present for the activation state of the molecule, in case of the other GPCRs much less is known about agonist induced conformational changes that occur during the activation of the receptor. (Wess, J., Han, S. J., Kim, S. K., Jacobson, K. A., & Li, J. H. 2008) The most similar parts for the GPCRs are the cytoplasmic ends of the TM segments adjacent to the second and third cytoplasmic domains which interact with G protein.

2.2 Dimerization of GPCRs

An important feature in GPCRs is formation of dimer which affects the receptors in terms of signal trafficking and pharmacology.(Marshall, F. H. et al. 2010). In many cases, the GPCR dimer can alter or regulate coupling or potency of other receptors. As an instance,

dimerization of κ-opioid receptor was shown to be in close relation with δ-opioid receptor dimerization. Another consequence of such dimerization is the augmented selectivity of some agonists such as 6-guanidinonal for the dimer with respect to any of the monomers. In addition, it was postulated that the binding of some drugs to more than one type of receptor is the result of dimerization. (Panetta, R. & Greenwood, M. T. 2008)

2.3 GPCRs and drug discovery

Most researches in the area of GPCRs focused on development of more selective or potent compounds of the orthostatic sites, which are apart from the binding sites of endogenous ligands. The allosteric modulators are also considered as promising therapeutic Agents. (Moro, S. et al. 2005)

While, the binding site for most small organic agonists is within TM segments, in case of peptide hormones and proteins the binding site is laid in the extracellular domain.(Kobilka, B. K. 2007) The intrinsic plasticity of GPCRs is a major problem in using their inactive state for agonist design in drug discovery. (Katritch, V. et al. 2011)

3. The role of experimental techniques in structural elucidation of GPCRs

The 3D structures of GPCRs have been identified using different techniques such as electron paramagnetic resonance spectroscopy (EPR), site directed mutagenesis, Fluorescence spectroscopy, cysteine cross-linking studies, Atomic Force Microscopy (AFM) and X-ray crystallography.

The first structure for GPCRs originated from cryoelectron microscopy of 2 dimensional crystals of bovine rhodopsin. Meanwhile, EPR has provided complementary evidence about photoactivation of rhodopsin including rotation and tilting of TM6 with respect to TM3.

It was clarified through electron paramagnetic resonance spectroscopy that photo activation of rhodopsin includes rotation and tilting of TM6 with respect to TM3.

By using some experimental techniques such as site directed mutagenesis data and cysteine scanning mutagenesis, it was possible to detect conformational changes in GPCRs.(Kobilka, B. K. 2007) As an instance, through site directed mutagenesis studies, it was proposed that the rotamer positions of the three residues including Cys 282, Trp 286 and Phe 290 of β2AR modulate the binding of TM6 around the highly conserved proline kink and lead to movement of cytoplasmic end of TM6 (Ahuja, S. et al. 2009;Deupi, X. et al. 2007) (Fig3).

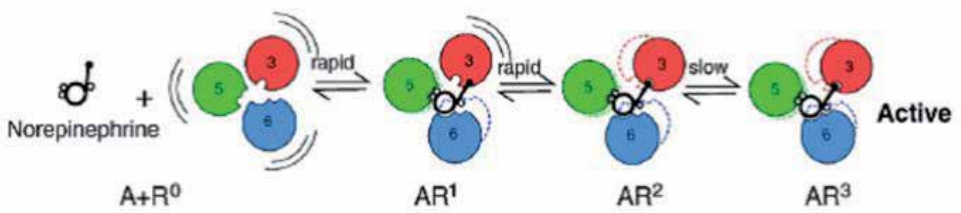


Fig. 3. Movement of TM6 (blue) in β2AR upon binding of an endogenous agonist (Norepinephrine) (Kobilka, B. K. 2007) with permission.

By using site directed mutagenesis data, binding sites of many receptors such as melatonin have also been discovered. (Panetta, R. et al. 2008)

Fluorescence spectroscopy and cross linking studies of some receptors such as muscarine (M3) has also revealed this fact that upon agonist binding, rotation or tilting of cytoplasmic end of transmembrane domains can take place in GPCRs. In cysteine cross-linking studies of M3 muscarinic receptor, it was suggested that the movements of TM5 and TM6 occurs upon agonist binding. (Kobilka, B. K. 2007) Through fluorescence spectroscopy and by tagging fluorophore to the intracellular end of TM6 in β2AR, the receptor revealed a single population of fluoroscence life time. Although, in the presence of antagonists the peak was reinforced, agonists showed an additional peak in the fluorescence life time. (Vaidehi, N. et al. 2011). Documents for the movements of TM6 in rhodopsin have been also provided by chemical reactivity measurements and fluorescence spectroscopy and Zinc cross linking studies of histidines. (Kobilka, B. K. 2007)

The most direct evidence for the structure of a GPCR oligomer comes from AFM (Atomic Force Microscopy). An advantage of AFM is that it provides a 3D profile of the protein. (Simpson, L. M., Taddese, B., Wall, I. D., & Reynolds, C. A. 2010)

Despite the great applicability of the described methods, X-ray crystal structures are normally the starting points in the studies of GPCRs. Since 2000 the high resolution X-ray structures of GPCRs began to emerge among which bovine rhodopsin was the first case study. (Topiol, S. & Sabio, M. 2009)

Crystallography of GPCRs has encountered some limitations due to low level of expression and instability of the proteins outside the membrane. Other problems are also attributed to the folding of the protein and homogeneity during purification. Another limitation in crystallography of the proteins is using detergents for dissociation of the protein from the membrane which might lead to some modifications in 3D structure of the protein. During the procedure for the preparation of crystal structures, a cost effective and straight forward to use system of protein expression is bacteria. The most important problem in this issue is the post translational limitations in the bacteria such as glycosylation which is required for the correct folding of GPCRs. To solve this problem many researchers have focused on using yeasts for the expression of GPCRs. The main problem with using yeasts is the differences observed in yeasts membranes in comparison with those observed in human. The most efficient system for the expression of the GPCRs includes the baculovirus expression system in the insects' cells. This method has been successfully adopted for the receptors including the β-adrenoceptors, the adenosine A2A receptor and the chemokine receptor CXCR4. (Congreve, M., Langmead, C., & Marshall, F. H. 2011)

In crystallography of GPCRs a solution for obtaining the crystal structure in single conformation is addition of the ligand with the prefential affinity to one conformation of the receptor. (Congreve, M. et al. 2011)

Although by X-ray crystallographic data, it is possible to obtain static of the protein complex at near atomistic resolutions biophysical and structural studies are still needed to complete the data of crystallography. (Jaakola, V. P. et al. 2010)

4. Homology modeling studies of GPCRs

Regarding the limitations for using X-ray crystallography in case of GPCRs, molecular modeling techniques such as homology modeling and docking studies are needed to fill the gaps between the primary sequence and secondary structures for drug design studies. There are two major types of drug design including ligand based drug design and structure based drug design. In structure based drug design, the 3D structure of the target molecules is necessary for the study. (Ostopovici-Halip, L., Curpan, R., Mracec, M., & Bologa, C. G. 2011) An important goal of molecular modeling is to provide a microscopic details of the membrane proteins where there is no way to obtain enough information through experimental approaches. (Henin, J., Maigret, B., Tarek, M., Escrieut, C., Fourmy, D., & Chipot, C. 2006)

The crystal structures for seven GPCRs have been represented so far. Most GPCRs are representing the inactive state of the receptor and are therefore suitable for the discovery of antagonists and reverse agonists. (Schneider, M. et al. 2011)

Homology modeling is a knowledge based approach relying upon the crystallographic structure of related receptor and experimental information. This method has limitations when targeting active receptors. (Henin, J. et al. 2006)

Homology modeling made it possible to align the protein of interest with the homologous structure and to subsequently evaluate the model with different scoring functions. Since the most important part of a homology modeling study is the alignment procedure, errors in predicting the structure protein with low homology are normally common. The alignment procedure is usually done by different methods such as cluatalW or T-cofee servers. In order to refine homology based models energy minimization or limited conformational sampling using molecular dynamic simulations are used. (Harterich, S., Koschatzky, S., Einsiedel, J., & Gmeiner, P. 2008)

Due to diversity in loop area of the proteins, a loop refinement is usually necessary in most homology modeling studies. The available loop modeling algorithms are limited to up to 13 residues long. Therefore in loop refinement step considerations must be taken with loops of the long size. (Ostopovici-Halip, L. et al. 2011)

By the emergence of 3D structure for bovine rhodopsin, this receptor was widely used as template structure homology modeling studies of GPCRs. For example, 3D structure of neurotensin, a neuropeptide distributed in the CNS was modeled based on Rhodopsin as template.(Harterich, S. et al. 2008) Two other GPCRs namely, the principal canabinoid receptors CB1 and CB2 are the components of endogenous endocanabinoid systems. The 3D structures of CB1 and CB2 have also been modeled using Rhodopsin as template. (Pei, Y., Mercier, R. W., Anday, J. K., Thakur, G. A., Zvonok, A. M., Hurst, D. et al. 2008)

Another GPCR which has been characterized by homology modeling techniques was melatonin receptor. This receptor is responsible for the effects of melatonin, a compound taking part in resynchronization of biological rhythms such as sleep. In case of MT1 and MT2 receptors, the helices of the receptor were supposed to be superposable with the experimentally known helices of bovine rhodopsin. It was also reported that the identity of MT1 towards rhodopsin is more in respect to MT2 (23% vs 19%).

In other studies the active state of opsin has been used as template to model active structures of β2-adrenergic receptor. Interaction fingerprint studies have been used for dynamic ligand binding study of the interaction of ligands in the active and inactive states. It was concluded that the active structure of opsin is suitable for modeling GPCR agonists. (Schneider, M. et al. 2011)

In spite of the many reports for the usefulness of rhodopsin in homology modeling studies, rhodopsin is merely suitable for antagonist design. The reason can be explained by the fact that rhodopsin is merely crystallized in its inactive state. Therefore, in order to design agonists of GPCRs, it is important to obtain information about the active states of the receptor.(Topiol, S. et al. 2009). There are some other limitations for using rhodopsin as template in homology modeling of GPCRs. One is that it has low homology (less than 25%) with family A GPCRs and no homology with other families of GPCRs such as secretin, adhesion and metabotropic receptors. The other limitation is the very complicated mechanism of activation in rhodopsin in comparison with the other GPCRs. The binding of the ligand to rhodopsin is covalent and signaling is conducted through activation of the ligand by photoisomerism. (Congreve, M. et al. 2011)

Another problem with using rhodopsin is that the binding domains are arranged clockwise in this receptor while sequentially oriented anticlockwise in case of others.(Claude Nofre 2001) Therefore, it is very difficult to obtain a reasonable overview for the activation of the GPCRs from rhodopsin antagonist binding.(Congreve, M. et al. 2011) Based on the findings it was claimed that rhodopsin might not be a suitable template for some GPCRs such as the cholecystokinin CCK1 receptor. (Kobilka, B. K. 2007) A revolution in the field of GPCRs has occurred after publication of the crystal structure of β1 and β2- adrenoceptor.(Congreve, M. et al. 2011) Afterwards, the crystal structures of adenosine A_2A (A_2AR), Chemokine CXCR4, dopamine D3 and histamine H1 in complex with antagonists have been reported which made a reasonable framework for the studies of GPCR functions and drug discovery. (Katritch, V. et al. 2011)

The β2AR is almost a good model for the studies of agonist binding since much information is obtained about the site of interaction between the receptor and catechol amine ligands. (Kobilka, B. K. 2007)

The crystal structures of ß1 and ß2 have been used for homology modeling of 5-HT_2C receptor. They showed similar homology rates of 41% and 62% with the regions of 5-HT_2C. (Renault, N., Gohier, A., Chavatte, P., & Farce, A. 2010)

Another successful example of homology modeling studies using β2-AR as template was in Alpha 2 adrenoreceptor (α_2ARs) receptors. These receptors with wide distributions are responsible for many activities such as the control of nervous system and cardiovascular systems. In this study, the resulted models have been minimized using the OPLS2005 force field implemented in schrodinger package. (Ostopovici-Halip, L. et al. 2011)

In many cases of homology modeling, the validity of the structures was verified by ramachandran plot. A common method for docking the homology based models is WHATIF algorithm which generates ramachandran plots to identify outliers in terms of torsion angles and also compares the quality of the model with reliable structures presented in the form of Z-scores. It is also possible to get consensus votes through WHATIF to select

between homologous structures. (Abu-Hammad, A., Zalloum, W. A., Zalloum, H., Abu-Sheikha, G., & Taha, M. O. 2009)

Another tool to assess the structural validity of the models is to use hydrophobic moments of the helices. By this method, it is possible to obtain the orientation of hydrophobic moment in transmembrane domains. (Panetta, R. et al. 2008)

5. Simulation studies of GPCRs

In modeling studies for keeping the receptors electrochemically sealed the interaction of the lipids and proteins are needed. (Escriba, P. V., Wedegaertner, P. B., Goni, F. M., & Vogler, O. 2007)

Different molecular dynamic (MD) simulation methods have been used to study GPCRs. While all atom MD simulations in lipid bilayer and water is used to study the dynamics of the membrane proteins, By using targeted MD simulation such as metadynamics, it was possible to study the process of activation in the receptors. In metadynamics a Gaussian term is added to the free energy which disallows the system from returning to previous state. A pitfall of this method is the bias used for forcing the system change its state. The requirement of this method is the primary knowledge needed about the active and inactive states of the receptor. By using this method it is possible to obtain the intermediates in the activation process of the receptor. (Vaidehi, N. et al. 2011)

Since the ligand induced conformational change in GPCRs happen in the range of microseconds, all atom MD simulations are not able to predict large scale simulations such as conformational change in GPCRs. (Vaidehi, N. et al. 2011)

Another simulation method is elastic network model (ENM) in which the protein is represented as a collection of beads connected by springs, where beads refer to protein residues and springs refer to connections. By using this method it is possible to study the micro second simulations. (Vaidehi, N. et al. 2011)

LITicon is a method in which the receptor conformations are permitted to have coarse grain degree of freedom to avoid the built in bias observed in targeted MD simulations. In this method the TMs are considered as rigid bodies connected to each other by flexible loops. The TM helices are rotated in a desired range of rotation angles and the side chain conformations are optimized for each backbone conformation using a rotamer library. Subsequently, the potential energy is minimized using all atom force field function. By this method it is possible to obtain an energy landscape for the GPCRs in the rotational span of the TM helices. After identifying the local minima in the landscape, the global minima state of energy landscape is chosen on the most stable state of the protein. In LITicon, the coarse grain simulation is used to forecast the ensemble of active and inactive states from the inactive crystal structure of the protein. A problem with coarse grain method is some significant barriers which might be missing during the activation pathway.

Monte Carlo (MC) simulations have been also used to calculate the pathway for the activation of some receptors. By MC simulation, it was possible to search the minimum energy from the inactive state towards the ligand stabilized states. An important note to be considered in the computational studies of GPCRs, is the role of water molecules in the

activation procedure which has been proposed by many researches to take role in conformational changes of GPCRs. It must be denoted that some water molecules in the crystal structure of the GPCRs might be either absent or not well resolved. It is known that multiscale methods with a combination of coarse grain and fine grain all atom methods are required for understanding the conformational changes of GPCRs. (Vaidehi, N. et al. 2011)

Although some studies have reported the usefulness of MD simulations in studying the changes during dimerization, it was normally difficult to study the GPCRs in dimer form by molecular dynamic simulation methods. (Simpson, L. M. et al. 2010) Recent developments in protein-protein docking made it possible to perform studies on dimer formation.

As an instance, the 5HT4 receptor was subjected to docking approach using GRAMM wherein the interface for dimerization was TM2.4- TM2.4. (Simpson, L. M. et al. 2010)

Docking simulation studies have also been taken to predict the binding mode in GPCRs and estimate the ligand- receptor affinities in case of many receptors such as cholecystokinin (CCK). (Henin, J. et al. 2006)

A further step in modeling based discovery of drugs for GPCRs is to identify potential binding sites in the receptors. The binding site for some GPCRs such as human sweet receptor (HSR) was modeled using ligand based approach. (Claude Nofre 2001)

A successful example for using computational methods was in melanin concentrating hormone (MCH1R). The 3D structure of melanin concentrating hormone which belongs to rhodopsin superfamily was predicted using homology based modeling studies. During the procedure the models were built by the web based model suit SWISS MODEL and scanned for the ligands binding site. The model was then subjected to docking studies of ligands with known activities. The result of the docking step was used for making comparative molecular force field analysis. The combinations of docking/scoring/COMFA were previously reported to be successful in predicting docked conformer/pose closed to that of cocrystalized ligand. In these types of studies, the validity of COMFA models can be verified using ligand based approaches. (Abu-Hammad, A., Zalloum, W. A., Zalloum, H., Abu-Sheikha, G., & Taha, M. O. 2009)

6. Conclusion

Different experimental and computational approaches proposed the role of molecular switches on structural and conformational changes of GPCRs.

By using experimental techniques such as site directed spin labeling, it was observed that in case of receptors such as rhodopsin, a well conserved salt bridge between TM3 and TM6 known as ionic lock is broken during activation. This cleavage leads to flexibility of TM 6 and its movement towards TM3.

Based on molecular modeling studies, it was suggested that in case of some receptors such as MCH, the binding site is a cleft inside the helical domain of the receptor including three hydrophobic regions and a hydrogen bonding polar region.(Abu-Hammad, A. et al. 2009) Other studies revealed that polar interactions of serines with agonists and the movement of TM5 in B_2AR pocket is resulted by shift of TM7 towards TM3 upon agonist binding. An

optimal confrontation for this was made based on virtual ligand screening of known ligands. (Katritch, V. et al. 2011)

Biochemical and mutagenesis of B2AR established two major interactions for full agonists in which the amine group forms a salt bridge in Aspll3 while the hydrogen groups of catechol interact with serine in TM5. Analysis of B2AR demonstrated that an inward shift (~2 A) of TM5 is needed for binding of full agonists. The same results have been observed in induced fit docking studies with flexible TM helices. The TM5 shift was caused by conformational freedom in this domain and strong H-binding between catechol OH and Ser 207. The modeling studies based on Adenosine A2A receptor was another example for modeling of agonists. In this case some interactions were revealed to be common for agonists and antagonists such as aromatic ring and amine core contacts. It was seen that adjustment of ligand in an optimal position and engagement of all polar interactions is needed for the shift of the conserved Trp 6.48. (Katritch, V. et al. 2011)

In case of M_3 receptor it was predicted that the binding of Ach to M_3 triggers conformational changes within the TM receptor core. Agonist binding causes the disruption of the existing interhelical interactions and promotes a set of interactions that leads to a new favorable conformational state for the receptor. (Wess, J. et al. 2008)

An important molecular switch is the ionic lock bottom highly conserved D/E motif found in all class A GPCRs. This ionic interaction holds together the cytoplasmic ends of TM3 and TM6 in many amine receptors. Another example for the role of ionic lock is in Angiotensin 1 receptors. The evidence shows that Asn111 interacts with Asn295 in TM7 to stabilize the inactive state of the receptor. (Ahuja, S. et al. 2009;Deupi, X. et al. 2007) In another study it was postulated that reduction of conserved disulfide bridge might be a molecular switch for the activation of the receptor. This study was based on molecular dynamic simulation and virtual screening of dopamine D_2 receptor. It was observed that a predictive model for the catechol binding cavity of D_2 had reduced disulfide bridge. The movement of TM6 towards TM5 was supposed to be the result of cleavage in the conserved disulfide bridge (Fig 4) (Sakhteman, A., Lahtela-Kakkonen, M., & Poso, A. 2011)

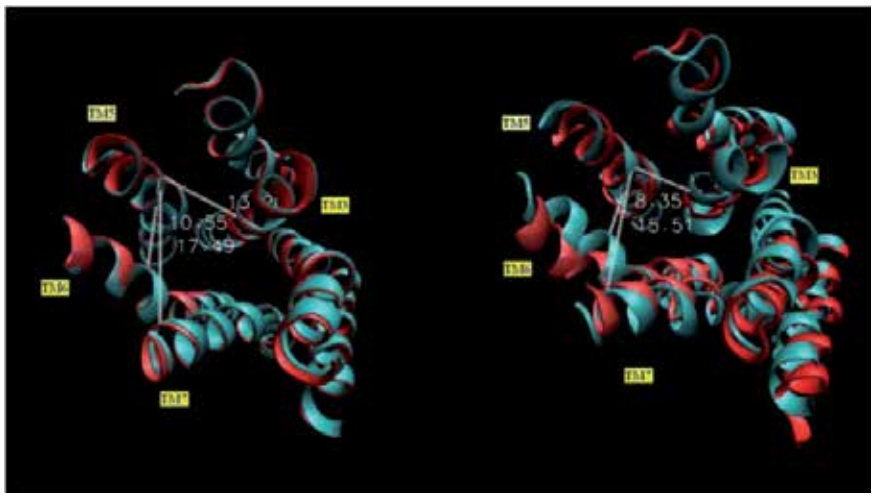


Fig. 4. Movement of TM6 towards TM5 in D_2 model with reduced disulfide bridge.

7. References

Abu-Hammad, A., Zalloum, W. A., Zalloum, H., Abu-Sheikha, G., & Taha, M. O. (2009). Homology modeling of MCH1 receptor and validation by docking/scoring and protein-aligned CoMFA. *Eur.J.Med.Chem.,* Vol.44, No.6, pp. (2583-2596)

Abu-Hammad, A., Zalloum, W. A., Zalloum, H., Abu-Sheikha, G., & Taha, M. O. (2009). Homology modeling of MCH1 receptor and validation by docking/scoring and protein-aligned CoMFA. *Eur.J.Med.Chem.,* Vol.44, No.6, pp. (2583-2596)

Ahuja, S. & Smith, S. O. (2009). Multiple switches in G protein-coupled receptor activation. *Trends Pharmacol.Sci.,* Vol.30, No.9, pp. (494-502)

Claude Nofre. (2001). New hypotheses for the GPCR 3D arrangement based on a molecular model of the human sweet-taste receptor. *Eur.J.Med.Chem.,* Vol.36, pp. (101-108)

Congreve, M., Langmead, C., & Marshall, F. H. (2011). The use of GPCR structures in drug design. *Adv.Pharmacol.,* Vol.62, pp. (1-36)

Deupi, X. & Kobilka, B. (2007). Activation of G protein-coupled receptors. *Adv.Protein Chem.,* Vol.74, pp. (137-166)

Escriba, P. V., Wedegaertner, P. B., Goni, F. M., & Vogler, O. (2007). Lipid-protein interactions in GPCR-associated signaling. *Biochim.Biophys.Acta,* Vol.1768, No.4, pp. (836-852)

Gilbert Vassart & Sabine Costagliola. (2003). The thyrotropin receptor, a GPCR with a built-in inverse agonist. *International Congress Series,* Vol.1249,

Harterich, S., Koschatzky, S., Einsiedel, J., & Gmeiner, P. (2008). Novel insights into GPCR-peptide interactions: mutations in extracellular loop 1, ligand backbone methylations and molecular modeling of neurotensin receptor 1. *Bioorg.Med.Chem.,* Vol.16, No.20, pp. (9359-9368)

Henin, J., Maigret, B., Tarek, M., Escrieut, C., Fourmy, D., & Chipot, C. (2006). Probing a model of a GPCR/ligand complex in an explicit membrane environment: the human cholecystokinin-1 receptor. *Biophys.J.,* Vol.90, No.4, pp. (1232-1240)

Jaakola, V. P. & Ijzerman, A. P. (2010). The crystallographic structure of the human adenosine A2A receptor in a high-affinity antagonist-bound state: implications for GPCR drug screening and design. *Curr.Opin.Struct.Biol.,* Vol.20, No.4, pp. (401-414)

Katritch, V. & Abagyan, R. (2011). GPCR agonist binding revealed by modeling and crystallography. *Trends Pharmacol.Sci.,* Vol.32, No.11, pp. (637-643)

Kobilka, B. K. (2007). G protein coupled receptor structure and activation. *Biochim.Biophys.Acta,* Vol.1768, No.4, pp. (794-807)

Kobilka, B. K. (2007). G protein coupled receptor structure and activation. *Biochim.Biophys.Acta,* Vol.1768, No.4, pp. (794-807)

Marshall, F. H. & Foord, S. M. (2010). Heterodimerization of the GABAB receptor-implications for GPCR signaling and drug discovery. *Adv.Pharmacol.,* Vol.58, pp. (63-91)

Moro, S., Spalluto, G., & Jacobson, K. A. (2005). Techniques: Recent developments in computer-aided engineering of GPCR ligands using the human adenosine A3 receptor as an example. *Trends Pharmacol.Sci.,* Vol.26, No.1, pp. (44-51)

Ostopovici-Halip, L., Curpan, R., Mracec, M., & Bologa, C. G. (2011). Structural determinants of the alpha2 adrenoceptor subtype selectivity. *J.Mol.Graph.Model.,* Vol.29, No.8, pp. (1030-1038)

Panetta, R. & Greenwood, M. T. (2008). Physiological relevance of GPCR oligomerization and its impact on drug discovery. *Drug Discov.Today,* Vol.13, No.23-24, pp. (1059-1066)

Pei, Y., Mercier, R. W., Anday, J. K., Thakur, G. A., Zvonok, A. M., Hurst, D., Reggio, P. H., Janero, D. R., & Makriyannis, A. (2008). Ligand-binding architecture of human CB2 cannabinoid receptor: evidence for receptor subtype-specific binding motif and modeling GPCR activation. *Chem.Biol.,* Vol.15, No.11, pp. (1207-1219)

Renault, N., Gohier, A., Chavatte, P., & Farce, A. (2010). Novel structural insights for drug design of selective 5-HT(2C) inverse agonists from a ligand-biased receptor model. *Eur.J.Med.Chem.,* Vol.45, No.11, pp. (5086-5099)

Sakhteman, A., Lahtela-Kakkonen, M., & Poso, A. (2011). Studying the catechol binding cavity in comparative models of human dopamine D2 receptor. *J Mol Graph.Model.,* Vol.29, No.5, pp. (685-692)

Schneider, M., Wolf, S., Schlitter, J., & Gerwert, K. (2011). The structure of active opsin as a basis for identification of GPCR agonists by dynamic homology modelling and virtual screening assays. *FEBS Lett.,* Vol.585, No.22, pp. (3587-3592)

Simpson, L. M., Taddese, B., Wall, I. D., & Reynolds, C. A. (2010). Bioinformatics and molecular modelling approaches to GPCR oligomerization. *Curr.Opin.Pharmacol.,* Vol.10, No.1, pp. (30-37)

Smith, B., Hill, C., Godfrey, E. L., Rand, D., van den Berg, H., Thornton, S., Hodgkin, M., Davey, J., & Ladds, G. (2009). Dual positive and negative regulation of GPCR signaling by GTP hydrolysis. *Cell Signal.,* Vol.21, No.7, pp. (1151-1160)

Topiol, S. & Sabio, M. (2009). X-ray structure breakthroughs in the GPCR transmembrane region. *Biochem.Pharmacol.,* Vol.78, No.1, pp. (11-20)

Vaidehi, N. & Bhattacharya, S. (2011). Multiscale computational methods for mapping conformational ensembles of G-protein-coupled receptors. *Adv.Protein Chem.Struct.Biol.,* Vol.85, pp. (253-280)

Wess, J., Han, S. J., Kim, S. K., Jacobson, K. A., & Li, J. H. (2008). Conformational changes involved in G-protein-coupled-receptor activation. *Trends Pharmacol.Sci.,* Vol.29, No.12, pp. (616-625)

13

Computational Approaches to Elucidating Transient Protein-Protein Interactions, Predicting Receptor-Ligand Pairings

Ernesto Iacucci[1], Samuel Xavier de Souza[2] and Yves Moreau[1]
[1]*K.U.Leuven*
[2]*Universidade Federal do Rio Grande do Norte*
[1]*Belgium*
[2]*Brazil*

1. Introduction

Protein-protein interactions (PPI) are one of the most important biological events which occur in the cell. As PPIs regulate almost all biological processes in the cell, aberrations in PPI may cause severe health problems. One specific area of PPI is receptor-ligand interactions. These interactions are transient yet account for a large part of cell-to-cell communication. As PPI is an important area of research, many groups have proposed methods to make computational predictions of PPI.

The basis of the majority of these methods rely largely on the phylogenetic profile analysis of candidate interactors. These methods determine the similarity of the phylogenetic history of a protein *A* and its putative protein partner *B*, examining the most accurate measure of similarity between the phylogenetic histories of *A* and *B* in order to predict interaction. As interacting proteins should co-adapt as they are under the same evolutionary pressures, it is self-evident that interacting receptors and ligands should be identifiable by application of the same methodology.

While several methods, described below, make use of phylogenetic information to predict protein-protein interaction (PPI), more contemporary work has been conducted in the area of data fusion and kernel learning. We describe one method [Iacucci et al. 2011] in detail which does both. In this work, the existing line of phylogenetic research is extended by using phylogenetic data to construct a kernel to train a least square support vector machines (LS-SVM) in order to classify candidate receptors and ligands as *interacting* or *non-interacting*.

In this chapter, we discuss the plethora of various methods for determining protein-protein interactions. In addition, we evaluate the application of LS-SVMs to the sub-problem of receptor-ligand interaction prediction.

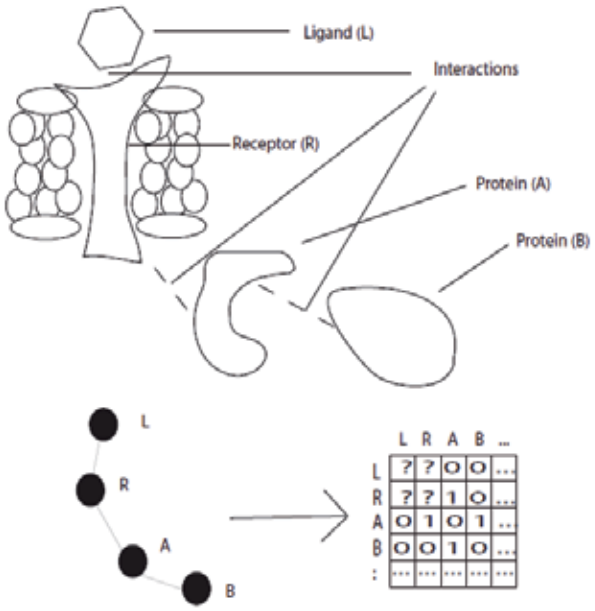


Fig. 1. The Receptor Ligand Schematic. Schematic of receptor-ligand and protein-protein interaction model. Top image is a representation of in-vivo interaction of proteins, receptors, and ligands while bottom image is the graph representation from which a PPI adjacency matrix may be derived. (Figure published in Iacucci *et al*. 2010)

2. Current computational approaches for predicting protein-protein interaction

During the past decade, many methods for prediction of interaction between proteins have been studied due to the crucial role that these interactions have in the understanding of the diverse cellular mechanisms of life forms. Many of these methods involve experimental analysis of specific protein pairs in a smaller scale or, in current high throughput methods [Uetz et al. 2000, Giot et al. 2003], a large amount of protein interactions. The later can be used to detect many interactions with reasonable sensitivity but rather low specificity. Another, relatively inexpensive, way to predict protein-protein interactions does not include wet lab analysis, using instead a variety of computational approaches. These approaches can complement experimental wet lab techniques and are often supported by either the hypothesis of protein co-evolution [Tan et al. 2004, Tillier et al. 2006, Izarzugaza et al. 2006], structural similarities [Gong et al. 2005, Ogmen et al. 2005] or amino-acids sequence conservation [Pitre et al. 2006].

While the entire genomes of many species are already completely sequenced, the interactone of these life forms is often many orders of magnitude larger and yet far from being fully mapped [Claverie et al. 2001, Rubin et al. 2001]. High throughput experimental techniques will certainly help to create this mapping and computational approaches can complement their results identifying false positive interactions, and therefore improving the specificity of these experimental techniques. Apart from the experimental techniques, computational methods are themselves a powerful and affordable alternative to contribute to interactome mapping.

Several computational approaches have been developed in recent years. Many of them are freely available as web tools offering a variety of services to biologists and bioinformatics that range from prediction of interactions between of proteins in pairs or in batch mode, through browsing of consolidated large scale analysis, up to visualization of binding sites and physical interactions in 3-dimensional images.

The methodologies of these many different approaches vary, but they all seem to be supported by the following findings: (a) evidences in favor of the hypothesis of protein co-evolution and the similarities observed in the phylogenetic trees of these proteins; and (b) datasets of already known protein-protein interactions verified by experimental techniques. Co-evolutionary methods find protein pairs with the highest co-evolutionary signal. This information is powerful to predict which members of interacting protein families are associated structurally or functionally although it is not specific enough to predict whether or not two protein families interact. On the other hand, methods supported by verified protein-protein interactions make use of the structural or amino-acid sequence similarities of interacting proteins partners to predict interaction between query protein pairs. This makes such methods more suitable to predict physical interactions rather than functional relationships.

We have reviewed 6 methods and their web tools for predicting protein-protein interactions. Three of them, supported by the protein co-evolution hypothesis, are: TSEMA [Izarzugaza et al. 2006], ADVICE [Tan et al. 2004], Codep [Tillier et al. 2006]. The other three, supported by datasets of verified interactions, are: PIPE [Pitre et al. 2006], PSIbase [Gong et al. 2005], and PRISM [Ogmen et al. 2005]. In the next Sections, we describe each one of these two types of methods.

2.1 Current co-evolutionary methods

Many studies of the problem of predicting protein-protein interactions investigate the similarity of the phylogenetic history of the interaction partners. Many examples of interaction between proteins have presented signs of co-evolution in such a way that members of different interacting protein families present similarity between their phylogenetic trees [Fryxell 1996, Goh et al. 2000, van Kesteren et al. 1996, Moyle et al. 1994, Pazos and Valencia 2001]. The core of co-evolutionary methods is based on measures of similarity for the phylogenetic trees of interacting protein partners.

There are several measures for similarity between phylogenetic trees. The trees can be compared directly [Goh et al. 2000], via distance matrices [Moyle et al. 1994, Goh and Cohen 2002, Ramani and Marcotte 2003, Gertz et al. 2003], or using multiple sequence alignments [Tillier et al. 2006]. In the following Sections, we present three co-evolutionary methods: TSEMA and ADVICE, which uses distance to compare the phylogeneitc trees, and Codep, which computes the correlation between co-evolving partners from their multiple sequence alignments.

2.1.1 Interactive prediction of protein pairing between interacting families TSEMA

TSEMA is a method and web tool to predict mappings between two families of homologous proteins. The probed protein families can either be inputted using the Newick format or in a format comparable with ClustalW, which is used to build the trees. The distances for all

pairs of proteins within both families are extracted from their phylogenetic trees by summing the length of the branches separating each pair of proteins in the trees. The algorithm of TSEMA finds the mapping between the two sets proteins which maximizes the matching between the sets of distances using a modified implementation of the Ramani and Marcotte's Monte Carlo Metropolis method [Ramani and Marcotte 2003].

Availability: http://tsema.bioinfo.cnio.es/

2.1.2 Automated Detection and Validation of Interaction by Co-Evolution – ADVICE

ADVICE predicts and validate protein-protein interactions using observed co-evolution between proteins. The web tool retrieves orthologous sequences of a list of input protein sequences and compute the similarities among the proteins evolutionary histories. The tool also provides visualization for the resulting network of co-evolved proteins.

The ADVICE algorithm infers interaction based on the correlation between distance matrices constructed from the evolutionary history using orthologous sequences of top 10 species. The tool uses BLAST [Altschul et al. 1990] to search the orthologous sequences from Swiss-Prot and TrEMBL databases [Boeckmann 2003]. The distance matrices are constructed using only pairs of orthologous sequences occurring together in the same species. By default, only the orthologous sequences of the top 10 species, based on the BLAST E-value, are used to construct the matrices, excluding those species where more than one orthologous sequence of the input sequence is found. The actual distance matrices are build from the respective multiple sequence alignments using ClustalW [Thompson et al. 1994]. The algorithm then calculates the correlation between pairs of matrices measuring the Pearson's correlation coefficients, which has values between -1, implying 100% anti-correlation, and 1,which representing 100% evolutionary history similarity, being velues above 0.8 good indicators of interaction and values below 0.3 a good cut-off value to detect potential spurious interaction.

Availability: http://advice.i2r.a-star.edu.sg

2.1.3 Maximizing co-evolutionary interdependencies to discover interacting proteins – Codep

Codep and the other co-evolutionary methods find proteins with the highest co-evolutionary signals, independent of physical or functional interaction. The main difference of Codep is that it uses multiple sequence alignments directly rather than distances obtained from the sequences. The user inputs two phylogenetic trees with orthologous sequences. The algorithm maximizes interdependency based on the maximal mutual information. It does this by fixing one of the multiple sequence alignments and varying the order of the other via exhaustive search or via simulated annealing.

The rationale to use directly multiple sequence alignments instead of the distance matrices, which provides a faster way to calculate correlation, is that character-state methods in the field of phylogenetic analysis are more powerful than distance method and some information can be lost in transforming character-state data into distance matrices.

Availability: http://www.uhnresearch.ca/labs/tillier/

2.2 Methods based on verified interactions

Another promising computational approach to predict new protein-protein interactions is to look at the physical structure and the conservation of amino-acid sequences in partners of interactions that are already reliably known to exist. Then, use the gathered information to find correlation with query protein partners of a probed interaction. Many methods apply this approach, which have delivered powerful tools for finding new interactions [Pitre et al. 2006] and even to corroborate with the protein co-evolution hypothesis [Kim et al. 2004]. In the next three Sections we describe three of these methods: PIPE, which compares amino-acid subsequences between probed protein partners and partners of verified protein interactions from a database; and PSIbase and PRISM, both which compare structural characteristics of probed and verified interactions.

2.2.1 Protein-Protein Interaction Prediction Engine – PIPE

PIPE is a computational tool that can effectively identify protein-protein interactions among *S. cerevisiae* protein pairs. It relies on previously determined *S. cerevisiae* protein interactions compiled from the DIP [Salwinski et al. 2004] and MIPS [Mewes et al. 2002] databases to construct a graph where the nodes are proteins and the edges represent the relationship of interacting proteins.

The working principle of the PIPE algorithm to probe interaction between the pair of proteins A-B is to compare sliding subsequences of amino-acids of size w from A to subsequences of the same size of all proteins in the graph of known interactions; then compare sliding subsequences of B to the neighbors of all matches of A. If protein pair C-D are connected in the graph, representing a verified interaction, and if A has subsequence matches with C and B has matches with D, then the pair A-B is more likely to present interaction. The accumulation of all matches of subsequence comparisons presented in form of a matrix indicates a predicted interaction when the higher values in this matrix is above a given threshold of M matches.

The algorithm has three tuning parameters: w, M, and S_{PAM}, which is the threshold value that indicates a match between two subsequences of amino-acids. The author of PIPE chose to fix w in 20, and tune the other two parameter either by trial and error or by statistical evaluation.

PIPE is reported to have success rate comparable to biochemical techniques, with a sensitivity of 61% , specificity of 89%, and overall accuracy of 75%. The main disadvantages of PIPE is its heavy computational burden and its limitation to yeast proteins.

Availability: http://pipe.cgmlab.org

2.2.2 Protein Structural Interactome Map – PSIMAP

PSIMAP is a map that describes the information about domain-domain and protein-protein interactions known to exist in the Protein Data Bank of structures. It is based on the principle that interaction between protein structures is conserved as closely as protein structures themselves [Park et al. 2001, Aloy and Rossell, 2002; Aloy et al. 2003]. It that predicts if domains or proteins structures interact calculating if every possible pair of

structures has an Euclidean distance below a certain threshold. There are three different methods to do this: Full Atom Contact (FAC); Sample Atom Contact (SAC); and Bounding Box Contact (BBC). FAC is the most accurate, whereas SAC and BBC [Dafas et al. 2004] are faster methods.

PSIMAP extract the molecular interaction information of proteins from the PDB. It associates this information to domains using the Structural Classification of Proteins (SCOP) to assign the domains to the structures.

Availability: http://psimap.org and http://psibase.kaist.ac.kr/

2.2.3 Protein Interactions by Structural Matching – PRISM

The PRISM tool allows the user to explore protein interfaces and predict protein-protein interactions by comparing the structure of query proteins to those of a structurally and evolutionarily subset of biological and crystal interactions present in the Protein Data Bank (PDB) [Berman 2000]. Interfaces are defined as the set of residues forming the region of the structure through which two different protein chains bind to each other. This set consists the contacting residues between the chains and the neighboring residues up to a certain distance threshold.

The interfaces in PRISM were obtained from all higher complexes of proteins available in the PDB [Keskin et al. 2004]. From the 49512 interfaces extracted form the PDB, 8205 clusters were obtained using a sequence order-independent computer vision-based algorithm to structurally compare the interfaces. From these 8205 clusters, PRISM considers only 158 template interfaces (Oct/2011) that were found to have evolutionary hotspots [Keskin et al. 2005].

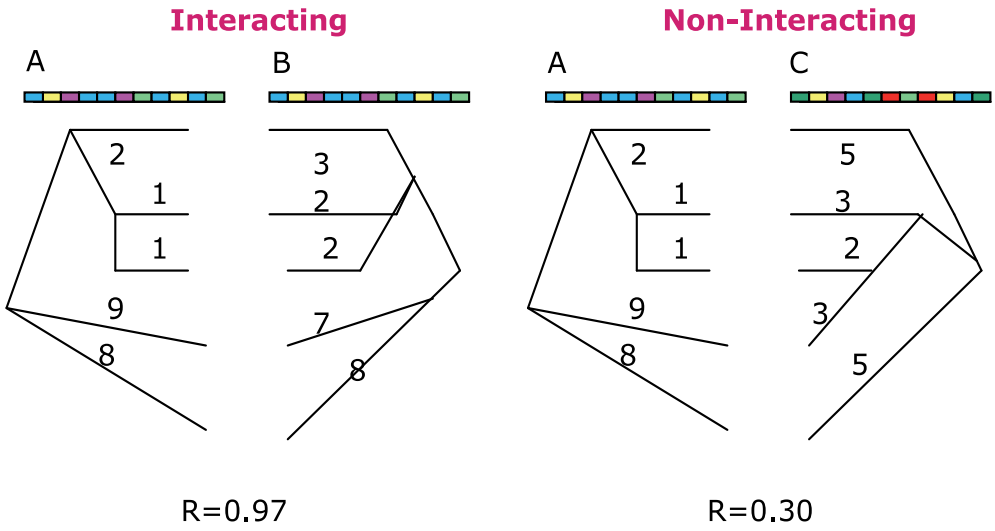


Fig. 2. Phylogenetic Analysis of Proteins

As proteins *A* and *B* are interacting proteins, they share a similar phylogenetic history and thus their phylogenetic profiles are highly correlated (R=0.97). Proteins A and C are non-interacting and are thus not strongly correlated (R=0.30).

PRISM algorithm compares the 158 template interfaces to a target dataset of 18698 structures obtained from passing the structures extracted form the PDB through a 50% sequence identity filter, splitting multimeric proteins into constituent chains, and counting homologous chains only once. The user can also probe a protein structure that is originally not present in the target dataset. To compare target proteins to template interfaces PRISM algorithm do as follow: (a) extract target protein surfaces; (b) compare the target surface with all interface complementary partners from the template dataset using MULTIPROT [Shatsky et al. 2004] in order to detect common geometrical cores in a sequence-order-independent way; (c) check for the presence of hotspots in the target structure. The final prediction score is calculated weighting the structural match ratio and the hotspot match ratio.

Availability: http://prism.ccbb.ku.edu.tr/prism/

3. Phylogenetics and beyond, how multiple kernel learning can improve predictions of receptor-ligand pairings

As seen in the sections above, there are several groups which have used phylogenetic analysis to predict PPI. Here we examine the use of multiple kernel learning in the task of PPI prediction. Kernel learning provides the ability to utilize directly and indirectly related data (such as expression measures, domain content, etc.) and perform classification in high dimensional space. When different data sources are used, separate kernel classifiers can be built and the combined output used to provide a final result.

One of the first groups to look at predicting PPI using multiple data sources was Bhardwaj et al. (2003). They use both phylogenetic information as well as expression data to make their predictions. The use of both data sources were proved, in their work, to provide results with greater accuracy than with using phylogenetic analysis alone. Co-expression is a logical source of information for use in this setting as proteins which interact for the purpose of performing a common function are likely to be co-expressed as they will need to be present at the same time in the cell [Bhardwaj et al. 2003, Grigoriev et al. 2001].

The idea of combining expression and phylogenetic information to predict PPI is clearly a step on a path which leads one to consider a wider variety of data integration. Other data sources include domain information as domains are known to interact and it is clear that this data would provide additional insight into the task of protein-protein interaction. Combining the above mentioned data sources can be carried out by using multiple kernel learning.

To examine the utility of multiple kernel learning with respect to this task, it is necessary to cite an example in which it performs better than other settings. One such example exists when one looks at the work of Gertz et al. (2003) and compare it with the work presented in Iacucci et al. (2011). Both groups look at the receptor-ligand prediction task and apply computational methods to the same dataset. The datasets consist of members of the chemokine and tgfβ ligand families with their respective receptor families. In the case of Gertz et al (2003), distances matrices are created for the families and are matched according to their similarity. Using a Metropolis Monte Carlo optimization algorithm, the Gertz et al. (2003) group explored and scored possible matches between the two matrices, until they reached optimal solutions. A limitation of this approach is that it relied on phylogenetic distance information alone.

Contrary to the work of Gertz et al (2003), the work presented by Iacucci et al 2011 proposes that the integration of multiple data sources results in more accurate matches. This work involved the creating of a combined kernel classifier to carry out the learning task. While other kernel-based works have been applied to the PPI task [Kim et al. 2010, Miwa et al. 2009], the work of Iacucci et al (2011) is unique as they apply multiple kernel learning to the receptor-ligand problem. More specifically, they apply the least-squares support vector machines (LS-SVM) method based on the conclusions by Suykens et al. (2001) which shows this implementation to be robust.

The ability of Iacucci et al. (2011) to predict candidate receptor-ligand pairs has been show to outpace that of Gertz et al. (2003) on the same dataset. This work involves using multiple data sources (expression, phylogenetic, and protein-domain content information), computing separate kernels for each data type, creating LS-SVM classifiers and combining the results to predict receptor-ligand pairs. The specifics of these steps will be discussed below.

3.1 Data sources

Several choices for data sources can be considered when addressing the PPI prediction task. While the studies, mentioned above, which use phylogenetic information rely on sequence data, other sources are available. Such sources include domain content data and expression data.

The phylogenetic data used in the Iacucci et al. (2011) study was derived through several steps. First, candidate receptor and ligand sequences were retrieved for seven species (*Rattus norvegicus*, *Mus* musculus, *Homo sapiens*, *Pan troglodytes*, *Canis familiaris*, *Cavia porcellus*, and *Bos taurus*) from ensemble build 51 [Hubbard et al. 2009]. Following this, the sequences were aligned using ClustalW [Thompson et al. 1994] Once aligned, the sequences were edited so as to eliminate the positions which were not conserved across the seven orthologous sequences. Finally, the pair-wise alignment score was then taken for each possible species to species comparison between the edited orthologous sequences (as seven species are used, a total of 21 pair-wise comparisons for each candidate are created). The distance scores form a phylogenetic vector which was then used to create the phylogenetic kernel.

The expression data used in the Iacucci et al. (2011) work was taken from the well-known GNF human expression atlas (79 tissues) [Su et al. 2004], the data was normalized (values were mean-zeroed and the standard deviation was set to one) and was further transformed into the expression kernel.

For the Iacucci et al. (2011) work, the domain content of each candidate protein (receptor or ligand) was taken from the Interpro Database [Hunter et al. 2009]. A vector for each candidate protein was created where the presence of a protein domain was indicated with a '1' and the absence of a domain was indicated by a '0'. This data was then transformed to create the domain content kernel.

The "Golden Standard" for the verification of the Gertz et al (2003) and the Iacucci (2011) et al. work is based on the Database of Ligand-Receptor Partners (DLRP) [Graeber et al. 2001]. This dataset is an experimentally derived dataset where known receptor-ligand pairs are stored. The information found here was used to train the LS-SVM described below. In addition, it was also used as the "Golden Standard" to determine which predictions, by both

groups, were true positives and false positive as well as false negatives and true negatives. These values were then used to calculate specificity and sensitivity of each groups' predictions to ultimately determine which approach provided better results.

3.2 Kernel creation and the LS-SVM

The creation of the kernels and the training of the least-squares support vector machine (LS-SVM) in the work presented by Iacucci et al. (2011) required multiples steps. First, the data sources, discussed above, were used to create data matrices (phylogenetic, expression, and domain content) which were then used to create three kernels for each receptor-ligand family. Following this, the LS-SVMs were trained using the three kernels to predict outcomes for receptor-ligand pairs known from the DLRP "Golden Standard".

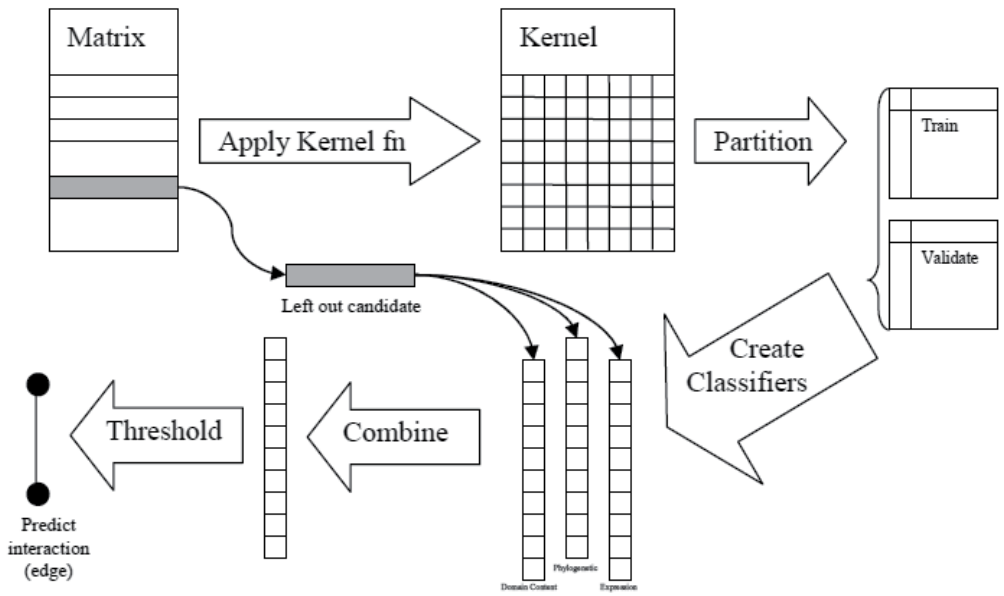


Fig. 3. Work flow of the combined kernel classifier

Data was partitioned into training and validation sets and parameters were tuned using a five fold validation strategy. The final output of the classifiers was achieved by a leave one out strategy. The classifier values were combined for a final result and a threshold was applied to determine which values are predicted edges (Figure published in Iacucci *et al.* 2011).

The kernel function used by Iacucci et al. (2011) measures the similarity between two proteins A and B (K(A,B)), one a candidate receptor A and the other a candidate ligand B. The LS-SVM classifier produced by Iacucci et al. (2011) is a binary predictor which assigns new examples in "interacting" or "non-interacting" classes. Creating the kernels from the various data matrices involved trials with different kernel functions, with linear functions ultimately being found to give the best performance in all cases. Data was partitioned into training and validation sets and parameters were tuned using a five fold validation strategy The final output of the classifiers was achieved by a leave-one-out strategy. The classifier values were scaled (minimum set to zero, maximum set to one). The values were then

combined, as defined in (1), for a final result. Figure 3 provides an overview of the workflow as described above.

$$g_{comb}(x) = \frac{g_{phylo}(x) + g_{\exp}(x) + g_{dom}(x)}{3} \quad (1)$$

3.3 Results and discussion

The comparison of the phylgenetic based method of Gertz et al. (2003) and the combined kernel classifer method of Iacucci et al. (2011) provides a clear perspective on the advantages of multiple kernel learning in the PPI prediction task. As both groups use the same dataset and have results which can be summarized and contrasted using recall, precision, and the *F*-measures.

The Iacucci et al. (2011) predictions for the tgfβ family accurately reconstructed over 76% of the supported edges (0.76 recall and 0.67 precision) of the know DLRP receptor-ligand pairs. In this case, the combined kernel classifier was able to relatively improved upon the Gertz et al. (2003) work by a factor of approximately two as the Gertz et al. (2003) work reconstructs 44% of the supported edges (0.44 recall and 0.53 precision) of the know DLRP receptor-ligand pairs. Comparing *F*-measures, we see that the combined kernel classifer method improved upon that of Gertz et al. (2003) significantly as the Iacucci et al. (2011) method has an *F*-measure of 0.71 while that of Gertz et al. (2003) has a value of 0.48.

The Iacucci et al. (20011) predictions for the chemokine family accurately reconstructed over 65% of the supported edges (0.65 recall and 0.23 precision) of the know DLRP receptor-ligand pairs. In this case, the combined kernel classifier was able to relatively improved upon the Gertz et al. (2003) work by a factor of approximately three as the Gertz et al. (2003) work reconstructs 22% of the supported edges (0.22 recall and 0.37 precision) of the know DLRP receptor-ligand pairs. Comparing *F*-measures, we see that the combined kernel classifer method improved upon that of Gertz et al. (2003) significantly as the Iacucci et al (2011) method has an *F*-measure of 0.33 while that of Gertz et al. (2003) has a value of 0.27.

Qualitatively, the performance of the Iacucci et al (2011) method also seems to be matching the performance of Gertz et al. (2003), as the novel interaction of CCR1 with SCY11 [Gao et al. 1996] reported in their work is also discovered using Iacucci et al (2011) method.

The comparison of the results of the two methods discussed here support the notion that kernel learning presents a useful methodology for elucidating receptor-ligand pairings. The benefits of the combined kernel classifier method over the Gertz et al. (2003) method are clear. Foremost in the advantages are the ability to predict multiple ligands for one receptor, which represents an necessary feature for receptor-ligand research. Also, as the classifier output is continuous, the results can be considered to be prioritized, this presents a major convenience to researchers as often the set of candidate ligands are large and financial and time resources to validate few.

4. Conclusion

The task of PPI prediction is a difficult and important area of bioinformatics research. As the number of possible interacting protein pairs in the cell is huge, wet-lab experimentation

validation of all of them is essentially impossible. In addition to being time consuming, in-vivo validation costs are also a consideration. Having a computational method for predicting PPI is therefore a necessary tool for researchers.

Several groups have addressed the PPI prediction task. While several have used phylogenetics to solve the problem, others have used physical protein structures and amino-acid sequence information to assist in making the predictions. We have reviewed these methods and discussed the key differences among them.

Methods, which rely on the physical structure and the conservation of amino-acid sequences in partners of interactions that are already reliably known to exist, also give researchers additional insight to function prediction as the methods are based on known examples. The drawback of these methods is that one has to have a known example for a comparison, which is not always the case when researching candidate receptor-ligand pairs.

Methods which rely on phylogenetic histories to determine PPI are based on a well-establish rational which holds that as interacting proteins co-evolve, there phylogenetic histories should be similar. This explains why the methods which rely on phylogenetic information are largely based on measures of similarity for the phylogenetic trees of interacting protein partners.

The advantage of using multiple kernel learning to predict PPI is apparent when using multiple sources of data. Many of the methods, mentioned above, rely on an ever growing amount of publicly available data. The ever expanding amount of high throughput data which continues to become available to the bioinformatics community represents an excellent opportunity to enhance the kernel classifier method presented in Iacucci et al. (2011).

A practical advantage of using multiple data sources allows one to extend the method as new and higher quality sources become available. For example, if better micro-array dataset becomes available in the future, it is an advantage to be able remove the existing expression-based kernel with one derived from the new dataset without having to the retrain a global classifier. Likewise, if additional data sources become available, adding an additional sub-classifier based on the new data source would take less time to train than adding the data source and retraining the global classifier.

Looking forward many exciting challenges remain to be addressed in this field. While the task of PPI is daunting and complex, the work reviewed above demonstrates that it is also rich with opportunities for improvement and further development.

5. Acknowledgments

Funding: The authors would like to acknowledge support from:

- Research Council KUL:
  ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys , START 1, several PhD/postdoc & fellow grants
- Flemish Government:
  FWO: PhD/postdoc grants, projects , G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR)
  IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3

FOD:Cancer plans
IBBT
- Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011) ;
-EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED

6. References

Aloy, P., & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. Proceedings of the National Academy of Sciences of the United States of America, 99(9), 5896-901. doi:10.1073/pnas.092147999

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. Journal of molecular biology, 332(5), 989-98. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14499603

Berman, H. M. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242. doi:10.1093/nar/28.1.235

Bhardwaj N, Lu H: Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics 2005, 21:2730-2738.

Bleakley K, Yamanishi Y: Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 2009, 25:2397-2403.

Boeckmann, B. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, 31(1), 365-370. doi:10.1093/nar/gkg095

Claverie, J. M. (2001). Gene number. What if there are only 30,000 human genes? Science (New York, N.Y.), 291(5507), 1255-7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11233450

Dafas, P., Bolser, D., Gomoluch, J., Park, J., & Schroeder, M. (2004). Using convex hulls to extract interaction interfaces from known structures. Bioinformatics (Oxford, England), 20(10), 1486-90. doi:10.1093/bioinformatics/bth106

Fryxell, K. J. (1996). The coevolution of gene family trees. Trends in Genetics, 12(9), 364-369. doi:10.1016/S0168-9525(96)80020-5

Gao JL, Sen AI, Kitaura M, Yoshie O, Rothenberg ME, Murphy PM, Luster AD: Identification of a mouse eosinophil receptor for the CC chemokine eotaxin. Biochem Biophys Res Commun 1996, 223:679-684.

Ge H, Liu Z, Church GM, Vidal M: Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 2001, 29:482-486.

Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, Rothschild B: Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 2003, 19:2039-2045.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., et al. (2003). A protein interaction map of Drosophila melanogaster. Science (New York, N.Y.), 302(5651), 1727-36. doi:10.1126/science.1090289

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. Journal of molecular biology, 299(2), 283-93. doi:10.1006/jmbi.2000.3732

Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H. H., et al. (2005). PSIbase: a database of Protein Structural Interactome map (PSIMAP). Bioinformatics (Oxford, England), 21(10), 2541-3.

doi:10.1093/bioinformatics/bti366

Graeber TG, Eisenberg D: Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nat Genet 2001, 29:295-300.

Grigoriev A: A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. Nucleic Acids Res 2001, 29:3513-3519.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L et al.: Ensembl 2009. Nucleic Acids Res 2009, 37:D690-D697.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L et al.: InterPro: the integrative protein signature database. Nucleic Acids Res 2009, 37:D211-D215.

Izarzugaza, J. M. G., Juan, D., Pons, C., Ranea, J. a G., Valencia, A., & Pazos, F. (2006). TSEMA: interactive prediction of protein pairings between interacting families. Nucleic acids research, 34(Web Server issue), W315-9. doi:10.1093/nar/gkl112

Jacob L, Vert JP: Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 2008, 24:2149-2156.

Keskin, O., Ma, B., & Nussinov, R. (2005). Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. Journal of molecular biology, 345(5), 1281-94. doi:10.1016/j.jmb.2004.10.077

Keskin, O., Tsai, C.-J., Wolfson, H., & Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein science : a publication of the Protein Society, 13(4), 1043-55. doi:10.1110/ps.03484604

Kim S, Yoon J, Yang J, Park S: Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics 2010, 11:107.

Kim, W. K., Bolser, D. M., & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics (Oxford, England), 20(7), 1138-50. doi:10.1093/bioinformatics/bth053

Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., et al. (2002). MIPS: a database for genomes and protein sequences. Nucleic acids research, 30(1), 31-4. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99165&tool=pmcentrez&rendertype=abstract

Miwa M, Saetre R, Miyao Y, Tsujii J: Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform 2009, 78:e39-e46.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y., & Wang, X. (1994). Co-evolution of ligand-receptor pairs. Nature, 368(6468), 251-5. doi:10.1038/368251a0

Nagamine N, Sakakibara Y: Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. Bioinformatics 2007, 23:2004-2012.

Ogmen, U., Keskin, O., Aytuna, a S., Nussinov, R., & Gursoy, a. (2005). PRISM: protein interactions by structural matching. Nucleic Acids Research, 33(Web Server), W331-W336. doi:10.1093/nar/gki585

Park, J., Lappe, M., & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. Journal of molecular biology, 307(3), 929-38. doi:10.1006/jmbi.2001.4526

Pazos, F., & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein engineering, 14(9), 609-14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11707606

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., et al. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC bioinformatics, 7, 365. doi:10.1186/1471-2105-7-365

Rubin, G. M. (2001). The draft sequences. Comparing species. Nature, 409(6822), 820-1. doi:10.1038/35057277

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic acids research, 32(Database issue), D449-51. doi:10.1093/nar/gkh086

Sato T, Yamanishi Y, Kanehisa M, Toh H: The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 2005, 21:3482-3489.

Shatsky, M., Nussinov, R., Wolfson, H., Guigó, R., & Gusfield, D. (2002). Algorithms in Bioinformatics. (R. Guigó & D. Gusfield, Eds.) (Vol. 2452, pp. 235-250). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/3-540-45784-4

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke,M.P., Walker,J.R., Hogenesch,J.B: A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004, 101:6062-6067.

Suykens JA, Vandewalle J, De MB: Optimal control by least squares support vector machines. Neural Netw 2001, 14:23-35.

Tan, S.-H., Zhang, Z., & Ng, S.-K. (2004). ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucleic acids research, 32(Web Server issue), W69-72. doi:10.1093/nar/gkh471

Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22:4673-4680.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673-80. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308517&tool=pmcentrez&rendertype=abstract

Tillier, E. R. M., Biro, L., Li, G., & Tillo, D. (2006). Codep : Maximizing Co-Evolutionary Interdependencies to Discover Interacting Proteins, 831(December 2005), 822- 831. doi:10.1002/prot

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403(6770), 623-7. Macmillian Magazines Ltd. doi:10.1038/35001009

van Kesteren, R. E., Tensen, C. P., Smit, A. B., van Minnen, J., Kolakowski, L. F., Meyerhof, W., Richter, D., et al. (1996). Co-evolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. The Journal of biological chemistry, 271(7), 3619-26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8631971

14

Finding Protein Complexes via Fuzzy Learning Vector Quantization Algorithm

Hamid Ravaee[1,2,*], Ali Masoudi-Nejad[1] and Ali Moeini[2]
[1]*Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran,*
[2]*Department of Algorithm and Computations, College of Engineering, University of Tehran, Tehran,*
Iran

1. Introduction

Protein–protein interactions (PPI) make up fundamentals of biological processes inside a cell. PPI has most important roles in cells such as post-translational regulation of protein activity, which is occurred by transient protein-protein interactions and participating in enzymatic complexes ensures substrate channelling which drastically increases fluxes through metabolic pathway (Lin et al., 2006). Metabolic pathways, for instance, consist of several proteins, called enzymes, organize a series of chemical reactions with the intent of altering a variety of chemical substance into the other forms, namely products. Proteins interactions happen in signalling pathways where a set of proteins, by an ordered sequence of reactions, try to convert a type of chemical signal to other form, enabling a cell to obtain environmental information quickly. Proteins interactions can be found in any sort of biological processes within cells. Indeed, existence of these interactions makes a cell function, to grow and more importantly survive (Bader & Hogue, a2003).

The objective in PPI network analysis is the discovering dense highly-connected subgraphs that represent functional modules and protein complexes. For understanding the cell function, it is essential first to find all functional modules in protein interaction networks (Bader & Hogue, b2003). Protein complexes are a group of proteins which have more interactions with each other at the same time and place (Chua et al. , 2008). On the other hand, the functional module consists of proteins that participate in a particular cellular process while interacting with each other at different time and place (Mirny & Spirin, 2003) . In order to simplify the terms, we used protein complex and functional modules as same. Since each protein could be involved in several protein complexes, the partitioning of PPI network to some disjoint groups of subgraphs could not explain the true nature of protein complexes occuring in PPI network. Hence, the finding of vertices group with overlapped boundary can be more useful in analyzing PPI network.

* Corresponding Author

In recently years, advances in the high-throughput PPI detection have produced a high volume of PPI datasets freely available to researchers. Therefore many methods and approaches have emerged to analyze experimental PPI data in various organisms. The experimental approaches for discovering protein complexes are more time consuming and expensive. Instead, computational methods which use PPI data are faster and cheaper (Ito et al., 2001).

The most common method of modelling PPI network is using graph theory, which in such a graph G=(V,E) where the nodes correspond to proteins and the edges correspond to interactions. Since the number of proteins and interactions between them in some organism such as yeast or human is remarkably high, the graph modelling PPI is called a complex graph. Partitioning of a complex graph to some disjoint subgraphs is called the graph clustering.

Clustering is the process of grouping data into sets (clusters) which shows more similarity between the objects in the same clusters than they are in different clusters (Schaeffer, 2007). Clustering analysis seeks a set of clusters based on similarity between pairs of elements. Graph clustering is the practice of distribution the vertices of the graph into the clusters taking into consideration the edge connectivity in the graph in such a way that many edges exist within each cluster and relatively few between the clusters. The result of this clustering can define the PPI network's structure and imply functions of proteins in the cluster which were previously uncharacterized (Lin et al., 2006).

Each complex graph modelling a system such as biological systems or social networks has specific properties and characteristics. The properties of graph could be fall into broad categories as the local properties and global properties (Przulj, 2005). The scale-free for distribution of degree and small world properties could be more affective on the result of graph clustering. A scale-free network has a vertex connectivity distribution that follows a power law, with relatively few highly connected vertices and many vertices having a low degree. Most biological networks such as PPI networks have the scale-free property (Pizzuti & Rombo, 2007). In this paper, we convert the normal scale-free PPI network to a non-scale free network by using line graph transformation. In the graph theory, line graph is produced by substituting edges and nodes in the graph. Each interaction is condensed into a node that includes the two interacting proteins. These nodes are then linked by shared protein content.

Important of results of the clustering in PPI network is illustration of structure of the PPI network which can be used to predict the functionality of uncharacterized protein based on other known proteins functions in the same cluster's elements. These clusters correspond to meaningful biological units such as protein complexes and functional modules.

Many clustering approaches (Gao, 2009; Bader & Hogue, b2003; Adamcsek, 2006; Wu et al., 2008 ;Vlasblom, 2009) could not place elements in multiple clusters, which can be unrealistic for biological systems, where proteins may participate in multiple cellular processes and pathways. Since each protein could participate in more than one protein complexes, in the clustering PPI graph, each protein probably have membership to more than one cluster. So in this paper, we present a clustering method that allows to having overlapping founded clusters. Disjoint clusters and overlapping clusters are illustrated in figure 1.

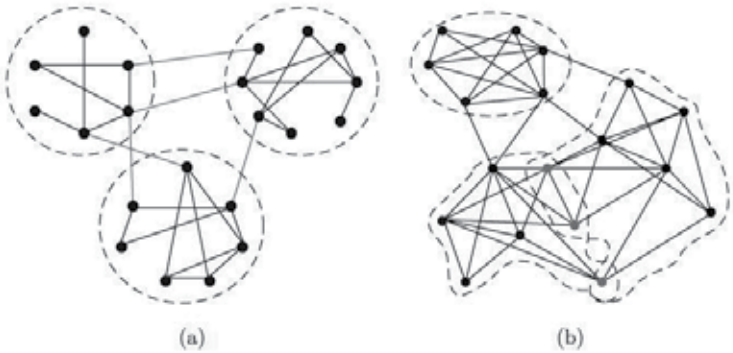


Fig. 1. Illustration of the concept of modules. (a) Disjoint modules; (b) Overlapping modules.

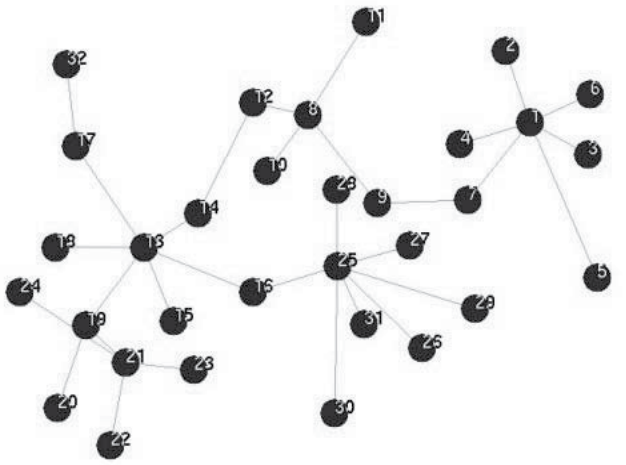
K-means (c-means) clustering (Hartigan, 1975) is applied on unlabeled data by partitioning them on predefined number of groups (k) based on the specifying the centers of groups. After each iteration in the k-means algorithm, the distances between each center of group and other data points are calculated and the center points are updated. Learning Vector Quantization uses k-means idea by defining some codebook vectors each of which represents a cluster for n-dimensional input data. The fuzzy clustering based on fuzzy set theory (Zadeh, 1965) is used to deal with indistinct boundaries between clusters. The most widely used fuzzy clustering method is the fuzzy c-means (FCM) algorithm (Bezdek, 1973) which is generalized from hard c-means algorithm. In this paper, extended FLVQ (Bezdek, 1995) as an intelligent computational method has been used for clustering PPIs. The results of this algorithm can be verified by biological and non-biological criteria and we showed that FLVQ technique is more effective and accurate for finding protein complexes in PPI network.

2. Primary definitions

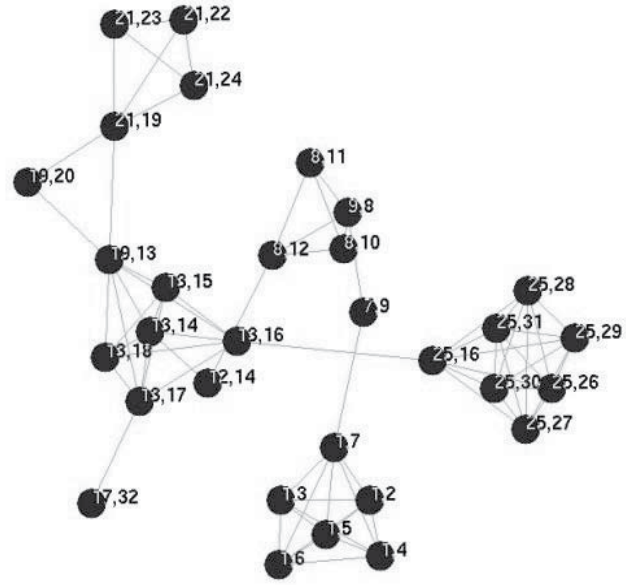
The problem of clustering of PPIs starts with a mathematical representation of PPI networks. A conventional way for representing PPI network is using graph theory concepts. PPI network could be illustrated by a graph G=(V,E) with a set of vertices V and a set of edges E in which each vertex is corresponded by a protein in PPI network and each edge connects to two vertices whose corresponding proteins have physical interaction with each other.

Clusters in the graph could be interpreted as dense subgraphs the number of edges within each subgraph is the maximum number and the number of edges between clusters is the minimum one. Therefore, the PPI clustering is an optimization problem and like other optimization problems, there is a need to an objective function to get optimum point.

PPI networks have scale-free property and finding the dense subgraphs is most difficult task in these networks. So using line graph we eliminate the scale-free property. In each node in the line graph is an edge in original network and every two nodes with common proteins are connected to each other. Figure 2 shows a scale free network and the generated line graph based on original graph.



(a)



(b)

Fig. 2. **a.** Orginal scale-free graph **b.** converted graph by line graph.

2.1 Learning Vector Quantization

Learning Vector Quantization (LVQ) is placed in the competitive learning category and it is closely related to Self-Organizing Map (SOM) (Kohonen, 1990). SOM is a well-developed neural network technique for data clustering and visualization. It can be used for projecting a large data set of a high dimension into a low dimension (usually one or two dimensions) while retaining the initial pattern of data samples. Indeed, SOM has two main principles:

vector quantization and vector projection. Vector quantization makes up a delegate set of vectors called output vectors (codebook vectors) from the input vectors. Let's denote the set of output vectors (codebook vectors) as $Y=\{y_1,y_2,..,y_c\}$ with the same dimension as input vectors. In general, vector quantization reduces the number of vectors, and this can be considered as a clustering process. The maximum number of clusters in a network is defined by user specified value, c. After learning process, it may be possible for some codebook vectors to correspond to empty clusters.

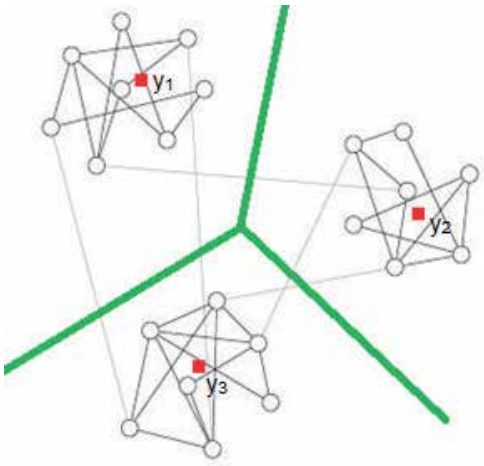


Fig. 3. The red points (y_1,y_2,y_3) corresponded to output vectors indicating a dense subgraph in the sample network.

The LVQ algorithm represents a set of input vectors $x_i \in X \subset \Re^n$ by a set of c prototypes $Y = \{y_1, y_2, .., y_c\} \subset \Re^n$ .. The LVQ is associated with a competitive network which consists of an input layer and an output layer. Each node in the input layer is connected directly to the cells, or units, in the output layer. A weight vector, also referred to as prototype, is assigned to each cell in the output layer (Ravuri & Karayiannis, 1995). The codebook vector having minimum distance with input vector x_i is called winner vector, k, and is defined as:

$$k = \arg \min_l \|y_l - x_i\| \tag{1}$$

Update equation of LVQ algorithm is:

$$y_j(t+1) = y_j(t) + \alpha_t h_{ij,k} \|x_i - y_j(t)\| \tag{2}$$

Here a_t is the scalar-valued learning rate, $0<a_t<1$, and decreases monotonically with time t. The neighborhood function $h_{ij,k}$ denotes the interaction between codebook vector i and j and winner vector k. The simple definition of $h_{ij,k}$ is:

$$h_{ij,k} = \begin{cases} 1 \; if \; j = k \\ 0 \; if \; j \neq k \end{cases} \tag{3}$$

In the LVQ algorithm, neighborhood radius is one and only the winner vector could be updated.

2.2 Fuzzy Learning Vector Quantization

While most typical clustering algorithms assigns each data point to exactly one cluster, fuzzy clustering allows for the extent of membership, to which a data point belongs to different clusters. The FLVQ may be seen as a learning fuzzy c-means using a fuzzification index m. Karayiannis et al (Ravuri & Karayiannis, 1995) presented a broad family of FLVQ algorithms, which were initially introduced on the basis of perceptive arguments. This derivation was based on the minimization of the average generalized distance between the input vectors and the prototype vectors. The fuzzy partitioning algorithm, FCM is run into by minimization problem that is solved by reformation of FCM algorithm to FLVQ algorithm (Bezdek, 1995).

The updated equation for the FLVQ involves the membership functions which are used to determine the strength adjacency between each prototype and input vectors.

$$\alpha_{ij,t} = (u_{ij,t})^{m_t} \tag{4}$$

$$u_{ij} = \left[\sum_{l=1}^{c} (\frac{D_{ij}}{D_{lj}})^{\frac{2}{m-1}} \right]^{-1} \tag{5}$$

$$D_{ij} = \left\| x_i - y_j \right\| \tag{6}$$

Where $m = m_t = m_0 - \Delta mt$ and $\Delta m = (m_0 - m_f) / MaxItr$ and D_{ij} is the distance and m_0 is some constant value greater than the final value (m_f) of the fuzzification parameter m. *MaxItr* is the constant parameter for limitation of iterations.

3. The FLVQ algorithm

The calculation of distances between network vertices and prototype vectors in the FLVQ is critically challenging. In the following algorithm, we used a new definition of vertices based on n-dimensional vectors and; we representing new scalar distance between input vectors and codebooks (output) vectors. Each vertex in PPI graph is modeled by a vector called input vector. Given *G=(V,E)* represents a PPI network including $|V|$ vertices and $|E|$ edges. An input vector is defined as :

$$x_m = \{x_{m1}, x_{m2}, .., x_{mn}\}$$

$$x_{ij} = \begin{cases} 0 + \varepsilon & if \quad i \neq j \quad and \quad e_{ij} = 0 \\ 1 - \varepsilon & if \quad e_{ij} = 1 \\ 1 - \varepsilon & if \quad i = j \end{cases} \tag{7}$$

Where $n=|V|$, e_{ij} is element *(i,j)* in adjacency matrix corresponding the graph G and ε is a real small value between *(0,1)*.

This definition makes possible to use scalar distance measure such as the dot product is possible. There are some distance criteria in vector space to measure similarity (distance) between two vectors. Correlation is a simple way for measuring distance between two vectors in the same dimension. If x_i and x_j are two vectors with the dimension of n, the equation (8) is the inner product of two vectors:

$$S_{ij} = X_i . X_j = \sum_{k=1}^{n} x_{ik} x_{jk} \tag{8}$$

$$D_{ij} = S_{ij}^{-1} \tag{9}$$

Where D_{ij} is the distance and S_{ij} is the inner product between X_i and X_j.

The FLVQ algorithm performs clustering of the input graph by training process. Training process consists of some iterations. The number of iteration depends on convergence criteria and can be limited by a user specified constant. Each iteration consists some epochs. The number of epochs is equal by c (number of prototype vectors and the maximum number of clusters). In each epoch, an input vector x_i is selected randomly. A selected input vector is not being selected in a same epoch again. The selected input vector x_i is compared with all the prototype vectors with a similarity measure *(ex. dot product)* and the prototype vector y_j with most similarity with x_i known as winner vector.

The implementation of the FLVQ algorithm is described as follows:

- **Step 1. Initialization**
  Initialize the c codebook's vectors $y=\{y_1,y_2,..,y_c\}$ *by randomly assigning each element of codebook vectors by a real number between* $(\varepsilon,1-\varepsilon)$. *Set iteration counter* $t=1$. *Give* $0\leq\varepsilon<1$. t_{max} *is the iteration limit.*
- **Step 2. Learning**
  Repeat until stopping criterion is satisfied:
- **Step 2.1** While there is a unselected input vector
  - *Randomly pick an input* x_i
  - *Compute winner vector based on distance measure of xi and every codebook vectors* $y_j : j=1..k$
  - *update winner vector* y_j *based on input vector* x_i *and learning ratio a*
- **Step 2.2** update learning ratio α

4. Data set

The PPI network is derived from the yeast subset in the Database of Interacting Proteins (DIP) (Xenarios et al., 2002). The dataset of yeast is composed of 4963 proteins and 17570 interactions. Most of these interactions have been derived by yeast two-hybrid screen. For evaluation of finding clusters, we use protein complex data from the MIPS database (Mewes et al., 2004). In the currated complex dataset, there are 404 protein complexes. The protein complex having most proteins is "cytoplasmic ribosomal large subunit" with 88 proteins and there are 169 protein complexes with just two proteins.

5. Experimental result

The FLVQ algorithm is applied on the PPI network of the *Saccharomyces Cerevisiae* (yeast) dataset downloaded from the DIP (Guldener, 2005). After using FLVQ on DIP protein-protein interaction, over than 300 clusters obtained frequency of each based on the number of vertices in is shown in figure (4). As the figure (4) shows most obtained clusters approximately include 9 and 12 vertices. In addition, the number of clusters with size of over 20 are also considerable. This means that the FLVQ algorithm could find larger dense subgraphs in the PPI network. When the cluster size became larger, few graph clustering methods could find these clusters with proper efficiency.

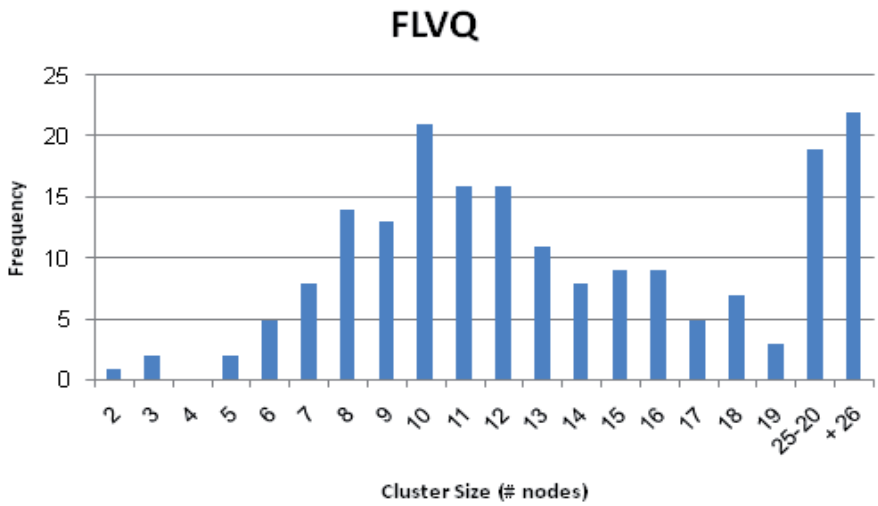


Fig. 4. Number of obtained clusters by FLVQ algorithm based on the cluster size.

The results of the FLVQ algorithm are evaluated by the clustering score used by (Bader & Hogue, a2003; Newman, M. & Girvan. M., 2004). The clustering score for each cluster is defined by the product of size and density of the cluster. The density of cluster is the ratio between number of edges in cluster $|E|$ and maximum number of possible edges in it $|E_{max}|$. The following equation (10) shows clustering score definition.

$$\sigma(\Gamma)=\delta(\Gamma).|V| \tag{10}$$

Where Γ is a cluster in the clustering result and $\delta(\Gamma)$ is the density of given subgraph Γ and is declared by equation (11) and $/V/$ shows the number of vertices in Γ subgraph.

$$\delta(\Gamma)=2|E|/(|V|(|V|-1)) \tag{11}$$

Where E is the set of edges that connects the existing vertices in V in given subgraph of Γ. The clustering score for each clusters is shown in figure (5). The cluster score for bigger

clusters is more elevated than smaller clusters proving that FLVQ is rather successful to find subgraphs with more higher number of vertices and with most density. Highest clustering score shows that the obtained clusters are more compact and larger.

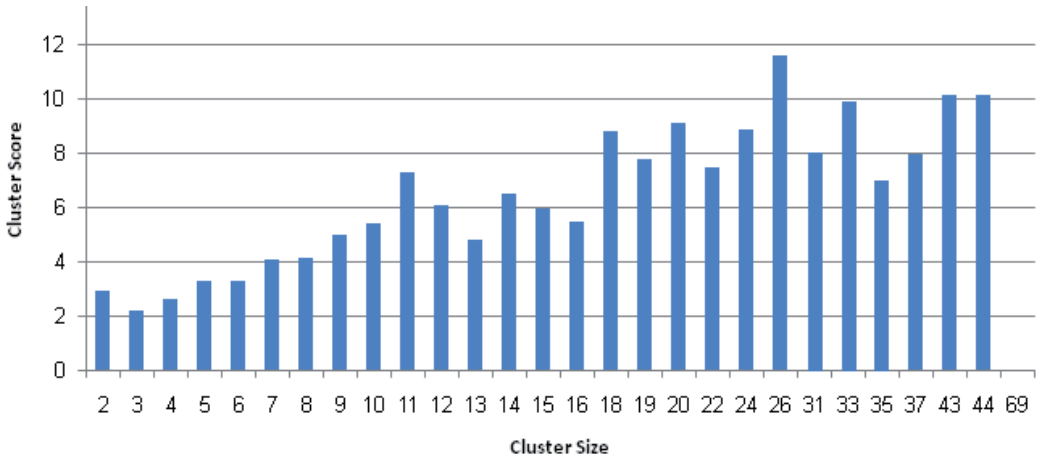


Fig. 5. Amount of clustering score for each obtained cluster in FLVQ algorithm.

The clustering results can be validated by ground truth with Precision and Recall. Assume a module (cluster) X is mapped to a functional module F_i. Recall, also termed the true positive rate or sensitivity, is the proportion of proteins common in both X and F_i to the size of F_i. Precision, which is also termed the positive predictive value, is the proportion of proteins common in both X and F_i to the size of X.

$$precision = \frac{|X \cap F_i|}{|X|} \tag{12}$$

$$recall = \frac{|X \cap F_i|}{|F_i|} \tag{13}$$

The accuracy of clusters is assessed by *f*-measure. The *f*-measure is defined as the harmonic mean of recall and precision:

$$f - measure = \frac{2(precision.recall)}{precision + recall} \tag{14}$$

Figure (6) shows the average of f-measure based of protein complex size for the FLVQ algorithm. In figure (6), the f-measure of each obtained cluster is measured based on experimental protein complexes MIPS. The value of f-measure could be between 0 and 1.

The highest f-measure value indicates the most conformity between experimental protein complex and obtained complex by the algorithm.

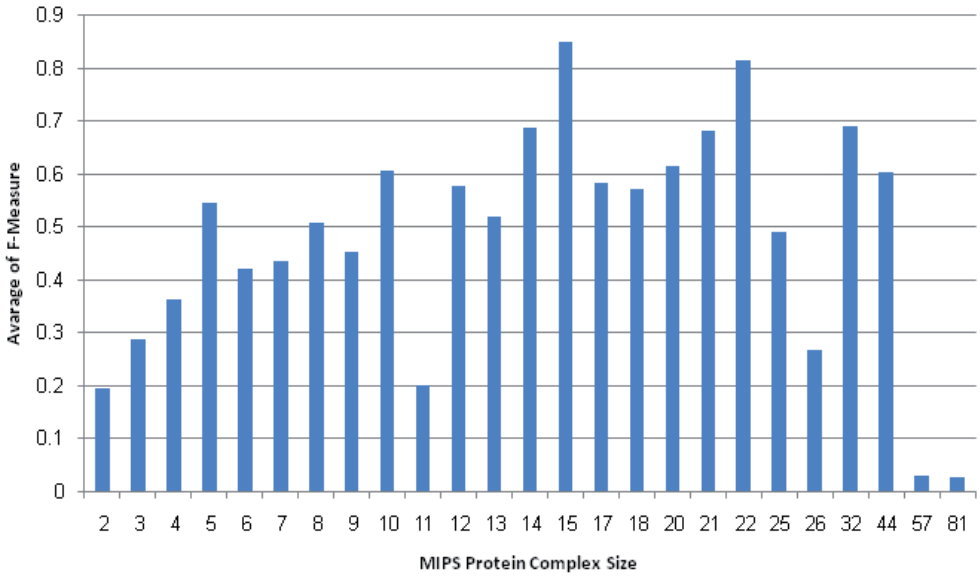


Fig. 6. f-measure between finding subgraphs and experimental protein complexes based on its size.

6. Conclusion

In this paper, we presented a FLVQ algorithm as a robust tolerable method to find dense subgraphs in PPI networks as protein complexes. The algorithm identifies more than 200 dense subgraphs having more overlap among experimentally known protein complexes. By clarifying the structure of protein interactions network, uncharacterized proteins could be predicted by the functions of other known proteins which belong to same clusters. By using line graph transformation, we eliminated the scale-free degree distribution in PPI network which caused larger number of dense highly connected subgraph revealed. There is overlapping between found subgraphs that express the results are more conforming with the reality nature of protein complexes.

7. References

Adamcsek, B. (2006). Cfinder: Locating Cliques And Overlapping Modules In Biological Networks, *Bioinformatics*, Vol. 22, pp. 1021-1023.

Bader, G. & Hogue, C. (2003). An Automated Method For Finding Molecular Complexes In Large Protein Interaction Networks, *BMC Bioinformatics*, Vol. 4

Bader, G. & Hogue, C. (2003). Analyzing Yeast Protein-Protein Interaction Data Obtained From Different Sources, *Nat. Biotechnol*, Vol. 20, pp. 991-997

Bezdek, C. &Hathaway,J. (1995). Optimization Of Clustering Criteria By Reformulation, *IEEE Transactions on Fuzzy Systems*, Vol. 2, pp. 241-246.

Bezdek, C.; (1973). Fuzzy Mathematics In Pattern Classification, *Ph.D. dissertation, Dept. Appl. Math.,Cornell Univ., Ithaca, NY*

Chua, H.; Ning, K.; Sung, W.; Leong. H., (2008). Using Indirect Protein–Protein Interactions For Protein Complex Prediction, *Journal of Bioinformatics and Computational Biology*, Vol. 6., pp. 435-466.

Gao, L.; Sun, p.; Song, j. (2009). Clustering Algorithms For Detecting Functional Modules In Protein Interaction Networks, *Journal of Bioinformatics and Computational Biology*

Guldener, U. (2005). CYGD: The Comprehensive Yeast Genome Database, *Nucleic Acids Res*, Vol. 33, pp. 364–368.

Hartigan, J.A. (1975). Clustering Algorithms. *New York : Wiley*

Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M. ; Hattori, M.; Sakaki, Y. (2001). A Comprehensive Two-Hybrid Analysis To Explore The Yeast Protein Interactome, *PNAS*, Vol. 98, pp. 4277-4278.

Kohonen, T. (1990). The Self Organizing Map, *IEEE Proc*, Vol. 78

Lin, C.; Cho, Y.; Hwang, W.; Pei, P.; Zhang, A. (2006). Clustering Methods In Protein-protein Interaction Network, In: *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*. John Wiley & Sons

Mewes, H.W. et al., (2004). MIPS: Analysis And Annotation Of Proteins From Whole Genomes, *Nucleic Acids Res*, Vol. 32, pp. D41-D44

Mirny, V. & Spirin, L. (2003). Protein Complexes And Functional Modules In Molecular Networks, *Proc. Natl Acad. Sci*, Vol. 100(21), pp. 12123–12126

Newman, M. & Girvan. M., (2004). Finding And Evaluating Community Structure In Networks, *Physical Review*, 2004.

Palla, G.; Der´enyi, I.; Farkas, I.; Vicsek, T., (2005). Uncovering The Overlapping Community Structure Of Complex Networks In Nature And Society. , *Nature*, Vol. 435, pp. 814–818

Pizzuti, C.; Rombo, S. (2007). Multi-functional Protein Clustering in PPI Networks, *International Conference on Intelligent Data Engineering and Automated Learning*

Przulj, N. (2005). *Knowledge Discovery in Proteomics Graph Theory Analysis of Protein-protein Interactions*, Department of Computer Science, University of Toronto

Ravuri, N. & Karayiannis, M. (1995). An Integrated Approach To Fuzzy Learning Vector Quantization And Fuzzy C-Means Clustering, *New York : Intelligent Engineering Syslems Through Artificial Neural Networks*, Vol. 4, pp. 247-252.

Schaeffer, S. (2007). Survey Graph Clustering, *Elsevier*

Vlasblom, J. & Wodak., S. (2009). Markov Clustering Versus Affinity Propagation For The Partitioning Of Protein Interaction Graphs, *BMC BIOINFORMATICS*, Vol. 10.

Wu, M.; Li, X.; Kwoh. C. (2008). Algorithms For Detecting Protein Complexes In PPI Networks: An Evaluation Study, *PRIB08*, pp. 135-146.

Xenarios, I.; Salwinski, L.; Duan, X.; Higney, P.; Kim, S.; Eisenberg, D. (2002) DIP, The Database Of Interacting Proteins: A Research Tool For Studying Cellular Networks Of Protein Interaction, *Nucleic Acids Res*, Vol. 30, pp. 303-305.

Zadeh L., A. (1965) Fuzzy Sets, *Inf. Control*, Vol. 8, pp. 338-353.

Zhang, A., (2009). Modularity Analysis of Protein Interaction Networks, *Protein Interaction Networks Computational Analysis* , pp. 66-77

Part 2

Experimental Approaches

15

In Vivo Imaging of Protein-Protein Interactions

Hao Hong[1], Shreya Goel[2] and Weibo Cai[1]
[1]*Departments of Radiology and Medical Physics, University of Wisconsin - Madison, Madison, WI,*
[2]*Centre of Nanotechnology, Indian Institute of Technology, Roorkee,*
[1]*USA*
[2]*India*

1. Introduction

Protein-protein interaction (PPI) plays a pivotal role in a wide variety of cellular events and physiological functions, such as enzymatic activity, signal transduction, immunological recognition, DNA repair/replication, among others (Valdar and Thornton, 2001). In addition, biological events that regulate proliferation, differentiation, and inflammation are also commonly mediated through PPI (Villalobos et al., 2007). Various techniques in molecular biology have been developed to understand the mechanism of these ubiquitous interactions, including qualitative methods such as yeast-two-hybrid screen (Fields and Song, 1989), immunoprecipitation (Williams, 2000), gel-filtration chromatography (Phizicky and Fields, 1995), etc. Meanwhile, quantitative biophysical methods have also been designed which include analytical ultracentrifugation (Hansen et al., 1994), calorimetry (Doyle, 1997), optical spectroscopy (Lakey and Raggett, 1998), etc. A decade ago, an assay for PPI based on β-galactosidase (gal) complementation was designed and successfully applied in cells (Wehrman et al., 2002).

Despite the success achieved by these techniques, none of them can be employed for interrogating PPI in living subjects due to several major limitations. First, traditional assays for measuring protein interactions require cell lysis, where the experimental conditions are inconsistent with the natural intracellular milieu. Second, these techniques may not be able to detect transient interactions that may have potent effects on cell signalling and intracellular processes. Lastly, the degree of false positives and false negatives vary from method to method, which significantly compromises the reproducibility and reliability of the data. With the tremendous expansion and evolution of the interdisciplinary field of molecular imaging over the last decade, many of these disadvantages have been or can be overcome.

Molecular imaging, "the visualization, characterization and measurement of biological processes at the molecular and cellular levels in humans and other living systems" (Mankoff, 2007), is an extremely powerful tool for imaging of PPI. The major molecular imaging modalities that have been applied for investigating PPI include bioluminescence, fluorescence, and positron emission tomography (PET) imaging. Quantitative and real-time molecular imaging of PPI can not only complement the already existing methodologies,

which are mostly used in vitro or in cell culture, but also provide invaluable insights on PPI that were unavailable or impossible to investigate previously. For example, non-invasive imaging of PPI can dramatically accelerate the evaluation of new drugs in living subjects that promote or inhibit homodimeric/heterodimeric protein assembly (Massoud et al., 2007; Villalobos et al., 2007).

In this chapter, we will summarize the current status of in vivo imaging of PPI with various techniques, including fluorescence, bioluminescence, and PET imaging. A schematic summary of the most commonly used strategies for imaging of PPI are shown in **Figure 1**. To the best of our knowledge, there is no literature available on fluorescence imaging of PPI in animal models. However, since this is an indispensible component of imaging PPI in cell culture, herein we will give a few representative examples on fluorescence imaging of PPI to provide a complete overview of this dynamic research area.

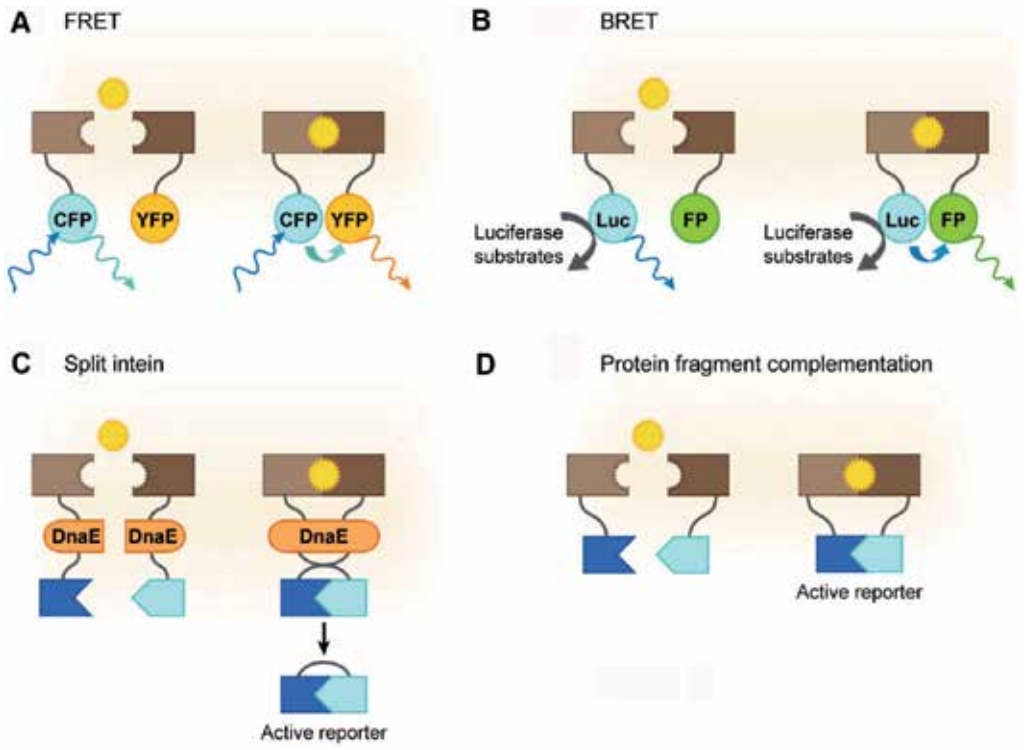


Fig. 1. Commonly used strategies for imaging of PPI. **A.** Fluorescence resonance energy transfer (FRET). **B.** Bioluminescence resonance energy transfer (BRET). **C.** Self-splicing split inteins (DnaE) can splice the two fragments of a reporter protein together into an intact and active reporter protein when they are brought within close proximity of each other. **D.** Protein fragment complementation. Brown fragments are proteins of interest and the yellow star represents an inducer of PPI. Adapted from (Villalobos et al., 2007).

2. Fluorescence imaging of PPI

The (imaging) techniques used to detect or quantify PPI need to be sensitive within the concentration ranges at which proteins are present in cells or tissues, where sometimes fewer than 10^4 protein molecules may be present. Furthermore, these techniques should be capable of identifying interactions of specific proteins against a background of more than 30,000 other proteins within a living cell. As a technology that has had an impact on almost all areas of biology, fluorescent imaging can meet these criteria under certain scenarios and has been widely used for imaging of PPI in vitro.

Fluorescence spectroscopy and fluorescence imaging have been demonstrated to be versatile tools for imaging of PPI. Fluorescent proteins (FPs), specifically variants of the green FP (GFP), are among the most frequently used for imaging of PPI (Giepmans et al., 2006; van Roessel and Brand, 2002). In a typical fluorescence process, an electron in the fluorophore within the FP absorbs photons from suitable excitation light (in the UV or visible range), which raises the energy level of the electron to an excited state. During this short excitation period, some of the energy is dissipated through molecular collisions or transferred to a proximal molecule, and the remaining energy is emitted as a photon to relax the electron back to the ground state (van Roessel and Brand, 2002). Since the energy is lower for the emission photon than the excitation photon, the emission wavelength is longer than the excitation wavelength which can be readily separated by applying a filter of specific wavelength range.

Fluorescence imaging of PPI in cell culture has the potential to provide information on the cellular and sub-cellular distribution of FPs with sub-second time resolution. Fluorescence microscopy techniques, primarily including fluorescence resonance energy transfer (FRET) and fluorescence correlation spectroscopy (FCS), are commonly used to quantify the activity, interaction, and dynamics of protein molecules within living cells (Yan and Marriott, 2003). Many protein interactions are transient, or energetically weak, thereby precluding their identification and analysis through traditional biochemical methods such as co-immunoprecipitation. In this regard, the genetically encodable FPs (GFP, yellow FP [YFP], cyan FP [CFP], red PP [RFP], etc.) and their associated overlapping fluorescence spectra have granted researchers the ability to monitor weak interactions in live cells using FRET.

2.1 Imaging of PPI with FRET

FRET requires the measurement of the relative intensity of the emission signal from a pair of fluorophores (Tsien, 2009). The underlying physics is attributed to a quantum mechanical effect between a given pair of fluorophores (i.e. a fluorescent donor and an acceptor) where, upon excitation of the donor, energy is transferred from the donor to the acceptor in a non-radiative manner by means of dipole-dipole coupling (Jares-Erijman and Jovin, 2003). Upon energy transfer, donor fluorescence is quenched and acceptor fluorescence is increased (sensitized), resulting in a decrease in donor excitation lifetime. The FRET efficiency is the quantum yield of the energy transfer transition, i.e. the fraction of energy transfer event occurring per donor excitation event, which is dependent upon several factors including the distance between the donor and the acceptor, the spectral overlap of the donor emission spectrum and the acceptor absorption spectrum, as well as the relative orientation of the donor emission dipole moment and the acceptor absorption dipole moment.

Since FRET is critically dependent upon molecular proximity, it has been described as a molecular ruler. FRET typically operates in a range of 1-10 nm, a distance that is relevant for most molecules engaged in complex formation or conformational changes. FRET from CFP to YFP is a commonly used strategy for monitoring protein interactions or conformational changes of individual proteins. For example, FRET-based assays involving CFP and YFP were designed and employed to monitor receptor interactions on endothelial cells in one report (Seegar and Barton, 2010). However, one disadvantage of FP-based FRET is that protein functions may be perturbed by fusion of FPs since they are quite large in size. In one study, G protein-coupled receptor (GPCR) activation in living cells was used as a model system to compare YFP with a small fluorescent agent (FlAsH), which was targeted to a short tetracysteine sequence (Hoffmann et al., 2005). It was found that FRET from CFP to FlAsH reports GPCR activation in living cells without disturbing receptor function, which is more advantageous than the use of YFP as the FRET acceptor.

FRET has also been employed to visualize the interaction between two FPs, enhanced GFP (EGFP) and mCherry (Albertazzi et al., 2009). One- and two-photon fluorescence lifetime imaging microscopy (FLIM) were used to determine the FRET efficiency values. It was found that this FP pair can be used for effective and quantitative FRET imaging of PPI. Since FLIM can produce images based on the differences in the exponential decay rate of the fluorescence signal from different fluorophores, advances in FRET and FLIM have enabled studies of PPI at the microscopic level. FLIM provides a promising and robust method of detecting molecular interactions via FRET by monitoring the variation of donor fluorescence lifetime, which is insensitive to many factors that can influence the conventional intensity-based measurements, such as fluorophore concentration, photobleaching, spectral bleed-through, donor-acceptor stoichiometry, light path length etc. (Pelet et al., 2006; Zhong et al., 2007). The fact that FRET can deplete the excited state population of the donor and cause a reduction in both its fluorescence intensity and lifetime makes this technique well suited for studies in intact cells.

Interrogating PPI deep inside living tissues requires precise fluorescence lifetime measurements to derive the FRET between two tagged fluorescent markers. In a recent study, FLIM was used in combination with a clinically licensed remote endoscopic cellular resolution imaging modality to map PPI in live cells embedded in a 3D matrix, which served as a model of a diseased organ structure in a patient (Fruhwirth et al., 2010). This strategy allowed accurate measurement of fluorescence lifetime changes on the order of 100 ps, which not only demonstrated the feasibility of studying PPI by FRET in cultured living cells within 3D matrices, but also provided potential instrumentation for other FRET-based assays.

The FRET/FLIM technique can also provide invaluable information for the mechanistic study of PPI in different types of diseases. In one study which investigated the mechanism of metastasis induction by the S100A4 protein, interactions of S100A4 with C-terminal recombinant fragment of non-muscle myosin heavy chain in living HeLa cells were mapped using confocal microscopy, FLIM, and time-correlated single-photon counting (Zhang et al., 2005). The findings indicated that not only there is direct interaction between S100A4 and its target in live mammalian cells, but also that such an interaction contributes to metastasis induction, thus shedding new light onto the mechanism of cancer metastasis. In another

report, FRET/FLIM enabled the study of the interaction between hypoxia-inducible factor-1α (HIF-1α) and HIF-2α with the aryl hydrocarbon receptor nuclear translocator in a hypoxia model, which provided new information about specific gene expression controlled by PPI in hypoxia (Konietzny et al., 2009). FRET/FLIM has also been employed to image dynamic PPI in neurons (**Figure 2**), which enhanced the understanding of nervous system development and function (Timm et al., 2011). Protein kinases of the microtubule affinity regulating kinase (MARK)/Par-1 family play important roles in the establishment of cellular polarity, cell cycle control, and intracellular signal transduction. Disturbance of their function is linked to cancer and various brain diseases. In this recent study, transfected Teal FP (TFP) and YFP were used as FRET donor and acceptor pairs in neurons and imaged by FLIM, which revealed that MARK was particularly active in the axons and growth cones of differentiating neurons (Timm et al., 2011).

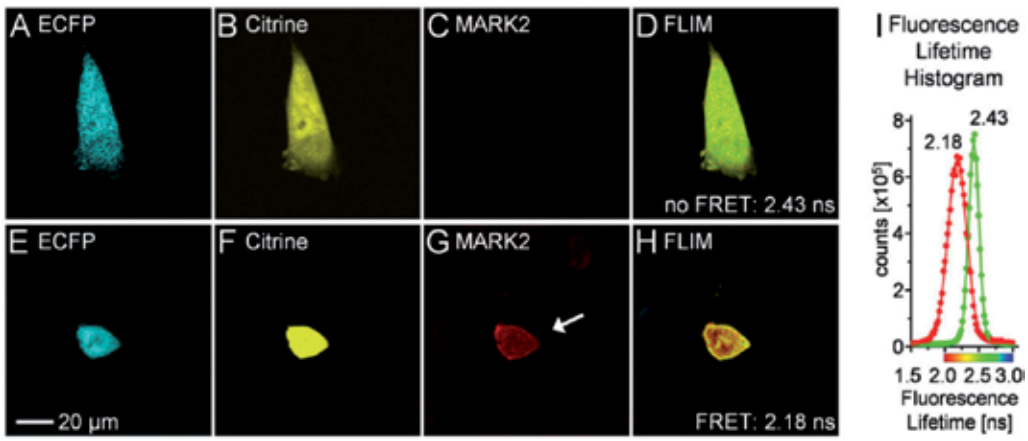


Fig. 2. The upper panel shows both channels of the fluorescence intensity image (**A, B**) of a cell transfected with a construct composed of ECFP (i.e. enhanced CFP) linked to Citrine (i.e. a stable variant of YFP), which does not exhibit FRET in the absence of fluorescently labeled MARK2 (i.e. the inducer of FRET) as indicated by a lack of fluorescence signal in **C**. The pseudo-colored FLIM image is shown in **D**, which has a long fluorescence lifetime of 2.43 ns. FRET between the two FPs (**E, F**) occurs when MARK2 is present, as indicated by the fluorescence signal in **G**. The short fluorescence lifetime of 2.18 ns is shown as red in **H** (high FRET). The graph **I** displays the averaged histograms of cells showing FRET (red dots) or no FRET (green dots) and gaussian fits of the data. Reprinted with permission from (Timm et al., 2011).

Not limited to the imaging of PPI, FRET can also be employed for imaging protein-DNA interactions, such as through the use of near-infrared fluorescent oligodeoxyribonucleotide reporters that can sense transcription factor NF-κB p50 protein binding (Zhang et al., 2008). Recently, a similar approach using hairpin-based FRET probes for the detection of human recombinant NF-κB p50/p65 heterodimer binding to DNA was reported (Metelev et al., 2011). Both of these studies demonstrated that FRET-based technique can give signal changes that are simple to interpret and stoichiometrically correct for detecting transcription factor-DNA interactions.

2.2 Imaging of PPI with FCS

Different from FRET, FCS detects the diffusion rate of single molecules which can give insights regarding whether a protein is part of a larger complex or not (Elson, 2004; Haustein and Schwille, 2007). Based on the analysis of intensity fluctuation of one or a few labeled protein conjugates at nanomolar concentration in a femtoliter volume, which depends on several factors such as the number of fluorescent species in the excitation volume, the diffusion constant of the conjugate, etc., FCS has been used to study PPI, protein-lipid/ligand-receptor interactions, dimerization of membrane receptors and proteins involved in the downstream signalling, DNA dynamics, among others (Elson, 2004; Haustein and Schwille, 2007). The high sensitivity and the possibility to monitor these dynamic interactions makes FCS a powerful tool to study signal transduction in cellular or even tissue environment at physiologically relevant conditions (Hink et al., 2002).

FCS is relatively insensitive to molecular mass. Therefore, species with similar molecular weight cannot be differentiated. Dual color fluorescence cross-correlation spectroscopy (FCCS) measures interactions by cross-correlating two or more fluorescent channels (one channel for each molecule/protein of interest), which can distinguish interactions and dynamics of biomolecules more sensitively than FCS, particularly when the mass change in the reaction/interaction is small. However, the inherent drawback of FCCS is that it suffers from non-ideal confocal volume overlap and spectral cross-talk which severely limits its applications. Fluorescence lifetime correlation/cross-correlation spectroscopy has the potential to resolve this issue, as demonstrated in a recent study (Chen and Irudayaraj, 2010). Interaction of a fluorescently-labeled antagonist antibody with the epidermal growth factor receptor (EGFR)-GFP construct in live HEK293 cells were monitored by both fluorescence lifetime cross-correlation measurements and FLIM, which not only opens up new opportunities in studying PPI in solutions and in live cells but also provides new biological insights in understanding how an antagonist influences EGFR through live cell imaging and quantification.

The field of plant sciences has also benefited from these techniques mentioned above. For example, FRET/FLIM was used to investigate CDC48A, a member of the AAA ATPases (i.e. ATPases associated with diverse cellular activities) family which has various functions in cell division, membrane fusion, as well as proteasome- and ER-associated degradation of proteins (Aker et al., 2007). With the use of FCS, it was shown that CDC48A hexamers are part of even larger complexes.

2.3 Imaging of PPI with other fluorescence techniques

Besides FRET/FLIM and FCS, enzyme complementation was also adopted for fluorescence imaging of PPI a decade ago (Spotts et al., 2002). A reporter technology based on the differential induction of β-lactamase (Bla) enzymatic activity was developed to function as a sensor for the interaction state of two target proteins within single neurons. Bla was split into two separate, complementary protein fragments which can be brought together by phosphorylation-dependent association of the kinase inducible domain of the cyclic adenosine monophosphate (cAMP) response element binding (CREB) protein and the KIX domain of the CREB binding protein (Spotts et al., 2002). Using an intracellular substrate whose fluorescence spectrum changes upon hydrolysis by Bla, time-lapse ratiometric

imaging measurements were achieved after association of CREB and CREB binding protein, which permits direct imaging of PPI in single cells with high signal discrimination.

To investigate the conformational changes of proteins in living cells when external force is applied, a genetically encoded fluorescent sensor was constructed and tested in a myosin-actin model system using the proximity imaging (PRIM) technique, which detects spectral changes of two GFP molecules that are in direct contact (Iwai and Uyeda, 2008). The developed PRIM-based strain sensor module (PriSSM), consisted of the tandem fusion of a normal and circularly permuted GFP, was inserted between two motor domains of dictyostelium myosin II to study the effect of strain. It was suggested that this technology may provide a general approach for studying force-induced protein conformational changes in cells.

2.4 A brief summary of fluorescence imaging of PPI

The FRET/FLIM technique can be used as a versatile tool to characterize the spatial distribution of various proteins and detect/quantify PPI in a living cell, which can measure intermolecular FRET through quite sophisticated mathematical algorithms. However, no in vivo fluorescence imaging of PPI has been reported so far since these techniques (in particular FP-based) cannot be readily used for in vivo imaging applications due to several major limitations.

First, FRET-based techniques require the use of incident light to activate the donor protein. Given that the excitation wavelength is typically in the green range, little excitation light will travel through tissue since most tissues have strong light absorption/attenuation below a wavelength of 600 nm (Frangioni, 2003). Therefore these techniques are intrinsically not suitable for non-invasive imaging studies in live animals. Second, there is strong auto-fluorescence signal from animal tissue which confounds the interpretation of the imaging data. Third, the sensitivity of fluorescence imaging is not very high. Fourth, the relative molar ratios of the FRET donor/acceptor pair are not always 1:1, which can cause significant problems in calibration, detection, and quantification, especially when the situation is exacerbated in vivo when compared to cell-based studies. Lastly, there is significant photobleaching when the FPs are exposed to excitation light for a prolonged period.

3. Bioluminescence imaging (BLI) of PPI

Because of very low background signal and high sensitivity, BLI can be a more suitable technique for in vivo imaging of PPI than fluorescence imaging. The fact that no additional excitation light will be needed in BLI is highly advantageous for reducing the background signal. Two major strategies have been adopted for BLI of PPI: bioluminescence resonance energy transfer (BRET) and enzyme complementation.

3.1 Imaging of PPI with BRET

BRET displays similar characteristics as FRET except the donor is a bioluminescent protein, typically a luciferase, which requires the presence of small molecule substrates but not incident light. Similar to FRET, BRET is also a quantum process in which energy is transferred over a distance, usually < 10 nm, from the donor (e.g. luciferase) to a FP

(Villalobos et al., 2007). However, BRET offers many distinct advantages over FRET because of its higher quantum yield and better detection sensitivity.

As a popular technique for studying PPI in live cells, BRET is particularly suitable for real-time monitoring of such interactions. For example, many cellular signal transduction can be visualized by this technique, such as agonist-induced GPCR/β-arrestin interaction (Pfleger et al., 2006), calcium sensing receptor homodimer formation (Jensen et al., 2002), β2-adrenergic receptor dimerization (Angers et al., 2000), interaction of circadian clock proteins (Xu et al., 1999), etc. Since the potential for studying the modulation of such interactions by agonists, antagonists, inhibitors, dominant negative mutants, and co-expressed accessory proteins is tremendous, high-throughput BRET-based screening system is an ever-expanding area of interest for the pharmaceutical industry. However, imaging PPI with BRET in animal models is very challenging and only a few successful examples are available in the literature (Massoud et al., 2007; Villalobos et al., 2007).

In one early study, a cooled charge-coupled device (CCD) camera-based spectral imaging strategy enabled simultaneous visualization and quantitation of BRET signal from live cells and cells implanted in living mice, where renilla luciferase (RLuc) and its substrate were used as an energy donor and a mutant GFP was used as the acceptor (De and Gambhir, 2005). As a proof-of-principle, the donor and acceptor proteins were fused to FKBP12 and FRB respectively, which are known to interact only in the presence of the small molecule mediator rapamycin (Banaszynski et al., 2005; Choi et al., 1996). Mammalian cells expressing these fusion constructs were imaged using a cooled-CCD camera either directly from culture dishes or after implanting them into mice, where the specific BRET signal was determined by comparing the emission photon yields in the presence and absence of rapamycin. Such CCD camera-based imaging of BRET signal is very appealing since it can seamlessly bridge the gap between in vitro and in vivo studies, thus validating BRET as a powerful tool for interrogating and detecting PPI directly at limited depths in living mice.

Subsequently, a highly photon-efficient and self-illuminating fusion protein, which combines a mutant RFP (mOrange) and a mutant RLuc (RLuc8), was constructed to improve the BRET efficiency/signal (De et al., 2009). This new BRET fusion protein, termed as "BRET3", exhibited several-fold improvement in light intensity when compared with the previous BRET fusion proteins. In addition, BRET3 also exhibits red-shifted light output, which can allow for deeper tissue imaging in small animals. At single cell level, the BRET3 construct (which contains FKBP12 and FRB) was demonstrated to only exhibit BRET signal in the presence of rapamycin. With increased photon intensity, red-shifted light output and good spectral resolution (approximately 85 nm), it was suggested that BRET3-based assays will allow imaging of PPI using a single assay that is directly scalable from living cells to small animals.

Recently, further improvement on the BRET3 construct was reported, which was termed "BRET6" (Dragulescu-Andrasi et al., 2011). Red light-emitting BRET-based reporter systems were developed to allow for assaying PPI both in cell culture and in deep tissues of small animals (**Figure 3**). These BRET systems consist of the newly developed RLuc variants (RLuc8 and RLuc8.6, which serve as BRET donors) and two RFPs (TagRFP and TurboFP635, which serve as BRET acceptors). In addition to the native coelenterazine substrate for RLuc, a synthetic derivative (coelenterazine-v) was also used which further red-shifted the

emission maxima of RLuc by 35 nm. Ratiometric imaging of PPI in the presence of rapamycin-induced FKBP12-FRB association was demonstrated in both cultured cells and small animal tumor models.

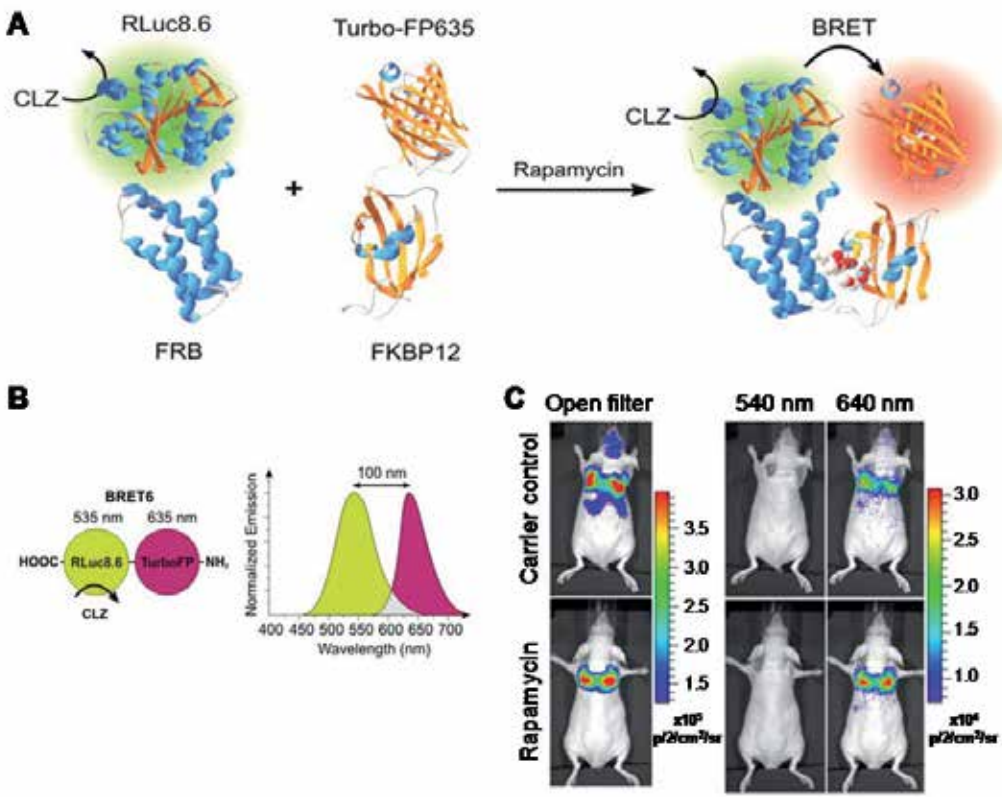


Fig. 3. Imaging of PPI with BRET6. **A.** Schematic illustration of the BRET6 construct for monitoring rapamycin-induced FRB-FKBP12 association. **B.** Schematic representation of the BRET6 fusion construct, the emission spectrum of the RLuc mutant, and the absorption spectrum of the acceptor protein. CLZ denotes coelenterazine (a substrate for RLuc). **C.** Bioluminescence images of cells stably expressing the BRET6 construct, accumulated in the lungs of nude mice after intravenous injection. Mice were also injected with both rapamycin (or control carrier which does not contain rapamycin) and CLZ before imaging. Adapted from (Dragulescu-Andrasi et al., 2011).

Currently, the number of BRET probes reported for the imaging of PPI is significantly lower when compared to FRET-based approaches. Much future work needs to be devoted to BRET-based imaging of PPI. The strategy of combining a fluorescent and a bioluminescent reporter to generate self-illuminated reporter proteins is advantageous to overcome the common problems associated with in vivo fluorescent imaging of PPI. As a genetically encodable approach for ratiometric imaging of PPI in cells and living subjects, light attenuation by tissue is the major challenge for ratiometric analysis of PPI with a BRET system. Since light attenuation varies with the wavelength of the emitted photons

and the tissue depth, red-shifted luciferases and FPs are clearly preferred choices. Meanwhile, consistency of the BRET ratio in different mice should also be monitored carefully to ensure sufficient spatial control to retain the ratiometric characteristics of a BRET sensor.

3.2 Imaging of PPI with complementation of split enzyme

Enzyme complementation assay depends on the division of a reporter enzyme (e.g. luciferase) into two separate inactive components that can regain function upon association (Massoud et al., 2007). When the two enzyme fragments are each fused to two interacting proteins, the enzyme can be reactivated upon PPI. For in vivo BLI applications, the split firefly luciferase (fLuc) with small overlapping sequences is a suitable choice because it consistently yields strong signal and excellent inducible complementation by a variety of PPIs. The reaction kinetics and ease of delivery of the substrate, D-luciferin, also allows for facile application of this technique in BLI assays. Besides fLuc, RLuc has also been investigated for BLI of PPI. However, although the split RLuc system functions quite efficiently, one major limitation of RLuc-based assay is its substrate, coelenterazine, which exhibits poor reaction profile for long-term kinetic experiments. In addition, the hydrophobicity of the molecule also makes it difficult to use for in vivo applications.

The first report on non-invasive BLI of PPI in living subjects based on a split luciferase was achieved a decade ago (Paulmurugan et al., 2002). In this study, split fLuc was designed and constructed for both intein-mediated reconstitution and complementation, where the two fLuc fragments could be brought together by the strong interaction between two proteins, MyoD and Id, both of which are members of the helix-loop-helix family of nuclear proteins. As a demonstration of the proof-of-principle, cells transiently transfected with the split reporter gene construct were used for imaging MyoD-Id interactions, both in cell culture and in cells implanted into living mice.

In a subsequent study, the split fLuc strategy was employed for imaging of PPI in hypoxia (Choi et al., 2008). HIF-1α is well known to regulate the activation of genes that promote malignant progression (Koh et al., 2010). HIF-1α is hydroxylated on prolines 402 and 564 under normoxia, which is targeted for ubiquitin-mediated degradation by interacting with the von Hippel-Lindau protein complex (pVHL). To study the interaction between HIF-1α and pVHL, the split fLuc-based system was used where HIF-1α and pVHL were fused to the amino-terminal and carboxy-terminal fragments of fLuc, respectively. Hydroxylation-dependent interaction between HIF-1α and pVHL led to complementation of the two fLuc fragments, resulting in bioluminescence in vitro and in vivo. Complementation-based bioluminescence was diminished when mutant pVHL with decreased binding affinity for HIF-1α was used. This strategy represents a useful approach for studying PPI involved in the regulation of protein degradation. In another study, split fLuc was also used for investigating epidermal growth factor (EGF)-induced Ras/Raf-1 interaction in mammalian cells (Kanno et al., 2006).

Similar strategy has been adopted to develop an inducible split RLuc-based bioluminescence assay for quantitative measurement of real time PPI in mammalian cells (Paulmurugan and Gambhir, 2003). In a follow-up study, the split RLuc construct was used to evaluate drug-modulated PPI in a cancer model in living mice (**Figure 4**)

(Paulmurugan et al., 2004). The heterodimerization of FRB and FKBP12, mediated by rapamycin, was also utilized in this study. The concentration of rapamycin needed for efficient dimerization, as well as the amount of ascomycin (a competitive binder of rapamycin) required for dimerization inhibition, were investigated. These studies demonstrated that such split reporter-based strategies can be used to efficiently screen small molecule drugs that modulate PPI, and further evaluate the effect of the drugs in living animals.

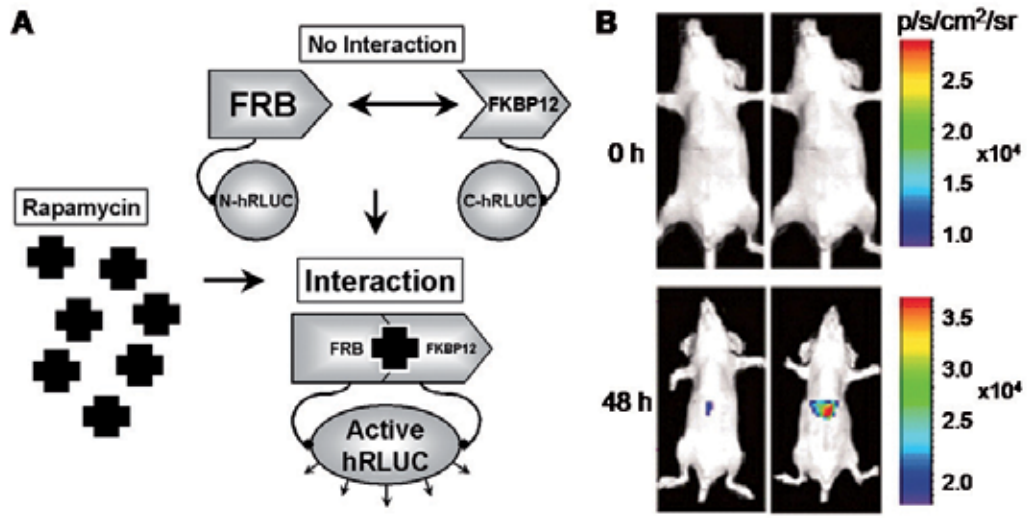


Fig. 4. In vivo imaging of drug-modulated PPI. **A.** Schematic diagram of rapamycin-mediated complementation of the two fragments of synthetic renilla luciferase (hRLUC). **B.** Non-invasive imaging of PPI in living mice, intravenously injected with human 293T embryonic kidney cancer cells that were transiently co-transfected with both split constructs. Mice not receiving rapamycin (left) showed only background signal, whereas the animals receiving repeated injections of rapamycin emitted higher signals originating from the 293T cells in the liver (right). Adapted from (Paulmurugan et al., 2004).

Homodimeric PPI, potent regulators of cellular functions and particularly challenging to study in vivo, can also be visualized by the split RLuc strategy. Split RLuc complementation-based bioluminescence assay was used to study the homodimerization of herpes simplex virus type 1 thymidine kinase (HSV1-TK) in mammalian cells and in living mice (Massoud et al., 2004). Homodimerization of HSV1-TK chimeras containing the N-terminal or C-terminal fragments of RLuc in the upstream and downstream positions, respectively, was visualized and quantified. A mutant of HSV1-TK was used to confirm the specificity of the RLuc complementation signal from HSV1-TK homodimerization. This generalizable assay to screen for molecules that promote or disrupt ubiquitous homodimeric PPI can not only serve as an invaluable tool to understand the biological signaling networks,

but will also be useful in drug discovery/validation in live animal disease models. In a cell-based study, the split RLuc strategy was shown to be useful beyond the visualization and confirmation of the existence of PPI. It also helped in identifying the critical amino acid residues involved in a specific PPI (Jiang et al., 2010).

Besides fLuc and RLuc complementation, split click beetle luciferase has been used to study the interaction between GPCR and β-arrestin (Misawa et al., 2010), whereas split Gaussia luciferase has been employed to image the interaction between calmodulin and other proteins (Kim et al., 2009). However, neither of these split luciferases has been demonstrated for in vivo visualization of PPI. Other split enzymes have also been explored for the imaging of PPI, such as the use of split β-gal for BLI of GPCR interactions in vivo (von Degenfeld et al., 2007). Currently, there is a paucity of sensitive and specific methods for quantitative comparison of the pharmacological properties of GPCRs in physiological and/or pathological settings in live animals. In this study, low affinity and reversible β-gal complementation was developed to quantify GPCR activation via interaction with β-arrestin, which enabled real time BLI of GPCR activity in live animals with high sensitivity and specificity (von Degenfeld et al., 2007). Imaging was achieved by using a recently developed luminescent β-gal substrate, which is a caged luciferin molecule that can be recognized by fLuc to generate light only after it has been cleaved by β-gal (Wehrman et al., 2006). Following implantation of the cells into mice, it was possible to monitor pharmacological GPCR activation and inhibition in their physiological context by non-invasive BLI, suggesting that this technology may have unique advantages to enable novel applications in the functional investigation of GPCR modulation in biological research and drug discovery.

4. PET imaging of PPI

Typically, PPI represents a low-level biological event and is therefore very challenging to detect, locate, and image in intact living subjects. When compared with BLI and fluorescence imaging, PET possesses very high sensitivity, while being quantitative and tomographic (Massoud and Gambhir, 2003). In addition, it is one of the few non-invasive imaging techniques that can be applied in humans for non-invasive monitoring of reporter gene expression (Kang and Chung, 2008). Although PET has enormous potential in imaging complex biological events such as PPI, to date only one example of PET imaging of PPI has been reported (Massoud et al., 2010).

The PET reported gene HSV1-TK was molecularly engineered and cleaved between Thr265 and Ala266, where the fragments were used in a protein-fragment complementation assay to quantify as well as to non-invasively image PPI in mammalian cells and living mice (Massoud et al., 2010). It was found that a point mutation (V119C) could be introduced to markedly enhance the HSV1-TK complementation modulated by several different PPIs such as the rapamycin-mediated FKBP12- FRB, HIF-1α-pVHL, etc. In vivo PET imaging of the FKBP12-FRB interaction modulated through rapamycin was successfully achieved (**Figure 5**). Future applications of this unique split HSV1-TK strategy are potentially far reaching, including accurate monitoring of immune and stem cell therapies, as well as allowing for fully quantitative and tomographic PET localization of PPI in preclinical small and large animal models of various diseases.

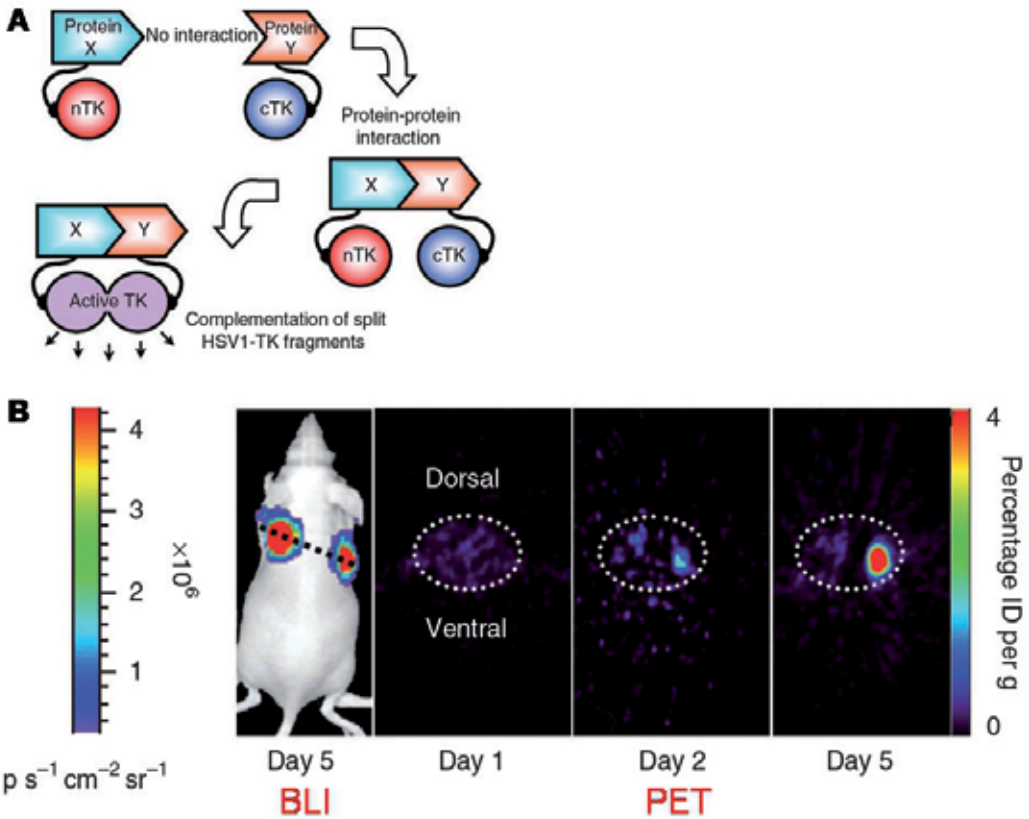


Fig. 5. Non-invasive PET imaging of PPI. **A.** Schematic diagram showing the use of split HSV1-TK to monitor the hypothetical X-Y heterodimeric PPI. **B.** Transaxial PET images of a mouse implanted subcutaneously with mock-transfected 293T cells (left) and 293T cells stably expressing both split constructs of HSV1-TK each fused to FRB and FKBP12 respectively (right). The serial images at different days were acquired after injection of the PET reporter probe for HSV1-TK (i.e. ^{18}F-FHBG). A BLI image of the mouse is also shown to delineate the two tumors. Adapted from (Massoud et al., 2010).

5. Conclusion

The interactions of specific cellular proteins form the basis of a wide variety of biological processes, including many signal transduction and hormone activation pathways involved in maintaining important biological functions. Accurate measurement of PPI can significantly help in deciphering the genetic and proteomic code. The tremendous complexity of cellular events requires assays that can measure different types of PPIs using an array of different methods. Molecular imaging, an extremely powerful tool to study molecular events in living subjects, can provide invaluable information and insight in elucidating the process of various PPIs.

To date, the major molecular imaging modalities used for visualization of PPI include fluorescence imaging (not suitable for in vivo studies), BLI, and PET imaging. All these techniques require extensive efforts in protein engineering due to the complex and challenging nature of imaging PPI in living cells and animals. Particularly for split reporter-based strategies, intensive efforts are needed to obtain better functioning split reporters that exhibit efficient PPI-induced complementation but not self-complementation. At the same time, sufficiently high reporter activity needs to be maintained upon PPI-induced complementation. For in vivo imaging of PPI, PET serves as a better choice over BLI and fluorescence due to its superb sensitivity, excellent tissue penetration, high quantification accuracy, and potential for clinical translation.

Future work on the imaging of PPI may include the design of second-generation complementation reporters with improved signal-to-noise ratios, inducibility, and red-shifted spectral properties for more wide spread use in vivo. The ideal reporter for imaging of PPI should not only serve as an "on/off" signal, but also give a graduated and quantitative response with minimal background signal and excellent induced signal output. Lastly, since no single imaging modality is perfect, combination of different imaging techniques to study the same PPI may provide complementary information.

6. References

Aker, J.; Hesselink, R.; Engel, R.; Karlova, R.; Borst, J.W.; Visser, A.J. & de Vries, S.C. (2007). In vivo hexamerization and characterization of the Arabidopsis AAA ATPase CDC48A complex using forster resonance energy transfer-fluorescence lifetime imaging microscopy and fluorescence correlation spectroscopy. *Plant Physiol*, 145, 339-350.

Albertazzi, L.; Arosio, D.; Marchetti, L.; Ricci, F. & Beltram, F. (2009). Quantitative FRET analysis with the EGFP-mCherry fluorescent protein pair. *Photochem Photobiol*, 85, 287-297.

Angers, S.; Salahpour, A.; Joly, E.; Hilairet, S.; Chelsky, D.; Dennis, M. & Bouvier, M. (2000). Detection of beta 2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *Proc Natl Acad Sci USA*, 97, 3684-3689.

Banaszynski, L.A.; Liu, C.W. & Wandless, T.J. (2005). Characterization of the FKBP.rapamycin.FRB ternary complex. *J Am Chem Soc*, 127, 4715-4721.

Chen, J. & Irudayaraj, J. (2010). Fluorescence lifetime cross correlation spectroscopy resolves EGFR and antagonist interaction in live cells. *Anal Chem*, 82, 6415-6421.

Choi, C.Y.; Chan, D.A.; Paulmurugan, R.; Sutphin, P.D.; Le, Q.T.; Koong, A.C.; Zundel, W.; Gambhir, S.S. & Giaccia, A.J. (2008). Molecular imaging of hypoxia-inducible factor 1 alpha and von Hippel-Lindau interaction in mice. *Mol Imaging*, 7, 139-146.

Choi, J.; Chen, J.; Schreiber, S.L. & Clardy, J. (1996). Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP. *Science*, 273, 239-242.

De, A. & Gambhir, S.S. (2005). Noninvasive imaging of protein-protein interactions from live cells and living subjects using bioluminescence resonance energy transfer. *FASEB J*, 19, 2017-2019.

De, A.; Ray, P.; Loening, A.M. & Gambhir, S.S. (2009). BRET3: a red-shifted bioluminescence resonance energy transfer (BRET)-based integrated platform for imaging protein-protein interactions from single live cells and living animals. *FASEB J*, 23, 2702-2709.

Doyle, M.L. (1997). Characterization of binding interactions by isothermal titration calorimetry. *Curr Opin Biotechnol*, 8, 31-35.

Dragulescu-Andrasi, A.; Chan, C.T.; De, A.; Massoud, T.F. & Gambhir, S.S. (2011). Bioluminescence resonance energy transfer (BRET) imaging of protein-protein interactions within deep tissues of living subjects. *Proc Natl Acad Sci USA*, 108, 12060-12065.

Elson, E.L. (2004). Quick tour of fluorescence correlation spectroscopy from its inception. *J Biomed Opt*, 9, 857-864.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340, 245-246.

Frangioni, J.V. (2003). *In vivo* near-infrared fluorescence imaging. *Curr Opin Chem Biol*, 7, 626-634.

Fruhwirth, G.O.; Ameer-Beg, S.; Cook, R.; Watson, T.; Ng, T. & Festy, F. (2010). Fluorescence lifetime endoscopy using TCSPC for the measurement of FRET in live cells. *Opt Express*, 18, 11148-11158.

Giepmans, B.N.; Adams, S.R.; Ellisman, M.H. & Tsien, R.Y. (2006). The fluorescent toolbox for assessing protein location and function. *Science*, 312, 217-224.

Hansen, J.C.; Lebowitz, J. & Demeler, B. (1994). Analytical ultracentrifugation of complex macromolecular systems. *Biochemistry*, 33, 13155-13163.

Haustein, E. & Schwille, P. (2007). Fluorescence correlation spectroscopy: novel variations of an established technique. *Annu Rev Biophys Biomol Struct*, 36, 151-169.

Hink, M.A.; Bisselin, T. & Visser, A.J. (2002). Imaging protein-protein interactions in living cells. *Plant Mol Biol*, 50, 871-883.

Hoffmann, C.; Gaietta, G.; Bunemann, M.; Adams, S.R.; Oberdorff-Maass, S.; Behr, B.; Vilardaga, J.P.; Tsien, R.Y.; Ellisman, M.H. & Lohse, M.J. (2005). A FlAsH-based FRET approach to determine G protein-coupled receptor activation in living cells. *Nat Methods*, 2, 171-176.

Iwai, S. & Uyeda, T.Q. (2008). Visualizing myosin-actin interaction with a genetically-encoded fluorescent strain sensor. *Proc Natl Acad Sci USA*, 105, 16882-16887.

Jares-Erijman, E.A. & Jovin, T.M. (2003). FRET imaging. *Nat Biotechnol*, 21, 1387-1395.

Jensen, A.A.; Hansen, J.L.; Sheikh, S.P. & Brauner-Osborne, H. (2002). Probing intermolecular protein-protein interactions in the calcium-sensing receptor homodimer using bioluminescence resonance energy transfer (BRET). *Eur J Biochem*, 269, 5076-5087.

Jiang, Y.; Bernard, D.; Yu, Y.; Xie, Y.; Zhang, T.; Li, Y.; Burnett, J.P.; Fu, X.; Wang, S. & Sun, D. (2010). Split Renilla luciferase protein fragment-assisted complementation (SRL-PFAC) to characterize Hsp90-Cdc37 complex and identify critical residues in protein/protein interactions. *J Biol Chem*, 285, 21023-21036.

Kang, J.H. & Chung, J.K. (2008). Molecular-genetic imaging based on reporter gene expression. *J Nucl Med*, 49 Suppl 2, 164S-179S.

Kanno, A.; Ozawa, T. & Umezawa, Y. (2006). Intein-mediated reporter gene assay for detecting protein-protein interactions in living mammalian cells. *Anal Chem*, 78, 556-560.

Kim, S.B.; Sato, M. & Tao, H. (2009). Split Gaussia luciferase-based bioluminescence template for tracing protein dynamics in living cells. *Anal Chem*, 81, 67-74.

Koh, M.Y.; Spivak-Kroizman, T.R. & Powis, G. (2010). HIF-1alpha and cancer therapy. *Recent Results Cancer Res*, 180, 15-34.

Konietzny, R.; Konig, A.; Wotzlaw, C.; Bernadini, A.; Berchner-Pfannschmidt, U. & Fandrey, J. (2009). Molecular imaging: into in vivo interaction of HIF-1alpha and HIF-2alpha with ARNT. *Ann N Y Acad Sci*, 1177, 74-81.

Lakey, J.H. & Raggett, E.M. (1998). Measuring protein-protein interactions. *Curr Opin Struct Biol*, 8, 119-123.

Mankoff, D.A. (2007). A definition of molecular imaging. *J Nucl Med*, 48, 18N, 21N.

Massoud, T.F. & Gambhir, S.S. (2003). Molecular imaging in living subjects: seeing fundamental biological processes in a new light. *Genes Dev*, 17, 545-580.

Massoud, T.F.; Paulmurugan, R.; De, A.; Ray, P. & Gambhir, S.S. (2007). Reporter gene imaging of protein-protein interactions in living subjects. *Curr Opin Biotechnol*, 18, 31-37.

Massoud, T.F.; Paulmurugan, R. & Gambhir, S.S. (2004). Molecular imaging of homodimeric protein-protein interactions in living subjects. *FASEB J*, 18, 1105-1107.

Massoud, T.F.; Paulmurugan, R. & Gambhir, S.S. (2010). A molecularly engineered split reporter for imaging protein-protein interactions with positron emission tomography. *Nat Med*, 16, 921-926.

Metelev, V.; Zhang, S.; Tabatadze, D. & Bogdanov, A. (2011). Hairpin-like fluorescent probe for imaging of NF-kappaB transcription factor activity. *Bioconjug Chem*, 22, 759-765.

Misawa, N.; Kafi, A.K.; Hattori, M.; Miura, K.; Masuda, K. & Ozawa, T. (2010). Rapid and high-sensitivity cell-based assays of protein-protein interactions using split click beetle luciferase complementation: an approach to the study of G-protein-coupled receptors. *Anal Chem*, 82, 2552-2560.

Paulmurugan, R. & Gambhir, S.S. (2003). Monitoring protein-protein interactions using split synthetic renilla luciferase protein-fragment-assisted complementation. *Anal Chem*, 75, 1584-1589.

Paulmurugan, R.; Massoud, T.F.; Huang, J. & Gambhir, S.S. (2004). Molecular imaging of drug-modulated protein-protein interactions in living subjects. *Cancer Res*, 64, 2113-2119.

Paulmurugan, R.; Umezawa, Y. & Gambhir, S.S. (2002). Noninvasive imaging of protein-protein interactions in living subjects by using reporter protein complementation and reconstitution strategies. *Proc Natl Acad Sci USA*, 99, 15608-15613.

Pelet, S.; Previte, M.J. & So, P.T. (2006). Comparing the quantification of Forster resonance energy transfer measurement accuracies based on intensity, spectral, and lifetime imaging. *J Biomed Opt*, 11, 34017.

Pfleger, K.D.; Dromey, J.R.; Dalrymple, M.B.; Lim, E.M.; Thomas, W.G. & Eidne, K.A. (2006). Extended bioluminescence resonance energy transfer (eBRET) for monitoring prolonged protein-protein interactions in live cells. *Cell Signal*, 18, 1664-1670.

Phizicky, E.M. & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59, 94-123.

Seegar, T. & Barton, W. (2010). Imaging protein-protein interactions in vivo. *J Vis Exp*, pii: 2149.

Spotts, J.M.; Dolmetsch, R.E. & Greenberg, M.E. (2002). Time-lapse imaging of a dynamic phosphorylation-dependent protein-protein interaction in mammalian cells. *Proc Natl Acad Sci USA*, 99, 15142-15147.

Timm, T.; von Kries, J.P.; Li, X.; Zempel, H.; Mandelkow, E. & Mandelkow, E.M. (2011). Microtubule affinity regulating kinase (MARK) activity in living neurons examined by a genetically encoded FRET/FLIM based biosensor: Inhibitors with therapeutic potential. *J Biol Chem*, Epub ahead of print.

Tsien, R.Y. (2009). Indicators based on fluorescence resonance energy transfer (FRET). *Cold Spring Harb Protoc*, 2009, pdb top57.

Valdar, W.S. & Thornton, J.M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42, 108-124.

van Roessel, P. & Brand, A.H. (2002). Imaging into the future: visualizing gene expression and protein interactions with fluorescent proteins. *Nat Cell Biol*, 4, E15-20.

Villalobos, V.; Naik, S. & Piwnica-Worms, D. (2007). Current state of imaging protein-protein interactions in vivo with genetically encoded reporters. *Annu Rev Biomed Eng*, 9, 321-349.

von Degenfeld, G.; Wehrman, T.S.; Hammer, M.M. & Blau, H.M. (2007). A universal technology for monitoring G-protein-coupled receptor activation in vitro and noninvasively in live animals. *FASEB J*, 21, 3819-3826.

Wehrman, T.; Kleaveland, B.; Her, J.H.; Balint, R.F. & Blau, H.M. (2002). Protein-protein interactions monitored in mammalian cells via complementation of beta -lactamase enzyme fragments. *Proc Natl Acad Sci USA*, 99, 3469-3474.

Wehrman, T.S.; von Degenfeld, G.; Krutzik, P.O.; Nolan, G.P. & Blau, H.M. (2006). Luminescent imaging of beta-galactosidase activity in living subjects using sequential reporter-enzyme luminescence. *Nat Methods*, 3, 295-301.

Williams, N.E. (2000). Immunoprecipitation procedures. *Methods Cell Biol*, 62, 449-453.

Xu, Y.; Piston, D.W. & Johnson, C.H. (1999). A bioluminescence resonance energy transfer (BRET) system: application to interacting circadian clock proteins. *Proc Natl Acad Sci USA*, 96, 151-156.

Yan, Y. & Marriott, G. (2003). Analysis of protein interactions using fluorescence technologies. *Curr Opin Chem Biol*, 7, 635-640.

Zhang, S.; Metelev, V.; Tabatadze, D.; Zamecnik, P.C. & Bogdanov, A., Jr. (2008). Fluorescence resonance energy transfer in near-infrared fluorescent oligonucleotide probes for detecting protein-DNA interactions. *Proc Natl Acad Sci USA*, 105, 4156-4161.

Zhang, S.; Wang, G.; Fernig, D.G.; Rudland, P.S.; Webb, S.E.; Barraclough, R. & Martin-Fernandez, M. (2005). Interaction of metastasis-inducing S100A4 protein in vivo by fluorescence lifetime imaging microscopy. *Eur Biophys J*, 34, 19-27.

Zhong, W.; Wu, M.; Chang, C.W.; Merrick, K.A.; Merajver, S.D. & Mycek, M.A. (2007). Picosecond-resolution fluorescence lifetime imaging microscopy: a useful tool for sensing molecular interactions in vivo via FRET. *Opt Express*, 15, 18220-18235.

16

NMR Investigations on Ruggedness of Native State Energy Landscape in Folded Proteins

Poluri Maruthi Krishna Mohan
Department of Chemistry & Chemical Biology,
Rutgers University, New Jersey,
USA

1. Introduction

The ability of proteins to adopt their functional, highly structured states in the intracellular environment during and after its synthesis is one of the most remarkable evolutionary achievements of biology. Deciphering the code of protein self-organization process has been an intellectual challenge for scientists over the past few decades. Although the structure-function paradigm about folded structures and functions remains valid, the role of internal dynamics and conformational fluctuations in protein function is becoming increasingly evident (Bhabha *et al* 2011; Boehr *et al* 2006; Eisenmesser *et al* 2005; Fraser *et al* 2009; Mittermaier and Kay 2006; Parak 2003b; Popovych *et al* 2006; Tzeng and Kalodimos 2009; Whitten *et al* 2005). Further, recent structural and genomic data have clearly shown that not all proteins have unique folded structures under normal physiological conditions. Hence, the way a protein exists, is bound to have a profound effect on its function.

A complete understanding of protein folding process requires characterization of all the species populating along the folding coordinate, these include the unfolded state, the partially folded intermediate states, low energy excited states and the fully folded native state. The most recent and widely accepted model is the 'funnel view' of protein folding (Bryngelson *et al* 1995; Dill and Chan 1997; Onuchic *et al* 1997; Shoemaker *et al* 1999; Wolynes 2005) also known as the 'Energy landscape model' **(Fig. 1)**, which is inclusive of the earlier concepts of 'folding pathways'. According to this model protein folding is a parallel, diffusion-like motion of conformational ensemble on the energy landscape biased towards the native state. This model is free from the Levinthal paradox (Dill and Chan 1997; Levinthal C 1969) as it envisages the process of reaching a global minimum in free energy as a rapid process occurring by multiple routes on a funnel like energy landscape **(Fig. 1)**. This view focuses on the rapid decrease of the conformational heterogeneity in the course of the folding reaction and is based on a statistical description of a protein's potential surface (Wolynes *et al* 1995; Wolynes 2005). The depth in the funnel represents the free energy of the polypeptide chain in fixed conformations and the width indicates the chain entropy **(Fig. 1)**. The funnel becomes narrower in the lower energy region because of the low chain entropy. The broad end of the funnel reflects the heterogeneous unfolded state, while the narrow end represents the supposedly homogeneous native state (Dill and Chan 1997; Dobson and Karplus 1999; Dyson and Wright 2005). Different members of the ensemble

may fold/unfold along independent pathways and their energy profiles could be different. Protein folding theories start from the unfolded state **(Fig. 1)** and encompass a range of topologies like the pre-molten globule, the molten globule and various other ordered or disordered forms as the protein folds down the funnel (Dunker *et al* 2002; Uversky 2002; Uversky 2003).

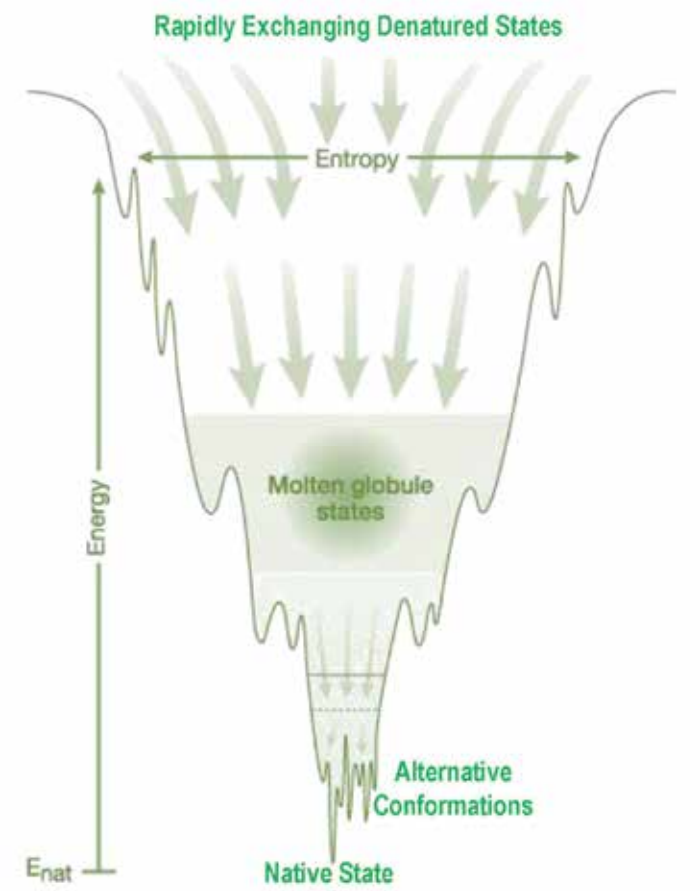


Fig. 1. A schematic energy landscape view of protein folding: The surface of the funnel represents a whole range from the multitude of denatured conformations to the unique native structure (Dill and Chan 1997). The ordered state is the natively folded structure of a protein that has a well defined secondary and tertiary structure. Alternative conformations are higher energy native state conformations and contain all the secondary and tertiary structural characteristics of folded state. Molten globule states are intermediates in the protein folding pathway with compact structures that exhibit a high content of secondary structure, nonspecific tertiary structure, and significant structural flexibility. Random coils are highly unstructured protein denatured states.

In a living cell, a polypeptide chain chooses between three potential fates - functional folding, potentially deadly misfolding and mysterious non-folding (Dobson 2003). This choice is dictated by the peculiarities of amino acid sequence and/or by the pressure of

environmental factors. The biological function of a protein arises as a result of interplay between specific conformational forms, namely, native state (ordered forms), low energy excited states, molten globules, pre-molten globules, and denatured state (random coils). In view of this, it will not be an exaggeration to assume an ensemble existence of all these states at any particular time, their relative abundance being governed by basic thermodynamics. Upon ligand binding or some signaling modification, concentration of one state may increase at the expense of the others. This can explain the fast regulatory steps involved in various biological functions.

Much of structural biology of proteins is so focused on studies of native state, providing detailed atomic descriptions and coordinates of static three-dimensional (3D) structures. A large body of evidence using a diverse spectrum of biophysical methods clearly establishes that proteins are dynamic over a broad range of timescales and such dynamics play critical roles in various biological processes, such as: initial formation of encounter complexes in macromolecular association, target searching in specific protein-protein/protein-DNA recognition, conformational preferences in ligand binding, conformational transitions associated with allostery, the course of enzyme catalysis, intermediates along the protein folding pathway, and early events in self-assembly processes (Bai *et al* 1995; Boehr *et al* 2006; Clore 2011; Dunker *et al* 2002; Eisenmesser *et al* 2005; Feher and Cavanagh 1999; Fraser *et al* 2009; Kitahara *et al* 2005; Korzhnev *et al* 2003; Kumar *et al* 2007; Lambers *et al* 2006; Mohan *et al* 2006; Piana *et al* 2002; Popovych *et al* 2006; Tang *et al* 2008; Villali and Kern 2010). The amplitudes and the timescales of motion that characterize the dynamics of a protein under a given set of conditions can be understood in terms of an 'energy landscape' as described above. Ground-state conformers that occupy the bottom of the energy landscape funnel and are separated from other conformational states by very small kinetic barriers that are easily overcome by thermal energy form the basis of structural studies by NMR (Nuclear Magnetic Resonance Spectroscopy) and X-ray diffraction for last few decades.

In general, the dynamic phenomenon involves the inter-conversion between ground state conformers with higher energy structures known as 'excited states'. The populations of these low-energy excited states/near native states/alternative conformations at equilibrium are very sparse and their lifetimes are short. Moreover, these transient states arising from rare but rapid excursions between the global free energy minimum and higher free energy local minima are extremely challenging to study at atomic resolution under equilibrium conditions since they are effectively invisible to most structural and biophysical techniques including crystallography and conventional NMR spectroscopy (Bhabha *et al* 2011; Boehr *et al* 2006; Eisenmesser *et al* 2005; Fraser *et al* 2009; Mittermaier and Kay 2006; Popovych *et al* 2006; Tzeng and Kalodimos 2009; Whitten *et al* 2005; Clore and Iwahara 2009; Clore 2011). However, a complete understanding of the conformational fluctuations these bio-molecules undergo is essential to gain an insight into their biochemical and biophysical properties. Hence, it is critical to characterize the structural ensembles that describe these functionally important states and the mechanisms by which they interconvert with the ground-state conformers.

Recent developments in NMR, however, have rendered short-lived, sparsely populated states accessible to spectroscopic analysis, yielding considerable insights into their kinetics, thermodynamics, and structures. Over the past decade, new and powerful NMR approaches

such as paramagnetic relaxation enhancement (PRE) (Clore and Iwahara 2009; Clore 2011), relaxation dispersion (RD) (Boehr *et al* 2006; Mittermaier and Kay 2006; Tzeng and Kalodimos 2009) and non-linear temperature dependence of amide proton chemical shifts (Krishna Mohan *et al* 2008; Mohan *et al* 2008b; Tunnicliffe *et al* 2005; Williamson 2003) have emerged and significantly contributed to our understanding of the relationship between structure, dynamics and function of proteins with respect to the excited-state conformers that are sparsely populated and often exist transiently.

In the present chapter the theoretical basis of NMR approach for the curved temperature dependence of amide proton chemical shifts will be discussed in detail. Theoretical simulations will be presented to understand the nature and extent of curvature of the chemical shifts. Further, experimental studies performed till date on different protein systems will be reviewed to demonstrate the curved temperature dependence of amide proton chemical shifts as a tool to detect the low populated near native states/ alternative conformations of the protein residues. Moreover, the significance of these conformational fluctuations will be evaluated with regard to protein function and folding.

2. Theory of curved temperature dependence of amide proton chemical shifts

NMR chemical shift is a sensitive indicator of the environment and molecular conformation. In proteins ^{1}H, ^{13}C and ^{15}N chemical shifts are sensitive to protein secondary structures and are used to deduce the preliminary structural information (Schwarzinger *et al* 2000; Schwarzinger *et al* 2001; Wishart and Sykes 1994; Wishart *et al* 1995; Wüthrich K 1986). The temperature dependence of amide proton chemical shifts in globular proteins has been investigated for over more than three decades by many researchers and continues to be investigated even today (Anderson *et al* 1997; Baxter and Williamson 1997; Cierpicki and Otlewski 2001; Krishna Mohan *et al* 2008). The amide proton chemical shifts are directly proportional to bond magnetic anisotropy (σ^{ani}) and this is crucially dependent on H-bonding, either intramolecular or intermolecular. In the former case, the carbonyl groups, the H-bond acceptors play a crucial role. The bond magnetic anisotropy is proportional to r^{-3} where r is the distance between the affected amide proton and the centre of the bond magnetic anisotropy, which lies close to the oxygen atom in the carbonyl groups (Krishna Mohan *et al* 2008). In case of solvent accessible groups, H-bonding with solvent molecules influences the amide proton chemical shifts. Thus the amide proton chemical shift is critically dependent on the length of the H-bond the proton is engaged in.

When the temperature of the solution is raised, thermal fluctuations increase which results in an increase in the average distance between atoms; X-ray crystallographic studies at several temperatures (98 – 320 K) on ribonuclease-A indicated that the protein volume increases linearly with temperature to an extent of about 0.4% per 100 K (Tilton, Jr. *et al* 1992). Such an increase in the distance between the atoms participating in a H-bond results in weakening of the H-bond. Consequently, chemical shifts of most amide protons move up field when the temperature is increased. Since bond magnetic anisotropy (σ^{ani}) is proportional to r^{-3} and molecular volume (V) is proportional to r^{3}, there is an inverse relationship of their variation with temperature ($(\sigma^{ani}) \alpha 1/V$). However, over a small temperature range, (σ^{ani}) may appear to decrease linearly with temperature, and

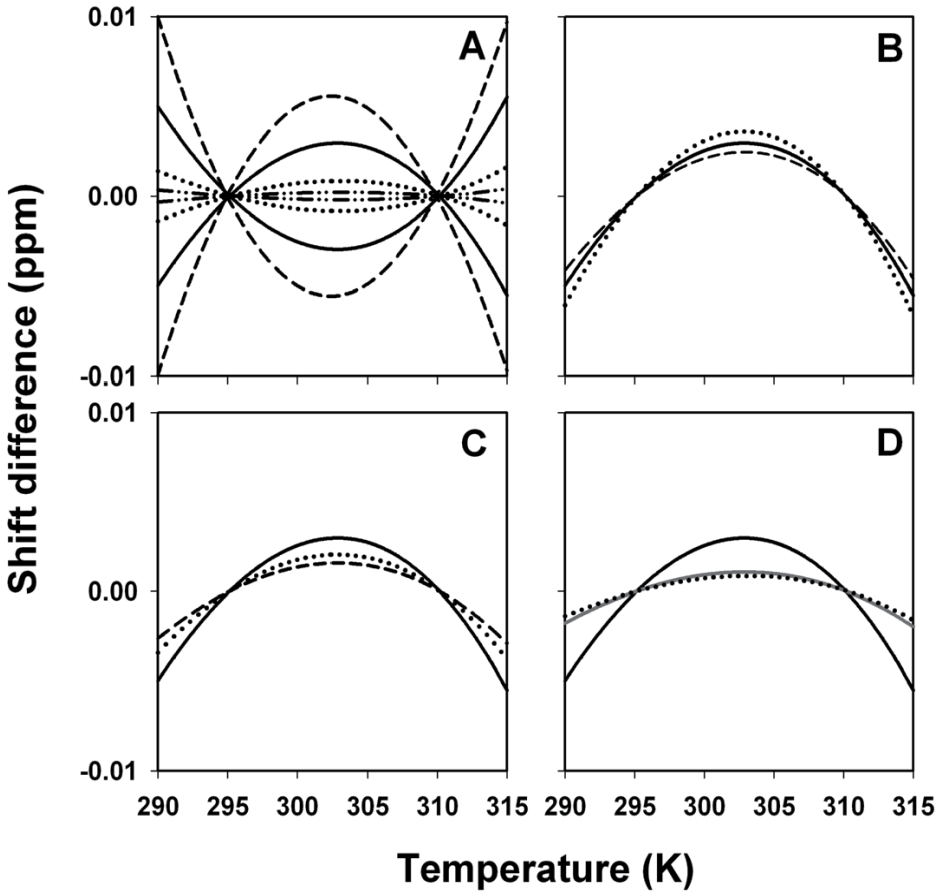


Fig. 2. Simulations of the dependence of H^N chemical shift variation with temperature (290 K – 315 K). In all the calculations shown chemical shifts are calculated and fitted to a straight line. Then the deviations from linearity are used to derive the residual curvatures (Krishna Mohan *et al* 2008; Williamson 2003). **(A)** Different curves show the dependence on free energy difference between the native and the higher energy alternate state: ΔG = 1 kcal/mol (dashed), ΔG = 2 kcal/mol (solid), ΔG = 3 kcal/mol (dotted) and ΔG = 4 kcal/mol (dash double dot dash); for these $T\Delta S$ at 298 K was fixed at 5.1 kcal/mol and ΔH varied as 6.1, 7.1, 8.1, and 9.1 kcal/mol respectively. The chemical shift and gradient parameters are: δ_1 = 8.5 ppm, δ_2 = 8.0 ppm, and g_1 = -2 ppb/K, g_2 = -7 ppb/K for convex shapes and δ_1 = 8.0 ppm, δ_2 = 8.5 ppm, and g_1 = -7 ppb/K, g_2 = - 2 ppb/K for concave shapes. **(B)** The solid curve is the same as in '**A**' (ΔG = 2 kcal/mol); Dashed curve, δ_1 = 8.1 ppm, δ_2 = 8.0 ppm, g_1 = -2 ppb/K, g_2 = -7 ppb/K, ΔG = 2 kcal/mol; dotted curve, δ_1 = 9.0 ppm, δ_2 = 8.0 ppm, g_1 = -2 ppb/K, g_2 = -7 ppb/K, ΔG = 2 kcal/mol **(C)** The solid curve is the same as in '**A**' (ΔG = 2 kcal/mol); dashed curve, δ_1 = 8.5 ppm, δ_2 = 8.0 ppm, g_1 = -2 ppb/K, g_2 = - 4 ppb/K, ΔG = 2 kcal/mol; dotted curve, δ_1 = 8.5 ppm, δ_2 = 8.0 ppm, g_1 = -4 ppb/K, g_2 = - 7 ppb/K, ΔG = 2 kcal/mol. **(D)** The solid black curve (ΔG = 2 kcal/mol) and the dotted curve (ΔG = 3 kcal/mol) are the same as in **'A'**; solid grey curve is for δ_1=8.1 ppm, δ_2=8.0 ppm, g_1 = -2 ppb/K, g_2 = -4 ppb/K with ΔG = 2 kcal/mol.

consequently amide proton chemical shifts would appear to vary linearly with temperature. In BPTI (basic pancreatic trypsin inhibitor) and lysozyme which are known to be extremely stable under a variety of extreme conditions, including temperature, it was indeed observed that the amide proton chemical shifts change linearly with temperature over the ranges, 279-359 K for BPTI and 278 – 328 K for Lysozyme (Baxter and Williamson 1997). Such measurements have been carried out on many other proteins (Cierpicki and Otlewski 2001; Cierpicki *et al* 2002) and the temperature coefficients or the gradients of temperature dependence of the amide protons have been found to span a wide range, -16 to + 4 ppb/K. For a strongly H-bonded amide this value is more positive than -4.5 ppb /K (Baxter and Williamson 1997). This is because the lengthening of the average H-bond distance will be greater for the intermolecular H-bond, such as those with bulk water, than for the intra-molecular H-bonds.

However, if the protein structure is not very rigid, as would be the case for many systems, the chemical shifts would also be influenced by local structural and dynamics changes, and then the temperature dependence of chemical shifts may deviate from linearity. Indeed, in certain situations the amide proton chemical shifts have been seen to be non linearly dependent on temperature, and this has been interpreted to indicate existence of alternative conformations the residues can access (Baxter *et al* 1998; Williamson 2003). Identification of such residues provides a description of the energy landscape of the protein in the native state. The observed curvatures can be theoretically deduced as described in the following paragraphs.

Consider a residue having two conformational states accessible to it i.e., a native state and a higher energy state. Following the discussion in the above paragraphs, each of them can be assumed to have a linear variation of chemical shift with temperature as, $\delta_1 = \delta^0_1 + g_1T$ and $\delta_2 = \delta^0_2 + g_2T$, where g_1 and g_2 are the gradients of temperature dependence, δ_1 and δ_2 are the chemical shifts of the native and the excited states respectively, and T is the temperature. If P_1 and P_2 are the corresponding populations of the native and the excited states, the observed chemical shift, δ_{obs}, of the amide proton will be given by,

$$\delta_{obs} = \delta_1 P_1 + \delta_2 P_2 \tag{1}$$

These populations depend on the free energy difference between the two states. If there are more states contributing, then the observed shift will be a weighted average over all the accessible states. It is this complex dependence of chemical shifts on many thermodynamic and other factors, which leads to non linear dependence of chemical shifts on temperature. To understand the influence of these factors, simulations of H^N chemical shift variation are performed with temperature in the range, 290 K – 315 K, using a two state model (Krishna Mohan *et al* 2008; Williamson 2003)

$$\delta_{obs} = \frac{(\delta^0{}_1 + g_1T) + [(\delta^0{}_2 + g_2T)e^{-(\Delta G/RT)}]}{1 + e^{-(\Delta G/RT)}} \tag{2}$$

where, ΔG is the free-energy difference between the two states, and $\Delta G = \Delta H - T\Delta S$, where, ΔH and ΔS are the enthalpy difference and the entropy difference respectively. The results of the simulations are shown in **Fig. 2**. **Fig. 2A** shows the curves for ΔG ranging from 1 - 4

kcal/mol keeping the gradients and chemical shifts of the native and the excited states constant. Here, it is worthwhile to note that in the chosen temperature range (290 K – 315 K) the curvature almost disappears above ΔG = 3 kcal/mol. **Fig. 2B** and **Fig. 2C** show the dependence of curvature on chemical shift differences and gradient differences respectively, between the native and the excited states, when ΔG is held constant (ΔG = 2 kcal/mol). In **Fig. 2B**, three values of δ_1: 8.5 (reference), 8.1 and 9.0 are considered, keeping the other parameters the same as in **Fig. 2A** for convex shape of curvature. Similarly, in **Fig. 2C**, three combinations of gradients: (g_1,g_2) (ppb/K) = (-2, -7) (reference), (-2, -4) and (-4, -7) are considered keeping the other parameters same as in **Fig. 2A** for convex shape of curvature. From these it is evident that neither the chemical shift difference nor the difference in gradients, by itself changes the curvature to a noticeable extent. A simulation carried out for a combination of changes in 'chemical shift difference' and 'gradient difference' (δ_1=8.1 ppm, δ_2=8.0 ppm, g_1 = -2 ppb/K, g_2 = -4 ppb/K) keeping the free energy constant (ΔG = 2 kcal/mol). This is shown by solid grey line in **Fig 2D**. Interestingly, this curve almost exactly overlaps with the curve for which ΔG = 3 kcal/mol in **Fig. 2A** which has a lower curvature compared to that of the curve with ΔG = 2 kcal/mol; for ease of comparison, the corresponding curve from **Fig. 2A** is redrawn in **Fig. 2D** as a dotted line. This clearly suggests that although the appearance of curvature confirms the presence of alternative states, the lack of curvature does not necessarily imply the absence of low energy excited states. These theoretical simulations will be of great help for interpreting the experimental results on temperature dependence of amide proton chemical shifts.

3. Investigations on native state ruggedness of complex protein systems

A deep well, the bottom of which corresponds to the native state would imply high stability of the native state **(Fig. 1)**. In contrast, a potential well with low lying excited states for the native state would be shallow, and this would have significant influence on the dynamics, structural adaptability, or susceptibility of the protein to various functions (Agarwal 2005; Boehr *et al* 2010; Eisenmesser *et al* 2005; Feher and Cavanagh 1999; Kitahara *et al* 2005; Korzhnev *et al* 2003; Parak 2003a; Piana *et al* 2002; Tobi and Bahar 2005). Application of small environmental perturbations such as small concentrations of chemical denaturants, change in pressure, pH change etc., is often useful to investigate the preferential sensitivities of different residues to external perturbations, while the protein itself remains entirely in the native state ensemble (Akasaka 2006; Baxter *et al* 1998; Chatterjee *et al* 2007; Kumar *et al* 2007; Mohan *et al* 2006; Piana *et al* 2002). In fact, these environment sensitive residues of polypeptide chains adopt unique 3D structures, they access various near native states which are structurally similar and energetically closer to the native state. These low populated alternative conformations dictate the ruggedness of the native structure and its biological function.

3.1 Differential native state conformational fluctuations in calcium sensor proteins

Calcium ion plays a crucial role in the regulation of various biological processes. To perform several of its functional activities, Ca^{2+} binds to different protein molecules, which are called as calcium binding proteins (CaBPs) (Heizmann and Schafer 1990). CaBPs with EF-hand motif (EF-CaBPs) belong to a growing sub family of CaBPs (Ababou and Desjarlais 2001; Bhattacharya *et al* 2004; Heizmann 1992; Nelson and Chazin 1998). The EF-hand motif

represents the canonical Ca^{2+} binding motif that consists of a contiguous 12 amino-acid residue long loop flanked by two helices (Kretsinger and Nockolds 1973; Strynadka and James 1989). The EF-CaBPs are broadly classified as Ca^{2+} sensors and Ca^{2+} buffers. Ca^{2+} sensors (Finn *et al* 1995; Hanley and Henley 2005; Hilge *et al* 2006; Shaw *et al* 1990; Vinogradova *et al* 2005) such as Calmodulin (CaM), Troponin C (TnC) etc., undergo huge conformational change upon binding to Ca^{2+} whereas, the Ca^{2+} buffers (Hackney *et al* 2005; Lambers *et al* 2006; Rosenbaum *et al* 2006; Vinogradova *et al* 2005) such as Calbindin D_{9k},

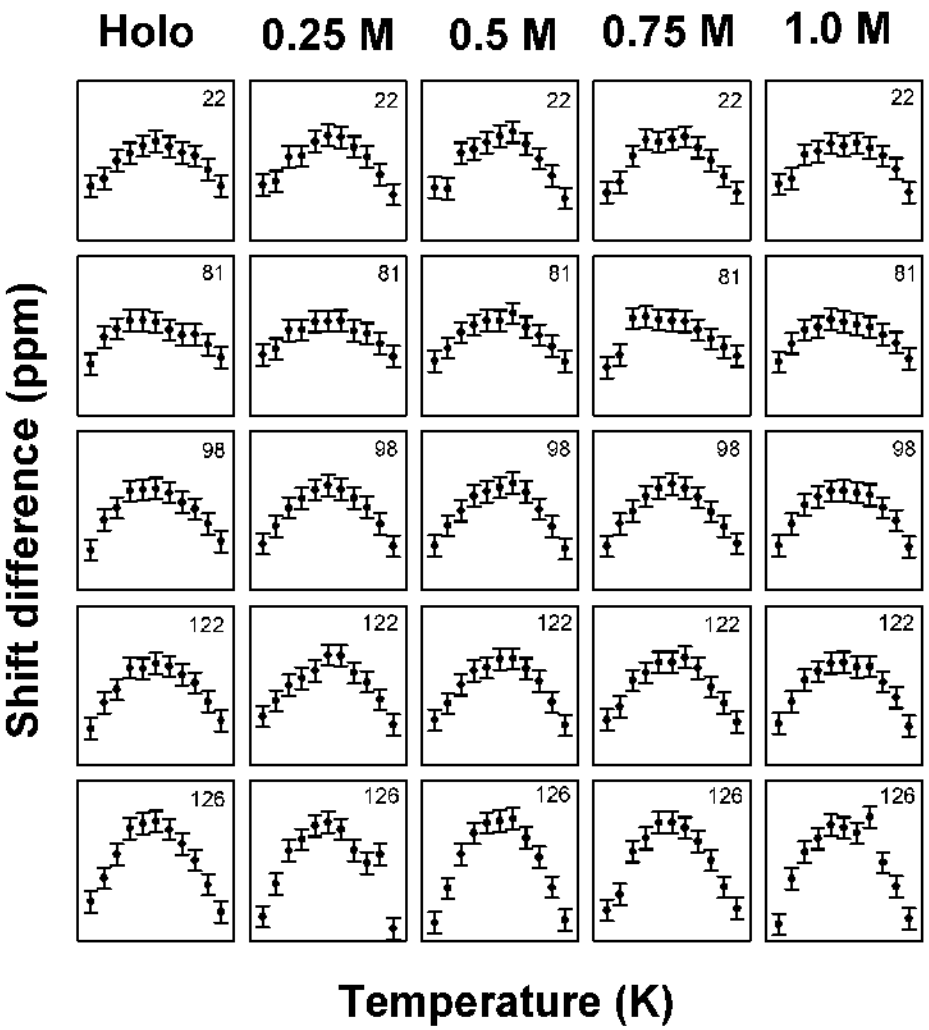


Fig. 3. Illustrative examples for the residues showing nonlinear temperature dependence of backbone $^1H^N$ chemical shifts in *Eh*CaBP as measured in native state and at different concentrations of GdmCl. The measured chemical shifts were fitted to a linear equation. The residuals (observed value – calculated value according to the linear fit) have been plotted against temperature; total scale of y-axis is 0.06 ppm: +0.03 to -0.03 centered at zero, and the temperature range is 280 K – 335 K. The error bars give an indication of the approximate error in measured chemical shifts (±0.004 ppm).

undergo modest conformational changes upon Ca^{2+} binding. In the current section the experimental evidence for the native state ruggedness on a Ca^{2+} sensor protein from the protozoan *Entamoeba histolytica* (*Eh*CaBP), an etiologic agent of amoebiasis has been demonstrated (Atreya *et al* 2001; Bhattacharya *et al* 2006).

As an illustration, **Fig 3** shows the experimentally measured temperature dependence of backbone proton ($^1H^N$) chemical shifts for few residues in the protein carried out in the temperature range 280 K – 335 K by recording temperature dependent HSQC spectra (Mohan *et al* 2008b). As evident from **Fig. 3**, the observed curvatures are different for different residues. These convex and the concave shapes (Mohan *et al* 2008b; Krishna Mohan *et al* 2008; Williamson 2003) reflect on different kinds of structural perturbations in the excited state compared to the native state as described above in theoretical simulations and illustrated in **Fig 2**. A summary of all the non-linear temperature dependences observed in *Eh*CaBP at different concentrations of GdmCl is given in **Fig. 4**.

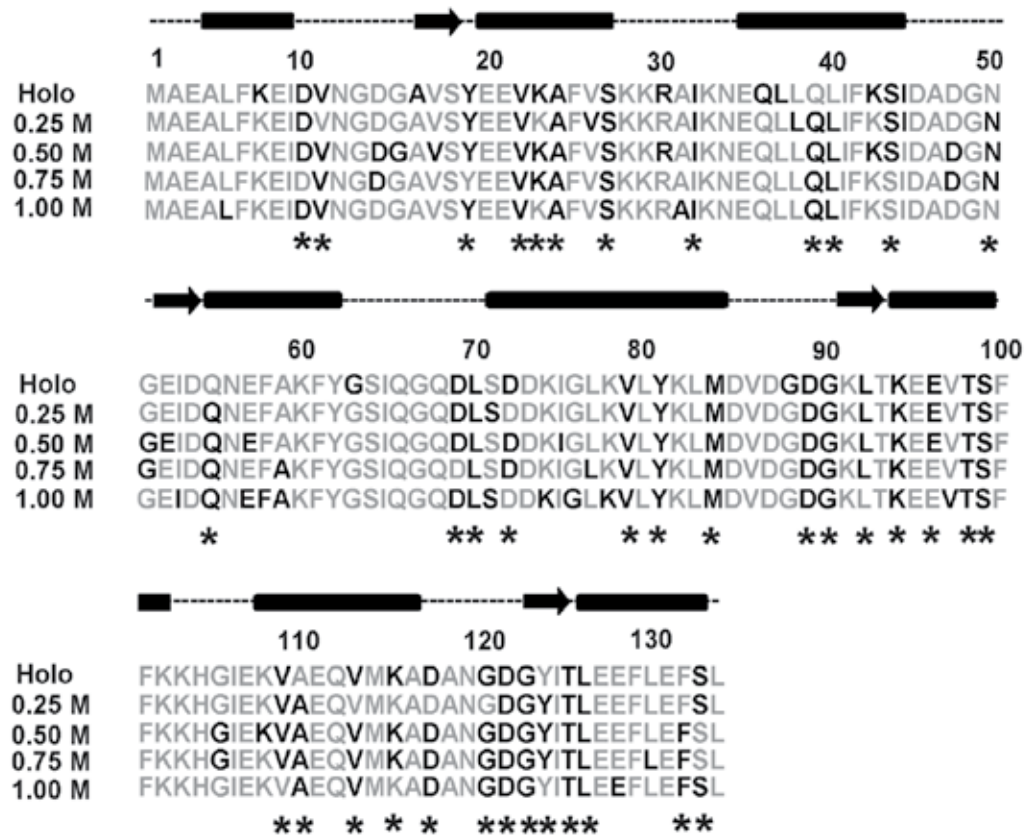


Fig. 4. Residues showing nonlinear temperature dependence of amide proton chemical shifts (black) in the native protein and at different concentrations of GdmCl were shown on the primary sequence of the polypeptide chain. The native secondary structures are shown by arrows (β strands) and cylinders (α helix). Residues accessing alternative conformations at least 3 out of 5 measured GdmCl concentrations are marked with asterisks along the polypeptide chain.

The number of residues accessing alternative states at different concentrations of GdmCl in *Eh*CaBP is (0 M (holo) – 41; 0.25 M – 35; 0.5 M – 50; 0.75 M – 39 and 1.0 M – 42) **(Fig 4)**. The total number of residues which access alternative conformations at least 3 out of 5 measured GdmCl concentrations turned out to be 39 (residues shown with asterisks in Fig. 3), implying that ~ 30 % (39 out of 134) of the residues are accessing alternative conformations. The theoretical simulations described above suggest that the observed curvatures are ~ 2-3 kcal/mol. All these residues are shown in **Fig. 5** on the 3D structure of the protein. Further, the extent of curvatures of individual residues increases or decreases with change in concentration of GdmCl **(Fig.3)** (Williamson 2003). The residues that become more curved with the increasing concentrations of GdmCl are most likely due to the presence of alternate states which are more similar in energy, or more different in shift or gradient to the corresponding native state. This is primarily an off-shoot of contracted unfolding energy landscape in the presence of GdmCl since the GdmCl is not expected to change the nature of the alternative state. Whereas the decrease of curvature can be explained by considering more than one alternative states within 5 kcal/mol, or have an alternative state that becomes very close to native state as GdmCl is increased.

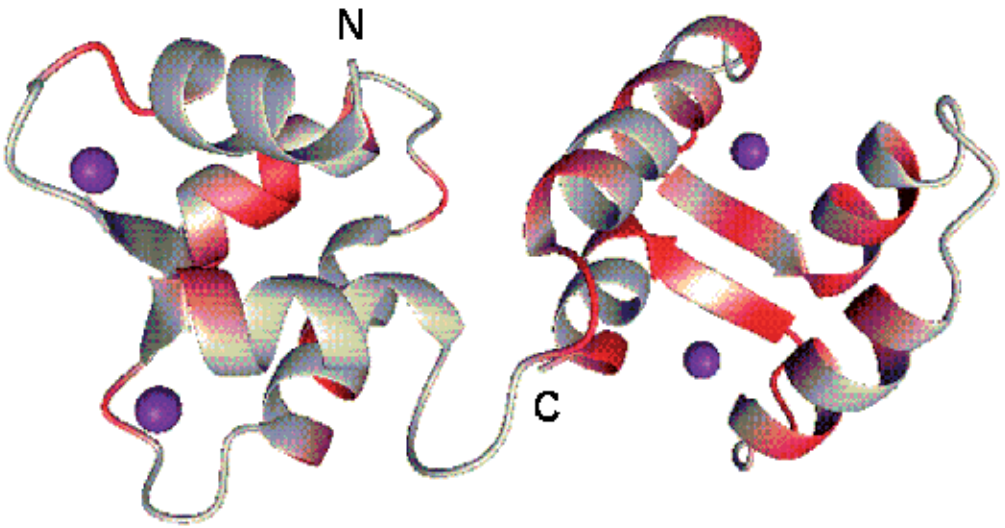


Fig. 5. Residues exhibiting curved temperature dependence at least three out of five concentrations of GdmCl measured in *Eh*CaBP are marked with red color on the 3D structure of the protein (PDB Id: 1JFK).

It is interesting to note that the low energy excited states detected in *Eh*CaBP are not uniformly distributed along the polypeptide chain; different segments of the protein have their own intrinsic preferences to access the alternative conformations. Out of the 39 residues which access low energy excited states, 7 residues belong to EF-hand I; 5 to EF-hand II; 11 to EF-hand III; 13 to EF-hand IV and 3 to the interconnecting loops **(Fig.4)**. It is evident from the data that the density of the conformational fluctuations in the C-terminal domain (24 residues) are twofold compared to the N-terminal counterpart (12 residues)

(**Fig. 4**). This suggests that the C-terminal domain is more flexible and susceptible to structural rearrangements. Further some novel features have been observed in the locations of alternative conformations. The residue at the 5th position of the calcium binding loop (Asp/Asn) that coordinates with Ca^{2+} shows alternative conformations consistently in all the EF-hands. The Gly-6 (the residue at 6th position in the calcium binding loop), which acts as a hinge in the calcium binding loop accesses alternative states in both EF III (Gly-90) and EF IV (Gly-122) hands. Among the calcium binding loops, the IV EF-loop is found to be the most dynamic with maximum number of residues (6 residues out of 12 residue loop) access low energy excited states. This loop has relatively low affinity towards Ca^{2+} compared to the other three loops as demonstrated earlier by the EGTA titration (Mukherjee *et al* 2005), though highly specific for Ca^{2+}. Moreover, recently it has been observed that IV EF-loop also differs with the remaining three EF-loops in the case of Mg^{2+} binding as evidenced by the Mn^{2+} titration (Mukherjee *et al* 2007a). Thus from all the discussion, it can be established that the native state of *Eh*CaBP is rugged due to accessing of various alternative states and the ruggedness is more in the C-terminal domain compared to that in the N-terminal domain. It is interesting to note that among the four EF hands, EF-hands I and II belonging to the N-terminal domain show different conformational dynamics from that of EF-hands III and IV belonging to the C-terminal domain (Mohan *et al* 2008b; Mukherjee *et al* 2007b).

Recently Chandra et al (Chandra *et al* 2011) measuring nonlinear temperature dependence of the backbone amide proton chemical shifts on non-myristoylated (non-myr) and myristoylated (myr) neuronal calcium sensor-1 (NCS-1). The authors reported that ~20% of the residues in the protein access alternative conformations in non-myr case, which increases to ~28% for myr NCS-1. These residues are spread over the entire polypeptide stretch and include the edges of α-helices and β-strands, flexible loop regions, and the Ca^{2+}-binding loops. Besides, residues responsible for the absence of Ca–myristoyl switch are also found accessing alternative states. The C-terminal domain is more populated with these residues compared to its N-terminal counterpart. Individual EF-hands in NCS-1 differ significantly in number of alternate states. Such differences in the conformational dynamics between the two domains and among the EF-loops have significant influence on the specificity and affinity of the metal binding properties and also have implications to domain dependent calcium signaling pathways of calcium sensor proteins (Mohan *et al* 2008b; Mukherjee *et al* 2007b).

3.2 Near native states and structure adaptability of dynein light chain protein

Dynein light chain protein (DLC8), a 10.3 kDa protein (89 residues) is the smallest subunit of the Dynein motor complex. DLC8 is a dimer at physiological pH and a stable monomer below pH 4.0 (Barbar and Hare 2004; Mohan *et al* 2006; Nyarko *et al* 2005). The differences between the monomeric and dimeric structures are, (i) the β3 strand in the dimer loses its secondary structure on dissociation to the monomer, and (ii) the helices α1 and α2 and the strands β1 and β2 get shortened by two residues (Liang *et al* 1999; Makokha *et al* 2004). DLC8 dimer acts as a cargo adapter and recognized as interactive "protein hub" (Barbar 2008). The dimer binds the target molecules in an anti-parallel β-strand fashion through its β3-strand, whereas the monomer form of DLC8 is not capable of binding to target proteins

(Alonso *et al* 2001; Fan *et al* 2001; Fan *et al* 1998; Fuhrmann *et al* 2002; Jaffrey and Snyder 1996; Lo *et al* 2001; Naisbitt *et al* 2000; Puthalakath *et al* 1999). This property is expected to have a regulatory role in the protein function.

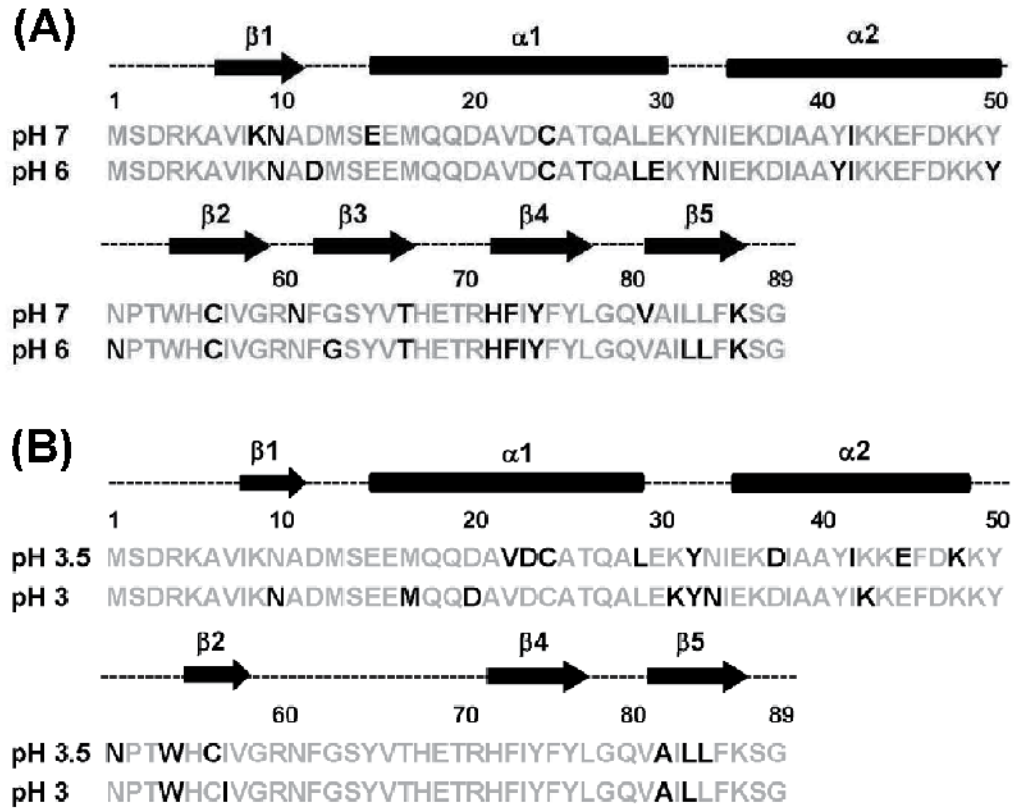


Fig. 6. Residues showing non linear temperature dependence of amide proton chemical shifts (black) in the temperature range is 290 K – 315 K along the polypeptide chain. The results are shown for pH 7.0 and 6.0 **(A)** and for pH 3.5 and 3.0 **(B)**. The arrows (β strands) and cylinders (α helix) indicate native secondary structures.

Temperature dependence of the amide proton chemical shifts in the DLC8 dimer (pH 7) and in the monomer (pH 3) has been measured in the temperature range 290-315 K (Krishna Mohan *et al* 2008). Among the above mentioned environment perturbations, pH variation is a mild perturbation and in general it changes the protonation states of the various residues depending on the chosen pH range. In order to identify the pH sensitive conformational dynamics in DLC8 protein the temperature dependence of the amide proton chemical shifts in both the dimer and the monomer were measured at slightly different pH conditions i.e., dimer at pH 6 and monomer at pH 3.5. A summary of all these results at various pH values i.e., pH 7 and 6 for the dimer and pH 3.5 and 3 for the monomer are shown in **Fig. 6**. Comparison of **Figs. 6A** and **6B** reveals that the residues accessing alternative conformations

have many differences between the dimer and the monomer. The number of residues accessing alternative conformations in the dimer at pH 7 and 6 are 13 and 21 respectively **(Fig.6A)**. **Figs. 7A** and **7B** display the locations of these residues on the native structures of the protein. Likewise, the number of residues accessing alternative conformations in the monomer at pH 3.5 is 15 and that at pH 3 is 11 **(Fig.6B)**. The locations of these residues are marked with red color on the native structure of the monomeric protein in **Figs. 7C** and **7D** respectively.

The differences observed in the positions of the residues accessing alternative conformations in the dimer and in the monomer due to small pH perturbations provide insights into the sensitivity of the conformational fluctuations due to environment perturbations in the two cases. In fact, the perturbation of the dimer landscape would have functional significance since small pH differences are known to exist in different parts of a cell (Spitzer and Poolman 2005; Stewart *et al* 1999; Swietach and Vaughan-Jones 2004; Swietach *et al* 2005; Vaughan-Jones *et al* 2002; Willoughby and Schwiening 2002; Zaniboni *et al* 2003). It is evident from **Figs. 7A** and **7B** that several of the residues that access low energy excited states are surrounding the dimer interface of the molecule which is also the cargo binding site (Krishna Mohan *et al* 2008; Liang *et al* 1999). It can be envisaged that the observed sensitivity of conformational dynamics at the dimer interface due to small environmental perturbations can significantly influence the cargo binding nature of the protein. Likewise, in the monomer **(Figs. 7C** and **7D)** (Krishna Mohan *et al* 2008; Liang *et al* 1999), noticeable differences have been observed in both α1 and α2 helices. Interestingly, the α2 helix participates in several inter-monomer contacts once the dimer is formed and hence its sensitivity to small perturbations may have a crucial role for the proper formation/folding of the functional dimer.

3.2.1 Relationship between sequence, structure and pH sensitivity of DLC8 landscapes

The roughness of the energy landscape and the consequent fluctuations in the native state of a protein is a reflection on the nature of the interactions between the side chains of the different amino acid residues in the three dimensional structure of the protein. While this is not generally predictable, some insights may be obtained in some cases by closely examining the structure and the properties of the amino acids along the sequence. For example, the behaviors of residues with titratable groups, which are likely to be affected by a pH perturbation, can provide useful clues. An observed perturbation at such locations would indicate that conformational fluctuations could be arising due to existence of species with different protonation states; a change in the protonation state of a side–chain causes a local change in the electrostatic potential, and thereby results in some population of an alternative conformation on energetic considerations. Inter-conversion between the major population and the minor population so created leads to the so-called conformational fluctuations.

In the above background it is interesting to note that most of the residues with titratable groups in the side chains in DLC8 **(Fig.7E)** are located in the regions which are exhibiting conformational fluctuations, and hence, their perturbation by small pH changes provides useful mechanistic insights. In the case of the dimer, the sensitivity of conformational

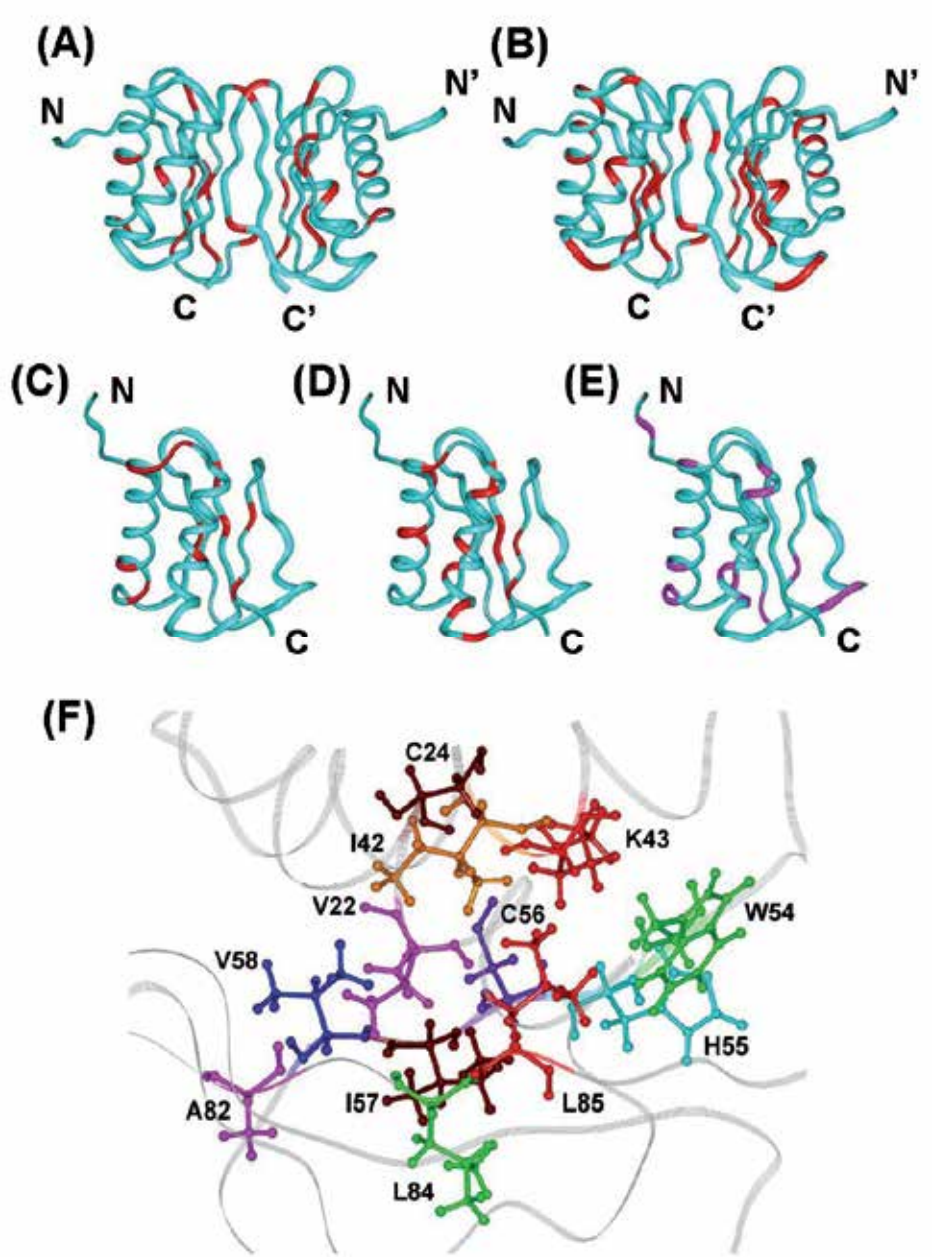


Fig. 7. Residues exhibiting curved temperature dependence in DLC8 dimer **(A)** pH 7.0, (**B)** pH 6.0 (PDB Id: 1f3c) and in monomer **(C)** pH 3.0, (**D)** pH 3.5 (PDB Id: 1hrw), are coloured red on the three dimensional structure of the protein. **(E)** Positions of all the titratable groups in the pH range 7.0 to 3.0 (Aspartates, Glutamates and Histidines) are marked with pink color on the monomer structure (PDB Id: 1rw). **(F)** Zooming in on a particular region surrounding β2 strand in the NMR structure of the monomer (PDB Id: 1rhw) to show the side chain interactions. Only a few residues in α1, α2 and β5 are shown for the sake of clarity.

dynamics can be readily traced to partial protonation of the His side chains as described earlier (Mohan *et al* 2006; Nyarko *et al* 2005). There will be inter-conversions between charged and neutral His and there will also be charge-charge repulsions. These will cause fluctuations in local electrostatic potentials and consequently in local side chain packing, which in turn will affect the main chain conformations. Among the three histidines, His 55 (pK 4.5,), His 68 and His 72 (both have pK of 6.0, (Mohan *et al* 2006; Nyarko *et al* 2005), the latter two would be the major contributors to the observed differences in the fluctuations of the native state in the pH range of 6-7.

In the case of DLC8 monomer His 68 and His 72 do not have any effect on the observed differences as they are completely protonated below pH 4.0. On the other hand His 55 (pK 4.5, (Mohan *et al* 2006; Nyarko *et al* 2005)), would have a significant effect. At pH 3.5 the side chain of His 55 will be protonated to the extent of 90 % and exchange between protonated and free His will contribute to a local dynamics. The environmental perturbation due to this dynamics would get relayed through the β2 strand and the α1, α2 helices and the β5 strand due to the close packing of the side chains in the protein structure **(Fig. 7F)**. The side chains for a few residues of α1, α2 and β5 are shown in the figure and all of these residues are seen to exhibit curved temperature dependence. At pH 3.0, the population of protonated His will increase and this results in the observed perturbation differences. Similarly, the perturbations at the other titratable groups such as aspartates and glutamates in the α1 and α2 helices **(see Fig. 7E)** would also cause local relays and contribute to the accessibility of different low energy excited states. All these influence the native energy landscape of the protein.

3.3 Conformational fluctuations at the phosphorylation site of dynein light chain protein

Recent studies on p21-activated kinase 1 (Pak1), revealed DLC8 as its physiological interacting substrate (binding sites aa 61-89) and the phosphorylation site at Ser 88 (Vadlamudi *et al* 2004). Pak1 phosphorylation of DLC8 on Ser 88 controls vesicle formation and trafficking functions, whereas mutation of Ser 88 to Ala (S88A) prevents macropinocytosis (Song *et al* 2008; Song *et al* 2007; Vadlamudi *et al* 2004; Yang *et al* 2005). Further, DLC8 phosphorylation by Pak1 prevents the interaction with apoptotic protein Bim and plays an essential role in cell survival (Vadlamudi *et al* 2004) and also promotes the dissociation from Intermediate chain (IC74) and hence regulates the assembly of the motor complex (Song *et al* 2007). All these results highlight any perturbation at or near the interface is likely to affect the biological function. Intuitively, the remote effects of any perturbation in a protein must be a consequence of a strong network of interactions which may cause rapid relay of perturbations from any one particular site on the protein structure. However, specific knowledge of how the perturbations travel will be essential in each case to understand the specificities of interactions. In general, perturbations are often introduced deliberately in the form of specific mutations in an attempt to understand the regulatory roles of specific residues involved in target recognition, structural architecture, stability, aggregation and folding features of the wild type protein (Buck *et al* 2007; Frankel *et al* 2007; Grant *et al* 2007; Ishibashi *et al* 2007; Piana *et al* 2008; Riley *et al* 2007; Stollar *et al* 2003).

The phosphorylation site Ser 88 represents an unusual behavior. To understand the conformational behavior phosphorylation site, the amide proton temperature coefficients of

Ser 88 in the WT dimer with those of Ala 88 in the S88A mutant and of Ser 88 in the DLC8 monomer (at pH 3) are measured (Mohan and Hosur 2008). The plots of temperature dependence for these residues are shown in **Fig. 8A**. The measured temperature coefficients are -19.8 ± 0.3, -9.3 ± 0.2 and -5.6 ± 0.1 ppb/^{o}C for Ser 88 in WT dimer, Ala 88 in S88A mutant and Ser 88 in DLC8 monomer respectively. In general the temperature coefficients range between (-2 to -4) ppb/K for a strongly H-bonded amide protons and between (-5 to -10) ppb/K for an exposed solvent accessible (random coil) amide protons (Baxter and Williamson 1997). The values obtained in the case of S88A mutant and DLC8 monomer clearly indicate that the aa 88 is solvent exposed and not strongly hydrogen bonded. On the other hand, a large value of -19.8 ± 0.3 ppb/^{o}C suggests that the environment of Ser 88 in WT dimer is highly susceptible to perturbation. None of the other residues, either in WT monomer or in S88A mutant, showed such a huge temperature coefficient value (Baxter and Williamson 1997; Mohan and Hosur 2008). Williamson et al reported such a large value of temperature coefficient in the herpes simplex virus glycoprotein D-1 antigenic domain (Williamson *et al* 1986) and the experiments demonstrated by Andersen et al (Andersen *et al* 1992) suggested that this amide proton was in fact involved in a transient hydrogen-bonded structure, and thus the large temperature coefficient could be attributed to a loss of secondary/tertiary structure on heating. If the large temperature coefficient of Ser 88 is a consequence of transient hydrogen-bonding and due to loss of secondary/tertiary structure, then Ser 88 should exhibit conformational fluctuations (alternative states). The presence of alternative states for Ser 88 has been tested on Ala 88 residue of S88A mutant and Ser 88 in DLC8 monomer and dimer **(Fig. 8B)**. It is evident that Ser 88 in WT dimer does show a curved temperature dependence of amide proton chemical shifts. Thus, it can be concluded that the amide proton in Ser 88 in WT DLC8 dimer is transiently H-bonded; that it is not a stable H-bond was independently inferred from deuterium exchange studies (Mohan and Hosur 2008; Mohan *et al* 2008a). Moreover, the dimeric structure suggests that the amide group of Gly 89 is very close (~ 2.6 Å) to the backbone nitrogen of Ser 88 suggesting a possibility of potential transient H-bonding.

The mechanism for the relay of perturbations from the Ser 88 site can be envisaged by understanding the close side chain packing. The side chain packing of perturbed residues and Ser 88 are shown in **Fig. 8C**. The crystal structure shows that Ser 88 OG atom is packed against the imidazole ring of His 55 and in addition forms a hydrogen bond with the backbone carbonyl of Ser 88′ (Ser 88 of other monomer) (Liang *et al* 1999). From **Fig. 8C** it is evident that Ser 88 is buried inside and packed over side chains of crucial residues at the dimer interface. A closer look at Ser 88 environment in **Fig. 8C** depicts that Ser 88 is closely packed against the side chains of Thr 67, His 68 and Glu 69 of both the monomers in the dimer. Furthermore, the distance measurements between the back bone and side chain atoms of Ser 88 and those of Thr 67, His 68, Glu 69 and Thr 70 indicated that several atoms of Glu 69 are very close (~ 2 – 4 Å), whereas, for residues Thr 67, His 68, Thr 70 there is at least one atom in the distance range of 4 -6 Å. All of these residues are perturbed by the S88A mutation as seen from the chemical shift data. The other perturbed residues are slightly farther (> 6 Å). All these are shown in a color coded manner in **Fig. 8C**. This qualitative analysis provides a mechanistic insight into the relay of perturbation from the phopshorylation site; Glu 69 is most easily perturbed and the disturbance then runs on both sides at the dimer interface. Then, from Tyr 65 the relay spreads to Lys 44 which is engaged in a side chain H-bond.

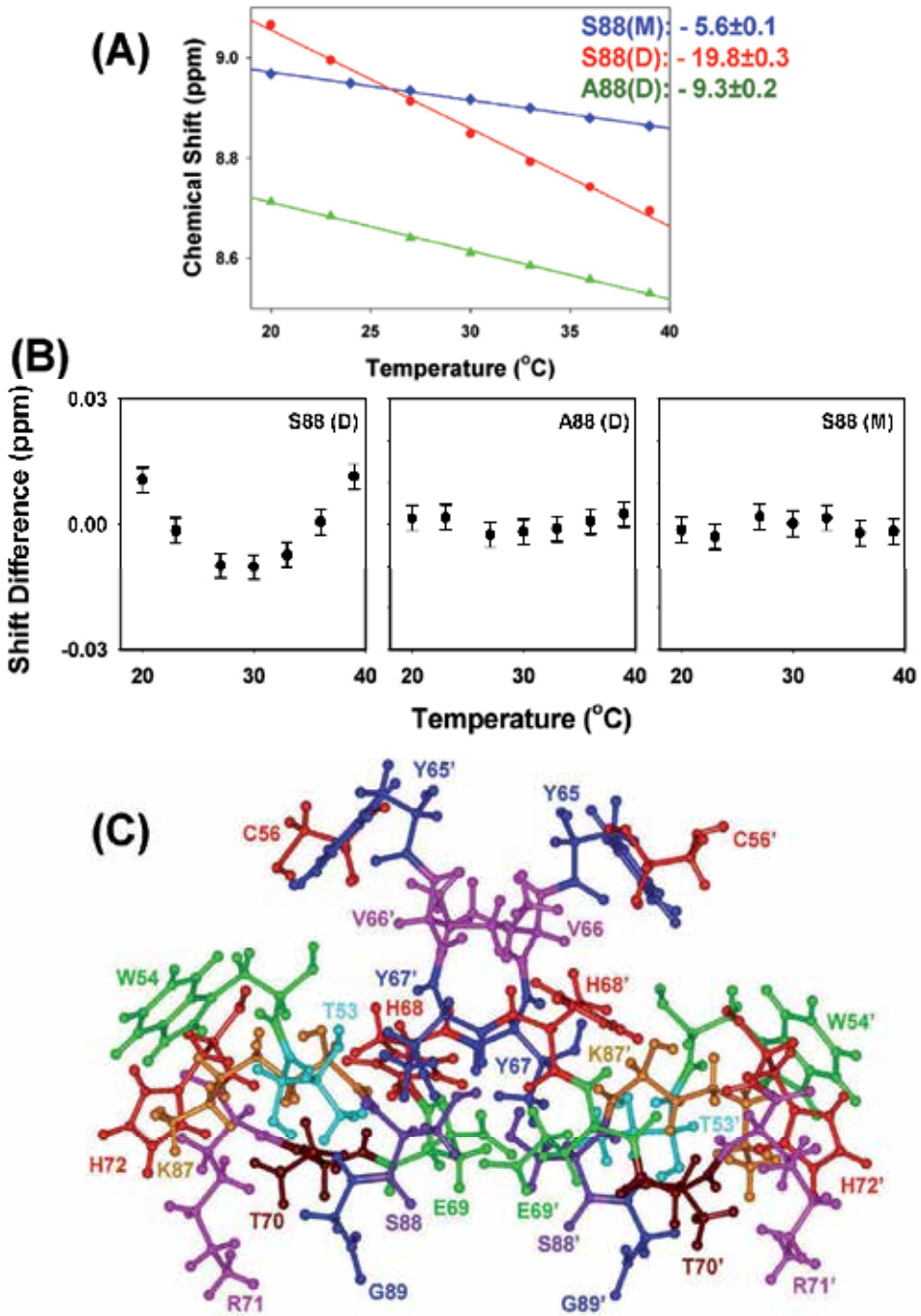


Fig. 8. **(A)** Graph depicting the temperature dependence of amide proton chemical shifts for Ser 88 in DLC8 WT-dimer (Circles, Red), Ala 88 in S88A mutant at pH 7 (Triangles, Green), and Ser 88 in DLC8 monomer at pH 3 (Diamonds, Blue). The solid lines line represents the best linear fit. **(B)** Comparison of non linear/linear temperature dependence of amide

proton chemical shifts at amino acid position 88 in DLC8 protein: Ser 88 in DLC8 WT-dimer [S88-(D)], Ala 88 in S88A mutant [A88-(D)] at pH 7 and Ser 88 in DLC8 monomer [S88-(M)] at pH 3. The measured chemical shifts were fitted to a linear equation. The residuals (observed value – calculated value according to the linear fit) have been plotted against temperature; total scale of y-axis is 0.06 ppm: +0.03 to -0.03 centered at zero, and the temperature range is 18 ^{o}C – 40 ^{o}C. **(C)** Zooming in on a particular region of the dimer interface around Ser 88 in the crystal structure (PDB Id: 1cmi), to show accessibility of the phosphorylation site (side chain –OH of Ser 88) and the interactions of side chains various other residues with Ser 88. Side chains of the residues which are perturbed due to S88A are only shown. The different residues are color coded to indicate the proximity of the side chain atoms of the residue to backbone NH of ser 88; Blue: at least one atom of the side chain is within 2-4 Å, Green: at least one atom of the side chain is within the range 4-6 Å, Red: all atoms are beyond 6 Å.

3.4 Alternative conformations in small globular proteins: Sources of fluctuations and implications to function/folding

Experiments have been performed by various research groups to detect the alternative conformations on different monomeric proteins in order to throw light either on the functional implications or on the folding trajectories. Investigations by Kumar et al (Kumar *et al* 2007) on SUMO-1 suggested that the alternative conformations span the length of the protein chain but are located at particular regions on the protein structure. The authors observed that several of the regions of the protein structure that exhibit such fluctuations coincide with the protein's binding surfaces with different substrate like GTPase effector domain (GED) of dynamin, SUMO binding motifs (SBM), E1 (activating enzyme, SAE1/SAE2) and E2 (conjugating enzyme, UBC9) enzymes of sumoylation machinery and speculated that these conformational fluctuations have significant implications for the binding of diversity of targets by SUMO-1. Another report by Srivastava et al (Srivastava and Chary 2011) on hahellin (a βγ-crystallin domain) in its Ca^{2+}-bound form depicted a large conformational heterogeneity with nearly 40% of the residues, some of which are part of Ca2+-binding loops. Further, they observed that out of the two Greek key motifs, the second Greek key motif is floppy as compared to its counterpart.

Extensive research investigations on theoretical and experimental aspects of different protein systems regarding low-energy excited states have been performed by Williamson and co-workers (Baxter *et al* 1998; Tunnicliffe *et al* 2005; Williamson 2003). Studies on conformational ensemble of cytochrome *c* revealed high structural entropy (Williamson 2003). The density of alternative states is particularly high near the heme ligand Met80, which is of interest because both redox change and the first identified stage in unfolding are associated with change in Met80 ligation. By combining theoretical and experimental approaches, it is concluded that the alternative states each comprise approximately five residues, have in general less structure than the native state and are therefore locally unfolded structures. The locations of the alternative states the global unfolding pathway of cytochrome *c*, hinted that they may determine the pathway. Similar experiments on B1 domains of streptococcal proteins G and L (Tunnicliffe *et al* 2005), which are structurally similar, but have different sequences and folding established that several of the residues

have curved amide proton temperature and indicated approximately 4–6 local minima for each protein. Further, reports on N-terminal domain of phosphoglycerate kinase, hen egg-white lysozyme, SUMO1 and BPTI (Baxter *et al* 1998; Kumar *et al* 2007) established that conformational heterogeneity arises from a number of independent sources such as, aromatic ring current effects, a minor conformer generated through disulphide bond isomerisation; an alternative hydrogen bond network associated with buried water molecules; alternative hydrogen bonds involving backbone amides and surface-exposed side-chain hydrogen bond acceptors; and the disruption of loops, ends of secondary structural elements and chain termini.

In conclusion, on one hand the ruggedness of the native energy landscape of the protein systems provide rationales for the adaptability of the protein structure to bind various target molecules in order to carry out the biological functions efficient manner. On the other hand, it throws light on many potential unfolding initiation sites in the protein. Furthermore, the origins of these conformational fluctuations provide mechanistic insights into the protein network of hydrophobic/H-bond interactions that dictate the protein stability.

4. Abbreviations

NMR - Nuclear magnetic resonance spectroscopy; HSQC - Hetero nuclear single quantum correlation spectroscopy; DLC8 - Dynein light chain protein; *Eh*CaBP - *Entamoeba histolytica* calcium binding protein; Gdmcl - Guanidine Hydrochloride; NCS - Neuronal calcium sensor protein; Myr - Myristoylated; Non-myr - Non-Myristoylated; Pak-1 - P21 activated kinease; SUMO - Small Ubiquitin-like Modifier; BPTI - Bovine pancreatic trypsin inhibitor; UBC9 - Ubiquitin carrier protein 9 .

5. Acknowledgements

The author thank Prof. Ramakrishna V Hosur (TIFR, Mumbai), Prof. K. V. R. Chary (TIFR, Mumbai) for their invaluable guidance, Dr. Sulakshana Mukherjee (UCSD, California) for critical suggestions, the NMR facility at Tata Institute of Fundamental Research (Mumbai, India) and the library facilities at Rutgers University (Piscataway, New Jersey) are greatly acknowledged.

6. References

Ababou A and Desjarlais J R 2001 Solvation energetics and conformational change in EF-hand proteins; *Protein Sci.* 10 301-312.

Agarwal P K 2005 Role of protein dynamics in reaction rate enhancement by enzymes; *J. Am. Chem. Soc.* 127 15248-15256.

Akasaka K 2006 Probing conformational fluctuation of proteins by pressure perturbation; *Chem Rev.* 106 1814-1835.

Alonso C, Miskin J, Hernaez B, Fernandez-Zapatero P, Soto L, Canto C, Rodriguez-Crespo I, Dixon L and Escribano J M 2001 African swine fever virus protein p54 interacts

with the microtubular motor complex through direct binding to light-chain dynein; *J. Virol.* 75 9819-9827.

Andersen N H, Chen C P, Marschner T M, Krystek S R, Jr. and Bassolino D A 1992 Conformational isomerism of endothelin in acidic aqueous media: a quantitative NOESY analysis; *Biochemistry* 31 1280-1295.

Anderson N H, Neidigh J W, Harris S M, Lee G M, Liu Z and Tong H 1997 Extracting information from the temperature gradients of polypeptide HN chemical shifts. 1. The importance of conformational averaging; *J. Am. Chem. Soc.* 119 8547-8561.

Atreya H S, Sahu S C, Bhattacharya A, Chary K V and Govil G 2001 NMR derived solution structure of an EF-hand calcium-binding protein from Entamoeba Histolytica; *Biochemistry* 40 14392-14403.

Bai Y, Sosnick T R, Mayne L and Englander S W 1995 Protein folding intermediates: native-state hydrogen exchange; *Science* 269 192-197.

Barbar E 2008 Dynein light chain LC8 is a dimerization hub essential in diverse protein networks; *Biochemistry* 47 503-508.

Barbar E and Hare M 2004 Characterization of the cargo attachment complex of cytoplasmic dynein using NMR and mass spectrometry; *Methods Enzymol.* 380 219-241.

Baxter N J, Hosszu L L, Waltho J P and Williamson M P 1998 Characterisation of low free-energy excited states of folded proteins; *J. Mol. Biol.* 284 1625-1639.

Baxter N J and Williamson M P 1997 Temperature dependence of 1H chemical shifts in proteins; *J. Biomol. NMR* 9 359-369.

Bhabha G, Lee J, Ekiert D C, Gam J, Wilson I A, Dyson H J, Benkovic S J and Wright P E 2011 A dynamic knockout reveals that conformational fluctuations influence the chemical step of enzyme catalysis; *Science* 332 234-238.

Bhattacharya A, Padhan N, Jain R and Bhattacharya S 2006 Calcium-binding proteins of Entamoeba histolytica; *Arch. Med. Res.* 37 221-225.

Bhattacharya S, Bunick C G and Chazin W J 2004 Target selectivity in EF-hand calcium binding proteins; *Biochim. Biophys. Acta* 1742 69-79.

Boehr D D, McElheny D, Dyson H J and Wright P E 2006 The dynamic energy landscape of dihydrofolate reductase catalysis; *Science* 313 1638-1642.

Boehr D D, McElheny D, Dyson H J and Wright P E 2010 Millisecond timescale fluctuations in dihydrofolate reductase are exquisitely sensitive to the bound ligands; *Proc. Natl. Acad. Sci. U. S. A* 107 1373-1378.

Bryngelson J D, Onuchic J N, Socci N D and Wolynes P G 1995 Funnels, pathways, and the energy landscape of protein folding: a synthesis; *Proteins* 21 167-195.

Buck T M, Wagner J, Grund S and Skach W R 2007 A novel tripartite motif involved in aquaporin topogenesis, monomer folding and tetramerization; *Nat. Struct. Mol. Biol.* 14 762-769.

Chandra K, Sharma Y and Chary K V 2011 Characterization of low-energy excited states in the native state ensemble of non-myristoylated and myristoylated neuronal calcium sensor-1; *Biochim. Biophys. Acta* 1814 334-344.

Chatterjee A, Krishna Mohan P M, Prabhu A, Ghosh-Roy A and Hosur R V 2007 Equilibrium unfolding of DLC8 monomer by urea and guanidine hydrochloride: Distinctive global and residue level features; *Biochimie* 89 117-134.

Cierpicki T and Otlewski J 2001 Amide proton temperature coefficients as hydrogen bond indicators in proteins; *J. Biomol. NMR* 21 249-261.

Cierpicki T, Zhukov I, Byrd R A and Otlewski J 2002 Hydrogen bonds in human ubiquitin reflected in temperature coefficients of amide protons; *J. Magn Reson.* 157 178-180.

Clore G M 2011 Exploring sparsely populated states of macromolecules by diamagnetic and paramagnetic NMR relaxation; *Protein Sci.* 20 229-246.

Clore G M and Iwahara J 2009 Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes; *Chem. Rev.* 109 4108-4139.

Dill K A and Chan H S 1997 From Levinthal to pathways to funnels; *Nat. Struct. Biol.* 4 10-19.

Dobson C M 2003 Protein folding and misfolding; *Nature* 426 884-890.

Dobson C M and Karplus M 1999 The fundamentals of protein folding: bringing together theory and experiment; *Curr. Opin. Struct. Biol.* 9 92-101.

Dunker A K, Brown C J, Lawson J D, Iakoucheva L M and Obradovic Z 2002 Intrinsic disorder and protein function; *Biochemistry* 41 6573-6582.

Dyson H J and Wright P E 2005 Elucidation of the protein folding landscape by NMR; *Methods Enzymol.* 394 299-321.

Eisenmesser E Z, Millet O, Labeikovsky W, Korzhnev D M, Wolf-Watz M, Bosco D A, Skalicky J J, Kay L E and Kern D 2005 Intrinsic dynamics of an enzyme underlies catalysis; *Nature* 438 117-121.

Fan J, Zhang Q, Tochio H, Li M and Zhang M 2001 Structural basis of diverse sequence-dependent target recognition by the 8 kDa dynein light chain; *J. Mol. Biol.* 306 97-108.

Fan J S, Zhang Q, Li M, Tochio H, Yamazaki T, Shimizu M and Zhang M 1998 Protein inhibitor of neuronal nitric-oxide synthase, PIN, binds to a 17-amino acid residue fragment of the enzyme; *J. Biol. Chem.* 273 33472-33481.

Feher V A and Cavanagh J 1999 Millisecond-timescale motions contribute to the function of the bacterial response regulator protein Spo0F; *Nature* 400 289-293.

Finn B E, Evenas J, Drakenberg T, Waltho J P, Thulin E and Forsen S 1995 Calcium-induced structural changes and domain autonomy in calmodulin; *Nat. Struct. Biol.* 2 777-783.

Frankel B A, Tong Y, Bentley M L, Fitzgerald M C and McCafferty D G 2007 Mutational analysis of active site residues in the Staphylococcus aureus transpeptidase SrtA; *Biochemistry* 46 7269-7278.

Fraser J S, Clarkson M W, Degnan S C, Erion R, Kern D and Alber T 2009 Hidden alternative structures of proline isomerase essential for catalysis; *Nature* 462 669-673.

Fuhrmann J C, Kins S, Rostaing P, El F O, Kirsch J, Sheng M, Triller A, Betz H and Kneussel M 2002 Gephyrin interacts with Dynein light chains 1 and 2, components of motor protein complexes; *J. Neurosci.* 22 5393-5402.

Grant M A, Lazo N D, Lomakin A, Condron M M, Arai H, Yamin G, Rigby A C and Teplow D B 2007 Familial Alzheimer's disease mutations alter the stability of the amyloid beta-protein monomer folding nucleus; *Proc. Natl. Acad. Sci. U. S. A* 104 16522-16527.

Hackney C M, Mahendrasingam S, Penn A and Fettiplace R 2005 The concentrations of calcium buffering proteins in mammalian cochlear hair cells; *J. Neurosci.* 25 7867-7875.

Hanley J G and Henley J M 2005 PICK1 is a calcium-sensor for NMDA-induced AMPA receptor trafficking; *EMBO J.* 24 3266-3278.

Heizmann C W 1992 Calcium-binding proteins: basic concepts and clinical implications; *Gen. Physiol Biophys.* 11 411-425.

Heizmann C W and Schafer B W 1990 Internal calcium-binding proteins; *Semin. Cell Biol.* 1 277-282.

Hilge M, Aelen J and Vuister G W 2006 Ca2+ regulation in the Na+/Ca2+ exchanger involves two markedly different Ca2+ sensors; *Mol. Cell* 22 15-25.

Ishibashi M, Tatsuda S, Izutsu K, Kumeda K, Arakawa T and Tokunaga M 2007 A single Gly114Arg mutation stabilizes the hexameric subunit assembly and changes the substrate specificity of halo-archaeal nucleoside diphosphate kinase; *FEBS Lett.* 581 4073-4079.

Jaffrey S R and Snyder S H 1996 PIN: an associated protein inhibitor of neuronal nitric oxide synthase; *Science* 274 774-777.

Kitahara R, Yokoyama S and Akasaka K 2005 NMR snapshots of a fluctuating protein structure: ubiquitin at 30 bar-3 kbar; *J. Mol. Biol.* 347 277-285.

Korzhnev D M, Karlsson B G, Orekhov V Y and Billeter M 2003 NMR detection of multiple transitions to low-populated states in azurin; *Protein Sci.* 12 56-65.

Kretsinger R H and Nockolds C E 1973 Carp muscle calcium-binding protein. II. Structure determination and general description; *J. Biol. Chem.* 248 3313-3326.

Krishna Mohan P M, Barve M, Chatterjee A, Ghosh-Roy A and Hosur R V 2008 NMR comparison of the native energy landscapes of DLC8 dimer and monomer; *Biophys. Chem* 134 10-19.

Kumar A, Srivastava S and Hosur R V 2007 NMR characterization of the energy landscape of SUMO-1 in the native-state ensemble; *J. Mol. Biol.* 367 1480-1493.

Lambers T T, Mahieu F, Oancea E, Hoofd L, de L F, Mensenkamp A R, Voets T, Nilius B, Clapham D E, Hoenderop J G and Bindels R J 2006 Calbindin-D28K dynamically controls TRPV5-mediated Ca2+ transport; *EMBO J.* 25 2978-2988.

Levinthal C 1969. Mössbauer Spectroscopy in Biological Systems. (eds. DeBrunner JTP and Munck E), pp 22-24. University of Illinois Press: Illinois.

Liang J, Jaffrey S R, Guo W, Snyder S H and Clardy J 1999 Structure of the PIN/LC8 dimer with a bound peptide; *Nat. Struct. Biol.* 6 735-740.

Lo K W, Naisbitt S, Fan J S, Sheng M and Zhang M 2001 The 8-kDa dynein light chain binds to its targets via a conserved (K/R)XTQT motif; *J. Biol. Chem.* 276 14059-14066.

Makokha M, Huang Y J, Montelione G, Edison A S and Barbar E 2004 The solution structure of the pH-induced monomer of dynein light-chain LC8 from Drosophila; *Protein Sci.* 13 727-734.

Mittermaier A and Kay L E 2006 New tools provide new insights in NMR studies of protein dynamics; *Science* 312 224-228.

Mohan P M, Barve M, Chatterjee A and Hosur R V 2006 pH driven conformational dynamics and dimer-to-monomer transition in DLC8; *Protein Sci.* 15 335-342.

Mohan P M, Chakraborty S and Hosur R V 2008a Residue-wise conformational stability of DLC8 dimer from native-state hydrogen exchange; *Proteins*.

Mohan P M and Hosur R V 2008 NMR characterization of structural and dynamics perturbations due to a single point mutation in Drosophila DLC8 dimer: functional implications; *Biochemistry* 47 6251-6259.

Mohan P M, Mukherjee S and Chary K V 2008b Differential native state ruggedness of the two Ca2+-binding domains in a Ca2+ sensor protein; *Proteins* 70 1147-1153.

Mukherjee S, Kuchroo K and Chary K V 2005 Structural characterization of the apo form of a calcium binding protein from Entamoeba histolytica by hydrogen exchange and its folding to the holo state; *Biochemistry* 44 11636-11645.

Mukherjee S, Mohan P M and Chary K V 2007a Magnesium promotes structural integrity and conformational switching action of a calcium sensor protein; *Biochemistry* 46 3835-3845.

Mukherjee S, Mohan P M, Kuchroo K and Chary K V 2007b Energetics of the native energy landscape of a two-domain calcium sensor protein: distinct folding features of the two domains; *Biochemistry* 46 9911-9919.

Naisbitt S, Valtschanoff J, Allison D W, Sala C, Kim E, Craig A M, Weinberg R J and Sheng M 2000 Interaction of the postsynaptic density-95/guanylate kinase domain-associated protein complex with a light chain of myosin-V and dynein; *J. Neurosci.* 20 4524-4534.

Nelson M R and Chazin W J 1998 Structures of EF-hand Ca(2+)-binding proteins: diversity in the organization, packing and response to Ca2+ binding; *Biometals* 11 297-318.

Nyarko A, Cochrun L, Norwood S, Pursifull N, Voth A and Barbar E 2005 Ionization of His 55 at the dimer interface of dynein light-chain LC8 is coupled to dimer dissociation; *Biochemistry* 44 14248-14255.

Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 Theory of protein folding: the energy landscape perspective; *Annu. Rev. Phys. Chem.* 48 545-600.

Parak F G 2003a Proteins in action: the physics of structural fluctuations and conformational changes; *Curr. Opin. Struct. Biol.* 13 552-557.

Parak F G 2003b Proteins in action: the physics of structural fluctuations and conformational changes; *Curr. Opin. Struct. Biol.* 13 552-557.

Piana S, Carloni P and Parrinello M 2002 Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease; *J. Mol. Biol.* 319 567-583.

Piana S, Laio A, Marinelli F, Van T M, Bourry D, Ampe C and Martins J C 2008 Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their Pro62Ala mutants; *J. Mol. Biol.* 375 460-470.

Popovych N, Sun S, Ebright R H and Kalodimos C G 2006 Dynamically driven protein allostery; *Nat. Struct. Mol. Biol.* 13 831-838.

Puthalakath H, Huang D C, O'Reilly L A, King S M and Strasser A 1999 The proapoptotic activity of the Bcl-2 family member Bim is regulated by interaction with the dynein motor complex; *Mol. Cell* 3 287-296.

Riley P W, Cheng H, Samuel D, Roder H and Walsh P N 2007 Dimer dissociation and unfolding mechanism of coagulation factor XI apple 4 domain: spectroscopic and mutational analysis; *J. Mol. Biol.* 367 558-573.

Rosenbaum E E, Hardie R C and Colley N J 2006 Calnexin is essential for rhodopsin maturation, Ca2+ regulation, and photoreceptor cell survival; *Neuron* 49 229-241.

Schwarzinger S, Kroon G J, Foss T R, Chung J, Wright P E and Dyson H J 2001 Sequence-dependent correction of random coil NMR chemical shifts; *J. Am. Chem. Soc.* 123 2970-2978.

Schwarzinger S, Kroon G J, Foss T R, Wright P E and Dyson H J 2000 Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView; *J. Biomol. NMR* 18 43-48.

Shaw G S, Hodges R S and Sykes B D 1990 Calcium-induced peptide association to form an intact protein domain: 1H NMR structural evidence; *Science* 249 280-283.

Shoemaker B A, Wang J and Wolynes P G 1999 Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble; *J. Mol. Biol.* 287 675-694.

Song C, Wen W, Rayala S K, Chen M, Ma J, Zhang M and Kumar R 2008 Serine 88 Phosphorylation of the 8-kDa Dynein Light Chain 1 Is a Molecular Switch for Its Dimerization Status and Functions; *J. Biol. Chem* 283 4004-4013.

Song Y, Benison G, Nyarko A, Hays T S and Barbar E 2007 Potential role for phosphorylation in differential regulation of the assembly of dynein light chains; *J. Biol. Chem* 282 17272-17279.

Spitzer J J and Poolman B 2005 Electrochemical structure of the crowded cytoplasm; *Trends Biochem. Sci.* 30 536-541.

Srivastava A K and Chary K V 2011 Conformational heterogeneity and dynamics in a betagamma-Crystallin from Hahella chejuensis; *Biophys. Chem.* 157 7-15.

Stewart A K, Boyd C A and Vaughan-Jones R D 1999 A novel role for carbonic anhydrase: cytoplasmic pH gradient dissipation in mouse small intestinal enterocytes; *J. Physiol* 516 (Pt 1) 209-217.

Stollar E J, Mayor U, Lovell S C, Federici L, Freund S M, Fersht A R and Luisi B F 2003 Crystal structures of engrailed homeodomain mutants: implications for stability and dynamics; *J. Biol. Chem* 278 43699-43708.

Strynadka N C and James M N 1989 Crystal structures of the helix-loop-helix calcium-binding proteins; *Annu. Rev. Biochem.* 58 951-998.

Swietach P, Leem C H, Spitzer K W and Vaughan-Jones R D 2005 Experimental generation and computational modeling of intracellular pH gradients in cardiac myocytes; *Biophys. J.* 88 3018-3037.

Swietach P and Vaughan-Jones R D 2004 Novel method for measuring junctional proton permeation in isolated ventricular myocyte cell pairs; *Am. J. Physiol Heart Circ. Physiol* 287 H2352-H2363.

Tang C, Louis J M, Aniana A, Suh J Y and Clore G M 2008 Visualizing transient events in amino-terminal autoprocessing of HIV-1 protease; *Nature* 455 693-696.

Tilton R F, Jr., Dewan J C and Petsko G A 1992 Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K; *Biochemistry* 31 2469-2481.

Tobi D and Bahar I 2005 Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state; *Proc. Natl. Acad. Sci. U. S. A* 102 18908-18913.

Tunnicliffe R B, Waby J L, Williams R J and Williamson M P 2005 An experimental investigation of conformational fluctuations in proteins G and L; *Structure. (Camb.)* 13 1677-1684.

Tzeng S R and Kalodimos C G 2009 Dynamic activation of an allosteric regulatory protein; *Nature* 462 368-372.

Uversky V N 2002 Natively unfolded proteins: a point where biology waits for physics; *Protein Sci.* 11 739-756.

Uversky V N 2003 Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?; *Cell Mol. Life Sci.* 60 1852-1871.

Vadlamudi R K, Bagheri-Yarmand R, Yang Z, Balasenthil S, Nguyen D, Sahin A A, den H P and Kumar R 2004 Dynein light chain 1, a p21-activated kinase 1-interacting substrate, promotes cancerous phenotypes; *Cancer Cell* 5 575-585.

Vaughan-Jones R D, Peercy B E, Keener J P and Spitzer K W 2002 Intrinsic H(+) ion mobility in the rabbit ventricular myocyte; *J. Physiol* 541 139-158.

Villali J and Kern D 2010 Choreographing an enzyme's dance; *Curr. Opin. Chem. Biol.* 14 636-643.

Vinogradova M V, Stone D B, Malanina G G, Karatzaferi C, Cooke R, Mendelson R A and Fletterick R J 2005 Ca(2+)-regulated structural changes in troponin; *Proc. Natl. Acad. Sci. U. S. A* 102 5038-5043.

Whitten S T, Garcia-Moreno E B and Hilser V J 2005 Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins; *Proc. Natl. Acad. Sci. U. S. A* 102 4282-4287.

Williamson M P 2003 Many residues in cytochrome c populate alternative states under equilibrium conditions; *Proteins* 53 731-739.

Williamson M P, Hall M J and Handa B K 1986 1H-NMR assignment and secondary structure of a herpes simplex virus glycoprotein D-1 antigenic domain; *Eur. J. Biochem.* 158 527-536.

Willoughby D and Schwiening C J 2002 Electrically evoked dendritic pH transients in rat cerebellar Purkinje cells; *J. Physiol* 544 487-499.

Wishart D S, Bigam C G, Holm A, Hodges R S and Sykes B D 1995 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects; *J. Biomol. NMR* 5 67-81.

Wishart D S and Sykes B D 1994 Chemical shifts as a tool for structure determination; *Methods Enzymol.* 239 363-392.

Wolynes P G 2005 Energy landscapes and solved protein-folding problems; *Philos. Transact. A Math. Phys. Eng Sci.* 363 453-464.

Wolynes P G, Onuchic J N and Thirumalai D 1995 Navigating the folding routes; *Science* 267 1619-1620.

Wüthrich K 1986. *NMR of protein and nucleic acids*. John Wiley and Sons: New York.

Yang Z, Vadlamudi R K and Kumar R 2005 Dynein light chain 1 phosphorylation controls macropinocytosis; *J. Biol. Chem* 280 654-659.

Zaniboni M, Swietach P, Rossini A, Yamamoto T, Spitzer K W and Vaughan-Jones R D 2003 Intracellular proton mobility and buffering power in cardiac ventricular myocytes from rat, rabbit, and guinea pig; *Am. J. Physiol Heart Circ. Physiol* 285 H1236-H1246.

17

Conformational and Disorder to Order Transitions in Proteins: Structure / Function Correlation in Apolipoproteins

José Campos-Terán[1], Paola Mendoza-Espinosa[2], Rolando Castillo[3] and Jaime Mas-Oliva[2]
[1]*Departamento de Procesos y Tecnología, DCNI, Universidad Autónoma Metropolitana,*
[2]*Instituto de Fisiología Celular, Universidad Nacional Autónoma de México,*
[3]*Instituto de Física, Universidad Nacional Autónoma de México*
México, D.F., México

1. Introduction

The concept of protein folding is directly related with the process of reversible disorder-to-order transitions, by which an unfolded polypeptide chain folds into a specific functional native structure (Eaton et al., 2000; Rose et al., 2006). For folding into a native state, unfolded polypeptide chains require the intervention of weak interactions. Driven by hydrophobic interactions, a polypeptide chain begins to fold when placed in an aqueous medium, and rapidly becomes a molten globule followed by an important release of latent heat. Stabilization of the molten globule is achieved mainly through the distribution of hydrophobic residues away from the water matrix. On the other hand, because the polar residues contained in a protein develop hydrogen bonds with the water network as well as with each other, α-helices and β-sheets can be formed when bonds switch between molecules. It has been calculated that such bonds might be in the order of 10^{-12} s, very similar to those we find in water itself. The random equilibrium can be shifted toward one of these conformations by means of two stages: a fast stage, during which the unfolded polypeptide becomes a molten globule; and a slow stage, in which the molten globule slowly transforms into a fully folded form or native state (Huang, 2005). These two stages in protein folding can be illustrated by a ''folding funnel'', during which due to a small change in entropy with a large loss of energy, a molten globule evolves into the native state (Fig. 1a) (Dobson, 2003; Gsponer & Vendruscolo, 2006).

Although the process is extremely efficient, there is always the possibility that this accurate mechanism might fail, and the possibility of finding a protein folded into a non-native state becomes a reality (Dobson, 1999). Proteins that follow this pathway might present transiently stable conformations, promoting their interaction with other molecules and facilitating the fact that they might form amorphous oligomers and end in a state of aggregation. Aggregation does not arise from a random coil state, but rather from a series of intermediates that—based on the type of secondary structure acquired during folding—

might or might not resemble the native state (Fig. 1b) (D. Eisenberg et al., 2006, Gsponer & Vendruscolo, 2006). It is well known now that primary polypeptide sequences become the key factor during this process, while the environment surrounding the protein is an important factor for explaining the folding process (Fink, 1998). On the other hand, natively unfolded proteins, known to lack the presence of permanent secondary and tertiary structures, have been recognized at least in the absence of other proteins, to present the tendency to organize themselves into amyloidogenic structures. Considering that the native state is located at the lowest minimum of the ''folding funnel'', it indicates that this region is the most thermodynamically stable configuration of the polypeptide chain under physiological conditions. For proteins, whose functional state is a tightly packed globular fold, a key step in fibril formation related to partial or complete unfolding is less likely to occur and therefore remains protected against aggregation (Dobson, 2004). In this respect, it has been proposed that the more transient structures thus formed in proteins, the better probability for key determinants in amyloid fibril formation to be found (Ohnishi & Takano, 2004). Thus, many of the known forms of amyloid diseases associated with genetic mutations that decrease protein stability and promote unfolding (Ohnishi & Takano, 2004), are both related to disorder-to-order conformational transitions.

The first experimental evidence about a specific disorder-to-order transition was presented over 30 years ago with the mechanism description for the conversion of trypsinogen to trypsin (Bode & Huber, 1976). This mechanism is characterized by the enzymatic removal of a hexapeptide from the N-terminal region of trypsinogen in order to form trypsin. This basic change promotes the transition from a disordered state of the "specificity pocket'' in trypsinogen to an ordered state in trypsin (Huber & Bode, 1978). Since it is known that several amino acids that make up a protein strongly favor a disordered state, at present this ''new view'' of folding is beginning to be further studied, in which the influence of external or environmental conditions sustains well-tested transitions between disordered and ordered states. Specific polypeptide chains contained in proteins or complete proteins

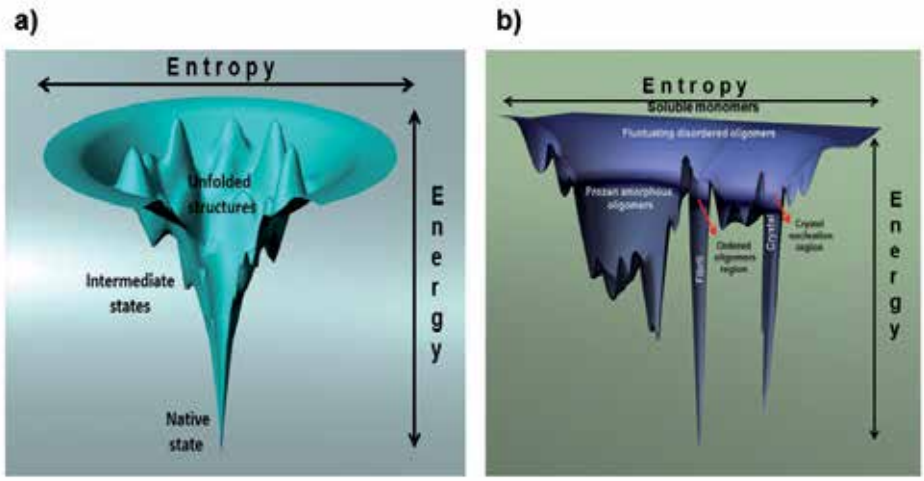


Fig. 1. a) Folding funnel energy landscape b) Protein aggregation energy landscape.

lacking defined tertiary structures are known to have the capacity to undergo disorder-to-order transitions upon binding to specific (Tompa, 2002) or multiple partners (James & Tawfik, 2003). It is precisely this ability that allows the concept of ''protein disorder'' to be proposed as an important feature in the capability of proteins to present regions with switching properties (Bustos & Iglesias, 2006; Dalal & Regan, 2000; Kriwacki et al., 1996).

From an evolutionary point of view, it appears that intrinsic disorder in proteins might have been the driving force behind many of the adaptability processes found in proteins (Dobson, 1999; Dunker et al., 1998). Taking into account that the number of proteins presenting disordered regions directly related with function and therefore with disease is increasingly growing, an interest to also generate accessible data banks for improving information management has increased. Therefore, the database of disordered proteins (DisProt) was created and released in August 2006 by the group of Dunker (Sickmeier et al., 2007) with extremely good results at present (Cortese et al., 2008). Since then, other systems for studying disorder in proteins have been released, such as the Integrated Protein Disorder Analyzer, which aims at identifying and predicting disordered region in proteins (Su et al., 2007), or algorithms for predicting and evaluating aggregation ''hot spots'' (AGGRESCAN) (Conchillo-Sole´ et al., 2007). According to Dunker's group and as predicted by the Predictor of Natural Disordered Regions (PONDR) server (Romero et al., 2001), a large percentage of all proteins involved with some sort of a disease have been identified as directly related with disordered regions in proteins closely associated with signaling. From a general point of view, disordered regions in proteins have been divided into the following two classes: the class in which proteins retain a low percentage of secondary structure together with unstable tertiary structures during a molten globule state, recognized as the collapsed class; and second, the extended class in which proteins with a highly extended backbone resemble a β-sheet conformation (Dunker et al., 2001; Uversky, 2002).

In general, proteins containing disordered regions have been recognized as associated with several human diseases, including cardiovascular disease, cancer, degenerative diseases, and diabetes. Interestingly, because in many of these cases cell signaling function has been involved, there is a strong possibility that disorder-to-order transitions in proteins playing normal switching roles in the cell might become distorted and therefore abolish or transform the normal protein–protein language into an aberrant one. Therefore, the basic properties of a switching mechanism must be based on the equilibrium between high specificity and weak affinities accompanied by a large conformational entropy decrease. This phenomenon is based principally on the fact that upon binding, disorder-to-order transitions can overcome steric restrictions and thereby enable larger interaction surfaces in protein–protein complexes than those that could be obtained for rigid partners. Despite the extraordinary importance of this type of transition, we continue to lack detailed biophysical studies that might demonstrate a close relationship between this type of disorder-to-order organization and protein function.

In an attempt to define the possibility that folding key features in proteins could provide us with the manner in which to explain basic issues such as receptor recognition, lipid transfer activity, and self-exchangeability carried out by several lipid transfer proteins including Apolipoproteins (Apos), our group has attempted to address these points by directly measuring molecular conformational changes of Apos at air/water and lipid/water

interfaces, in order to approach the possible mechanisms that might explain these phenomena (Xicohtencatl-Cortes et al., 2004a, 2004b). As described below, this has been achieved employing Langmuir monolayers in conjunction with Brewster angle microscopy (BAM), atomic force microscopy (AFM) of Apos LB films (Bolaños-García et al., 1999, 2001; Mas-Oliva et al., 2003; Xicohtencatl-Cortes et al., 2004a), grazing incidence X-ray diffraction on protein monolayers (Ruíz-García et al., 2003), and surface force measurements (SFA) (Campos-Terán et al., 2004; Ramos et al., 2008). Because at that time, we were unable to define whether the secondary structure of specific segments of Apolipoprotein CI (ApoCI) and AII (Apo AII) remained stable independently of their position at air/water and lipid/water interfaces, recently we have addressed the possibility that these segments responding to specific environmental changes and following disorder-to-order transitions might function as molecular switches that trigger function (Mendoza-Espinosa et al., 2008, 2009). Moreover, following the same approach with specific peptides synthesized from the reported structure of Apolipoprotein AI (Apo AI), we have found that when left in water at 4°C a very slow disorder-to-order transition develops over the course of days, from a fully disordered state to a well-developed β-sheet secondary structure. This behavior further supports the fact that the physicochemical characteristics of the environment must be considered as a key factor in the equilibrium displacement within the secondary structure of a protein or specific segments toward α-helices or β-sheets (Andreola et al., 2003). Here, the result that specific segments of Apo AI slowly develop fibril-like structures indicates the possibility that pathological processes such as atherogenesis might be also considered as an amyloidotic-related process (Fig. 2) (Mendoza-Espinosa et al., 2009; Westermark et al., 1995). New results related to these studies are also described in this chapter.

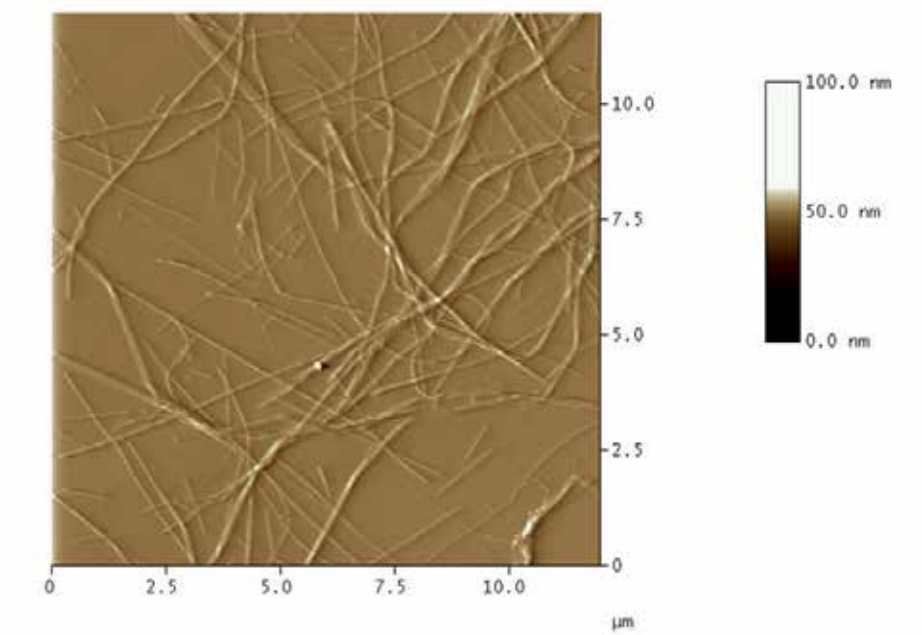


Fig. 2. Atomic force microscopy image (12 x 12 µm) of apolipoprotein AI-peptide DRV fibrils (amino acids 9–24). Fibrils show an average length of 300 nm and 25 nm in height.

2. Structural characteristics of Apolipoproteins CI, AII and AI

Apolipoproteins (Apos) are membrane active proteins that are constituents of high-density lipoproteins (HDL), which are related to the reverse cholesterol transport (Despres et al., 2000). These proteins have an amphiphilic character, since a polar protein face is formed by charged amino acid residues clustered on one side of the α-helices, whereas a hydrophobic surface composed of non-polar residues is formed at the opposite face (Bolaños-García et al., 1997). When Apos are in contact with a polar/non-polar media, their natural tendency is to anchor the hydrophilic and hydrophobic regions in the polar and in the non-polar media, respectively. Thus, a hydrophobic/hydrophilic interface tends to induce a specific orientation on the adsorbed molecules. Some lipoprotein-bound Apos are able to dissociate from the lipoprotein surface in a lipid-poor form, and then transferred through the plasma serum to other lipoproteins (Castro & Fielding, 1984; Clay et al., 1999; Liang et al., 1995; Wang, 2002; Weinberg & Spector, 1985). Although, this mechanism is poorly understood, it is known to be conducted by interactions between Apolipoproteins located at the lipid surface. Apo CI, AII and AI are members of this family of proteins that apparently give lipoproteins directionality and the ability to interact with receptors at the surface of cells.

Apo CI is composed of 57 amino acid residues in length, with a molecular mass of 6.63 KDa. This protein plays a key role in the chylomicron uptake (S. Eisenberg, 1990) and in the regulation of apolipoprotein-E/β-VLDL (very low-density lipoproteins) particle interaction (Swaney & Weisgraber, 1994). Secondary structure predictions, nuclear magnetic resonance, and circular dichroism studies made on Apo CI have revealed a high α-helix content, distributed in two α-helices (Bolaños-García et al., 1999). The first α-helix (residues 4-30) presents approximately 7.5 periods, while the second one (residues 35-53) consists of 5.2 periods (see Fig. 3). In addition both α-helices present important hydrophobic moments (μH) (Bolaños-García et al., 1999). Apo AII is the second major apolipoprotein of high-density lipoproteins (HDL) and it is synthesized in the liver (Eggerman et al., 1991). This protein has been suggested as a modulator of reverse cholesterol transport rather than a strong determinant of lipid metabolism (Tailleux et al., 2002). Apo AII is formed by two identical polypeptide chains connected by a disulfide bridge at position 6, where each chain corresponds to 77 amino acid residues in length and a molecular mass of 8.708 kDa (Brewer et al., 1972, 1986). Predictive and circular dichroism studies (Bolaños-García et al., 1997, 2001), as well as high-resolution crystal structure studies (Kumar et al., 2002) have shown that each chain of the Apo AII presents two α-helix motifs (segments encompassing 7-27 and 32-67) as its main secondary structure (see Fig. 3). These α-helices present an important hydrophobic moment, have approximately 31.5 and 54 Å in length and they are connected by a short peptide chain as a loose hinge (Bolaños-García et al., 2001). Correlation between protein stability to thermal denaturation and secondary structure content has also been investigated (Bolaños-García et al., 2001).

Apo AI has been studied in its free state and membrane models due to the importance that involves understanding the processes that give rise to nascent HDL, as well as the precise mechanisms that support these phenomena in relationship with the process of reverse cholesterol transport. The 243 amino acid polypeptide chain of the Apo AI is organized in blocks of 22 and 11 residues, which are predicted to form helix-type amphipathic

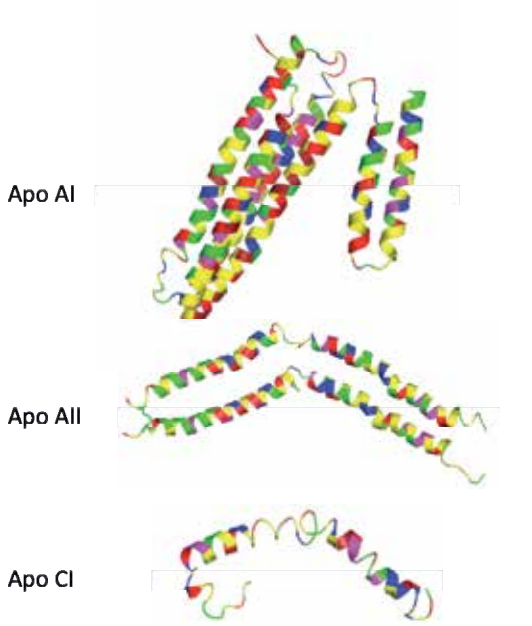


Fig. 3. Secondary structure images of Apolipoproteins showing their α-helical conformation. The color code for the residues is as follows: aromatic-magenta, aliphatic-yellow, polar non charged-green, positively charged-blue, negatively charged-red.

segments (see Fig. 3). The helices that make up the Apo AI, have been classified as follows: (1-45 aa) G*, (44-65,66-87,121-142,143-164,165-186,187-208 aa) A_1, (88-98,99 - 120,209-219,220-241) Y. Helices classified as type G correspond to amphipathic helices that form the interior of globular proteins, reason why amino acids they contain correspond to a hydrophobic type character. Amphipathic helices of the A_1 type have as a characteristic the presence of positively charged amino acids at the hydrophobic/hydrophilic interface, while the negative residues are in the center of the polar face. On the other hand, Y-type helices present the characteristic of having positive charged aminoacids separated by negative ones (Segrest et al., 1992). Currently, there are two crystal structures of the lipid-free Apo AI in different conformations. In the crystal structure obtained by Borhani et al. (Δ1-43) (Borhani et al., 1997), the N-terminal segment is truncated. This structure is unique in presenting a conformation similar to the one that would be in the presence of lipids. Also, the Apo AI (1-243) structure obtained by Ajees et al. (Ajees et al., 2006), presents two domains formed by four α-helices in the N-terminal and 2 α-helices in the C-terminus. Spectroscopic techniques have shown that the lipid-free Apo AI in solution presents a three-dimensional arrangement in two domains similar to that observed in the crystal structure, but with much less organization (Tanaka et al., 2008).

3. Monolayer behavior of Apolipoproteins

When Apolipoproteins are in contact with a polar/non-polar media, they will anchor the hydrophilic and hydrophobic regions in the polar and in the non-polar media, respectively. Thus, a hydrophobic/hydrophilic interface tends to induce a specific orientation on the

adsorbed proteins. As mentioned, Apo CI, AII and AI are associated with lipoproteins particles that are modeled (Borhani et al., 1997) as spheres with a shell of a phospholipid monolayer, with the polar head groups oriented towards the aqueous phase, and the core consists of triglycerides and cholesterol esters (hydrophobic region). In these models, Apos are usually oriented parallel to the surface of the lipoprotein particles. A way to understand the behavior of Apos on the lipoprotein surface is to deposit them on an interface that models the lipoprotein surface, which could be increasingly complex as needed.

The first attempts in this direction have used Apo CI and AII Langmuir monolayers deposited at the air/water interface (Bolaños-García et al., 1999, 2001). For both proteins the compression isotherm showed two first-order phase transitions (see Fig. 4). The first one corresponds to the coexistence between a liquid (L) and a gaseous (G) phase where the proteins have low interaction. The second transition involves two condensed phases; the liquid phase, L, and a condensed phase denoted by LC. For the case of Apo CI, this second transition occurs at a surface pressure (Π) of approximately 33 mNm^{-1} and at an area (A) between 350 and 600 $Å^2$ /molecule. For Apo AII it was found at Π~30-35 mNm^{-1} and

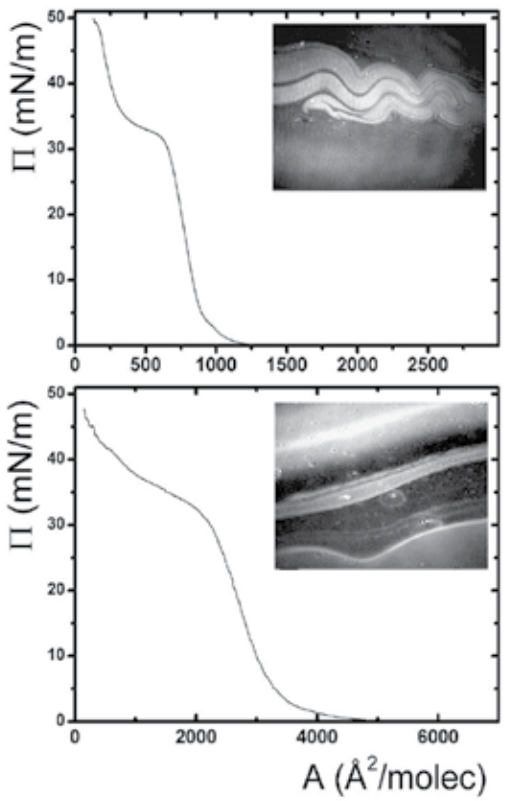


Fig. 4. Langmuir monolayer isotherms of Apo CI (upper panel) and Apo AII (lower panel) at 25.1 °C. Both proteins were dispersed over a phosphate buffer subphase (pH=8.0) containing 3.5 M KCl. Insets show BAM images at the L-LC coexistence. Adapted from (Xicohtencatl-Cortes et al., 2004a).

A~1000-2500 Å^2/molecule. Brewster angle microscopy (BAM) images taken at this transition showed the L phase as dark regions while the LC phase was clearly observed as very bright domains. In the liquid phase, the protein configurations are restricted to a horizontal orientation at the interface due to the amphiphilic character of these proteins. As the surface area is decreased on isothermal compression, one of the α-helix segments for the case of Apo CI and two for Apo AII are expelled from the interface. Direct evidence of this conformational change, as well as of the α-helix structure of Apo CI and AII, have been shown using grazing incidence X-ray diffraction and atomic force microscopy (AFM) of Langmuir-Blodgett (LB) films of transferred protein monolayers (Ruíz-García et al., 2003). It is important to mention that a similar behavior was observed for Apo AI (Bolaños-García et al., 2001).

Experiments on more complex interfaces that are closer to the lipoprotein surface have been prepared adsorbing Apo CI and AII on rac-1,2-dipalmitoyl-sn-glycero-3-phosphocholine (DPPC) monolayers, which indicate that Apolipoproteins can penetrate the DPPC monolayer to form part of the monolayer at the air/water interface (Xicohtencatl-Cortes et al., 2004a). These monolayers also present two clear phase transitions between condensed phases, as well as one between a condensed phase and a gas phase. In this case, the Langmuir monolayer and BAM observations revealed that below surface pressures of 10 mN/m it was possible to have a 2D isotropic mixture where the surface area of the monolayer was approximately the sum of the area occupied respectively by the protein and DPPC molecules as if they were pure components. As the surface pressure is increased and it reaches the condensed phase transition at Π~24-31 mNm^{-1} there is a important loss of monolayer area with an increasing brightness in one of the condensed phases, as seen with BAM images (see Fig. 5). Taking into account this observations and that the Π values for this condensed transition of the

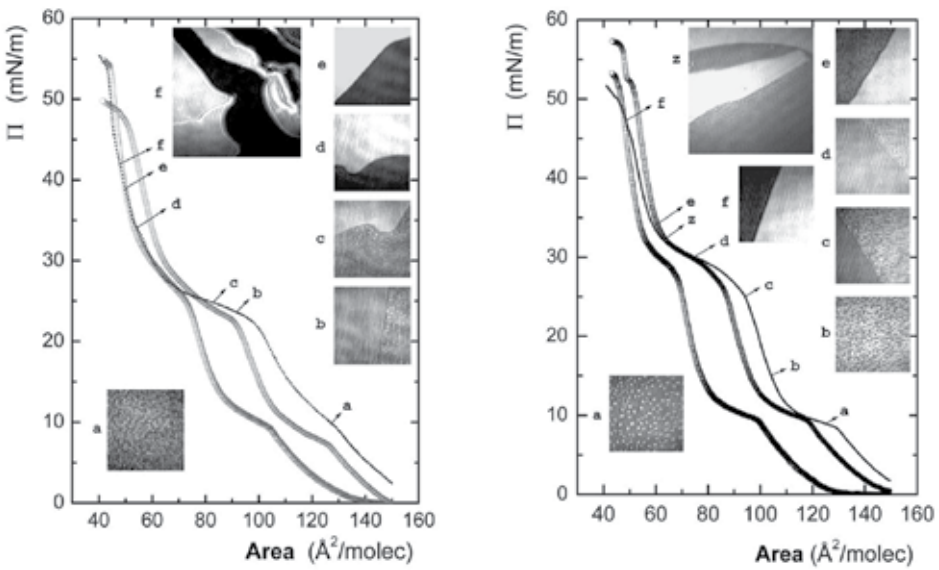


Fig. 5. Langmuir isotherms for Apo CI/DPPC (left, nominal protein mole fraction, from left to right x=0.04, 0.05, 0.12) and Apo AII/DPPC (right, nominal protein mole fraction, from left to right x=0.01, 0.02, 0.03). Insets shows BAM images at different lateral pressures. Adapted from (Xicohtencatl-Cortes et al., 2004a).

binary system are similar to the ones found for the proteins as single components, it was proposed that here there was also a conformational change of the Apo where α-helix segments desorb from the interface, aligning and following the DPPC tails inclination (Xicohtencatl-Cortes et al., 2004a).

4. Forces between adsorbed layers of Apolipoproteins

4.1 The surface force apparatus technique

Forces that control the interaction between proteins, proteins and surfaces and surfaces with adsorbed proteins, are the result of different contributions as hydrophobic interaction, entropy gain due to counterion release, van der Waals force, and to a large extent electrostatic interactions, where the latter is governed by variables like pH and salt concentration. All these interactions depend on the kind of surface and solution where the proteins are immersed, as well as on their charge, shape, and conformation. The surface force apparatus (SFA) (Israelachvili, 1973; Parker et al., 1989) offers the possibility of measuring long-range and contact forces between two mica surfaces covered with adsorbed proteins (Claesson et al., 1995), as well as, measuring the absorbed layer thickness and its compressibility. The latter parameter can give information about the conformational structure and size of the adsorbed protein.

The SFA instrument and experimental procedures have been described by Israelachvili (Israelachvili & McGuiggan, 1990) and Parker (Parker et al., 1989). In general, the force is measured between two curved molecularly smooth mica surfaces (typically 1 cm^2 of area with 2-5 μm constant thickness) where a silver layer of about 520 Å thick was deposited through evaporation on one side of each surface. After that the mica pieces are glued with an epoxy resin, with the silver side down, onto optically polished half-cylindrical silica disks (mean radius of curvature, R, ~ 1-2 cm) that are finally mounted in a crossed cylinder configuration on the SFA. Here, one of the disks is mounted on a double cantilever spring (spring constant, k, ~105 N/m) and the second one on a piezoelectric crystal. This setup produces an optical interferometer. The separation between the two surfaces, d, is controlled by the piezoelectric crystal and the absolute distance is measured interferometrically using fringes of equal chromatic order (FECO) with an accuracy of 2 Å. The magnitude of the force, F, as a function of the surface separation, normalized with respect to the mean radius of curvature, can be determined from the spring deflection measured down to ca. 10^{-7} N. Usually, a SFA experiment starts with the measurement of a standard force curve of water or a buffer solution. If the surfaces contact position is clean and the force curve measured is consistent with the theoretical Derjaguin, Landau, Verwey and Overbeek (DLVO) theory predictions, a known amount of protein is added to the SFA chamber to allow a slow adsorption to the surfaces from the surrounding solution. Then the force curves are usually measured at different times to evaluate this protein adsorption process.

With the SFA, as with other force measurement techniques, one has to consider that the comparison between theoretical and experimental force curves is not straightforward, since the measured force is the sum of different contributions, which are interrelated and therefore not easy to separate. In general, the electrostatic-double layer and the van der Waals forces are considered the most important contributions. However, an absolute

determination of the magnitude of each of these forces is complex, due to factors as protein and surface charge density, protein concentration and solution ionic strength, contribution from steric interactions at short distances, etc. In addition, the location of the plane of charge and the dielectric properties of the adsorbed protein layer usually cannot be determined unambiguously. Nevertheless, the results from SFA studies of the interaction between layers of globular proteins, like insulin and lysozyme, and of proteins with disordered structures have increased our knowledge on the proteins adsorbed layer structure (Claesson et al., 1995). This also includes our SFA studies with proteins formed mainly by α-helices, which will be described below.

4.2 Force measurements with Apolipoproteins deposited on hydrophilic surfaces

In general, the force curves measured between hydrophilic surfaces with adsorbed layers of Apos are mainly composed of electrostatic double layer forces at large surface separations and of steric repulsive forces at small distances. These steric forces are quite interesting since they give some insights of the preferred Apos conformations and the interaction produced by them. Apos amphiphatic structure produces a directional adsorption where the hydrophilic faces of the protein α-helices prefer to be adsorbed onto the mica leaving the hydrophobic faces of the α-helices in contact with water. As an example, figure 6 shows the force curves measured, using a SFA, between two mica surfaces adsorbed with Apo AII. In this case, the adsorption was produced from the protein buffer solution (acetic acid-sodium acetate, pH=4) that surrounds the surfaces. Also, a sequential increment of the protein concentration from 0.002 to 0.004 mg/mL was produced to observe the effect in the surface adsorption (Ramos et al., 2008).

As it can be observed, no forces were found until a surface separation of 700 Å was reached. From there, if the surfaces are brought together, a long-range repulsive force is observed until it is overcome by an attractive force (inward jump), which brings the surfaces from a surface separation of about 130-200 Å into a closer contact. The surface separation where the attractive force drives the surfaces close together decreases with adsorption time and it disappears if the protein concentration is increased (see the force curves with 0.004 mg/mL after 8 hrs of adsorption). In some cases, a small repulsive force was found before reaching a repulsive hard wall. The hard wall at the lower concentration was found to be at a surface separation, d, of 11 Å. Interestingly, when the protein concentration was increased (to 0.004 mg/mL) the surface separation value for the hard wall increases with protein adsorption time with approximately 10 Å increments. Also, an adhesive pull-force was found when the surfaces are taken apart, which decreases with protein adsorption. It was also observed that the force curves were the same on compression and on separation if the surfaces are not brought closer than the inward jump. Similar force curves were found for Apo CI (Campos-Terán et al., 2004).

In this case, the long-range repulsive force and the attractive force can be fitted using DLVO theory including additive contributions of non-retarded van der Waals forces and the electrostatic double layer force (see Fig. 6). Calculations of double layer force were performed with the algorithm of Chan et al. (Chan et al., 1980) bringing into play both constant surface potential and constant surface charge. In practice, it is most likely that both potential and surface charge vary as the surfaces approach, where the actual double layer

force, as it is in this case, falls between these two limits due to the proteins charge regulation. Although DLVO theory does not take into account additional forces occurring between the surfaces, e. g., hydration forces, hydrophobic forces, and steric forces, etc., the fitting is quite good, and the attractive force measured is close to what theory suggests, at constant surface potential.

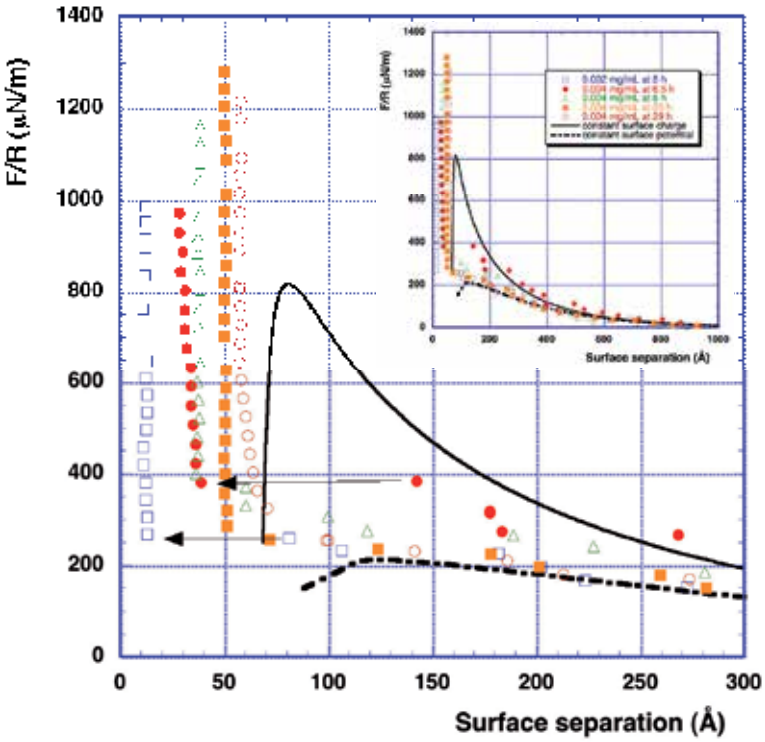


Fig. 6. Force normalized by the radius of curvature, F/R, as a function of surface separation and total adsorption time between mica surfaces adsorbed with Apo AII. The total protein concentration was increased at two times during the experiment and it is: 0.002 mg/mL at 5 h (□), 0.004 mg/ml: at 6 ½ h (●), at 8 h (Δ), and 25 h (■), 29 h (○). Lines indicate DLVO fitting (0.004 mg/ml) with constant surface charge and dashed lines with constant surface potential. Arrows indicate attractive jumps. Adapted from (Ramos et al., 2008).

4.2.1 Short-range forces probe the orientation of the adsorbed protein

As mention before, the analysis of the measured force curves can give an insight of the protein conformation at the surfaces. For the case described above it was seen that the average distance where attractive force appears, d~150 Å, is close to double of the maximum length of this proteins (~85 Å), which suggests that entire protein is oriented perpendicular to the surface or that individual protein segments, i. e, α-helices, protrudes from them. These protein segments could take part in bridging between the two surfaces and thus be responsible for the attractive force. This kind of attractive force was also observed in adsorbed surfaces with Apo CI (Campos-Terán et al., 2004). In addition, studies of Apo AII

and Apo CI monolayers have shown that it is possible to form a layer with protruding segments at an interface (Bolaños-García et al., 2001; Ruíz-García et al., 2003). The fact that at short adsorption time, it was observed a weak repulsive force, suggests a more extended conformation of the adsorbed proteins. Such protruding segments could be compressed, bearing in mind the relative flexibility of the polypeptide chains connecting the α-helices. Given enough time for adsorption, the protein molecule preferentially will be oriented parallel to the surface and hence, this repulsive force disappears. The driving force for protein reorientation is to avoid the exposure of hydrophobic segments to the aqueous environment, as well as to promote the electrostatic attractive interactions between the protein and the surface. At low concentrations, the attractive interaction is reduced with time, which implies a higher protein surface coverage confined in a thin layer. This thin protein layer was found experimentally since it was found a surface separation of just 11 Å at the hard wall position. This final layer thickness value is between 5 to 6 Å on each surface, which is similar to the estimated value for α-helices diameter. Previously, it has been shown that structural changes on adsorption are not enough to disturb the α-helix structure (Burkett & Read, 2001).

4.2.2 Sequential addition of Apos builds up protein multilayers

Experiments conducted at higher concentrations suggest the build up of more than one layer on each surface since each curve shown in figure 5 represents an increase in hard-wall separation of ~10 Å or approximately 5 Å of thickness on each surface (see Fig. 6). Confirmation of this process was obtained by ellipsometry measurements done by our group (Ramos et al., 2008), which showed that sequential addition of protein (at least at high ionic strength) leads to an increase in the adsorbed amount of protein, as well as the protein layer thickness. However, for the case of Apo AII, the presence of a repulsive interaction showed that protein adsorption do not lead to charge neutralization of the mica surface charge as it was found for Apo CI (Campos-Terán et al., 2004). This is most likely due to the structural difference between both proteins, where the Apo CI monomer can more efficiently arrange so that it better match the surface charge compare to the Apo AII dimer. However, since the apparent surface potential has a small change when the protein concentration is increased from 0.002 mg/ml to 0.004 mg/ml, a charge regulation mechanism involving small ions during the adsorption of the proteins cannot be discarded. This mechanism has been observed to occur in the surface adsorption of other proteins (Claesson et al., 1995). Protein multilayer adsorption has been observed in proteins with amphiphilic or flexible segments, as observed at SFA experiments with cytochrome c (Kekicheff et al., 1990) and β-casein (Nylander & Wahlgren, 1997). A multilayer protein adsorption requires attractive protein-protein interaction, which often is weaker than the protein-surface interactions. Confirmation of this statement was obtained by diluting the solution surrounding the surfaces. Here, the hard wall separation decreased from d ≈ 58 Å to d ≈ 26 Å and the apparent surface potential has increased from ~37 mV to ~53 mV 1 ½ h after dilution, which indicates protein desorption. Even more protein has desorbed after 18 h, and the hard-wall separation reaches d ≈ 11 Å, corresponding to one monolayer on each surface. In addition, an attractive jump appears. No further desorption occurs, which is mostly likely due to the strong interaction between the negatively charged mica and the cationic protein as well as the entropy gain due to counter ion release. This experiment

showed the reversibility of Apo AII adsorption process that produces in each stage different protein conformations. Also, it is noteworthy that quite similar force curves were observed in Apo CI (Campos-Terán et al., 2004), which also has a similar secondary structure but with a different peptide sequence, different net charge, and it is only monomeric. Therefore, the observed force curves seem to be a consequence of the particular features of the amphiphilic α-helices.

5. Lipid dependant disorder-to-order conformational transitions in Apolipoproteins

5.1 Apo CI derived peptides-lipid interaction

As mentioned in sections 3 and 4, it has been observed for exchangeable Apolipoproteins that when they are subjected to lateral pressures then several helical segments were placed directly in the hydrophobic phase of the interface. In the case of Apo CI, our data showed an interesting new property since we observed that injecting it into the subphase allows the protein to go to the water/lipid interface quickly and when lateral pressure is increased the C-terminal helical segment penetrates the monolayer. In addition, when lateral pressure is released, this segment is again incorporated into the water/lipid interface (Bolaños-García et al., 1999). With these results we were interested to know if the secondary structure of the C-terminal segment of Apo CI remained stable regardless of their position in the different hydrophilic/hydrophobic interfaces. To solve this question we conducted studies of peptides derived from the C-terminal segment of Apo CI in different environments.

The peptides were designated according to the first three letters of their amino acid sequence and called ALDO (A7-E24), ARELI (A22-M38) and SAK (S35-L53). Apo CI in solution shows a clear circular dichroism (CD) signal associated with a high degree of α-helix structure (Bolaños-García et al., 1999). However, when peptides ALDO, ARELI and SAK (Mendoza-Espinosa et al., 2008) were tested under the same experimental conditions, they showed no defined secondary structure and remain non-structured independently of pH, temperature and ionic strength. Interestingly, despite that these peptides have an amphipathic character and high hydrophobic moment values (μH > 0.315 kcal/mol), they remain completely unfolded in solution (see Fig. 7a).

Nevertheless, when peptides ALDO, SAK and ARELI are placed in aqueous solution with 40% v/v trifluoroethanol (TFE) or sodium dodecylsulphate (SDS, cmc of 8.5 mM) they show a CD signal clearly associated with an α-helical structure. If SDS was used at different concentrations (1.5-20 mM), each of the peptides acquire secondary structures in a differentiated way, where the lowest percentage of α-helix structure corresponds to ARELI and the highest to SAK peptide, which corresponds to the C-terminal segment of Apo CI (see Figs. 7b and 7c). Then in order to test the possibility that specific lipids on the surface of lipoproteins and plasma membrane induce an α-helix conformation as in the case of TFE and SDS, we tested a series of phospholipids above and below its critical micelle concentration and with different acyl long chain to probe their hydrophobic effect. L-α-Phosphatidylcholine (PC) was used above its critical micellar concentration (cmc <0.005 mM) and 1,2-dihexanoyl-sn-glycero-3-phosphocholine (DHPC) slightly below its cmc (~15 mM), because concentrations above the cmc of DHPC generate solutions that prevent the

determination of the CD signal (data not shown). Under these conditions, only peptide SAK showed a well-defined disorder to order type transition. Since medium hydrophobicity seems to be critical for the transition to be observed, we tested if these lipids mixed with small amounts of cholesterol altered these low percentages of α-helix, finding no changes (data not shown).

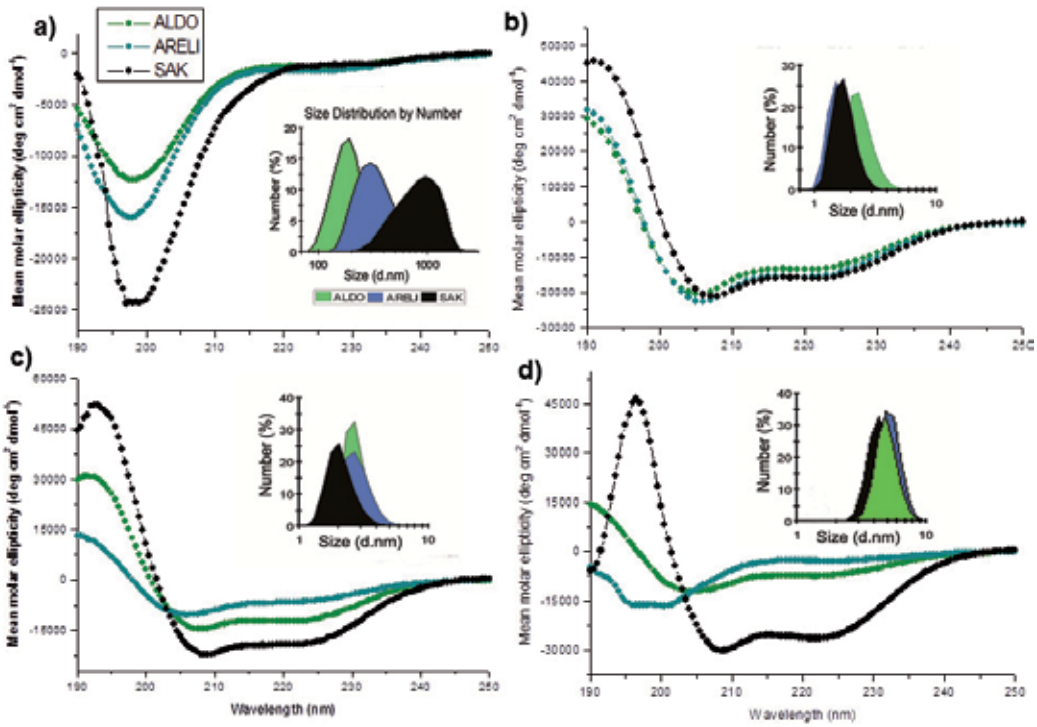


Fig. 7. Far-UV CD data of Apo CI-derived peptides. a) Spectra recorded in water for peptides ALDO, ARELI, and SAK. b) In the presence of 40% v/v TFE. c) In the presence of SDS (20 mM). d) In the presence of lyso-C_{12}PC (20 mM). Insets: DLS analysis of the same corresponding peptide solutions employed for CD experiments. Adapted from (Mendoza-Espinosa et al., 2008).

Tests performed with 1-hexanoyl-2-hydroxy-sn-glycero-3-phosphocholine (lyso-C_6PC) do not promote any change in the secondary structure of all peptides studied (data not shown). However, in the presence of 1-lauroyl-2-hydroxy-sn-glycero-3-phosphocholine (lyso-C_{12}PC), peptides corresponding to the N-terminal (ALDO) and C-terminal (SAK) segments acquired different percentages of α-helix structure as a function of the lysophospholipid concentration. ALDO was the peptide less sensitive to lyso-C_{12}PC, however, the effect of this lysophospholipid is greater than that generated by SDS. On the other hand, SAK presented disorder-to-order transitions from the lowest levels of lyso-C_{12}PC. Greater effect was observed for this lipid on this peptide compared to the one observed with SDS and TFE molecules (see Fig. 7d). Interestingly DLS experiments showed that peptide solutions with

pure water and lyso-C_6PC in which there was no disorder-order transitions, presented aggregates in solution. In contrast, for the peptides-lyso-C_{12}PC solutions that generated disorder-to-order transition and the promotion of a well-defined α-helical conformation, allows the association of lipid/peptide molecules in such an orderly fashion that the system avoids aggregation. It is interesting to note that while lyso-C_6PC aggregates increase in size in the presence of peptides for the case of lyso-C_{12}PC the size does not change with or without the same peptides (see Fig. 8).

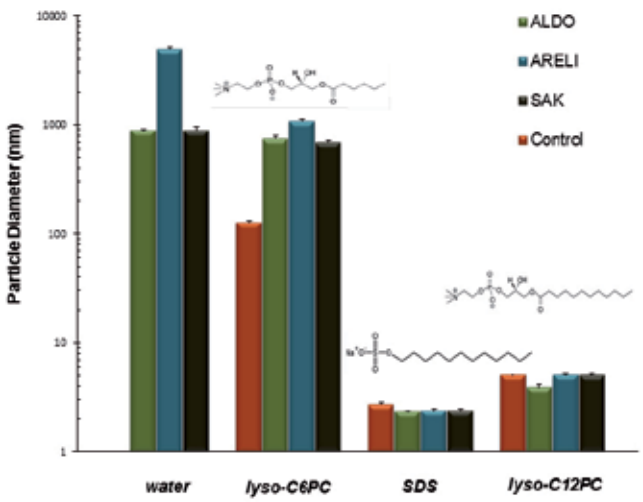


Fig. 8. Dynamic light scattering of Apo CI peptides associated with lysophospholipids of different acyl-chain lengths. a) Quantification of particle diameters given by peptide/lipid aggregates in the presence of SDS (20mM), lyso-C_6PC (20 mM) and lyso-C_{12}PC (20 mM). Adapted from (Mendoza-Espinosa et al., 2008).

Based on data obtained in this study, we have elucidated a mechanism by which the Apo CI could be functioning as a molecular switch on the surface of HDL. In this scenario we propose that Apo CI responds to a decrease in lateral pressure on the surface of HDL, which is given by an increase of cholesterol ester in the nucleus (Frank & Marcel, 2000), by promoting its C-terminal segment to the polar/nonpolar interface of the lipoprotein particle with a concomitant change from a disordered structure to an α-helix. The fact that the surface of lipoproteins and certain types of membranes are associated with the presence of molecules such as lyso-C_{12}PC could generate dramatic disorder-to-order transitions in the C-terminal segment of Apo CI. In consequence, these conformational changes generated by Apo CI could be related to the biological activity of molecules such as esphingosine 1 phosphate that when associated with HDL particles it has been observed that promotes an anti-inflammatory effect and therefore presents a potential role as atheroprotective (Jerzy-Roch & Assman, 2005).

Also, since cholesterol esters formed by the enzyme lecithin-cholesterol acyltransferase (LCAT) located at the surface of HDL particles promotes the transfer of a fatty acyl group from position two of phosphatidylcholine to cholesterol, with the consequent synthesis of lysophosphatidylcholine, it is possible that the presence of new OH groups at the polar/non-polar interface change the electrostatic properties of the interface and the way water is displaced from the interface during peptide folding. In fact, it has been proposed that in the presence of lipids, the process of peptide folding corresponds to an enthalpy driven process supported by the energy employed for water displacement (Rozek et al., 1997). Localized changes in secondary structure of a number of proteins have been found to be physiologically relevant (Chakrabartty & Baldwin, 1995; Meador et al., 1992). Therefore, a series of conformational switches have been proposed, in specific cases, to promote protein activation (Wei et al., 1994) and folding (Hamada et al., 1996). In order to find out the mechanism by which lyso-C_{12}PC is required to induce an important conformational change, further investigation is needed. These important changes might be important in the understanding of the mechanisms Apo CI employs to modulate protein/protein recognition directly related to enzyme activation and modulation of Apo E and the cholesterol ester transfer protein (CETP) function when associated to the surface of HDL particles. Our proposal of a lipid dependant disorder-to-order conformational transition in Apo CI might be considered a conformational switch mediating enzyme activation and lipid transport. This possibility opens new ways to visualize the concert of events that take place at the surface of HDL during their transformation from early protein/lipid discoidal aggregates to spherical particles, ready to be taken up by liver cells. Further investigation of this potential mechanism designed to recognize and promote localized secondary structure conformations in proteins, undoubtedly will provide an improvement to better comprehend the protein function at the surface of lipoproteins.

5.2 Apo AI-lipid interactions

Several studies have evaluated the lipid-binding propensity of each of the helices composing Apo AI, noting that the N-terminal domain determines the open or closed structure of the protein when modulated by the presence of cholesterol obtained by interacting with the ATP-binding cassette (ABC) A1 (ABCA1) giving rise to the nascent HDL. Likewise, due to its high hydrophobicity, the C-terminal domain of Apo AI facilitates anchoring to lipid membranes (Fang et al., 2003; Kono et al., 2008). These studies are based on the widely accepted model for discoidal HDL, which corresponds to a disk made of a lipid bilayer surrounded by two Apo AI helices with its long axis perpendicular to the acyl chains of phospholipids (Garda, 2007). These properties can be easily observed in the hydrophobicity profile of Apo AI obtained with the use of the EMBOSS Pepinfo algorithm, employing a window of 9 amino acids and the scale of Kyte J. & Doolittle R. F. (Kyte & Doolittle, 1982) (see Fig. 9a). While its negative profile is characteristic of membrane proteins at the N and C-terminal regions of the protein (10-17, 213-229), the positive profile indicates the hydrophobic ones. On the other hand, the use of the Hydrophobic Cluster Analysis (HCA) server, which predicts hydrophobic blocks depending on the secondary structure of the polypeptide chain, shows three highly hydrophobic segments (aa 13-22, 45-49, 216-232) (Fig. 9a, hydrophobic clusters) (Callebaut et al., 1997). The distribution in the helix of negatively charged, positive or neutral aminoacids, generates the different types of helices

present in the Apo AI. This change in the distribution of amino acids is important in the understanding of the way the protein associates with lipids (Segrest et al., 1992). For example, segments corresponding to helices 1-2 (A-type helices) and 9-10 (Y-type helices) are those with the greatest affinity for lipids, which are particularly high in the latter (Mishra et al., 1998).

Interestingly, in two SDSL-EPR spectroscopic studies (electron paramagnetic spin-label resonance spectroscopy), β-type segments were also detected in the N-and C-terminal domains of Apo AI (Lagerstedt et al., 2007; Oda et al., 2003). The possibility of having secondary structure conformational changes has been also observed in other proteins. For instance, the fusogenic HA2 unit of hemagglutinin of the influenza virus has been shown to present these types of conformational transitions. HA2 corresponds to a segment containing 36 amino acids, that presents the ability to carry out transitions from a random coil structure to an α-helix domain. The presence of these secondary structure conformational changes in Apo AI in the presence of lipid could serve as a mechanism to decrease the energy barrier in their interaction with these molecules, a crucial step in the flow of cholesterol and assembly of HDL (Oda et al., 2003; Tamm, 2003).

5.3 Intrinsic disorder in Apo AI

On the other hand, Apo AI is considered within the group of natively unstructured proteins (Uversky et al., 2000). Recently, this type of protein has taken a major importance when giving rise to the term "unfoldomics". A highly dysfunctional group of proteins has been associated with a number of conditions such as amyloidosis, cancer, diabetes, neurodegenerative diseases and others. The altered sites contained in many disordered proteins have been shown to be highly susceptible to proteolysis. In the lipid-free Apo AI, specifically the N-terminal segment has been observed by various techniques such as NMR, EPR that mobility presents a great variability in their secondary structure (Kono et al., 2008; Lagerstedt et al., 2007; Okon et al., 2001, 2002; Wang et al., 1996, 1997). Using the PONDR server with the native sequence of the native Apo AI it was estimated a high percentage of disorder for five segments of Apo AI (Fig. 9b). The first segment (aa 1-10) corresponds to a site that could serve as a lipid sensor when the Apo AI is in the discoidal particle (Kono et al., 2008; Wu et al., 2007). The second site (aa 69-89) has a particularly negative charge distribution compared to the other un-structured sites (Fig 9c). The site also includes a transition between a helix type A to a type Y, which has been postulated could be a destabilizing factor in the continuity of the secondary structure of an α-helix. Wu Z. et al (Wu et al., 2007) have proposed a discoidal HDL model where the third and longest disordered segment (aa 116-150) presents a region that could be considered as a hinge. The same model includes a loop (159-180) that corresponds to the fourth disordered segment (aa 172-194). Also, this site is adjacent to the segment postulated to be the one that interacts with LCAT (aa 159-170). The fifth disordered segment is located close to a transition region from a helix type A to a type Y. Although the latter site presents an α-helix structure in the crystal structure of Apo AI, by EPR tests this segment shows a β-structure that could serve as a mechanism to facilitate the interaction with lipids. Interestingly, very low concentrations of amyloid fibrils formed by a segment of 10 kD N-terminus of native Apo AI, have been found in vivo (Schmidt et al., 1997). These structures are constituted by β-cross structures

that in turn produce β-strands oriented in a perpendicular way with respect to the long axis of the fiber, resulting in its increased spreading capacity. Subsequently, protofibrils associate laterally or rotate together to form fibers of larger diameter (DuBay et al., 2004).

In our laboratory by their structural and perhaps biological importance, we have analyzed three peptides designed according to the sequence reported for the native Apo AI and its crystal structure (Borhani et al., 1997). These peptides are DRV (D9-D24) and KLL (K45-Q63) located within its N-terminal helical segments, and VLES (V221-K239) located in a C-terminal segment. By sequence analysis of native Apo AI employing the Zyggregator server (Tartaglia et al., 2008), it has been observed that this protein presents several sites with the propensity to form amyloid fibrils (A15-D20, W50-F57 and S224-L230), which interestingly enough are included in peptide sequence DRV, KLL and VLES (Fig. 9d, Zagg Propensity). This server uses an algorithm that considers patterns of hydrophobicity, polar amino acids

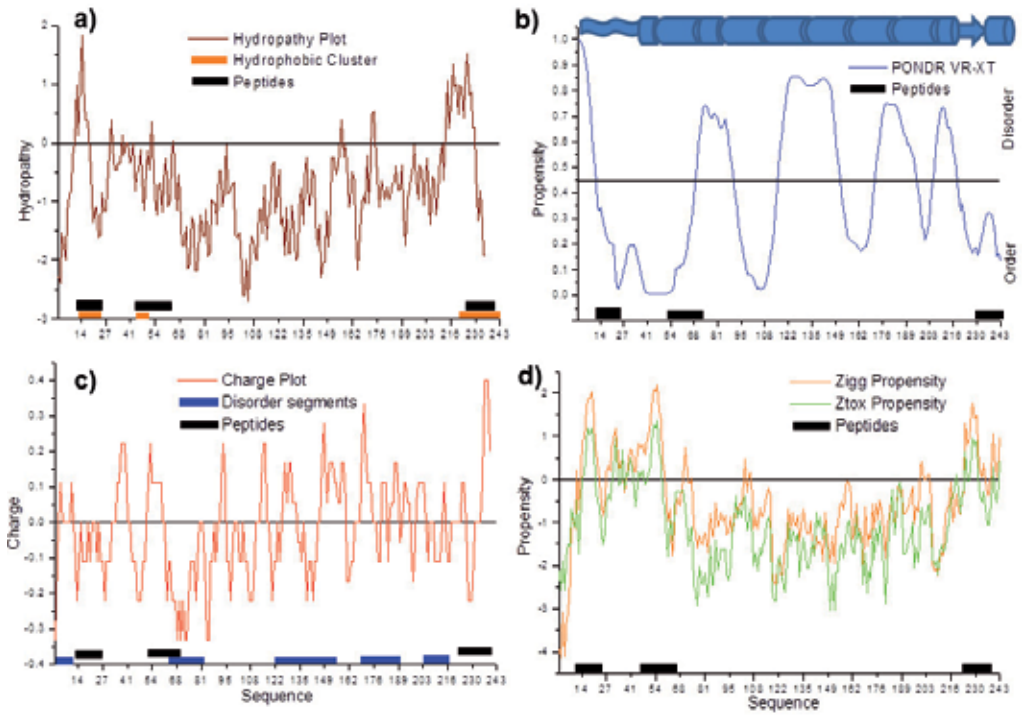


Fig. 9. Methods employed in the prediction of disorder, aggregation and propensity of amyloid fiber formation based on the sequence of native Apo AI. a) Hydrophobicity profile calculated with the EMBOSS server: Pepinfo, using a window of 9 aa and scale of Kyte & Doolittle hydrophobicity, red line. Hydrophobic segments calculated with the HCA server, orange box. Peptide position DRV, KLL and VLES in the sequence of Apo AI, black box. b) Profile of disorder determined by the PONDR server (PONDR VL-XT). c) Load profile calculated with the EMBOSS server: Charge using a window of 9 aa. d) Propensity of amyloid fiber formation with the Zyggregator server (Zagg Propensity) and formation of globular structures (Ztox Propensity).

and aromatic content observed in polypeptide chains of amyloidogenic proteins. The segments prone to aggregation in the Apo AI present sequences of highly hydrophobic blocks composed of six to seven amino acids flanked by negative and/or positive charges. On the other hand, Zyggregator calculates the tendency to form globular structures, which have been observed to be a step in the formation of amyloidogenic fibers. It has been observed with in vitro experiments that these globular structures formed by amyloid peptides are involved in a process of cellular cytotoxicity due the formation of pores in membranes (Lashuel & Lansbury, 2006). The patterns to form amyloid structures in Apo AI calculated by Zyggregator, show the same one that is observed during the formation of amyloid fibers (Fig. 9d, Ztox propensity).

5.4 Amyloidogenic Apo AI

Interestingly, within the existing mutants of Apo AI, there are 4 isoforms associated with systemic forms of hereditary amyloidosis. Mutations correspond to Gly26-Arg, Leu60-Arg, Trp50-Arg and a deletion/insertion of segment (Leu60-Phe71) - (Val-Val-Thr). Considering the structure for the lipid-free Apo AI proposed by Ajees et al., these mutations generally provide positive charges to the hydrophobic interface formed between the two pairs of helices at the N-terminal segment (aa 1-188). The introduction of a polar amino acid residue by the amyloidogenic mutations in the hydrophobic interface of the lipid-free Apo AI, probably prevents the formation of the cluster of α-helices in the N-terminal structure. This also hinders the formation of hydrophilic patches located in different areas of the protein (see Fig. 10) (Oram, 2002). These hydrophilic patches have been postulated to interact with ABCA1 for the transfer of phospholipid and cholesterol. One consequence of this obstacle might be that the formation of a properly sized discoidal HDL needed to interact with the enzyme LCAT, as observed in these mutations, is not properly achieved (Fang et al., 2003; Genschel et al., 1998). This interaction is crucial in the transition from discoidal to spherical HDL (Calabresi & Franceschini, 2010). Discoidal HDL formed by one of the several isoforms of the amyloidogenic Apo AI known nowadays, are rapidly catabolized and do not become spherical HDL (Genschel et al., 1998). At this stage, it is interesting to mention that the metabolic pathway of these Apo AI isoforms that cause deposition of amyloid fibrils, has not yet been clarified. However any of these mutations found in helix 1 and helix G * of Apo AI could be generating a loop susceptible to proteolysis between helices 2 and 3 (Apo AI disordered second site with low affinity to lipid profile and with a distinctly negative charge) (Figs. 9b and 9c). In all cases, amyloid fibers are generated with polypeptides from the first 83-94 amino acids at the N-terminus of Apo AI. Mutations at amino acids 26, 50 and 60 also generate charge changes, characteristic that favors the formation of extended β-sheets (García-González & Mas-Oliva, 2011). These peptides released regardless of the origin of the mutation, always have the same net charge, indicating the conservation of the hydrophilic profile. Likewise, the hydrophobic moment value and the average total hydrophobicity decrease in the mutations with respect to native sequences.

It is remarkable that several theoretical and experimental data related to several N-terminal and C-terminal segments of Apo AI indicate that both have a high propensity for aggregation. However, only the first one was found in the amyloidogenic plaques isolated from familial amyloidosis or Alzheimer-affected people. Segments of Apo AI with a tendency to be maintained in a disordered state, together with their low affinity sites for

lipid, could be the key for the understanding of HDL particles formation. Due to these characteristics, it seems Apo AI has the ability to modulate its secondary structure based on the presence of hydrophobic/hydrophilic interfaces that in turn might activate or inhibit the function of proteins that regulate the metabolism of HDL.

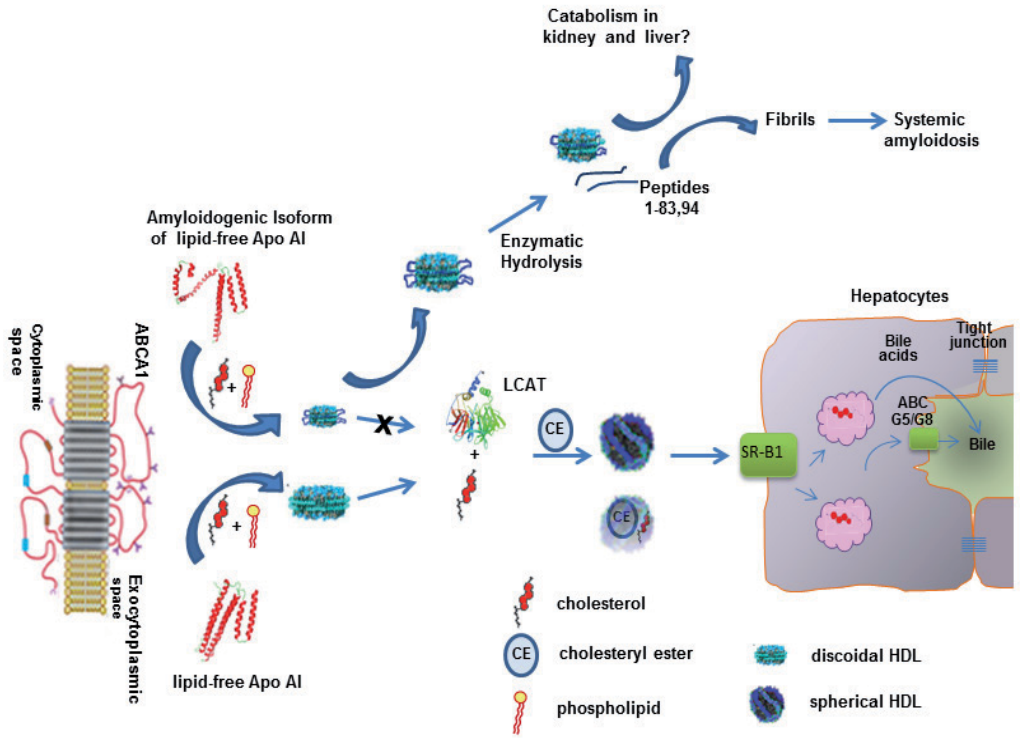


Fig. 10. The ABCA1 receptor promotes the transfer of phospholipids to a lipid-poor form of native Apo AI native, the major component of HDLs. The mechanism by which this process occurs is not fully understood; probably ABCA1 translocates phospholipids and cholesterol from the plasma membrane to HDL Apo AI producing a discoidal particle. These discoidal particles are transformed into spherical HDLs in the blood by the action of the enzyme LCAT. Spherical HDLs associate with the SRB1 receptor in the plasma membrane of hepatocytes and transfer free and esterified cholesterol to the liver for excretion into the bile as free cholesterol (via ABCG5/G8) or subsequent conversion to bile salts (Oram, 2002). In the case of the formation of amyloidogenic isoforms of Apo AI, a proper interaction with ABCA1 is impeded and therefore the formation of discoidal HDLs becomes a deficient process, with the consequent problems in the recognition of LCAT and also SR-B1. It is known that abnormal HDLs are rapidly catabolized releasing different peptides that in turn generate peptides that form amyloid fibers that might initiate a systemic amyloidosis.

6. Conclusions

We believe conformational changes observed in monolayers of Apo CI and AII during lateral compression could be of direct relevance to changes in surface tension at the surface

of lipoproteins. Thus, in response to changes in surface tension of the lipoprotein particle, Apolipoproteins could present structural changes. Interestingly, our working hypothesis is of relevance when extrapolated to the conformational changes observed at the lipid/plasma interface at the surface of a lipoprotein particle. In this case, because of the small size and rich protein composition of discoidal HDL (pre-2HDL β) it is proposed that the lateral pressure in the phospholipid monolayer of these particles is most likely to be high and only decreases in parallel with changes in size and shape of these lipoproteins when they begin to accumulate cholesterol esters to form HDL spherical (α-HDL). The results of our studies employing Apos allow us to postulate that the lateral pressure of the phospholipids monolayer associated with proteins on the surface of the different HDL particles may be very different depending on their size and shape.

In addition, the fact that these Apos could present a different conformation from newly synthesized lipoproteins with a discoidal form to a mature state with a spherical shape could be an important factor in the understanding of their physiological properties such as directionality and receptor recognition. However, the ways in which structural changes induced by lipid interaction modulate the functionality of these Apos are still to be clarified, since the formation of amyloidogenic forms for several segments of these Apos, as presented in this chapter, have also been found to play a critical role in their structure/function relationship.

Currently, the understanding of the mechanisms by which segments, entire native or mutated proteins get transformed into amyloid-like structures, has resulted to be a challenge. Since several disorder-to-order transitions in proteins have been found to be reversible, this phenomenon has been frequently associated with important signaling events in the cell. Due to the central role of this phenomenon in cell biology, protein misfolding and aberrant conformational transitions have been at present associated with an important number of diseases. Nevertheless, differences between ''functional'' and ''pathological'' disorder-to-order or order-to-disorder transitions that might lead to the formation of amyloids, might simply reside in the modulatory pathways involved along their synthesis and the environment proteins or protein segments are placed into.

7. Acknowledgment

Research described in the present chapter has been supported by Consejo Nacional de Ciencia y Tecnología (CONACYT), Red Temática de Materia Condensada Blanda (CONACYT), DGAPA-UNAM (Universidad Nacional Autónoma de México) and UAM (Universidad Autónoma Metropolitana).

8. References

Ajees, A.A., Anantharamaiah, G.M., Mishra, V.K., Hussain, M.M. & Murthy, H.M. (2006). Crystal Structure of Human Apolipoprotein A-I: Insights into Its Protective Effect against Cardiovascular Diseases. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 103, No. 7, (February 2006), pp. 2126-2131, ISSN 0027-8424

Andreola, A., Bellotti, V., Giorgetti, S., Mangione, P., Obici, L., Stoppini, M., Torres, J., Monzani, E., Merlini, G. & Sunde, M. (2003). Conformational Switching and

Fibrillogenesis in the Amyloidogenic Fragment of Apolipoprotein A-I. *The Journal of Biological Chemistry*, Vol. 278, No. 4, (January 2003), pp. 2444-2451, ISSN 0021-9258

Bode, W. & Huber, R. (1976). Induction of the Bovine Trypsinogen-Trypsin Transition by Peptides Sequentially Similar to the N-Terminus of Trypsin. *Federation of European Biochemical Societies Letters*, Vol. 68, No. 2, (October 1976), pp. 231-236, ISSN 0014-5793

Bolaños-García, V.M., Mas-Oliva, J., Ramos, S. & Castillo, R. (1999). Phase Transitions in Monolayers of Human Apolipoprotein C-I. *Journal of Physical Chemistry B,* Vol. 103, No. 30, (July 1999), pp. 6236–6242, ISSN 1089-5647

Bolaños-García, V.M., Ramos, S., Xicohtencatl-Cortes, J., Castillo, R. & Mas-Oliva, J. (2001). Monolayers of Apolipoproteins at the Air/Water Interface. *Journal of Physical Chemistry B,* Vol. 105, No. 24, (June 2001), pp. 5757–5765, ISSN 1089-5647

Bolaños-Garcia, V.M., Soriano-Garcia, M. & Mas-Oliva, J. (1997). CETP and Exchangeable Apoproteins: Common Features in Lipid Binding Activity. *Molecular and Cellular Biochemistry*, Vol. 175, No. 1-2, (October 1997), pp. 1-10, ISSN 0300-8177

Borhani, D.W., Rogers, D.P., Engler, J.A. & Brouillette, C.G. (1997). Crystal Structure of Truncated Human Apolipoprotein A-I Suggests a Lipid-Bound Conformation. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 94, No. 23, (November 1997), pp. 12291-12296, ISSN 0027-8424

Brewer, H.B., Jr., Lux, S.E., Ronan, R. & John, K.M. (1972). Amino Acid Sequence of Human ApoLp-Gln-II (ApoA-II), an Apolipoprotein Isolated from the High-Density Lipoprotein Complex. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 69, No. 5, (May 1972), pp. 1304-1308, ISSN 0027-8424

Brewer, H.B., Jr., Ronan, R., Meng, M. & Bishop, C. (1986). Isolation and Characterization of Apolipoproteins A-I, A-II, and A-IV. *Methods In Enzymology*, Vol. 128, (1986), pp. 223-246, ISSN 0076-6879

Burkett, S.L. & Read, M.J. (2001). Adsorption-Induced Conformational Changes of α-Helical Peptides. *Langmuir: The ACS Journal of Surfaces and Colloids,* Vol. 17, No. 16, (July 2001), pp. 5059-5065, ISSN 0743-7463

Bustos, D.M. & Iglesias, A.A. (2006). Intrinsic Disorder is a Key Characteristic in Partners that Bind 14-3-3 Proteins. *Proteins*, Vol. 63, No. 1, (April 2006), pp. 35-42, ISSN 1097-0134

Calabresi, L. & Franceschini, G. (2010). Lecithin:Cholesterol Acyltransferase, High-Density Lipoproteins, and Atheroprotection in Humans. *Trends in Cardiovascular Medicine*, Vol. 20, No. 2, (February 2010), pp. 50-53, ISSN 1873-2615

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. & Mornon, J.P. (1997). Deciphering Protein Sequence Information through Hydrophobic Cluster Analysis (HCA): Current Status and Perspectives. *Cellular and Molecular Life Sciences : CMLS*, Vol. 53, No. 8, (August 1997), pp. 621-645, ISSN 1420-682X

Campos-Terán, J., Mas-Oliva, J. & Castillo, R. (2004). Interactions and Conformations of α-Helical Human Apolipoprotein CI on Hydrophilic and on Hydrophobic Substrates. *Journal of Physical Chemistry B,* Vol. 108, No. 52, (November 2004), pp. 20442-20450, ISSN 1089-5647

Castro, G.R. & Fielding, C.J. (1984). Evidence for the Distribution of Apolipoprotein E between Lipoprotein Classes in Human Normocholesterolemic Plasma and for the Origin of Unassociated Apolipoprotein E (Lp-E). *Journal of Lipid Research*, Vol. 25, No. 1, (January 1984), pp. 58-67, ISSN 0022-2275

Chakrabartty, A. & Baldwin R.L (1995). Stability of α-Helices. *Advances in Protein Chemistry,* Vol. 46, (1995), pp. 141-176, ISSN 0065-3233

Chan, D.Y.C., Pashley, R.M. & White, L.R. (1980). A Simple Algorithm for the Calculation of the Electrostatic Repulsion between Identical Charged Surfaces in Electrolyte. *Journal of Colloid and Interface Science,* Vol. 77, No. 1, (September 1980), pp. 283-285, ISSN 0021-9797

Claesson, P.M., Blomberg, E., Fröberg, J.C., Nylander, T. & Arnebrant, T. (1995). Protein Interactions at Solid Surfaces. *Advances in Colloid and Interface Science,* Vol. 57, (May 1995), pp. 161-227, ISSN 0001-8686

Clay, M.A., Cehic, D.A., Pyle, D.H., Rye, K.A. & Barter, P.J. (1999). Formation of Apolipoprotein-Specific High-Density Lipoprotein Particles from Lipid-Free Apolipoproteins A-I and A-II. *Biochemical Journal*, Vol. 337 (February 1999), pp. 445-451, ISSN 0264-6021

Conchillo-Solé, O., de Groot, N.S., Aviles, F.X., Vendrell, J., Daura, X. & Ventura, S. (2007). Aggrescan: A Server for the Prediction and Evaluation of "Hot Spots" of Aggregation in Polypeptides. *BMC Bioinformatics*, Vol. 8, (February 2007), pp. 65, ISSN 1471-2105

Cortese, M.S., Uversky, V.N. & Dunker, A.K. (2008). Intrinsic Disorder in Scaffold Proteins: Getting More from Less. *Progress in Biophysics and Molecular Biology*, Vol. 98, No. 1, (September 2008), pp. 85-106, ISSN 0079-6107

Dalal, S. & Regan, L. (2000). Understanding the Sequence Determinants of Conformational Switching Using Protein Design. *Protein Science : A Publication of the Protein Society*, Vol. 9, No. 9, (September 2000), pp. 1651-1659, ISSN 0961-8368

Despres, J.P., Lemieux, I., Dagenais, G.R., Cantin, B. & Lamarche, B. (2000). HDL-Cholesterol as a Marker of Coronary Heart Disease Risk: The Quebec Cardiovascular Study. *Atherosclerosis*, Vol. 153, No. 2, (December 2000), pp. 263-272, ISSN 0021-9150

Dobson, C.M. (1999). Protein Misfolding, Evolution and Disease. *Trends in Biochemical Sciences*, Vol. 24, No. 9, (September 1999), pp. 329-332, ISSN 0968-0004

Dobson, C.M. (2003). Protein Folding and Misfolding. *Nature,* Vol. 426, No. 6968, (December 2003), pp. 884-890, ISSN 1476-4687

Dobson, C.M. (2004). Protein Chemistry. In the Footsteps of Alchemists. *Science,* Vol. 304, No. 5675, (May 2004), pp. 1259-1262, ISSN 1095-9203

Dubay, K.F., Pawar, A.P., Chiti, F., Zurdo, J., Dobson, C.M. & Vendruscolo, M. (2004). Prediction of the Absolute Aggregation Rates of Amyloidogenic Polypeptide Chains. *Journal of Molecular Biology*, Vol. 341, No. 5, (August 2004), pp. 1317-1326, ISSN 0022-2836

Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C. & Villafranca, J.E. (1998). Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions and Observations. *Pacific Symposium on Biocomputing*, Vol. 3, (August 1998), pp. 473-484, ISSN 1793-5091

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. & Obradovic, Z.J. (2001). Intrinsically Disordered Protein. *Journal of Molecular Graphics and Modelling*, Vol. 19, No. 1, (February 2001), pp. 26-59, ISSN: 1093-3263

Eaton, W.A., Muñoz, V.,Hagen, S.J., Jas, G.S., Lapidus, L.J. & Henry, E.R. (2000). Fast Kinetics and Mechanisms in Protein Folding. *Annual Review of Biophysics and Biomolecular Structure*, Vol. 29, (June 2000), pp. 327–359, ISSN: 1056-8700

Eggerman, T.L., Hoeg, J.M., Meng, M.S., Tombragel, A., Bojanovski, D. & Brewer, H.B., Jr. (1991). Differential Tissue-Specific Expression of Human ApoA-I and ApoA-II. *Journal of Lipid Research*, Vol. 32, No. 5, (May 1991), pp. 821-828, ISSN 0022-2275

Eisenberg, D., Nelson, R., Sawaya, M.R., Balbirnie, M., Sambashivan, S., Ivanova, M.I., Madsen, A.O. & Riekel, C. (2006). The Structural Biology of Protein Aggregation Diseases: Fundamental Questions and Some Answers. *Accounts of Chemical Research*, Vol. 39, No. 9, (September 2006), pp. 568-575, ISSN 0001-4842

Eisenberg, S. (1990). Metabolism of Apolipoproteins and Lipoproteins. *Current Opinion in Lipidology*, Vol. 1, No. 3, (June 1990), pp. 205-215, ISSN 0957-9672

Fang, Y., Gursky, O. & Atkinson, D. (2003). Lipid-Binding Studies of Human Apolipoprotein A-I and its Terminally Truncated Mutants. *Biochemistry*, Vol. 42, No. 45, (November 2003), pp. 13260-13268, ISSN 0006-2960

Fink, A.L. (1998). Protein Aggregation: Folding Aggregates, Inclusion Bodies and Amyloid. *Folding & Design*, Vol. 3, No. 1, (February 1998), pp. R9-23, ISSN 1359-0278

Frank, P.G. & Marcel, Y.L. (2000). Apolipoprotein A-I: Structure-Function Relationships. *Journal of Lipid Research,* Vol. 41, No. 6, (June 2000), pp. 853-872, ISSN 0022-2275

García-González, V. & Mas-Oliva, J. (2011). Amyloidogenic Properties of a D/N Mutated 12 Amino Acid Fragment of the C-Terminal Domain of the Cholesteryl-Ester Transfer Protein (CETP). *International Journal of Molecular Sciences*, Vol. 12, No. 3, (March 2011), pp. 2019-2035, ISSN 1422-0067

Garda, H.A. (2007). Structure–Function Relationships in Human Apolipoprotein A-I: Role of a Central Helix Pair. *Future Lipidology,* Vol. 2, No. 1, (February 2007), pp. 95-104, ISSN 1746-0875

Genschel, J., Haas, R., Pröpsting, M.J. & Schmidt, H.H.J. (1998). Hypothesis. Apolipoprotein A-I Induced Amyloidosis. *Federation of European Biochemical Societies Letters*, Vol. 430, No. 3, (July 1998), pp. 145-149, ISSN 0014-5793

Gsponer, J. & Vendruscolo, M. (2006). Theoretical Approaches to Protein Aggregation. *Protein and Peptide Letters*, Vol. 13, No. 3, (March 2006), pp. 287-293, ISSN 0929-8665

Hamada, D., Segawa, S. & Goto, Y. (1996). Non-Native Alpha-Helical Intermediate in the Refolding of Beta-Lactoglobulin, a Predominantly Beta-Sheet Protein. *Nature Structural Biology*, Vol. 3, No. 10, (October 1996), pp. 868-873, ISSN 1072-8368

Huang, K. (2005). *Lectures on Statistical Physics and Protein Folding,* World Scientific Publishing Company, ISBN 978-981-256-143-5, Singapore.

Huber, R. & Bode, W. (1978). Structural Basis of the Activation and Action of Trypsin. *Accounts of Chemical Research,* Vol. 11, No. 3, (March 1978), pp. 114-122, ISSN 0001-4842

Israelachvili, J.N. (1973). Thin Film Studies Using Multiple-Beam Interferometry. *Journal of Colloid and Interface Science*, Vol. 44, No. 2, (August 1973), pp. 259-272, ISSN 0021-9797

Israelachvili, J.N. & McGuiggan, P.M. (1990). Adhesion and Short Range Force between Surfaces. Part I: New Apparatus for Surface Force Measurements. *Journal of Materials Research*, Vol. 5, No. 10, (October 1990), pp. 2223-2231, ISSN 0884-2914

James, L.C. & Tawfik, D.S. (2003). Conformational Diversity and Protein Evolution--a 60-Year-Old Hypothesis Revisited. *Trends in Biochemical Sciences*, Vol. 28, No. 7, (July 2003), pp. 361-368, ISSN 0968-0004

Jerzy-Roch, N. & Assmann, G. (2005). Atheroprotective Effects of High-Density Lipoprotein-Associated Lysosphingolipids. *Trends in Cardiovascular Medicine,* Vol. 15, No. 7, (October 2005), pp. 265–271, ISSN 1050-1738

Kekicheff, P., Ducker, W.A., Ninham, B.W. & Pileni, M.P. (1990). Multilayer Adsorption of Cytochrome c on Mica around Isolectric pH. *Langmuir*, Vol. 6, No. 11 (November 1990), pp. 1704-1708, ISSN 0743-7463

Kono, M., Okumura, Y., Tanaka, M., Nguyen, D., Dhanasekaran, P., Lund-Katz, S., Phillips, M.C. & Saito, H. (2008). Conformational Flexibility of the N-Terminal Domain of Apolipoprotein A-I Bound to Spherical Lipid Particles. *Biochemistry*, Vol. 47, No. 43, (October 2008), pp. 11340-11347, ISSN 1520-4995

Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I. & Wright, P.E. (1996). Structural Studies of P21waf1/Cip1/Sdi1 in the Free and Cdk2-Bound State: Conformational Disorder Mediates Binding Diversity. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 93, No. 21, (October 1996), pp. 11504-11509, ISSN 0027-8424

Kumar, M.S., Carson, M., Hussain, M.M. & Murthy, H.M. (2002). Structures of Apolipoprotein A-II and a Lipid-Surrogate Complex Provide Insights into Apolipoprotein-Lipid Interactions. *Biochemistry*, Vol. 41, No. 39, (October 2002), pp. 11681-11691, ISSN 0006-2960

Kyte, J. & Doolittle, R.F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, Vol. 157, No. 1, (May 1982), pp. 105-132, ISSN 0022-2836

Lagerstedt, J.O., Budamagunta, M.S., Oda, M.N. & Voss, J.C. (2007). Electron Paramagnetic Resonance Spectroscopy of Site-Directed Spin Labels Reveals the Structural Heterogeneity in the N-Terminal Domain of ApoA-I in Solution. *The Journal of Biological Chemistry*, Vol. 282, No. 12, (March 2007), pp. 9143-9149, ISSN 0021-9258

Lashuel, H.A. & Lansbury, P.T., Jr. (2006). Are Amyloid Diseases Caused by Protein Aggregates That Mimic Bacterial Pore-Forming Toxins? *Quarterly Reviews of Biophysics*, Vol. 39, No. 2, (May 2006), pp. 167-201, ISSN 0033-5835

Liang, H.Q., Rye, K.A. & Barter, P.J. (1995). Cycling of Apolipoprotein A-I between Lipid-Associated and Lipid-Free Pools. *Biochimica et Biophysica Acta*, Vol. 1257, No. 1, (June 1995), pp. 31-37, ISSN 0006-3002

Mas-Oliva, J., Moreno, A., Ramos, S., Xicohtencatl-Cortes, J., Campos, J. & Castillo, R. (2003). Monolayers of Apolipoprotein AII at the Air/Water Interface, In: *Frontiers in Cardiovascular Health*, Dhalla, N.S., Chockalingam, A., Berkowitz, H.I. & Singal,

P.K., pp. (341-352), Kluwer Academic Publishers, ISBN 978-1-4020-7451-6, Boston, U.S.A.

Meador, W.E., Means, A.R. & Quiocho, F.A. (1992). Target Enzyme Recognition by Calmodulin: 2.4 a Structure of a Calmodulin-Peptide Complex. Vol. 257, No. 5074, (August 1992), pp. 1251-1255, ISSN 0036-8075

Mendoza-Espinosa, P., García-González, V., Moreno, A., Castillo, R. & Mas-Oliva, J. (2009). Disorder-to-Order Conformational Transitions in Protein Structure and Its Relationship to Disease. *Molecular and Cellular Biochemistry*, Vol. 330, No. 1-2, (October 2009), pp. 105-120, ISSN 1573-4919

Mendoza-Espinosa, P., Moreno, A., Castillo, R. & Mas-Oliva, J. (2008). Lipid Dependant Disorder-to-Order Conformational Transitions in Apolipoprotein CI Derived Peptides. *Biochemical and Biophysical Research Communications*, Vol. 365, No. 1, (January 2008), pp. 8-15, ISSN 1090-2104

Mishra, V.K., Palgunachari, M.N., Datta, G., Phillips, M.C., Lund-Katz, S., Adeyeye, S.O., Segrest, J.P. & Anantharamaiah, G.M. (1998). Studies of Synthetic Peptides of Human Apolipoprotein A-I Containing Tandem Amphipathic Alpha-Helixes. *Biochemistry*, Vol. 37, No. 28, (July 1998), pp. 10313-10324, ISSN 0006-2960

Nylander, T. & Wahlgren, M.N. (1997). Forces between Adsorbed Layers of Beta-Casein. *Langmuir: The ACS Journal of Surfaces and Colloids,* Vol. 13, No. 23, (November 1997), pp. 6219-6225, ISSN 0743-7463

Oda, M.N., Forte, T.M., Ryan, R.O. & Voss, J.C. (2003). The C-Terminal Domain of Apolipoprotein A-I Contains a Lipid-Sensitive Conformational Trigger. *Nature Structural Biology*, Vol. 10, No. 6, (June 2003), pp. 455-460, ISSN 1072-8368

Ohnishi, S. & Takano, K. (2004). Amyloid Fibrils from the Viewpoint of Protein Folding. *Cellular and Molecular Life Sciences : CMLS*, Vol. 61, No. 5, (March 2004), pp. 511-524, ISSN 1420-682X

Okon, M., Frank, P.G., Marcel, Y.L. & Cushley, R.J. (2001). Secondary Structure of Human Apolipoprotein A-I(1-186) in Lipid-Mimetic Solution. *Federation of European Biochemical Societies Letters*, Vol. 487, No. 3, (January 2001), pp. 390-396, ISSN 0014-5793

Okon, M., Frank, P.G., Marcel, Y.L. & Cushley, R.J. (2002). Heteronuclear NMR Studies of Human Serum Apolipoprotein A-I. Part I. Secondary Structure in Lipid-Mimetic Solution. *Federation of European Biochemical Societies Letters*, Vol. 517, No. 1-3, (April 2002), pp. 139-143, ISSN 0014-5793

Oram, J.F. (2002). ATP-Binding Cassette Transporter A1 and Cholesterol Trafficking. *Current Opinion in Lipidology*, Vol. 13, No. 4, (August 2002), pp. 373-381, ISSN 0957-9672

Parker, J.L., Christenson, H.K. & Ninham, B.W. (1989). Device for Measuring the Force and Separation between Two Surfaces Down to Molecular Separation. *Review of Scientific Instruments,* Vol. 60, No. 10, (October 1989), pp. 3135-3138, ISSN 0034-6748

Ramos, S., Campos-Teran, J., Mas-Oliva, J., Nylander, T. & Castillo, R. (2008). Forces between Hydrophilic Surfaces Adsorbed with Apolipoprotein AII Alpha Helices. *Langmuir : The ACS Journal of Surfaces and Colloids*, Vol. 24, No. 16, (August 2008), pp. 8568-8575, ISSN 0743-7463

Romero, P., Obradovic, Z. & Dunker, A.K. (2001). Intelligent Data Analysis for Protein Disorder Prediction. *Artificial Intelligence Review*, Vol. 14, No. 6, (December 2000), pp. 447-484, ISSN 0269-2821

Rose, G.D., Fleming, P.J., Banavar, J.R. & Maritan, A. (2006). A Backbone-Based Theory of Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 103, No. 45, (November 2006), pp. 16623-16633, ISSN 0027-8424

Rozek, A., Buchko, G.W., Kanda, P. & Cushley, R.J. (1997). Conformational Studies of the N-Terminal Lipid-Associating Domain of Human Apolipoprotein C-I by Cd and 1H NMR Spectroscopy. *Protein Science : A Publication of the Protein Society*, Vol. 6, No. 9, (September 1997), pp. 1858-1868, ISSN 0961-8368

Ruíz-García, J., Moreno, A., Brezesinski, G., Möhwald, H., Mas-Oliva, J. & Castillo, R. (2003). Phase Transitions and Conformational Changes in Monolayers of Human Apolipoprotein CI and AII. *Journal of Physical Chemistry B,* Vol. 107, No. 40, (September 2003), pp. 11117–11124, ISSN 1089-5647

Schmidt, H.H., Haas, R.E., Remaley, A., Genschel, J., Strassburg, C.P., Buttner, C. & Manns, M.P. (1997). In Vivo Kinetics as a Sensitive Method for Testing Physiologically Intact Human Recombinant Apolipoprotein A-I: Comparison of Three Different Expression Systems. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, Vol. 268, No. 1-2, (December 1997), pp. 41-60, ISSN 0009-8981

Segrest, J.P., Jones, M.K., De Loof, H., Brouillette, C.G., Venkatachalapathi, Y.V. & Anantharamaiah, G.M. (1992). The Amphipathic Helix in the Exchangeable Apolipoproteins: A Review of Secondary Structure and Function. *Journal of Lipid Research*, Vol. 33, No. 2, (February 1992), pp. 141-166, ISSN 0022-2275

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z. & Dunker, A.K. (2007). Disprot: The Database of Disordered Proteins. *Nucleic Acids Research*, Vol. 35, Suppl. 1, (January 2007), pp. D786-793, ISSN 1362-4962

Su, C.T., Chen, C.Y. & Hsu, C.M. (2007). Ipda: Integrated Protein Disorder Analyzer. *Nucleic Acids Research*, Vol. 35, Suppl. 2, (July 2007), pp. W465-472, ISSN 1362-4962

Swaney, J.B. & Weisgraber, K.H. (1994). Effect of Apolipoprotein C-I Peptides on the Apolipoprotein E Content and Receptor-Binding Properties of Beta-Migrating Very Low Density Lipoproteins. *Journal of Lipid Research*, Vol. 35, No. 1, (January 1994), pp. 134-142, ISSN 0022-2275

Tailleux, A., Duriez, P., Fruchart, J.C. & Clavey, V. (2002). Apolipoprotein A-II, HDL Metabolism and Atherosclerosis. *Atherosclerosis*, Vol. 164, No. 1, (September 2002), pp. 1-13, ISSN 0021-9150

Tamm, L.K. (2003). Hypothesis: Spring-Loaded Boomerang Mechanism of Influenza Hemagglutinin-Mediated Membrane Fusion. *Biochimica et Biophysica Acta*, Vol. 1614, No. 1, (July 2003), pp. 14-23, ISSN 0006-3002

Tanaka, M., Koyama, M., Dhanasekaran, P., Nguyen, D., Nickel, M., Lund-Katz, S., Saito, H. & Phillips, M.C. (2008). Influence of Tertiary Structure Domain Properties on the Functionality of Apolipoprotein A-I. *Biochemistry*, Vol. 47, No. 7, (February 2008), pp. 2172-2180, ISSN 0006-2960

Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., Chiti, F. & Vendruscolo, M. (2008). Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*, Vol. 380, No. 2, (July 2008), pp. 425-436, ISSN 1089-8638

Tompa, P. (2002). Intrinsically Unstructured Proteins. *Trends in Biochemical Sciences*, Vol. 27, No. 10, (October 2002), pp. 527-533, ISSN 0968-0004

Uversky, V.N. (2002). What Does It Mean to Be Natively Unfolded? *European Journal of Biochemistry / Federation of European Biochemical Societies*, Vol. 269, No. 1, (January 2002), pp. 2-12, ISSN 0014-2956

Uversky, V.N., Gillespie, J.R. & Fink, A.L. (2000). Why Are "Natively Unfolded" Proteins Unstructured under Physiologic Conditions? *Proteins*, Vol. 41, No. 3, (November 2000), pp. 415-427, ISSN 0887-3585

Wang, G. (2002). How the Lipid-Free Structure of the N-Terminal Truncated Human ApoA-I Converts to the Lipid-Bound Form: New Insights from NMR and X-Ray Structural Comparison. *Federation of European Biochemical Societies Letters*, Vol. 529, No. 2-3, (October 2002), pp. 157-161, ISSN 0014-5793

Wang, G.,Sparrow, J.T. & Cushley, R.J. (1997). The Helix-Hinge-Helix Structural Motif in Human Apolipoprotein A-I Determined by NMR Spectroscopy. *Biochemistry*, Vol. 36, No. 44, (November 1997), pp. 13657-13666, ISSN 0006-2960

Wang, G., Treleaven, W.D. & Cushley, R.J. (1996). Conformation of Human Serum Apolipoprotein A-I(166-185) in the Presence of Sodium Dodecyl Sulfate or Dodecylphosphocholine by 1H-NMR and CD. Evidence for Specific Peptide-SDS Interactions. *Biochimica et Biophysica Acta*, Vol. 1301, No. 3, (June 1996), pp. 174-184, ISSN 0006-3002

Wei, A., Rubin, H., Cooperman, B.S. & Christianson, D.W. (1994). Crystal Structure of an Uncleaved Serpin Reveals the Conformation of an Inhibitory Reactive Loop. *Nature Structural Biology*, Vol. 1, No. 4, (April 1994), pp. 251-258, ISSN 1072-8368

Weinberg, R.B. & Spector, M.S. (1985). Human Apolipoprotein A-IV: Displacement from the Surface of Triglyceride-Rich Particles by HDL2-Associated C-Apoproteins. *Journal of Lipid Research*, Vol. 26, No. 1, (January 1985), pp. 26-37, ISSN 0022-2275

Westermark, P., Mucchiano, G., Marthin, T., Johnson, K.H. & Sletten, K. (1995). Apolipoprotein A1-Derived Amyloid in Human Aortic Atherosclerotic Plaques. *The American Journal of Pathology*, Vol. 147, No. 5, (November 1995), pp. 1186-1192, ISSN 0002-9440

Wu, Z., Wagner, M.A., Zheng, L., Parks, J.S., Shy, J.M., 3rd, Smith, J.D., Gogonea, V. & Hazen, S.L. (2007). The Refined Structure of Nascent HDL Reveals a Key Functional Domain for Particle Maturation and Dysfunction. *Nature Structural & Molecular Biology*, Vol. 14, No. 9, (September 2007), pp. 861-868, ISSN 1545-9993

Xicohtencatl-Cortes, J.,Castillo, R. & Mas-Oliva, J. (2004b). In Search of New Structural States of Exchangeable Apolipoproteins. *Biochemical and Biophysical Research Communications*, Vol. 324, No. 2, (November 2004), pp. 467-470, ISSN 0006-291X

Xicohtencatl-Cortes, J.,Mas-Oliva, J. & Castillo, R. (2004a). Phase Transitions of Phospholipid Monolayers Penetrated by Apolipoproteins. *Journal of Physical Chemistry B,* Vol. 108, No. 22, (April 2004), pp. 7307-7315, ISSN 1089-5647

18

Protein-Protein Interactions in Salt Solutions

Jifeng Zhang
Department of Analytical and Formulation Sciences, Amgen Inc.,
Thousand Oaks, California
USA

1. Introduction

Protein-protein interactions drive many biophysical processes of proteins in solutions, such as aggregation, solubility, and phase transitions including crystallization, gelation, and amorphous precipitation. Many of these processes are of significant research interest because of their practical importance. In the biopharmaceutical industry, it is crucial to prevent therapeutic proteins from aggregation during the manufacturing process and storage in order to maintain safety and efficacy (1). In addition, protein crystallization and precipitation are used for industrialized recombinant protein purification process (2). In the field of structure biology, it is still a daunting task to produce diffractive quality protein crystals for determining protein 3-D structures because there is lack of clear understanding of the mechanisms for protein crystallization (3). Furthermore, studying protein-protein interactions could shed light on the mechanism of protein condensation (or phase transition) diseases, such as cataract and sickle cell disease (4). Finally, protein-protein interactions may play essential roles in many human neurodegenerative diseases attributed to protein aggregation, such as Parkinson and Alzheimer diseases (5).

In solutions, salts are ubiquitously used to control pH, ionic strength and osomlality in scientific research and industry applications. It is important to understand how salts modulate protein-protein interactions so that solution behavior, such as protein crystallization, precipitation, and solution stability, can be controlled and manipulated. However, the exact interaction mechanisms between salt ions and proteins are poorly understood (6, 7). As a consequence, modulations on protein-protein interactions by salt ions and their implications for protein solution behavior cannot be completely rationalized. The challenges rise because of (i) the sheer complexity of physical and chemical properties for both salt ions and proteins and (ii) the wide range of salt concentrations, which can be varied up to 1000 fold from millimolar to molar. It cannot be emphasized better than how Kunz and Neueder mentioned in their book with regards to salt solutions: "In total, it is still a fact that over the last decades, it was still easier to fly to the moon than to describe the free energy of even the simplest salt solutions beyond a concentration of 0.1 M or so"(6). Proteins probably belong to the most complex colloidal system in terms of variations in surface charge, surface chemistry, and size. Specifically, a protein could be net positive-charged, neutral, or negative-charged at pH conditions below, near, and above its pI (Isoelectric point), respectively. Additionally, protein surfaces are heterogeneously composed of

positive and negative charged, polar and nonpolar amino acid residues. Finally, the size of proteins in the range of 1-5 nm (estimated by the minimal radius of a sphere containing a given mass) would significantly impact the surface charge density(8).

Intermolecular interactions between protein molecules can have different origins, such as electrostatic, hydrophobic, van der waals, and hydrogen bonding (9). It is difficult to pinpoint the exact relative contributions from each type of interaction to the (overall) protein-protein interactions. In this review, I focus on explaining the modulations of electrostatic protein-protein interactions by the simple salt ions (shown in Figure 1) through their specific interactions (or binding) from both cation and anion with protein surface at salt concentrations below 0.5- 1 M. In addition, the complete picture of salt ion's effects on the intermolecular interactions may be better understood by considering the following biophysical properties of proteins and salt ions: (i) the net charge, surface charge density and hydrophobicity of a protein; (ii) hydration, size, polarizability and valency of salt ions. The discussion is based on the recent experimental results reported in literature and findings from Amgen using the following experimental techniques, such as protein solubility measurement, phase transition temperature of $T_{critical}$ (critical temperature) or T_{cloud} (cloud temperature) for liquid-liquid phase separation and small angle X-ray scattering (SAXS) (10-13). It has been demonstrated that there is a strong correlation between protein solubility and protein-protein interactions: protein solubility decreases when the protein-protein interactions become less repulsive or more attractive (for a protein for which its solubility increases with temperature)(12, 13). Also it is generally accepted that for a protein solution with an upper consolute point, an increase in phase transition temperature, as a result of change in the solution condition, indicates that protein-protein interactions become less repulsive or more attractive.

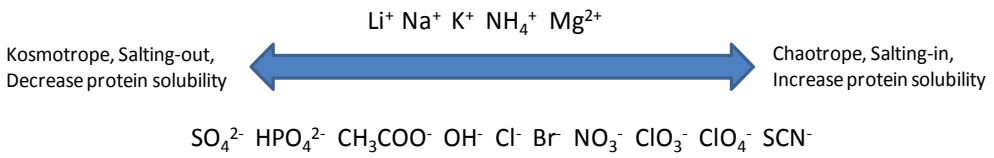


Fig. 1. Hofmeister series adapted from (14).

2. Historical background

2.1 Direct and reverse Hofmeister series

The most important experimental work on protein-protein interactions in salt solutions can be traced back more than 100 years ago when Franz Hofmeister and his coworkers studied salt effects at high salt concentrations on protein precipitation of hen egg white proteins whose main component is ovalbumin (pI=4.6). At that time, he hypothesized that the protein precipitating (salting-out) capability for the salts was dependent on their ion hydration properties (6). Later on, an empirical ranking for both cations and anions in their effectiveness, as shown in Figure 1, for precipitating proteins was named as (direct) Hofmeister series (14). Typically, the anions' effects are more dramatic than cation (14). In 1989, a surprising and complete reverse Hofmeister series was discovered by Ries-Kautt and

Ducruix in solubility measurement of lysozyme in salt solutions at pH below its pI where the protein was net positively charged (15).

2.2 Protein-protein interactions for a net charge neutral protein in salt solutions

A protein is net-charge neutral at its pI with the equal numbers of positive and negative charges. This is the most distinctive difference between proteins and the peptides with neutral side chains/small nonpolar molecules, for which extensive and detailed solubility experiments were conducted in salt solutions (16-19). However, there is lack of systematic protein solubility studies in salt solutions near their pIs. It is generally accepted that near the pI an increase in protein solubility (salting-in) is expected when salts are initially added and then a decrease occurs at high salt concentrations (salting-out by kosmotropic salts)(20). Although the mechanism of protein-protein interactions near its pI remains to be determined, it can be inferred from the observation above that the protein-protein interactions may initially become less attractive and then more attractive with increasing salt concentrations.

2.3 Protein-protein interactions for a net positive-charged protein in salt solutions

Lysozyme is a small globular protein with a Molecular Weight (MW) of 14.4 kilo-Dalton (kD) with a high pI value of ~11(12). Despite the fact that the experiments can mostly be conducted at pH conditions below its pI, lysozyme was frequently used as a model protein for studying both protein-protein interactions and protein-salt ion interactions in salt solutions probably due to its availability and easy crystallization propensity. Numerous experiments revealed very complex relationships between intermolecular interactions and salt concentration, salt type and pH; different theories were put into place to interpretate the trends (12, 21, 22).

In monovalent salt solutions under 1.0 M, the intermolecular interactions for lysozyme generally became monotonically more attractive as the salt concentration increased at pH conditions far below its pI(12, 21). These findings are consistent with the no salting-in event, i. e. protein solubility decrease, for lysozyme by NaCl in a pH range from 3 to 9 under the salt concentration up to 1.2 M (23). Acting as counter-ions to the net positively-charged lysozyme and following the reverse Hofmeister series, these monovalent anions imposed profound effects on the intermolecular interactions. But at pH 9.4 closer to pI, a nonmonotonic transition was discovered for SCN^- where the intermolecular interactions initially became more attractive and then less attractive when the phase transition temperature was measured (22). For γD-crystallins, a 20-kD protein, the same reverse Hofmeister series for anions was observed at pH 4.5 below its pI of ~7.0 by using SAXS (13).

Despite the dominant effect of the counter-ions (or anions), the co-ions (or cations) can still significantly perturb the protein-protein interactions. Specifically, comparing the effect by different cation in the salt solutions with the same anion, the intermolecular interactions for positive-charged lysozyme were less attractive and even perturbed nonmonotonically by the strongly hydrated divalent cation (Mg^{2+} and Ca^{2+}) , in comparison to the monotonic effect by the monovalent cations of Na^+ and K^+(12, 21). These findings are consistent with the findings from lysozyme solubility measurement in the multivalent cation salt solutions (12, 24).

2.4 Protein-protein interactions for a net negative-charged protein in salt solutions

Recently, many experiments were conducted to study protein-protein interactions for a net negatively-charged protein in salt solutions where a cation-dominant effect was expected. But the experimental findings were not straight-forward to interpret. Using SAXS and neutron scattering for studying protein-protein interactions of ovalbumin (MW=45 kD) in NaCl and YCl_3 solutions at pH conditions above its pI of 5.2, it was found that NaCl was ineffective in screening the electrostatic repulsive interactions between the proteins while YCl_3 not only suppressed the electrostatic repulsive interactions initially but also raised the repulsive interactions at higher concentrations (25). The ineffectiveness of Na^+ salts to screen the electrostatic repulsion was also confirmed for α-crystallins, a 800-kDa protein, at pH conditions above its pI of 4.5 by using SAXS (13). Similar behaviour was observed for BSA at pH conditions above its pI of 4.6 (26). Interestingly, Petsev et al found that NaAcetate was effective at screening the electrostatic repulsions (protein-protein interactions become more attractive) and then rendered the intermolecular interactions more repulsive for negatively-charged Apoferrtin (MW=450kD) (27).

2.5 Protein-protein interactions for an antibody at different pH conditions

Protein-protein interactions in salt solutions for an antibody with an experimentally determined pI of 7.2 were systematically explored through the measurements of protein solubility and phase transition temperature of $T_{critical}$ in liquid-liquid phase separation (11). The advantage of using this antibody is that the intermolecular interactions can be systemically assessed for the positive-charged and neutral for the same protein, allowing comprehensive experimental investigations of how salts modulate intermolecular interactions. Also, the antibody (MW=147 kD) is a much larger protein than lysozyme, which provides an opportunity for evaluating the surface charge density as a variable in protein-protein interactions(10). These approaches could help us understand how salt ions interact with proteins of different size.

At pH 7.1 close to its pI of 7.2, antibody solubility measurement revealed a general salting-in effect by all the anions as shown in Figure 2. More importantly, the specific anion

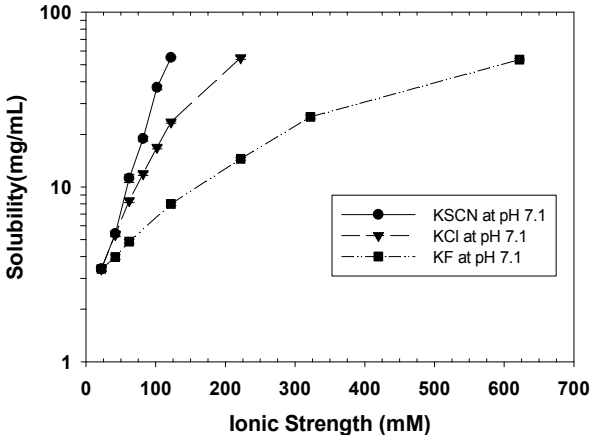


Fig. 2. Antibody solubility at pH 7.1 in KSCN, KCl and KF solutions [reprint with permission from ref (11)].

effect was observed in which SCN^- was the most effective at raising the antibody solubility, following the direct Hofmeister series. These observations are consistent with the ranking of these anions for disrupting the attractive intermolecular interactions as revealed by the results of $T_{critical}$ measurement (10).

At pH 5.3 below its pI, nonmonotonic behavior where protein solubility decreased and then increased with salt concentrations (in Figure 3) was observed for all the salts studied, suggesting that intermolecular interactions became less repulsive and then more. In addition, the effectiveness of the anions for reducing the protein solubility followed the reverse Hofmeister series, in which SCN^- was the most effective at reducing the antibody solubility. Then strikingly, the effectiveness for the anion to increase the protein solubility reverted back to the direct Hofmeister series as the salt concentration further increased. The above nonmonotonic transitions are in agreement with the protein-protein interactions pattern revealed by the measurement of $T_{critical}$ for liquid-liquid phase separation in the same salt solutions (10).

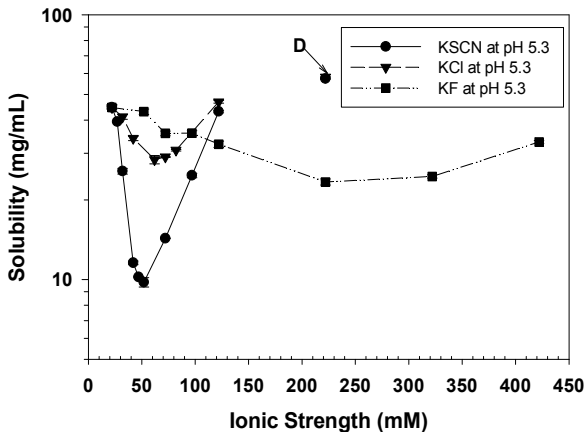


Fig. 3. Antibody solubility at pH 5.3 in in KSCN, KCl and KF solutions [reprint with permission from ref (11)].

It should be interesting to further study how salts affect the antibody solubility at pH values above its pI. Currently, experiments are on-going to do that.

3. Some theoretical explanations for protein-protein interactions in salt solutions

Recently Curtis and Lue wrote a comprehensive review of different theoretical treatments for understanding protein-protein interactions in salt solutions, pointing out that there is no single unified theoretical framework to rationalize the specificity of salt ion effects on protein intermolecular interactions (14). One of the important theories is the DLVO theory, in which proteins are treated as colloidal particles because their sizes are in the nanometer

range (9). The DLVO theory was named after the scientists: Derjaguin and Landau, and Verwey and Overbeek (9). This theory lays the foundation for explaining the interparticle electrostatic interactions in low salt concentrations below 0.1 M in the most simplified way when the protein is net-charged. Specifically, the intermolecular interactions between two protein molecules in low salt concentrations can be described by the following equation (28):

$$w_2(r) = w_{ex}(r) + w_{\text{disp}}(r) + w_{elec}(r) \quad (1)$$

Where r is the center-to-center distance from two molecules; w_{ex}(r) is the repulsive protein hard-sphere (excluded-volume) potential; w_{disp} (r) is the attractive dispersion potential; w_{elec} (r) is the electric double-layer repulsion potential, which can be further described by Debye-Huckel theory as the following:

$$w_{\text{elec}}(r) = \frac{(ze)^2 \exp[-\kappa (r-d)]}{4\pi r \varepsilon_0 \varepsilon_r (1+\frac{\kappa d}{2})^2} \quad \text{for } r > \sigma \quad (2)$$

Where ze is the net charge of a protein, e is the elementary charge, ε_0 is the dielectric permittivity of vacuum, ε_r is the dielectric constant of water, and κ is the inverse Debye length calculated by

$$\kappa^2 = \frac{2e^2 N_A I}{kT\varepsilon_0 \varepsilon_r} \quad (3)$$

Where I is the ionic strength of the solution, k is the Boltzmann's constant, T is the absolute temperature, and N_A is the Avogadro's number.

As presented in Equation 2, it is obvious that the more net charges a protein carries, the stronger the electrostatic double-layer repulsive force becomes. Also, Equation 2 indicates the addition of the salts monotonically decreases (or screens) the double-layer repulsion, and then reaches a plateau (the exponential term approach zero). The general screening effect is consistent with the initial drop in protein solubility and rise in liquid-liquid phase transition temperature as described above for the charged proteins. The DLVO theory was used to explain the protein solubility decrease of lysozyme (23). It should be pointed out that it is difficult to differentiate between the direct binding of salt ions to their opposite-charged partners on the protein surface and the screening by the salt-ion layer near the protein surface. The reason is that the first type of interaction decreases the double layer repulsion through balancing out the "ze" term in Equation 2 while the second type of interaction work through κ, the inverse Debye length. One of the major limitations of the DLVO theory is lack of ion-specificity as presented in Equation 2 and both cation and anion contribute equally as far as they have the same valency. Therefore, the DLVO theory cannot explain the anion-specific modulations on protein-protein interactions, i.e. the direct or reverse Hofmeister series at pH 5.3 for the antibody (3). In addition, the DLVO theory suggests that the double-layer repulsion decreases and levels off with salt addition, in contrary to the numerous nonmonotonic behavior mentioned above in Historical Background.

For a charge-neutral species (i.e. proteins at their pI), many other theoretical considerations were developed to explain the initial salting-in and later salting-out behavior (19, 29, 30). They

can be used to explain the general pattern of protein-protein interactions. In essence, the electrostatic interactions and hydrophobic interactions are the two major types of intermolecular forces (20, 31). The effects from the electrostatic interactions on the free energy of a protein in a low salt concentration solution may be described by Debye-Huckel theory in combination with Kirkwood's expression of the protein dipole moment as follows (20, 31):

$$\Delta G_{e.s.} = A - \frac{B(I^{1/2})}{1 + C(I^{1/2})} - DdI \tag{4}$$

Where A, B, C and D are constants, I is the ionic strength of the solution, d is the dipole moment for the protein. This theory predicts the salting-in effect: as the ionic strength increases, protein solubility rises. This idea is consistent with the observations of salting-in of proteins near pI. The main limitation of this theory is that it does not consider ion-specificity.

The free energy change for a protein involving the hydrophobic interactions may be illustrated by the cavity theory as follows(20):

$$\Delta G_{cav} = \left[N * Area + 4.8N^{1/3}(\kappa^e - 1)V^{2/3}\right]\left(\frac{\partial \sigma}{\partial m_3}\right)m_3 \tag{5}$$

where N is Avogadro's number, $Area$ is the surface area of a protein molecule, κ^e corrects the macroscopic surface tension of the solvent to molecular dimensions, V is the protein's molar volume, $\left(\frac{\partial \sigma}{\partial m3}\right)$ is the molal surface tension increment of the salt, and m_3 is the molality of the salt. This cavity theory describes how much free energy is needed to form a cavity in the solution to accommodate a hydrophobic protein molecule. Therefore, the surface tension of the solution is an important parameter and its modulation by salts impacts protein solubility and therefore protein-protein interactions. It predicts that the addition of kosmotropic salts, which increase the solution surface tension, will result in the salting-out effect and effectively strengthening of attractive protein-protein interactions. Therefore, these salting-in and salting-out effects in combination modulate protein solubility and protein-protein interactions in salt solutions (20, 31). Specifically, near the pI the salting-in effect dominates initially (protein solubility increases) and the addition of salts disrupts attractive protein-protein interactions. Then, further increase in (kosmotropic) salt concentration results in strengthening attractive protein-protein interactions as the salting-out effect begins to dominate (protein solubility decreases).

4. Molecular mechanism for protein-ion interactions

The simple ions shown in Figure 1 have different sizes, diverse hydration properties and polarizabilities (32). The interaction strength between an ion and water molecule in comparison to that between water-water determine the ion hydration property: an ion is strongly hydrated when it interacts with water molecules more strongly than the water-water interaction while the opposite makes an ion less hydrated (33-36). Shown in Figure 4 is the ranking of hydration property for the selected salt ions. Specifically, the large and more polarizable anion, i.e. SCN^-, is less hydrated while the small and less polarizable anion, i.e. F^-, is strongly hydrated.

The law of matching water affinities is the hallmark theory for defining the interaction strength between salt ions and proteins thermodynamically, in which the hydration and size properties of the ions and their counterparts on the protein surface are the key for explaining the protein-protein interaction behavior (33-36). Specifically according to the law of matching water affinities, oppositely charged ions in solutions form inner sphere ion pairs spontaneously when they have similar water affinities (36).

The chemistry of protein surface is heterogeneous, composed of both positive and negative-charged residues, and polar and nonpolar groups. As shown in Figure 4, monovalent anions of SCN^- and halides, except F^-, were weakly hydrated because of their large size, in comparison to the small-size monovalent cations being reasonably hydrated. On the protein surface, the positive-charged side chains on Arg, Lys and His are all derivatives of ammonium and therefore they are all weakly hydrated, matching well with the weakly hydrated SCN^-. According to the law of matching water affinity, the weakly hydrated anions, such as SCN^-, have the strongest interactions with the positive-charged side chains from the protein and neutralize them, followed by CI^- and F^-. On the other hand, the negative-charged side chains from Asp and Glu are strongly hydrated carboxylate, mismatching with Na^+ and K^+ whose interaction strengths are similar to that between water molecules (33-36). To the contrary, the divalent cation, i.e. Mg^{2+}, interacts with water molecules more strongly than Na^+ and K^+ and is strongly hydrated. It is then expected that the divalent cation interacts with the carboxylate more strongly than both Na^+ and K^+.

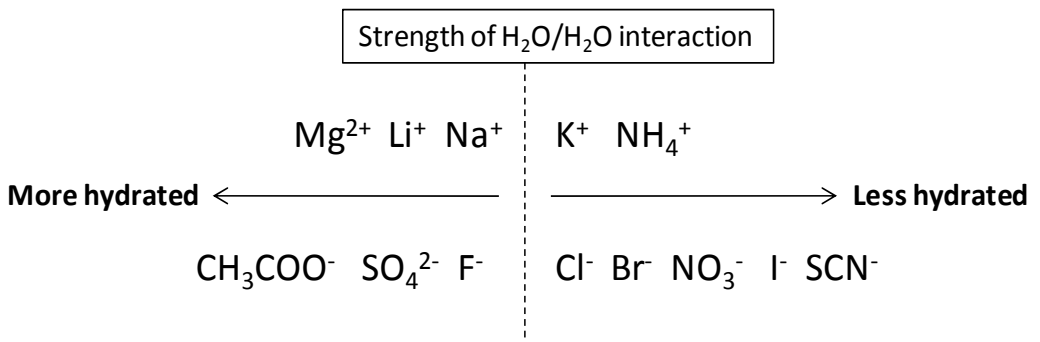


Fig. 4. Hydration properties of selected salt ions (34, 36).

Protein surface is composed of not only polar functional groups from the amide bonds of the exposed peptide backbone and the side chains of Asn and Gln, but also non-polar functional groups from the side chains of Phe, Ile and other amino acids. Both the polar and non-polar groups can be considered as weakly hydrated (37). Collins proposed that the weakly hydrated anions could also interact with both of the groups, besides the charged side chains (33-36). Recently, it was demonstrated, through a molecular dynamics (MD) study of lysozyme in a mixed aqueous solution of potassium chloride and iodide (0.4 M), that weakly hydrated anions, i.e. I^-, preferred to interact with the nonpolar groups besides the positive-charged residues on lysozyme (38). Furthermore, the interaction between

weakly hydrated anions and the amide bonds was also proposed based on the solubility study on poly(N-isopropylacrylamide) in salt solutions (39). For cations, it has been shown that both Ca^{2+} and Mg^{2+} can interact strongly with proteins through the diopolar amide bond (40) (18, 41).

The electroselectivity theory deserves attention when considering salt ion-protein interactions. Developed based on the anions' affinity for the anion exchanger, the electroselectivity theory proposed, purely based on the electrostatic interaction, that the ions with higher valency, such as SO_4^{2-} , interact with the positive-charge residues on the protein surface more strongly than those with a single valence, such as SCN^-(42, 43). The strong electrostatic interactions imparted by SO_4^{2-} were recently demonstrated by exploring specific ion effects on interfacial water structure adjacent to a bovine serum albumin at pH conditions below its pI using vibrational sum frequency spectroscopy (VSFS) (44).

5. From protein-ion interactions to protein-protein interactions

The complexity of protein-protein interactions as modulated by salt ions at low concentrations might be explained from the framework of dominance of specific electrostatic interactions from both cation and anions for the protein surface, concomitantly considering the following biophysical properties including net charge, surface charge density and hydration of a protein, and hydration, size, polarizability and valency of salt ions.

The first key property is the macroscopic net charge (considering the protein as a particle) as modulated by pH. First, a protein is net charge neutral, positively-charged and negatively-charged at pH near, below, or above its pI, respectively. Furthermore, patches of protein surface could be macroscopically weakly-hydrated because of the abundantly exposed nonpolar and polar groups, regardless of whether a protein surface is overall hydrophobic or hydrophilic. It was pointed out that in general 1/3 of the protein surface is hydrophobic, resulting in a partially weakly-hydrated surface(45). Although the net charge of the protein is dictated by the solution pH, its nonpolar or polar surface might maintain its property of weak hydration when the native folding structure is not drastically affected by pH and low salt concentrations. As pH decreases below its pI, the increasingly net positive-charges, from the weakly hydrated side chains of Arg, His and Lys, might render the protein surface even more weakly hydrated. At pH above its pI, the strongly hydrated carboxylates, from the strongly hydrated side chains of Asp and Glu, bring more water onto the protein surface, which results in the surface becoming more hydrated.

5.1 pH near pI

A protein is net charge neutral at pI with the equal number of positive and negative-charged residues. Therefore the protein molecules may approach each other and fully explore complementary interaction configurations (46). It is well-known that a protein has the lowest solubility near its pI and easily precipitates, suggesting the presence of strong intermolecular attractive interactions. The interactions can be highly anisotropic due to ionic-pair interactions, cation-π interaction, hydrophobic interaction and others types of interactions. It is difficult to dissect which type of interaction contributes most to the intermolecular interactions, which might be sequence dependent and protein-specific.

Our previous experiment of antibody liquid-liquid phase separation near its pI suggests that the intermolecular interactions were attractive and sensitive to salts, indicating that there were electrostatic interactions between the antibodies. Our observations of the general salting-in trends in the solubility measurement and disruption of intermolecular electrostatic attractive interactions in the LLPS are in agreement of the solubility data at low salt concentrations for other proteins near their respective pI, i.e. carboxyhemoglobin (47). The idea of attractive electrostatic interactions is especially supported by the salting-in behavior near its pI by KF. Typically, KF only salts out neutral peptides without charged side chains and nonpolar small molecules (16, 17). The general salting-in trend is also consistent with the electrostatic interaction theory as described by Equation 4. However, this theory cannot explain the ranking of the anion's effectiveness for raising the antibody solubility.

In the monovalent K^+ salt solutions, K^+ does not match well with the strongly hydrated carboxylate as discussed above. In contrast, the water affinity of the weakly hydrated positive-charge side chains, polar and nonpolar groups match well with those weakly hydrated anions from SCN^- to Cl^-. It is then expected that K^+ interacts with protein surface fairly weakly and anion could specifically binds to the protein surface in which their specificities are determined by their binding constants for the protein. This idea is consistent with the specific anion's effect, as described by a direct Hofmeister series, of raising the antibody solubility and disruption of the intermolecular attractive interactions at pH 7.1. In addition, this idea is in agreement with the recent findings where a chaotropic monovalent anion bound more strongly to a net-charge neutral macromolecule, like BSA near its pI and polar Poly-(N-isopropylacrylamide), than a kosmotropic monovalent anion(44) (48).

On the other hand, strongly hydrated multivalent cation, such as Mg^{2+} and Ca^{2+}, could bind to the strongly-hydrated carboxylate. In addition, there are strong interactions between the amide bond and multivalent cation (17). The above two modes of binding could make multivalent cations strong salting-in reagents (just like the anions) at low salt concentrations, overshadowing the possible salt-outing of the nonpolar residues on a protein by the multivalent cations.

In short, the electrostatic attractive interactions may dominate at protein-protein interactions in low salt solutions at pH near its pI, where the binding strengths between the protein surface for both cation and anions, working in synergy, determines the salting-in effectiveness of the salts as they are initially added.

5.2 pH below pI

When a protein is net charged at pH above and below its pI, the aforementioned observations of protein-protein interactions initially becoming more attractive or drop in protein solubility suggest that (i) the electrostatic repulsion dominates the protein-protein interactions and (ii) the initial addition of the salts to a charged protein effectively neutralizes the net charge of the protein and reduces the electrostatic repulsion.

Below pI, the positive-charges on proteins are from the weakly hydrated side chains of Arg, Lys or His. In addition, polar and nonpolar sites on the protein surface are also

weakly hydrated. As results, the more weakly hydrated a monovalent anion is, the more strongly it interacts with the positive-charged protein, and the more effectively it neutralizes the protein's net charge. The monovalent anions then follow the reverse Hofmeister series for their effectiveness of weakening the electrostatic repulsive intermolecular interactions and decreasing the protein solubility. This idea is consistent with the solubility measurement and phase transition data for both lysozyme and the antibody. The ranking for the binding strength between the anions and this antibody is also in agreement with what has been observed in monovalent salt solutions for other positive-charged proteins including other antibodies, BSA and lysozyme(49) (44) (22, 50). The binding of SO_4^{2-} to the positive-charged lysozyme and BSA, consistent with the electroselectivity theory, provides convincing experimental evidence that there is strong electrostatic interaction between a positive-charged protein and divalent anions, despite the mismatching water affinity.

The competitive interactions of co-ions against the counter-ions for a positive-charged protein become apparent for the strongly hydrated multivalent cation, i.e. Mg^{2+}. For example, Mg^{2+} may interact strongly at the strongly hydrated carboxylate or peptide groups in comparisons to Na^+ and K^+, effectively raising the positive-charges of the protein and hindering the anion's charge neutralization effect. Then, it appears that $MgCl_2$ will be less effective at weakening the electrostatic repulsive interactions and decreasing the protein solubility than NaCl (with the same molar concentration of Cl^-). Therefore, the protein-protein interactions are expected to be more repulsive in the $MgCl_2$ solutions than in the NaCl solutions, following the direct Hofmeister series. This notion is in agreement with the measurement of the phase transition temperature for lysozyme(21). Similarly, solubility of lysozyme in multivalent cation salt solutions was higher than that in the monovalent cation salt solutions with the same anion(24).

When anions complete their charge neutralization process as suggested by the minimum of protein solubility in Figure 3, the protein can be considered as pseudo charge-neutral. The salt's effect on protein-protein interactions then is expected to follow the direct Hofmeister series, as described above for a protein near its pI. This is the reason for why we observed the nonmonotonic behavior in the aforementioned proteins at pH below their pI.

5.3 pH above pI

On the other hand, at pH above its pI, the protein is negatively charged. Although the net negative charges are from the strongly hydrated carboxylate side chains on Asp and Glu, its surface still has significant presence of polar and nonpolar residues, attracting weakly hydrated anions. It is anticipated that the competitive bindings of cation and anion for protein surface determine the final effect on protein-protein interactions and solubility. The counterions with strong electrostatic interactions with the proteins, i.e. multivalent cations, can neutralize the net charge, weaken the repulsive electrostatic intermolecular interactions and decrease the protein solubility more effectively than the monovalent cations of Na^+, following the reverse Hofmeister series. Furthermore, in the Na^+ salt solutions, the anion's binding to the weakly hydrated sites, possibly stronger than that between Na^+ and the

carboxylate, may effectively increase the repulsive interactions. This is consistent with the experimental observation of the experimental findings for protein-protein interactions of ovalbumin in NaCl and YCl_3 solutions at pH conditions above its pI. Specifically, in the NaCl solution Cl^-'s binding to ovalbumin preempted that of Na^+, effectively raising the intermolecular repulsive interactions. On the other hand, the trivalent Y^{3+} could bind to the carboxylate strongly, neutralize the net negative-charges and weaken the repulsive intermolecular interactions. After charge neutralization, the salting-in effect by YCl_3 followed.

However, when either strongly hydrated F^- or acetate was used, they mismatched for both the positive-charged side chains and weakly hydrated polar and nonpolar residues on the net negative-charged protein surface. Possibly, Na^+ now might interact with the protein stronger than F^- or acetate and neutralize the negative charges. This could be a reasonable explanation for the nonmonotonic behavior mentioned for Apoferrtin in NaAcetate solution, but not in the NaCl solution.

5.4 Surface charge density

The surface charge density of a protein could dramatically change the above nonmonotonic behavior. At pH close to the pI or a large-size protein with small number of either positive or negative net charges, where the surface charge density is low, only the monotonic salting-in behavior could be observed because the charge neutralization process is less dramatic. On the other hand, when a protein has high surface charge density due to either a small size or a large number of positive charges, the anions might not completely neutralize the positive charges even at molar concentration and therefore only a decrease in protein solubility can occur. As a matter of fact, this might be for the case of lysozyme solubility at pH 4 and 7, especially when a weak chaotropic anion, i.e. Cl^-, was used(22). The reason is that Cl^- could bind to the protein surface less strongly and effectively at weakening the electrostatic repulsive interactions than a strong chaotropic anion, such as SCN^-. But at pH 9.4 where the surface charge density was smaller than at pH 4 and 7, the weakly hydrated SCN^- could neutralize the net charges completely, and as a result the nonmonotonic behavior appeared.

As proteins transition from a high surface charge density system to low, the interaction between a co-ion and charged surface could be explained through the smeared surface charge model and discrete surface charge model, respectively. In a low surface charge density system (discrete charge surface), such as a large-size antibody, the co-ion binding probably becomes more significant, in comparison to a small globular protein, i.e. lysozyme, of a high surface-charged density system. The reason is that the co-ion can approach the surface without experiencing the repulsive electrostatic force. This idea of co-ion adsorption to a low or medium negative-charged hydrophobic surface is supported by the recent molecular simulation for a self-assembled monolayer (51). The simulation results shows that even at a high surface charge density of -2.0×10^{-2} C/m^2, there was significant co-ion adsorption. Therefore, significant presence of co-ion adsorption is expected for a typical protein surface with a surface charge density in the low range of mC/m^2 (10, 52),.

5.5 Additional attractive interaction by polarizable anions

Another important feature of protein-protein interactions in salt solutions is the presence of possible additional protein-protein attractive force caused by the weakly hydrated anions for a positive-charged protein, although the exact mechanism remains to be defined. A recent Monte Carlo simulation reveals that the presence of chaotropic (or polarizable) ions, like SCN^-, introduced this additional interaction of dispersion force in nature between protein molecules (53). More importantly, liquid-liquid phase separation of the antibody at different pHs in a KSCN solution at a pH below its pI indicates that this attractive protein-protein interaction became stronger as the pH dropped and the protein carried more positive charges.

6. Conclusions

Despite the complexity of salt ion and protein interactions and their effects on protein-protein interactions, the rich salt-specific effect at low salt concentrations may be qualitatively explained based on the specific binding of both anions and cations for protein surface with heterogeneous surface chemistry as illustrated in Figure 5. In the future, it would be beneficial to have a quantitative description for the salt ions' effect on protein-protein interactions.

As shown in Figure 5, protein surface may always have hydrophobic patches, which are weakly hydrated and matches well with the weakly hydrated anions. Additionally, the exposed dipolar amide bond of the peptide backbone is the potential site for the divalent cation and weakly hydrated anions. Furthermore, pH change not only modulates the net charge property of the protein but also modifies the degree of surface hydration. Specifically, as the pH decreases away from their pIs, proteins become net positively-charged and even more weakly hydrated because the positive-charges are from the weakly hydrated side chains of Arg, Lys, and His. At pH values close to their pIs, proteins are net-charge neutral. Then as pH increases away from their pI, proteins become becomes net negatively-charged and less weakly hydrated because the negative charges are from strongly hydrated carboxylate from Asp and Glu.

At a pH close to the pI of a protein, both cations and anions can access the neutral protein and may work in synergy to disrupt the attractive intermolecular protein interactions and result an increase of protein solubility. On the other hand, they work competitive for a sufficiently charged protein (in Figure 5). Specifically, the counter-ion from the salt tends to neutralize the net charge of the protein, weakening the electrostatic repulsive intermolecular interactions while the co-ion is likely to hinder the charge-neutralization effect by the counter-ion, effectively strengthening the repulsive intermolecular interactions. The interaction strength between the ions and protein surface is dependent on both electrostatic and hydration properties for both ions and protein. The final outcome of protein-protein interactions is then determined by a combination of the protein surface charge density and the relative binding strength of both ions for the protein surface. When the counter-ions interact with the charge protein more strongly than the co-ions, the charge neutralization step dominates, resulting in protein-protein interactions becoming less repulsive, after which there could be the salting-in effect as if the protein-counter-ion complex is pseudo

charge-neutral. In the opposite situation, the strong interaction from the co-ions effectively renders the protein-protein interactions more repulsive.

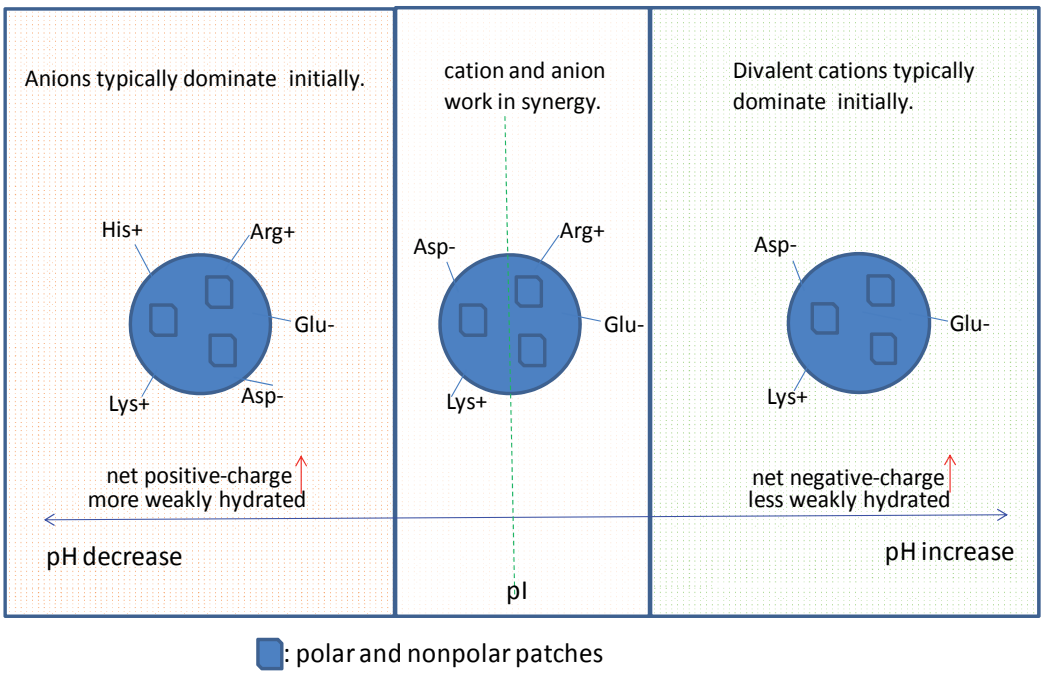


Fig. 5. Schematic illustration of the changes in net charge and hydration properties of a protein as pH varies.

7. Acknowledgements

The author would like to thank Dr. Izydor Apostol for reviewing the manuscript and providing valuable suggestions.

8. References

[1] Wei, W. 2005. Protein aggregation and its inhibition in biopharmaceutics. International Journal of Pharmaceutics 289:1-30.

[2] Schmidt, S., D. Havekost, K. Kaiser, J. Kauling, and H. J. Henzler. 2005. Crystallization for the Downstream Processing of Proteins. Engineering in Life Sciences 5:273-276.

[3] Chayen, N. E., editor. 2007. Protein Crystallization Strategies for Structural Genomics International University Line

[4] Gunton, J. D., A. Shiryayev, and D. L. Pagan. 2007. Protein condensation: kinetic pathways to crystallization and disease. Cambrige University Press, Cambridge.

[5] Uversky, V., and A. L. Fink. 2006. Protein Misfolding, Aggregation, and Conformational Disease Part A: Protein Aggregation and Conformational Diseases. Springer Science+ Business Media, Inc., Singapore.

[6] Kunz, W., editor. 2010. Specific Ion Effects. World Sceintific Publishing Co., Singapore.

[7] Zangi, R. 2010. Can salting-in/salting-out ions be classified as chaotropes/kosmotropes? J. Phys. Chem. B 114:643-650.

[8] Erickson, H. P. 2009. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. Biol Proced Online 11:32-51.

[9] Israelachvili, J. 1991. Intermolecular & Surface Forces. Academic Press, London.

[10] Mason, B. D., J. Zhang-van Enk, L. Zhang, R. L. Remmele, and J. Zhang. 2010. Liquid-liquid phase separation of a monoclonal antibody and influence of Hofmiester anions. Biophysical Journal 99:3792-3800.

[11] Zhang, L., Tan, H., Fesinmeyer, R. Matthew, Li, C., Catrone, D., Le, D., Remmele, R.L., and Zhang, J. 2011. Antibody Solubility Behavior in Monovalent Salt Solutions Reveals Specific Anion Effects at Low Ionic Strength J. Pharm. Sci. accepted.

[12] Broide, M. L., T. M. Tomine, and M. D. Saxowsky. 1996. Using phase transitions to investigate the effect of salts on protein interactions. Physical Review E 53:6325-6335.

[13] Finet, S., D. Vivarès, F. Bonneté, A. Tardieu, Charles W. Carter, Jr., and M. S. Robert. 2003. Controlling Biomolecular Crystallization by Understanding the Distinct Effects of PEGs and Salts on Solubility. In Methods in Enzymology. Academic Press. 105-129.

[14] Curtis, R. A., and L. Lue. 2006. A molecular approach to bioseparations: protein-protein and protein-salt intractions. Chemical Engineering Science 61:907-923.

[15] Ries-Kautt, M. M., and A. F. Ducruix. 1989. Relative effectiveness of various ions on the solubility and crystal growth of lysozyme. J. Biol. Chem. 264:745-748.

[16] Robinson, D. R., and W. P. Jencks. 1965. The effect of concentrated salt solutions on the activity coefficient of acetyltetraglycine ethyl ester. J. Am. Chem. Soc. 87:2470-2479.

[17] Nandi, P. K., and D. R. Robinson. 1972. Effects of salts on the free energies of nonpolar groups in model peptides. Journal of the American Chemical Society 94:1308-1315.

[18] Nandi, P. K., and D. R. Robinson. 1972. The effects of salts on the free energy of the peptide group. J. Am. Chem. Soc. 94:1299-1308.

[19] Baldwin, R. L. 1996. How Hofmeister ion interactions affect protein stability. Biophysical Journal 71:2056-2063.

[20] Arakawa, T., Timasheff, S. N. 1985. Theory of protein solubility. Meth. Enzymol. 114:49-77.

[21] Grigsby, J. J., H. W. Blanch, and J. M. Prausnitz. 2001. Cloud-point temperatures for lysozyme in electrolyte solutions: effect of salt type, salt concentration and pH. Biophysical Chemistry 9:231-243.

[22] Zhang, Y., and P. S. Cremer. 2009. The inverse and direct Hofmiester series for lysozyme. Proc. Natl. Acad. Sci USA 106:15249-15253.

[23] Retailleau, P., M. Ries-Kautt, and A. Ducruix. 1997. No salting-in of lysozyme chloride observed at low ionic strength over a large range of pH. Biophysical Journal 73:2156-2163.

[24] Benas, P., L. Legrand, and M. Riess-Kautt. 2002. Strong and specific effects of cations on lysozyme chloride solubility. Acta Cryst D 58:1582-1587.

[25] Ianeselli, L., F. Zhang, M. W. A. Skoda, R. M. J. Jacobs, R. A. Martin, S. Callow, S. PreÌ vost, and F. Schreiber. 2010. Proteinâ'Protein Interactions in Ovalbumin Solutions Studied by Small-Angle Scattering: Effect of Ionic Strength and the Chemical Nature of Cations. The Journal of Physical Chemistry B 114:3776-3783.

[26] Zhang, F., M. W. A. Skoda, R. M. J. Jacobs, S. Zorn, R. A. Martin, C. M. Martin, G. F. Clark, S. Weggler, A. Hildebrandt, O. Kohlbacher, and F. Schreiber. 2008. Reentrant Condensation of Proteins in Solution Induced by Multivalent Counterions. Physical Review Letters 101:148101.

[27] Petsev, D. N., B. R. Thomas, S. T. Yau, and P. G. Vekilov. 2000. Interactions and Aggregation of Apoferritin Molecules in Solution: Effects of Added Electrolytes. Biophysical Journal 78:2060-2069.

[28] Liu, W., D. Bratko, J. M. Prausnitz, and H. W. Blanch. 2004. Effect of Alcohols on Aqueous Lysozyme-Lysozyme Interactions from Static Light-Scattering Measurements. Biophysical Chemistry 107:289-298.

[29] Timasheff, S. N. 1985. Theory of protein solubility. Meth. Enzymol. 114:49-77.

[30] Melander, W., and C. Horvath. 1977. Salt effect on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic serie. Arch Biochem Biophys 183:200-215.

[31] Tanford, C., 1966. Physical Chemistry of Macromolecules. John Wiley & Sons, Inc., New York.

[32] Collins, K. D., and M. W. Washabaugh. 1985. The Hofmeister effect and the behaviour of water at interfaces. Quarterly Reviews of Biophysics 18:323-422.

[33] Collins, K. D., G. W. Neilson, and J. E. Enderby. 2007. Ions in water: characterizing the forces that control chemical processes and biological structure. Biophysical Chemistry 128:95-104.

[34] Collins, K. D. 2006. Ion hydration: Implications for cellular function, polyelectrolytes, and protein crystallization. Biophysical Chemistry 119:271-281.

[35] Collins, K. D. 2004. Ions from the Hofmiester series and osmolytes: effect on proteins in solution and in the crystallization process. Methods 34:300-311.

[36] Collins, K. D. 1997. Charge density-dependent strength of hydration and biological structure. Biophysical Journal 72:65-75.

[37] Collins, K. D. 1995. Sticky ions in biological systems. Proceedings of the National Academy of Sciences 92:5553-5557.

[38] Lund, M., L. Vrbka, and P. Jungwirth. 2008. Specific Ion Binding to Nonpolar Surface Patches of Proteins. Journal of the American Chemical Society 130:11582-11583.

[39] Zhang, Y., S. Furyk, D. E. Bergbreiter, and P. S. Cremer. 2005. Specific Ion Effects on the Water Solubility of Macromolecules: PNIPAM and the Hofmeister Series. Journal of the American Chemical Society 127:14505-14510.

[40] Baldwin, R. L. 1996. How Hofmeister ion interactions affect protein stability. Biophysical Journal 71:2056-2063.

[41] Zhang, Y., and P. S. Cremer. 2006. Interactions between macromolecules and ions: the Hofmiester series. Curr. Opin. Chem. Biol. 10:658-663.

[42] Gjerde, D. T., G. Schmuckler, and J. S. Fritz. 1980. Anion Chromatography with low-conductivity eluents. II. Journal of Chromatography A 187:35-45.

[43] Gregor, H. P., J. Belle, and R. A. Marcus. 1954. Studies on Ion Exchange Resins. IX. Capacity and Specific Volumes of Quaternary Base Anion Exchange Resins1. Journal of the American Chemical Society 76:1984-1987.

[44] Chen, X., S. C. Flores, S.-M. Lim, Y. Zhang, T. Yang, J. Kherb, and P. S. Cremer. 2010. Specific Anion Effects on Water Structure Adjacent to Protein Monolayers Langmuir 26:16447-16454.

[45] Schwierz, N., D. Horinek, and R. R. Netz. 2010. Reversed Anionic Hofmeister Series: The Interplay of Surface Charge and Surface Polarity. Langmuir 26:7370-7379.

[46] Leckband, D., and J. Israelachvili. 2001. Q. Rev. Biophys.

[47] Green, A. 1932. Studies in the physical chemistry of the proteins X: the solubility of hemoglobin in solutions of chlorides and sulfates of varying concentration. J. Biol. Chem. 95:47-66.

[48] Chen, X., T. Yang, S. Kataoka, and P. S. Cremer. 2007. Specific Ion Effects on Interfacial Water Structure near Macromolecules. Journal of the American Chemical Society 129:12272-12279.

[49] Fesinmeyer, R., S. Hogan, A. Saluja, S. Brych, E. Kras, L. Narhi, D. Brems, and Y. Gokarn. 2009. Effect of Ions on Agitation- and Temperature-Induced Aggregation Reactions of Antibodies. Pharmaceutical Research 26:903-913.

[50] Gokarn, Y. R., R. M. Fesinmeyer, A. Saluja, V. Razinkov, S. F. Chase, T. M. Laue, and D. N. Brems. 2011. Effective charge measurements reveal selective and preferential accumulation of anions, but not cations, at the protein surface in dilute salt solutions. Protein Science 20:580-587.

[51] Lima, E. R. A., M. Bostrom, D. Horinek, E. C. Biscaia, W. Kunz, and F. W. Tavares. 2008. Co-Ion and Ion Competition Effects: Ion Distributions Close to a Hydrophobic Solid Surface in Mixed Electrolyte Solutions. Langmuir 24:3944-3948.

[52] Sivasankar, S., S. Subramaniam, and D. Leckband. 1998. Direct molecular level measurements of the electrostatic properties of a protein surface. Proceedings of the National Academy of Sciences 95:12961-12966.

[53] Tavares, F. W., D. Bratko, H. W. Blanch, and J. M. Prausnitz. 2004. Ion-specific effects in the colloid-colloid or protein-protein potential of mean force: role of salt-macroion van der Waals interactions. J. Phys. Chem. B. 108:9228-9235.

Part 3

Others

19

Computational Tools and Databases for the Study and Characterization of Protein Interactions

Jose Ramon Blas[1], Joan Segura[2] and Narcis Fernandez-Fuentes[2,3]
[1]*Universidad de Castilla-La Mancha*
[2]*University of Leeds*
[3]*Aberystwyth University*
[1]*Spain*
[2,3]*United Kingdom*

1. Introduction

One of the most pressing challenges in the post genomic era is the characterization and charting of protein-protein interactions (PPIs) in living organisms, as these are essential in the shaping of normal and pathological behaviours in cells. It is for this reason that unravelling the nature of PPIs has been the pursuit of many experimental techniques, ranging from high-throughput to high-detail approaches (Shoemaker and Panchenko 2007), as well as a wide spectrum of computational prediction methods. Current estimations of human interactome size range from 100,000 to more than 600,000 interactions (Bork et al. 2004; Stelzl and Wanker 2006; Stumpf et al. 2008; Venkatesan et al. 2009). Experimental strategies have reached their best at describing around 50,000 interactions by collating a large number of small and very focused experiments with high-throughput ones, such as massive yeast two-hybrid (Rual et al. 2005; Stelzl et al. 2005), or mass spectrometry coupled to affinity purification experiments (Ewing et al. 2007; Hubner et al. 2010). The smallest gap between experimentally validated and theoretically predicted PPIs amounts to around 50% of total interactions, being probably much higher. When it comes to to studying PPIs in other species on which, even having been sequenced, experimental data is even more scarce, the need for PPI-map completeness is even more notorious. Computational prediction and characterization of PPIs, with its drawbacks, successes and challenges, constitutes a valuable aid in the way to a complete description of interactomes, hence being a promising research field that has enriched our image of living cells for some time now.

Computational tools can provide useful information at different levels of resolution and this chapter seeks to present an up-to-date and comprehensive review of these. The first part of the chapter presents the theoretical basis of computational tools designed to predict PPIs. The main aim of these tools is to predict whether two proteins A and B can interact, either directly or indirectly (functional associations), but without dwelling on the molecular details of the interaction, i.e. which proteins interact. These predictions are useful as complement to large-scale experimental analyses, either to confirm observed interactions or discard false

positives, and also to uncover novel interactions. The second part of the chapter is devoted to the computational methods developed to predict protein interfaces. At this level, predictions identify specific regions and residues of the protein that are likely to mediate PPIs. Thus, these methodologies uncover a higher level of detail, i.e. *how* proteins interact, and have a number of applications in experimental work such as guiding the mapping of protein interfaces by mutagenesis or structural modelling of protein complexes. A special emphasis will be given to a novel and highly accurate tool: VORFFIP(Segura et al. 2011). The concluding part of the chapter describes computational tools developed to predict the important regions or *hot spots* in protein interfaces. Recent successes in the quest for finding new therapeutic agents to modulate PPIs have been aided by the realization, following the pioneering work by Clackson and Wells(Clackson and Wells 1995), that the binding energy of many PPIs can be ascribed to a small and complementary set of interfacial residues: a *hot spot* of binding energy. Thus, identifying these critical residues by computational means has clear applications in drug discovery and in some aspects of protein design. PCRPi(Assi et al. 2010), a novel and highly precise tool will be discussed.

2. Prediction of protein-protein interactions

With the aim of detailing a complete protein interaction map that agglutinates the rising amount of genomic data, high-throughput experimental techniques have walked in parallel with computational approaches. There are six basic computational approaches to predict PPIs depending on the nature of the information used for the prediction. These include PPIs inferred from: (i) genomic context including phylogenetic profiles, gene neighbouring analyses and gene fusion events; (ii) co-evolution events; (iii) protein domain co-occurrence (or signatures) between pair of proteins; (iv) text mining; (v) transference of annotation between species: protein-protein interologs; and (vi) structural annotation including homology-based or *ab initio*. Figure 1 depicts an overall diagrammatic description of these basic approaches and tables 1 and 2 compile a number of on line databases and computational tools respectively.

2.1 Genomic context methods

Biological processes subjected to evolutionary pressure tend to cluster together all interrelated molecular actors in single units to simplify control mechanisms and thus avoid the lost of any essential component. This principle, which operates from maintaining bacterial operon systems to more sophisticated co-regulation strategies found in eukaryotes, is on the basis of genomic context based methods for the detection of functional PPIs.

The first group of genome context methods are based in the comparison of phylogenetic profiles. A phylogenetic profile is the presence or absence of a given gene across N species that can be expressed as an N-dimensional array of ones and zeroes. Originally, functional relationship was assumed if having similar phylogenetic profiles(Pellegrini et al. 1999); however, positive results were limited to very strong interactions and many relations between analogous proteins were missing. Further improvements were made by discarding overlaps given by chance(Wu et al. 2003), using protein domains instead of full length proteins(Pagel et al. 2004), through a concurrent search of multiple independent phylogenetic events of gain/loss of pairs of genes to discard spurious correlated

patterns(Barker and Pagel 2005), or the use of enhanced representation of phylogenetic trees(Ta et al. 2011). The second group of genome context methods is based on gene closeness among different genomes, considering closeness as a sign of functional relatedness. After some initial successes(Koonin et al. 2001; Evguenieva-Hackenberg et al. 2003) and despite some improvements such as allowing for changes in gene order and orientation(Szklarczyk et al. 2011), large-scale predictions should be considered cautiously. Finally, gene fusion approaches are a group of computational tools based on the evidence that some interacting proteins have orthologous where both proteins appear fused in a single protein. Thus, it has been observed that many of these pairs, fused in single proteins in other organisms, correspond to binding partners or at least functionally related proteins(Marcotte et al. 1999; Yanai et al. 2001). Rosetta method(Marcotte et al. 1999) and other implementations(Enright et al. 1999) exploit gene fusion events as predictors of PPIs.

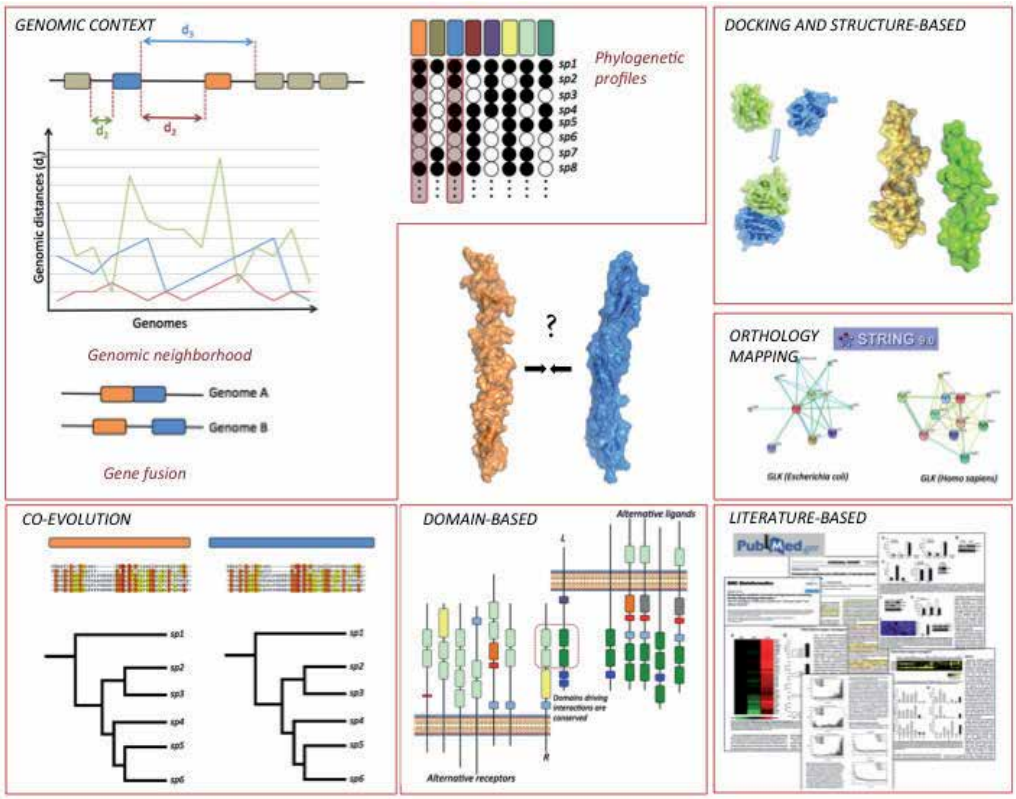


Fig. 1. Graphical description of the main strategies of PPIs prediction described in sections 2.1 to 2.6.

2.2 Co-evolution methods

Two proteins that share a functional relationship, either through direct interaction or functional association, may present evidences of co-evolution. Since the seminal work of Altschuh identifying correlated amino acid changes in Tobacco mosaic virus(Altschuh et al. 1987), studies recognizing co-evolution as an indicative, albeit subtle, signal of PPIs have

being reported in the literature (Travers and Fares 2007; Chao et al. 2008; Presser et al. 2008). Co-evolutionary information may be divided into three groups: the simultaneous loss or gain of orthologous genes(Marcotte et al. 1999), correlated changes affecting both interacting partners at whole sequence level (explored by *mirrortree*-based approaches)(Goh and Cohen 2002; Hakes et al. 2007; Juan et al. 2008) or single amino acids changes(Mintseris and Weng 2005; Madaoui and Guerois 2008).

In the case of mirrortree-based methods (e.g.(Ochoa and Pazos 2010)), the likelihood of interaction is measured as a correlation value between the phylogenetic trees of two families of proteins. Although these approaches have been successfully applied in PPIs prediction(Labedan et al. 2004; Dou et al. 2006; McPartland et al. 2007; Juan et al. 2008), it is still a major problem distinguishing between co-evolution arising from a direct PPI, what has been termed as *co-adaptation* (Pazos and Valencia 2008), from non-specific changes and thus not necessary driven by a functional relatedness(Lovell and Robertson 2010). Recent advances in this area include MatrixMatchMaker algorithm(Tillier and Charlebois 2009) and a faster implementation suitable for large-scale analyses(Rodionov et al. 2011).

The detection of site-specific co-evolution events reflecting PPIs, despite being more intuitive and informative, is even more challenging given the complexity of the mixed evolutionary-structural scenario involved. A single point mutation might ease or complicate each imaginable path of mutation at any other position in the complex, regardless of its distance from the interface(Lovell and Robertson 2010). In fact, co-evolution events have been detected affecting sites that are distant structurally (Gobel et al. 1994; Clarke 1995; Gloor et al. 2005; Fares and McNally 2006). On the other hand, the probability of correlated amino acid changes is closely related to the chemical nature of changes. In this sense, volume variations seem to strongly affect fitness, and so they are frequently balanced by evolution machinery (up to almost 50% of the cases)(Williams and Lovell 2009). Moreover, interface residues in obligate complexes evolve at a slower rate than those in transient interactions(Mintseris and Weng 2005).Taken together, all these particulars illustrate the challenges encountered when looking for site-specific co-evolution events related to PPIs. Recent developments have looked at improving the discrimination between direct and indirect correlations(Burger and van Nimwegen 2010), or including amino acid background distribution information and the mutual information of residues physicochemical properties(Gao et al. 2011). However, new, more discriminative, approaches are required to better understand co-evolution at residue-centred level.

2.3 Domain-based methods

There are strong evidences supporting the idea that the range of different PPIs can be accounted for by considering a more reduced set of specific domain-domain interactions, domain signatures, that are even conserved across different species(Finn et al. 2006; Itzhaki et al. 2006; Stein et al. 2011). Thus, the basis of domain-based methods is presence/absence of given domain signatures between pairs of proteins that can be used to infer interaction. An early method exploiting domain signatures was an association method where domain interactions were assumed if the frequency of association was higher than the expected frequency(Kim et al. 2002). Further improvements have been devised to improve predictions including the domain pair exclusion analysis, which implemented a new scoring

scheme(Riley et al. 2005), the use of Random Forest ensemble classifiers to deal with the pairing of multi-domain proteins(Chen and Liu 2005) or the use of Gene Ontology(Lee et al. 2006) or co-evolution data(Jothi et al. 2006).

2.4 Literature-based data mining methods

Numerous research efforts have been focused on automatically extracting and analysing information from the scientific literature in order to infer putative PPIs(Blaschke et al. 2001; Fundel et al. 2007; Airola et al. 2008). These include, the search for the co-occurrence of terms(Blaschke et al. 2001) or the presence of similar Gene Ontology terms(Pesquita et al. 2009) or kernel-based methods including subsequence kernels, tree kernels, shortest path kernels and graph kernels(Tikk et al. 2010). The most recent approaches use multiple kernels to maximize the information extracted from scientific papers(Kim et al. 2008; Miwa et al. 2009), the combination of multiple kernels and machine learning algorithms to improve the scoring(Yang et al. 2011), or the more recent neighbourhood hash graph kernels that are substantially faster than previous text-mining approaches(Zhang et al. 2011).

Name	URL	Reference
STRING	http://string-db.org	(Szklarczyk et al. 2011)
BioGRID	http://thebiogrid.org/	(Stark et al. 2011)
IntAct	http://www.ebi.ac.uk/intact/	(Aranda et al. 2010)
HPRD	http://www.hprd.org/	(Prasad et al. 2009)
HitPredict	http://hintdb.hgc.jp/htp/	(Patil et al. 2011)
DIP	http://dip.doe-mbi.ucla.edu/dip	(Salwinski et al. 2004)
MINT	http://mint.bio.uniroma2.it/mint/	(Chatr-aryamontri et al. 2007)
TAIR	www.arabidopsis.org/portals/proteome/	(Swarbreck et al. 2008)
iPFAM	http://ipfam.sanger.ac.uk/	(Finn et al. 2005)
3DID	http://3did.irbbarcelona.org/	(Stein et al. 2011)
DIMA 3.0	http://webclu.bio.wzw.tum.de/dima/	(Luo et al. 2011)
DOMINE	http://domine.utdallas.edu/cgi-bin/Domine	(Yellaboina et al. 2011)
GWIDD	http://gwidd.bioinformatics.ku.edu	(Kundrotas et al. 2010)
IsoBase	http://isobase.csail.mit.edu/	(Park et al. 2011)
I2D	http://ophid.utoronto.ca/ophidv2.201	(Brown and Jurisica 2007)
DroID	http://www.droidb.org	(Murali et al. 2011)
HCPIN	http://nesg.org:9090/HCPIN	(Huang et al. 2008)
HIV1,HPID	http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions	(Fu et al. 2009)
MPIDB	http://www.jcvi.org/mpidb/about.php	(Goll et al. 2008)

Table 1. List of major databases compiling experimentally determined or computationally predicted PPIs.

2.5 Orthology mapping (Interologs) methods

The basis of these methods is the transference of annotated interactions between organisms; hence the term *interologs* to refer to predicted homologous interactions(Walhout et al. 2000; Shoemaker and Panchenko 2007; Lewis et al. 2010). Interolog annotations have been successfully applied to transfer experimentally known interactions in yeast to predicted ones in worm(Matthews et al. 2001) and between mouse and human(Huang et al. 2007). Although some improvement has been devised such as scoring schemes that depend on the sources of experimental data(Jonsson and Bates 2006), the applicability of orthology mapping is limited. Firstly, accurate predictions require high sequence similarities between interologs (~70%)(Mika and Rost 2006) thus limiting its range of applicability. Secondly, even at high sequence identity level, in some cases small variations in protein sequence at the interface have been shown to dramatically change PPI specificity, thus redefining complex protein networks and leading to important phenotypic differences(Panni et al. 2002; Kiemer and Cesareni 2007).

2.6 Structure-based methods

A final category of computational methods includes those based in structural information. The structure of a protein complex formed by two or more proteins can be modelled using the structure of a known protein complex as template either by homology modelling or threading(Lu et al. 2002; Aloy et al. 2004; Hue et al. 2010). Even in the absence of a suitable template, the structure of the complex can be modelled by using protein docking(Wass et al. 2011) and selecting the protein complex based on predicted binding energy, i.e. *ab initio* modeling. Despite being a promising strategy, and without considering the high computational cost, the correlation between predicted and experimentally measured binding affinities, such as K_d, is very low thus greatly impairing its predictive power (Kastritis and Bonvin 2010; Stein et al. 2011). Other strategies combine structural data, docking and evolutionary conservation(Tuncbag et al. 2011).

Name	Methodology	URL	Reference
MirrorTree	Co-evolution	http://csbg.cnb.csic.es/mtserver/	(Ochoa and Pazos 2010)
MatrixMatchMaker	Co-evolution	http://www.uhnresearch.ca/labs/tillier/	(Tillier and Charlebois 2009)
iHOP	Text mining	http://www.ihop-net.org/	(Hoffmann and Valencia 2004)
PathBLAST	Orthology	http://www.pathblast.org/	(Kelley and Ideker 2005)
InterPreTS	Structure-based	http://www.russelllab.org/	(Aloy and Russell 2003)
IBIS	Structure-based	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	(Shoemaker et al. 2010)

Table 2. List of on line resources for the prediction of PPIs.

3. Prediction of protein binding sites

As indicated by its name, binding site prediction methods seek to define the regions in proteins that are more likely to mediate PPIs. The level of resolution is therefore higher and the starting point is either the sequences or structures of proteins that are known to interact (i.e. experimental evidence) but for which no structural details of the interaction are known.

3.1 Distinctiveness of interface residues

Large-scale analyses of the structures of protein complexes have shown that residues located in interfaces present a number of differential physicochemical and structural qualities. In general, hydrophobic residues are overrepresented in the interfaces of permanent complexes(Lo Conte et al. 1999; Glaser et al. 2001) and charge residues, Arg in particular, are also commonly found in interfaces and often define the lifetime of complexes(Zhou and Shan 2001). A higher accessibility to the solvent than exposed residues not located in interfaces is also a differential trait of interface residues(Chen and Zhou 2005), being the most effective feature to predict interfaces in homodimeric complexes(Jones and Thornton 1997). On the other hand and in agreement with earlier observations that found interface residues have lower crystallographic B-factors(Neuvirth et al. 2004), the side chains of interface residues are less likely to sample alternative rotamers, i.e. more rigid, to decrease the entropic cost upon complex formation(Cole and Warwicker 2002; Fleishman et al. 2011). Sequence conservation has also proved to be a predictor(Lichtarge et al. 1996; Wang et al. 2006), although it remains a contentious issue as some works have shown that interfaces are not more conserved than the rest of the protein(Grishin and Phillips 1994; Caffrey et al. 2004). Finally, it has been shown that interfaces are richer in β-strands and long loops while α-helical conformations are disfavoured(Neuvirth et al. 2004).

3.2 Prediction methods

Prediction methods rely on sequence and/or structural information that is unique to interface residues (see before). Hence, prediction methods can be divided into two groups: sequence-based methods, which rely only on the primary sequence of the protein and structure-based methods that require the three-dimensional structure of the protein. Table 3 compiles a list of on line computational tools to predict protein binding sites.

3.2.1 Structure-based prediction methods

One of the first structure-based prediction methods, later updated(Murakami and Jones 2006), was based on surface patch analysis(Jones and Thornton 1997). Surface patches were defined by grouping neighbouring exposed residues that were subsequently ranked using a scoring function that included the solvation potential, interface propensity, hydrophobicity, protrusion and accessible surface area of each of the residues within the patch. A probabilistic approach, ProMate, also based on patch analysis, was developed for heteromeric transient protein complexes by combining secondary structure content, hydrophobicity and crystallographic B-factors information(Neuvirth et al. 2004). The combination of ProMate's predictions and a parametric scoring function based of sequence conservation and structural features resulted in an improvement of the accuracy of the predictions(de Vries et al. 2006). Other implementations of prediction methods include an

empirical scoring function composed of side chain energy score, residue conservation and interface propensity(Liang et al. 2006), the search of structural interaction templates extracted from protein complexes(Chang et al. 2006) and a clustering algorithm that identifies residues with a high propensity of being located in interfaces(Negi et al. 2007).

In order to combine and integrate heterogenous data, i.e. sources of information of a different nature (e.g. hydrophobicity indexes and solvent accessibility surface) into a common and coherent scoring framework, a number of machine learning methods have been proposed including Neural Networks (NN)(Fariselli et al. 2002; Chen and Zhou 2005; Porollo and Meller 2007), Support Vector Machines (SVM)(Bradford and Westhead 2005), Random Forests (RF)(Sikic et al. 2009; Segura et al. 2011) and Bayesian Networks (BN)(Bradford et al. 2006; Ashkenazy et al. 2010). Thus, the commonality of these approaches is the use of a machine-learning algorithm (NN, SVM, RF or BN) to combine a set of sequence- and structural-based measures into an unified score or probability. The nature of the combined features used by the prediction methods includes: evolutionary conservation and surface disposition(Fariselli et al. 2002); sequence conservation, electrostatic potentials, SASA, hydrophobicity, protusion and interface propensity(Bradford and Westhead 2005; Bradford et al. 2006); properties taken from the AAIndex database(Kawashima et al. 2008) (e.g. expected number of contacts within 14 Å sphere), multiple sequence alignment-derived features (e.g. amino acid frequency), and structural features(Porollo and Meller 2007); structure-based, energy terms, sequence conservation and crystallographic B-factors(Segura et al. 2011); structural features, sequence and secondary structure(Sikic et al. 2009); or more complex approaches that combine several prediction methods in the form of a meta-prediction(Qin and Zhou 2007; Ashkenazy et al. 2010).

3.2.2 Sequence-based prediction methods

Even if the structure of the protein is not available, there are still a number of prediction methods that are based solely on sequence information. Early examples of approaches in this category include a NN (Ofran and Rost 2003) that uses local sequence information, which was subsequently improved by including a post-neural network filtering step(Ofran and Rost 2007). Other approaches include SVMs that combine sequence profiles and other sequence-based information such as spatially neighbouring residues(Koike and Takagi 2004; Res et al. 2005; Chen and Li 2010), a RF that integrates physicochemical properties of residues, evolutionary conservation and amino acid distances(Chen and Jeong 2009), and a naive Bayesian classifier trained to integrate position-specific scoring matrix and predicted accessibility(Murakami and Mizuguchi 2010). Finally, other sequence-based methods have been developed to improve prediction by tacking issues such as the problem of unbalanced data in protein sets(Yu et al. 2010), i.e. the interface accounts for a small proportion of the exposed residues so the number of negative cases (non-interface residues) is much larger than the number of positive cases (interface residues) or improving the sampling(Engelen et al. 2009) in evolutionary trace-based(Lichtarge et al. 1996) methodologies.

3.3 VORFFIP, a holistic approach to predict protein binding sites in protein structures

VORFFIP is a novel, structure-based, method that integrates a wide range of residue-based features and environment information using a 2-step Random Forest ensemble

classifier(Segura et al. 2011). Residue-based features include structural-based, energy terms, evolutionary conservation and crystallographic B-factors information. VORFFIP implements a novel definition of local environment by means of Voronoi Diagrams (see next and Fig. 2) that complements residue-based information improving the accuracy of predictions.

Residue-based information characterizes individual residues. Structure-based features account from 16 different features and define the local geometry of the protein at residue level. Structural features include, among others, the absolute and relative accessibility surface area, the protrusion index that is a measure of the local concavity/convexity and a deepness index(Vlahovicek et al. 2005). The energetic state of exposed residue is characterized by 10 energy terms including electrostatic potential, solvent exposure energy, entropy and hydrogen bond energy among others(Guerois et al. 2002). The sequence conservation of residues consist of the regional conservation score that defines the conservation for each residue and its neighbourhood in the 3D space(Landgraf et al. 2001) and a sequence positional score calculated from multiple sequence alignment profiles(Pei and Grishin 2001). Finally, crystallographic B-factors, which are a measure of thermal motion, are converted to Z-score as described previoulsy(Yuan et al. 2003).

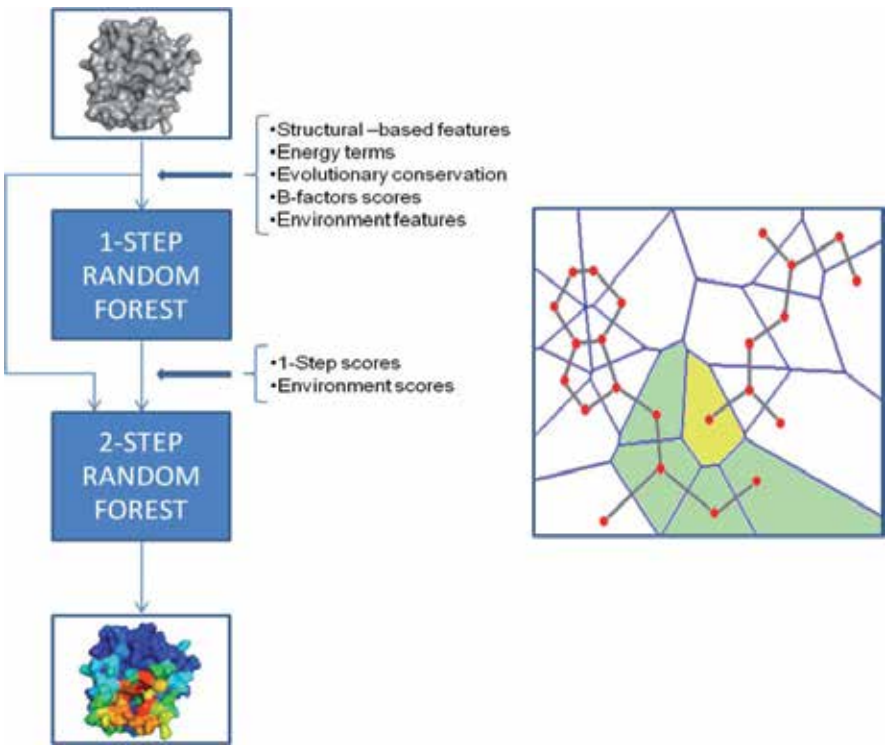


Fig. 2. Overview of prediction process in VORFFIP and a Voronoi Diagram of a interacting pair. The left side of the figure illustrates the 2-step prediction approach in VORFFIP. The right side of the figure shows the Voronoi Diagram of two neighbouring residues; heavy atoms are represented by red dots and coloured cells illustrate interaction between atoms of neighbouring residues.

Environment-based information accounts the local structural environment of residues. Interfaces tend to form contiguous patches on the surface and thus, the environment of a residue can provide valuable information for predictions. Several methods have been used to account for the local environment of residues including sliding window (e.g.(Ofran and Rost 2003)) and Euclidian distances (e.g. (Porollo and Meller 2007)). VORFFIP however uses a novel definition of environment by means of Voronoi Diagrams (VD). VD is computed using the heavy atoms coordinates as seeds and as a result the 3D space is partitioned into polyhedral cells where each single cell contains one of the atoms (Barber et al. 1996). Atoms sharing a common facet in the VD are said to be in contact or neighbours, i.e. part of the local environment. Figure 2 shows a 2D representation of a VD diagram depicting the interaction between atoms of two neighbouring residues. The number of contacts between

Name	Input	Method	URL	Reference
VORFFIP	Structure	RF	http://www.bioinsilico.org/VORFFIP	(Segura et al. 2011)
ProMate	Structure	Scoring function	http://bioinfo.weizmann.ac.il/promate	(Neuvirth et al. 2004)
ISIS	Sequence	NN	http://rostlab.org/cms/resources/web-services/	(Ofran and Rost 2007)
WHISCY	Structure	Scoring function	http://nmr.chem.uu.nl/Software/whiscy	(de Vries et al. 2006)
PPI-pred	Structure	SVM	http://www.bioinformatics.leeds.ac.uk/ppi_pred	(Bradford and Westhead 2005)
SPPIDER	Structure	NN	http://sppider.cchmc.org	(Porollo and Meller 2007)
PINUP	Structure	Scoring function	http://sparks.informatics.iupui.edu	(Liang et al. 2006)
meta-PPISP	Structure	Meta-server	http://pipe.scs.fsu.edu/meta-ppisp.html	(Qin and Zhou 2007)
Protemot	Structure	Scoring function	http://bioinfo.mc.ntu.edu.tw/protemot	(Chang et al. 2006)
InterProSurf	Structure	Scoring function	http://curie.utmb.edu	(Negi et al. 2007)
cons-PPISP	Structure	NN	http://pipe.scs.fsu.edu/ppisp.html	(Chen and Zhou 2005)
PSIVER	Sequence	BN	http://tardis.nibio.go.jp/PSIVER/	(Murakami and Mizuguchi 2010)
SHARP	Structure	Scoring function	http://www.bioinformatics.sussex.ac.uk/SHARP2	(Murakami and Jones 2006)

Table 3. List of online resources for protein binding site prediction.

neighbouring residues is used to derive weights that will be then used to normalize residue-based features among residues within the local environment. The advantage of using VD over other definition of local environment is that there are no requirements with regards cut-off to define the local environment (e.g. a distance cut-off) and that a weighting system can be easily implemented based on the number of interactions (i.e. neighbouring residues) in the VD. When the performance of VORFFIP was assessed in term of type of methods used to define local environment, VD were superior to Euclidean distances and sliding window approaches(Segura et al. 2011).

The final stage of the method is the integration of residue- and environment-based features using a machine learning approach: a 2-steps RF ensemble classifier (Fig. 2), which is also a novel feature as most machine learning methodologies use a single step classifier. In the first-step RF, residue and residue-environment features are calculated and used as input variables. The scores yielded by the first-step RF are then decomposed into a number of new input variables including VD-derived environment scores. Residue and environment scores together with the previously calculated features form the new set of input variables to the second-step RF that will output the final scores. The logic behind using a second-step RF relates to the observation that residues belonging to the same interface tend to form contiguous patches on the surface, i.e. high scoring residues are expected to be neighbouring mainly high scoring residues unless located at the boundaries of the interface. Thus, the second-step RF harmonizes outliers and generates more homogenous scores for interface residues resulting in better predictions as shown by the competitive results obtained(Segura et al. 2011) when comparing to other methods(de Vries et al. 2006; Porollo and Meller 2007; Sikic et al. 2009).

4. Prediction and charting of hot spots in protein interfaces

The final part of the chapter describes the current state in computational prediction of hot spots in protein interfaces. The goal of these methods is the prediction of the region of a given interface that contributes the most to the binding energy of the complex, i.e. the hot spot of the interaction. These methods are a good complement to highly intensive and costing experimental techniques, in particular in large-scale analyses, and have clear applications in drug discovery and protein engineering.

4.1 Distinctiveness of hot spot residues

As in the case of interface residues, hot spot residues present a number of structural and physicochemical properties unique to them and these are exploited by the prediction methods. The first is the type of residues that are commonly found in hot spots: while the proportion of Trp, Arg and Tyr is higher, Leu, Ser and Val are disfavoured(Bogan and Thorn 1998). Likewise, Asn and Asp are more commonly found in hot spots than chemically comparable (but bulkier) Gln and Glu(Bogan and Thorn 1998). Hot spot residues are optimally packed, structurally conserved and usually located in the central part of the interface(Keskin et al. 2005; Yogurtcu et al. 2008). One more characteristic of hot spot residues is that they are often located in complemented pockets, i.e. hot spot residues in one protein interact with hot spot residues of cognate protein(s)(Li et al. 2004). Finally, hot spot residues usually have a higher evolutionary conservation than the rest of the residues in the interface(Guharoy and Chakrabarti 2005).

4.2 Prediction algorithms

A number of computational methods have been developed for the prediction of hot spots in protein interfaces. An important part of these is represented by energy-based methods that predict changes in binding energy upon mutations, i.e. *in silico* alanine scanning. These methodologies range from scoring function derived from simple physical models(Guerois et al. 2002; Kortemme and Baker 2002; Kruger and Gohlke 2010) to more complex, time consuming atomistic simulations to model effect of mutations in the binding energy(Almlof et al. 2006; Lafont et al. 2007; Moreira et al. 2007; Benedix et al. 2009; Diller et al. 2010). Other methods exploit individual features (or combination of them) that are characteristic to hot spots such as solvent accessibility(Landon et al. 2007; Tuncbag et al. 2009; Xia et al. 2010; Li et al. 2011), atomic contacts(Li et al. 2006), structural conservation(Li et al. 2004), restricted mobility(Yogurtcu et al. 2008), relative location of residues in the interface(Keskin et al. 2005), sequence conservation(Hu et al. 2000; Ma and Nussinov 2007) and pattern mining(Hsu et al. 2007). Other examples include a number of machine learning approaches (Darnell et al. 2007; Ofran and Rost 2007; Cho et al. 2009; Lise et al. 2009; Assi et al. 2010) such as PCRPi (see next) that integrate a range of structural- and sequence-based information and a docking-based approach(Grosdidier and Fernandez-Recio 2008).

4.3 PCRPi: *P*resaging *C*ritical *R*esidues in *P*rotein *i*nterfaces, a novel and highly accurate prediction algorithm

While the attributes described in section 4.1 have predictive power, it has been found that individually cannot unambiguously define hot spot residues(DeLano 2002). To overcome this limitation, PCRPi(Assi et al. 2010), a novel computation tool for the prediction of hot spots residues, integrates seven different variables that account for structural, evolutionary conservation and predicted binding energy (Fig.3).

The structural information of interface residues is described by two different variables: the interaction engagement (IE) and the topographical (TOP) indexes. The IE index gauges for the number of inter-chain atomic interactions of the given residue normalized by total number of atoms that can potentially interact. An IE index of 1.0 would indicate that all atoms are actively engaged in atomic interactions with groups of cognate protein(s). The TOP index describes the structural environment of residues and is ratio between the number of neighbouring residues of cognate proteins and the average number neighbouring residues. Neighbouring residues are any residues of cognate protein(s) whose carbon alpha is enclosed in a sphere of 10 Angstroms of radius centered on the carbon alpha of the residue of interest. Thus, TOP index quantifies whether residues are intimately interacting with cognate proteins or are located in a more flat or unprotected region.

The second group of variables used by PCRPi relates to evolutionary conservation. Evolutionary conservation is quantified by looking at the sequence conservation and the 3D regional conservation (i.e. structural conservation of patches) in both target (ANCCON and ANC3DCON) and cognates proteins (CON and 3DCON). To calculate ANCCON and CON values, sequence profiles are derived as described(Fernandez-Fuentes et al. 2007). Next, ANCCON corresponds to conservation scores as calculated by al2co(Pei and Grishin 2001) and the CON variable is the ratio between residues with and al2co scores above 1.0 an the number of cognate residues in the interface. Likewise, the ANC3DCON and 3DCON values

are calculate but instead of using al2co scores, the normalize regional conservation scores as defined by Landgrad et al(Landgraf et al. 2001) are used. The last input used by PCRPi is the BE index, which represents the predicted binding energy change upon mutation, i.e. *in silico* Alanine scanning, as calculated using FoldX(Guerois et al. 2002).

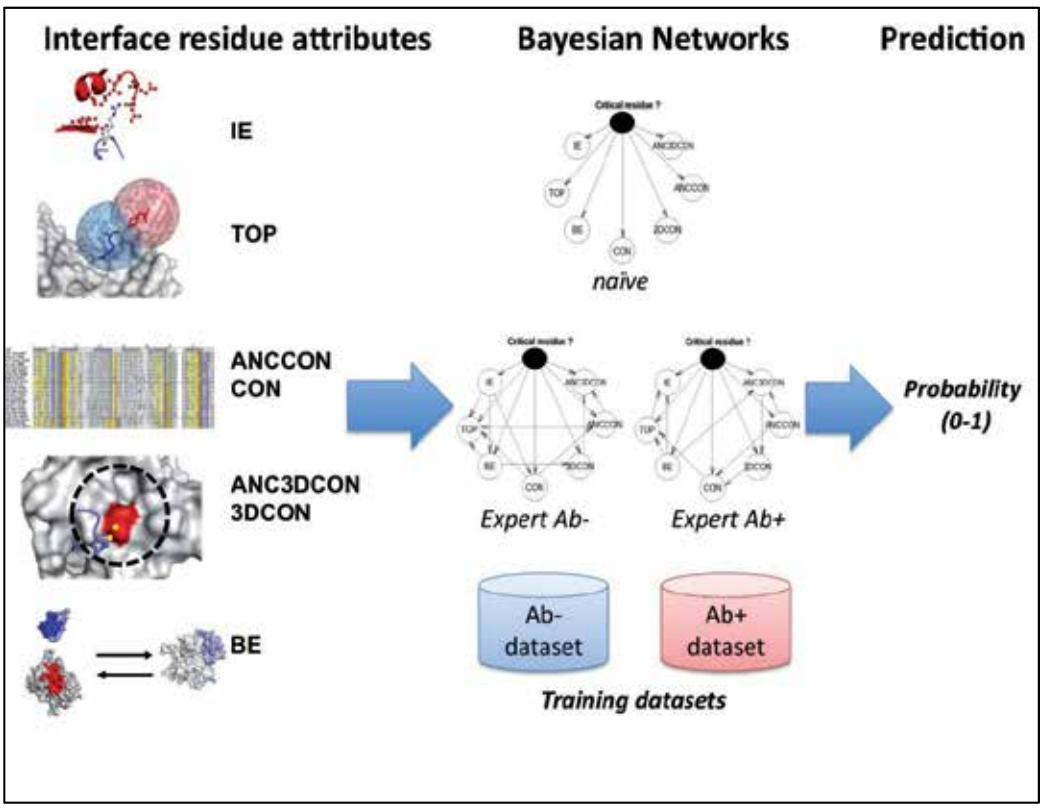


Fig. 3. Overview of the prediction process. PCRPi integrates seven features characterizing interface residues that are used as input variables to three different Bayesian networks, two experts and one naïve, that can be trained with protein complexes including (Ab+) or excluding (Ab-) Antigen-Antibodies complexes. PCRPi outputs a probability where the higher the probability the more likely the residues to be critical, i.e. hot spot residues, for the interaction.

The final part of the prediction is the integration of the data, i.e. IE, TOP, ANCCON, CON, ANC3DCON, 3DCON and BE, into a common probabilistic framework by using BN. PCPRi features three different types BN, two experts and one naïve (Fig. 3). The difference between them is the relationship of dependence between input variables; while naïve BN assumes independence, an expert BN allows conditional dependence between variables (Fig. 3). Both expert and naïve BNs are trained using two specific sets of protein complexes: Ab+ and Ab- (Fig. 3). The Ab+ set corresponds to protein complexes that can include non-evolutionary related complexes such as Antigen-Antibodies complexes while Ab- does not include the latter. The reason being is the lack of sequence conservation in the complementary determining regions of Antibodies, i.e. regions that mediate interaction, which renders

evolutionary information meaningless for prediction purposes and thus special BNs were devised to cope with this problem. In terms of performance, PCRPi delivers highly consistent and competitive predictions as shown in the study of the protein complex formed by RAS and VH-HRAS antibody(Tanaka et al. 2007) and a comprehensive comparative study(Assi et al. 2010). Moreover, PCPRi is a central part of a database that compiles and annotates hot spot in protein interfaces: PCRPI-DB (Segura and Fernandez-Fuentes 2011).

Name	Input	Method	URL	Reference
PCRPi	Structure	Machine learning	http://www.bioinsilico.org/PCRPi	(Assi et al. 2010)
Robetta	Structure	Energy-based	http://robetta.bakerlab.org	(Kortemme and Baker 2002)
FoldX	Structure	Energy-based	http://foldx.crg.es	(Guerois et al. 2002)
DrugScorePPi	Structure	Energy-based	http://cpclab.uni-duesseldorf.de/dsppi	(Kruger and Gohlke 2010)
CC/PBSA server	Structure	Energy-based	http://ccpbsa.biologie.uni-erlangen.de/ccpbsa	(Benedix et al. 2009)
KFC	Structure	Machine learning	http://kfc.mitchell-lab.org	(Darnell et al. 2008)
HotPoint	Structure	Scoring function	http://prism.ccbb.ku.edu.tr/hotpoint	(Tuncbag et al. 2009)
ISIS	Sequence	Neural Network	http://rostlab.org/cms/resources/web-services/	(Ofran and Rost 2007)

Table 4. List of online resources for prediction of hot spots.

5. Conclusions and outlook

During the last years, scientists aiming at understanding living organisms at a molecular level have seen their benches become swapped with the sheer amount of information and this burst of data being mirrored by the development of a wide and miscellaneous set of computational tools designed to unveil biologically relevant information from the noisy background. PPIs are among the most crucial events that define the behaviour of a living system and that explains the rise of research efforts and strategies to describe the nature of PPIs. This chapter presents a summary and extensive view on computational methods devoted to predict which proteins participate in PPIs (section 2), which are the regions involved in the interaction (section 3) and which are the most important regions or residues in the interaction (section 4).

In general the prediction tools achieve a high rate of prediction success and are important tools for scientists. However, there are still a number of unmet needs and challenges to be solved. In the case of prediction of PPIs, genome context approaches would benefit from improved definitions of phylogenetic profiles and the masking effect of gene fusion events. Text-mining approaches require further development to reduce false positive rates and increase efficiency. A deeper understanding of the complex interlink between (bio)chemistry,

structure and genetics that governs the evolution of protein interfaces would certainly benefit co-evolution-based methods. The correct detection of remote homology between interologs is a major challenge as is the lack of correlation between predicted and observed binding affinities in structure-based methods.

Protein binding site prediction methods have also their own limitations and challenges. The physical forces and chemical properties that drive the interaction between proteins are not fully understood and thus current models do not reflect the binding process accurately. However, the increasing amount of experimental data that is being generated in an important factor that plays in favour of developing novel and more accurate computational tools. Some specific challenges in the field are the prediction of binding sites in proteins that recognize multiple partners (hub proteins) and the distinction between each of the interfaces that are relevant to each of the interacting partners. Current methods cannot properly handle binding events that involve conformational changes in any of the intervening components, including those mediated by intrinsic disordered regions, and thus future efforts need to be directed to tackle this very important question. Finally, the main challenge in the prediction of hot spots is the development of new approaches to bridge the gap between highly computationally expensive methods and those based on simplified models by finding the right balance between the accuracy of the former and the speed of the latter.

6. Acknowledgments

NFF acknowledges support from the Research Councils United Kingdom (RCUK) under the Academic Fellowship scheme. JS acknowledges support from the Leeds Institute of Molecular Medicine (PhD scholarship). JR is supported by a postdoctoral grant awarded by the Consejeria de Educacion y Cultura of the Junta de Comunidades de Castilla La Mancha anb by the European Social Fund. NFF also thanks Dr Gendra for critical reading and insightful comments to the manuscript, and Ms Martina and Ms Daniela G Fernandez for continuing inspiration and motivation. Publication costs were funded by The Biomedical and Health Research Centre (BHRC).

7. References

Airola, A., S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter &T. Salakoski (2008). "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning." *BMC Bioinformatics* 9 Suppl 11: S2.

Almlof, M., J. Aqvist, A. O. Smalas &B. O. Brandsdal (2006). "Probing the effect of point mutations at protein-protein interfaces with free energy calculations." *Biophys J* 90(2): 433-42.

Aloy, P., B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano &R. B. Russell (2004). "Structure-based assembly of protein complexes in yeast." *Science* 303(5666): 2026-9.

Aloy, P. &R. B. Russell (2003). "InterPreTS: protein Interaction Prediction through Tertiary Structure." *Bioinformatics.* 19(1): 161.

Altschuh, D., A. M. Lesk, A. C. Bloomer &A. Klug (1987). "Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus." *Journal of Molecular Biology* 193(4): 693-707.

Aranda, B., P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk &H. Hermjakob (2010). "The IntAct molecular interaction database in 2010." *Nucleic Acids Res* 38(Database issue): D525-31.

Ashkenazy, H., E. Erez, E. Martz, T. Pupko &N. Ben-Tal (2010). "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." *Nucleic Acids Res* 38(Web Server issue): W529-33.

Assi, S. A., T. Tanaka, T. H. Rabbitts &N. Fernandez-Fuentes (2010). "PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces." *Nucleic Acids Res* 38(6): e86.

Barber, C. B., D. P. Dobkin &H. Huhdanpaa (1996). "The Quickhull algorithm for convex hulls." *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE* 22(4): 469-483.

Barker, D. &M. Pagel (2005). "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes." *PLoS Comput Biol* 1(1): e3.

Benedix, A., C. M. Becker, B. L. de Groot, A. Caflisch &R. A. Bockmann (2009). "Predicting free energy changes using structural ensembles." *Nat Methods* 6(1): 3-4.

Blaschke, C., R. Hoffmann, J. C. Oliveros &A. Valencia (2001). "Extracting information automatically from biological literature." *Comp Funct Genomics* 2(5): 310-3.

Bogan, A. A. &K. S. Thorn (1998). "Anatomy of hot spots in protein interfaces." *J Mol Biol* 280(1): 1-9.

Bork, P., L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee &E. M. Marcotte (2004). "Protein interaction networks from yeast to human." *Curr Opin Struct Biol* 14(3): 292-9.

Bradford, J. R., C. J. Needham, A. J. Bulpitt &D. R. Westhead (2006). "Insights into protein-protein interfaces using a Bayesian network prediction method." *J Mol Biol* 362(2): 365-86.

Bradford, J. R. &D. R. Westhead (2005). "Improved prediction of protein-protein binding sites using a support vector machines approach." *Bioinformatics* 21(8): 1487-94.

Brown, K. R. &I. Jurisica (2007). "Unequal evolutionary conservation of human protein interactions in interologous networks." *Genome Biol* 8(5): R95.

Burger, L. &E. van Nimwegen (2010). "Disentangling direct from indirect co-evolution of residues in protein alignments." *PLoS computational biology* 6(1): e1000633.

Caffrey, D. R., S. Somaroo, J. D. Hughes, J. Mintseris &E. S. Huang (2004). "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Sci* 13(1): 190-202.

Chang, D. T., Y. Z. Weng, J. H. Lin, M. J. Hwang &Y. J. Oyang (2006). "Protemot: prediction of protein binding sites with automatically extracted geometrical templates." *Nucleic Acids Res* 34(Web Server issue): W303-9.

Chao, J. A., Y. Patskovsky, S. C. Almo &R. H. Singer (2008). "Structural basis for the coevolution of a viral RNA-protein complex." *Nature Structural and Molecular Biology* 15(1): 103-105.

Chatr-aryamontri, A., A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli &G. Cesareni (2007). "MINT: the Molecular INTeraction database." *Nucleic Acids Res* 35(Database issue): D572-4.

Chen, H. &H.-X. Zhou (2005). "Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data." *Proteins* 61(1): 21-35.

Chen, P. &J. Li (2010). "Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information." *BMC Bioinformatics* 11: 402.

Chen, X.-w. &J. C. Jeong (2009). "Sequence-based prediction of protein interaction sites with an integrative method." *Bioinformatics* 25(5): 585-591.

Chen, X. W. &M. Liu (2005). "Prediction of protein-protein interactions using random decision forest framework." *Bioinformatics* 21(24): 4394-400.

Cho, K. I., D. Kim &D. Lee (2009). "A feature-based approach to modeling protein-protein interaction hot spots." *Nucleic Acids Res* 37(8): 2672-87.

Clackson, T. &J. A. Wells (1995). "A hot spot of binding energy in a hormone-receptor interface." *Science* 267(5196): 383-386.

Clarke, N. D. (1995). "Covariation of residues in the homeodomain sequence family." *Protein Science* 4(11): 2269-2278.

Cole, C. &J. Warwicker (2002). "Side-chain conformational entropy at protein-protein interfaces." *Protein Sci* 11(12): 2860-70.

Darnell, S. J., L. LeGault &J. C. Mitchell (2008). "KFC Server: interactive forecasting of protein interaction hot spots." *Nucleic Acids Res* 36(Web Server issue): W265-9.

Darnell, S. J., D. Page &J. C. Mitchell (2007). "An automated decision-tree approach to predicting protein interaction hot spots." *Proteins* 68(4): 813-23.

de Vries, S. J., A. D. J. van Dijk &A. M. J. J. Bonvin (2006). "WHISCY: what information does surface conservation yield? Application to data-driven docking." *Proteins* 63(3): 479-89.

DeLano, W. L. (2002). "Unraveling hot spots in binding interfaces: progress and challenges." *Curr Opin Struct Biol* 12(1): 14-20.

Diller, D. J., C. Humblet, X. Zhang &L. M. Westerhoff (2010). "Computational alanine scanning with linear scaling semiempirical quantum mechanical methods." *Proteins* 78(10): 2329-37.

Dou, T., C. Ji, S. Gu, J. Xu, K. Ying, Y. Xie &Y. Mao (2006). "Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species." *Frontiers in bioscience : a journal and virtual library* 11: 380-8.

Engelen, S., L. A. Trojan, S. Sacquin-Mora, R. Lavery &A. Carbone (2009). "Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling." *PLoS Comput Biol* 5(1): e1000267.

Enright, A. J., I. Iliopoulos, N. C. Kyrpides &C. A. Ouzounis (1999). "Protein interaction maps for complete genomes based on gene fusion events." *Nature* 402(6757): 86-90.

Evguenieva-Hackenberg, E., P. Walter, E. Hochleitner, F. Lottspeich &G. Klug (2003). "An exosome-like complex in Sulfolobus solfataricus." *EMBO Rep* 4(9): 889-93.

Ewing, R. M., P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore,

S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J. P. Lambert, H. S. Duewel, Stewart, II, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S. L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou &D. Figeys (2007). "Large-scale mapping of human protein-protein interactions by mass spectrometry." *Mol Syst Biol* 3: 89.

Fares, M. A. &D. McNally (2006). "CAPS: Coevolution analysis using protein sequences." *Bioinformatics* 22(22): 2821-2822.

Fariselli, P., F. Pazos, A. Valencia &R. Casadio (2002). "Prediction of protein--protein interaction sites in heterocomplexes with neural networks." *Eur J Biochem* 269(5): 1356-61.

Fernandez-Fuentes, N., B. K. Rai, C. J. Madrid-Aliste, J. E. Fajardo &A. Fiser (2007). "Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments." *Bioinformatics* 23(19): 2558-65.

Finn, R. D., M. Marshall &A. Bateman (2005). "iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions." *Bioinformatics* 21(3): 410-2.

Finn, R. D., J. Mistry, B. Schuster-B√∂ckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer &A. Bateman (2006). "Pfam: clans, web tools and services." *Nucleic Acids Res* 34(Database issue): D247-51.

Fleishman, S. J., S. D. Khare, N. Koga &D. Baker (2011). "Restricted sidechain plasticity in the structures of native proteins and complexes." *Protein Sci* 20(4): 753-7.

Fu, W., B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt &R. G. Ptak (2009). "Human immunodeficiency virus type 1, human protein interaction database at NCBI." *Nucleic Acids Res* 37(Database issue): D417-22.

Fundel, K., R. Kuffner &R. Zimmer (2007). "RelEx--relation extraction using dependency parse trees." *Bioinformatics* 23(3): 365-71.

Gao, H., Y. Dou, J. Yang &J. Wang (2011). "New methods to measure residues coevolution in proteins." *BMC Bioinformatics* 12: 206.

Glaser, F., D. M. Steinberg, I. A. Vakser &N. Ben Tal (2001). "Residue frequencies and pairing preferences at protein-protein interfaces." *Proteins* 43(2): 89.

Gloor, G. B., L. C. Martin, L. M. Wahl &S. D. Dunn (2005). "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions." *Biochemistry* 44(19): 7156-7165.

Gobel, U., C. Sander, R. Schneider &A. Valencia (1994). "Correlated mutations and residue contacts in proteins." *Proteins: Structure, Function and Genetics* 18(4): 309-317.

Goh, C.-S. &F. E. Cohen (2002). "Co-evolutionary analysis reveals insights into protein-protein interactions." *Journal of Molecular Biology* 324(1): 177-192.

Goll, J., S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb &P. Uetz (2008). "MPIDB: the microbial protein interaction database." *Bioinformatics* 24(15): 1743-4.

Grishin, N. V. &M. A. Phillips (1994). "The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences." *Protein Sci* 3(12): 2455-8.

Grosdidier, S. &J. Fernandez-Recio (2008). "Identification of hot-spot residues in protein-protein interactions by computational docking." *BMC Bioinformatics* 9: 447.

Guerois, R., J. E. Nielsen &L. Serrano (2002). "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." *J Mol Biol* 320(2): 369-87.

Guharoy, M. &P. Chakrabarti (2005). "Conservation and relative importance of residues across protein-protein interfaces." *Proc Natl Acad Sci U S A* 102(43): 15447-52.

Hakes, L., S. C. Lovell, S. G. Oliver &D. L. Robertson (2007). "Specificity in protein interactions and its relationship with sequence diversity and coevolution." *Proceedings of the National Academy of Sciences of the United States of America* 104(19): 7999-8004.

Hoffmann, R. &A. Valencia (2004). "A gene network for navigating the literature." *Nat Genet* 36(7): 664.

Hsu, C. M., C. Y. Chen, B. J. Liu, C. C. Huang, M. H. Laio, C. C. Lin &T. L. Wu (2007). "Identification of hot regions in protein-protein interactions by sequential pattern mining." *BMC Bioinformatics* 8 Suppl 5: S8.

Hu, Z., B. Ma, H. Wolfson &R. Nussinov (2000). "Conservation of polar residues as hot spots at protein interfaces." *Proteins* 39(4): 331-42.

Huang, T. W., C. Y. Lin &C. Y. Kao (2007). "Reconstruction of human protein interolog network using evolutionary conserved network." *BMC Bioinformatics* 8: 152.

Huang, Y. J., D. Hang, L. J. Lu, L. Tong, M. B. Gerstein &G. T. Montelione (2008). "Targeting the human cancer pathway protein interaction network by structural genomics." *Mol Cell Proteomics* 7(10): 2048-60.

Hubner, N. C., A. W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman &M. Mann (2010). "Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions." *J Cell Biol* 189(4): 739-54.

Hue, M., M. Riffle, J. P. Vert &W. S. Noble (2010). "Large-scale prediction of protein-protein interactions from structures." *BMC Bioinformatics* 11: 144.

Itzhaki, Z., E. Akiva, Y. Altuvia &H. Margalit (2006). "Evolutionary conservation of domain-domain interactions." *Genome Biol* 7(12): R125.

Jones, S. &J. M. Thornton (1997). "Prediction of protein-protein interaction sites using patch analysis." *J.Mol.Biol.* 272(1): 133.

Jonsson, P. F. &P. A. Bates (2006). "Global topological features of cancer proteins in the human interactome." *Bioinformatics* 22(18): 2291-7.

Jothi, R., P. F. Cherukuri, A. Tasneem &T. M. Przytycka (2006). "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions." *J Mol Biol* 362(4): 861-75.

Juan, D., F. Pazos &A. Valencia (2008). "High-confidence prediction of global interactomes based on genome-wide coevolutionary networks." *Proc Natl Acad Sci U S A* 105(3): 934-9.

Kastritis, P. L. &A. M. Bonvin (2010). "Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark." *J Proteome Res* 9(5): 2216-25.

Kawashima, S., P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama &M. Kanehisa (2008). "AAindex: amino acid index database, progress report 2008." *Nucleic Acids Res* 36(Database issue): D202-5.

Kelley, R. &T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." *Nat Biotechnol* 23(5): 561-6.

Keskin, O., B. Ma &R. Nussinov (2005). "Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues." *J Mol Biol* 345(5): 1281-94.

Kiemer, L. &G. Cesareni (2007). "Comparative interactomics: comparing apples and pears?" *Trends Biotechnol* 25(10): 448-54.

Kim, S., J. Yoon &J. Yang (2008). "Kernel approaches for genic interaction extraction." *Bioinformatics* 24(1): 118-26.

Kim, W. K., J. Park &J. K. Suh (2002). "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair." *Genome Inform* 13: 42-50.

Koike, A. &T. Takagi (2004). "Prediction of protein-protein interaction sites using support vector machines." *Protein Engineering Design and Selection* 17(2): 165-173.

Koonin, E. V., Y. I. Wolf &L. Aravind (2001). "Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach." *Genome Res* 11(2): 240-52.

Kortemme, T. &D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." *Proc Natl Acad Sci U S A* 99(22): 14116-21.

Kruger, D. M. &H. Gohlke (2010). "DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions." *Nucleic Acids Res* 38(Web Server issue): W480-6.

Kundrotas, P. J., Z. Zhu &I. A. Vakser (2010). "GWIDD: Genome-wide protein docking database." *Nucleic Acids Res* 38(Database issue): D513-7.

Labedan, B., Y. Xu, D. G. Naumoff &N. Glansdorff (2004). "Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase." *Molecular biology and evolution* 21(2): 364-73.

Lafont, V., M. Schaefer, R. H. Stote, D. Altschuh &A. Dejaegere (2007). "Protein-protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies "efficient amino acids"." *Proteins* 67(2): 418-34.

Landgraf, R., I. Xenarios &D. Eisenberg (2001). "Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins." *J.Mol.Biol.* 307(5): 1487.

Landon, M. R., D. R. Lancia, Jr., J. Yu, S. C. Thiel &S. Vajda (2007). "Identification of hot spots within druggable binding regions by computational solvent mapping of proteins." *J Med Chem* 50(6): 1231-40.

Lee, H., M. Deng, F. Sun &T. Chen (2006). "An integrated approach to the prediction of domain-domain interactions." *BMC Bioinformatics* 7: 269.

Lewis, A. C. F., R. Saeed &C. M. Deane (2010). "Predicting protein-protein interactions in the context of protein evolution." *Molecular BioSystems* 6: 55-64.

Li, L., B. Zhao, Z. Cui, J. Gan, M. K. Sakharkar &P. Kangueane (2006). "Identification of hot spot residues at protein-protein interface." *Bioinformation* 1(4): 121-6.

Li, X., O. Keskin, B. Ma, R. Nussinov &J. Liang (2004). "Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking." *J Mol Biol* 344(3): 781-95.

Li, Z., L. Wong &J. Li (2011). "DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts." *BMC Syst Biol* 5 Suppl 1: S5.

Liang, S., C. Zhang, S. Liu &Y. Zhou (2006). "Protein binding site prediction using an empirical scoring function." *Nucleic Acids Res* 34(13): 3698-707.

Lichtarge, O., H. R. Bourne &F. E. Cohen (1996). "An evolutionary trace method defines binding surfaces common to protein families." *J.Mol.Biol.* 257(2): 342.

Lise, S., C. Archambeau, M. Pontil &D. T. Jones (2009). "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods." *BMC Bioinformatics* 10: 365.

Lo Conte, L., C. Chothia &J. Janin (1999). "The atomic structure of protein-protein recognition sites." *J.Mol.Biol.* 285(5): 2177.

Lovell, S. C. &D. L. Robertson (2010). "An integrated view of molecular coevolution in protein-protein interactions." *Molecular Biology and Evolution* 27(11): 2567-2575.

Lu, L., H. Lu &J. Skolnick (2002). "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading." *Proteins* 49(3): 350-64.

Luo, Q., P. Pagel, B. Vilne &D. Frishman (2011). "DIMA 3.0: Domain Interaction Map." *Nucleic Acids Res* 39(Database issue): D724-9.

Ma, B. &R. Nussinov (2007). "Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design." *Curr Top Med Chem* 7(10): 999-1005.

Madaoui, H. &R. Guerois (2008). "Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking." *Proceedings of the National Academy of Sciences of the United States of America* 105(22): 7708-7713.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates &D. Eisenberg (1999). "Detecting protein function and protein-protein interactions from genome sequences." *Science* 285(5428): 751-3.

Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent &M. Vidal (2001). "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." *Genome Res* 11(12): 2120-6.

McPartland, J. M., R. W. Norris &C. W. Kilpatrick (2007). "Coevolution between cannabinoid receptors and endocannabinoid ligands." *Gene* 397(1-2): 126-35.

Mika, S. &B. Rost (2006). "Protein-protein interactions more conserved within species than across species." *PLoS Comput Biol* 2(7): e79.

Mintseris, J. &Z. Weng (2005). "Structure, function, and evolution of transient and obligate protein-protein interactions." *Proceedings of the National Academy of Sciences of the United States of America* 102(31): 10930-10935.

Miwa, M., R. Saetre, Y. Miyao &J. Tsujii (2009). "Protein-protein interaction extraction by leveraging multiple kernels and parsers." *Int J Med Inform* 78(12): e39-46.

Moreira, I. S., P. A. Fernandes &M. J. Ramos (2007). "Computational alanine scanning mutagenesis--an improved methodological approach." *J Comput Chem* 28(3): 644-54.

Murakami, Y. &S. Jones (2006). "SHARP2: protein-protein interaction predictions using patch analysis." *Bioinformatics* 22(14): 1794-5.

Murakami, Y. &K. Mizuguchi (2010). "Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites." *Bioinformatics* 26(15): 1841-8.

Murali, T., S. Pacifico, J. Yu, S. Guest, G. G. Roberts, 3rd &R. L. Finley, Jr. (2011). "DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila." *Nucleic Acids Res* 39(Database issue): D736-43.

Negi, S. S., C. H. Schein, N. Oezguen, T. D. Power &W. Braun (2007). "InterProSurf: a web server for predicting interacting sites on protein surfaces." *Bioinformatics* 23(24): 3397-9.

Neuvirth, H., R. Raz &G. Schreiber (2004). "ProMate: a structure based prediction program to identify the location of protein-protein binding sites." *J Mol Biol* 338(1): 181-99.

Ochoa, D. &F. Pazos (2010). "Studying the co-evolution of protein families with the Mirrortree web server." *Bioinformatics* 26(10): 1370-1.

Ofran, Y. &B. Rost (2003). "Predicted protein-protein interaction sites from local sequence information." *FEBS Lett* 544(1-3): 236-9.

Ofran, Y. &B. Rost (2007). "ISIS: interaction sites identified from sequence." *Bioinformatics* 23(2): e13-6.

Ofran, Y. &B. Rost (2007). "Protein-protein interaction hotspots carved into sequences." *PLoS Comput Biol* 3(7): e119.

Pagel, P., P. Wong &D. Frishman (2004). "A domain interaction map based on phylogenetic profiling." *J Mol Biol* 344(5): 1331-46.

Panni, S., L. Dente &G. Cesareni (2002). "In vitro evolution of recognition specificity mediated by SH3 domains reveals target recognition rules." *J Biol Chem* 277(24): 21666-74.

Park, D., R. Singh, M. Baym, C. S. Liao &B. Berger (2011). "IsoBase: a database of functionally related proteins across PPI networks." *Nucleic Acids Res* 39(Database issue): D295-300.

Patil, A., K. Nakai &H. Nakamura (2011). "HitPredict: a database of quality assessed protein-protein interactions in nine species." *Nucleic Acids Res* 39(Database issue): D744-9.

Pazos, F. &A. Valencia (2008). "Protein co-evolution, co-adaptation and interactions." *Embo J* 27(20): 2648-55.

Pei, J. &N. V. Grishin (2001). "AL2CO: calculation of positional conservation in a protein sequence alignment." *Bioinformatics* 17(8): 700-12.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg &T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proc.Natl.Acad.Sci.U.S.A* 96(8): 4285.

Pesquita, C., D. Faria, A. O. Falcao, P. Lord &F. M. Couto (2009). "Semantic similarity in biomedical ontologies." *PLoS Comput Biol* 5(7): e1000443.

Porollo, A. &J. Ç. Meller (2007). "Prediction-based fingerprints of protein-protein interactions." *Proteins* 66(3): 630-45.

Prasad, T. S., K. Kandasamy &A. Pandey (2009). "Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology." *Methods Mol Biol* 577: 67-79.

Presser, A., M. B. Elowitz, M. Kellis &R. Kishony (2008). "The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication." *Proceedings of the National Academy of Sciences of the United States of America* 105(3): 950-954.

Qin, S. &H. X. Zhou (2007). "meta-PPISP: a meta web server for protein-protein interaction site prediction." *Bioinformatics* 23(24): 3386-7.

Res, I., I. Mihalek &O. Lichtarge (2005). "An evolution based classifier for prediction of protein interfaces without using protein structures." *Bioinformatics* 21(10): 2496-2501.

Riley, R., C. Lee, C. Sabatti &D. Eisenberg (2005). "Inferring protein domain interactions from databases of interacting proteins." *Genome Biol* 6(10): R89.

Rodionov, A., A. Bezginov, J. Rose &E. R. Tillier (2011). "A new, fast algorithm for detecting protein coevolution using maximum compatible cliques." *Algorithms for molecular biology : AMB* 6: 17.

Rual, J. F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth &M. Vidal (2005). "Towards a proteome-scale map of the human protein-protein interaction network." *Nature* 437(7062): 1173-8.

Salwinski, L., C. Miller, A. Smith, F. Pettit, J. Bowie &D. Eisenberg (2004). "The Database of Interacting Proteins: 2004 update." *Nucleic Acids Res* 32: D449 - D451.

Segura, J. &N. Fernandez-Fuentes (2011). "PCRPi-DB: a database of computationally annotated hot spots in protein interfaces." *Nucleic Acids Res* 39(Database issue): D755-60.

Segura, J., P. F. Jones &N. Fernandez-Fuentes (2011). "Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams." *BMC Bioinformatics* 12: 352.

Shoemaker, B. &A. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS Comput Biol* 3(4): 595 - 601.

Shoemaker, B. A. &A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS Comput Biol* 3(4): e43.

Shoemaker, B. A., D. Zhang, R. R. Thangudu, M. Tyagi, J. H. Fong, A. Marchler-Bauer, S. H. Bryant, T. Madej &A. R. Panchenko (2010). "Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites." *Nucleic Acids Res* 38(Database issue): D518-24.

Sikic, M., S. Tomic &K. Vlahovicek (2009). "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests." *PLoS Comput Biol* 5(1): e1000278.

Stark, C., B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski &M. Tyers (2011). "The BioGRID Interaction Database: 2011 update." *Nucleic Acids Res* 39(Database issue): D698-704.

Stein, A., A. Ceol &P. Aloy (2011). "3did: identification and classification of domain-based interactions of known three-dimensional structure." *Nucleic Acids Res* 39(Database issue): D718-23.

Stein, A., R. Mosca &P. Aloy (2011). "Three-dimensional modeling of protein interactions and complexes is going -omics." *Current Opinion in Structural Biology* 21(2): 200-208.

Stelzl, U. &E. E. Wanker (2006). "The value of high quality protein-protein interaction networks for systems biology." *Curr Opin Chem Biol* 10(6): 551-8.

Stelzl, U., U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach &E. E. Wanker (2005). "A human protein-protein interaction network: a resource for annotating the proteome." *Cell* 122(6): 957-68.

Stumpf, M., T. Thorne, E. de Silva, R. Stewart, H. An, M. Lappe &C. Wiuf (2008). "Estimating the size of the human interactome." *Proceedings of the National Academy of Sciences of the United States of America* 105(19): 6959 - 6964.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang &E. Huala (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." *Nucleic Acids Res* 36(Database issue): D1009-14.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen &C. von Mering (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." *Nucleic Acids Res* 39(Database issue): D561-8.

Ta, H. X., P. Koskinen &L. Holm (2011). "A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees." *Bioinformatics* 27(5): 700-6.

Tanaka, T., R. L. Williams &T. H. Rabbitts (2007). "Tumour prevention by a single antibody domain targeting the interaction of signal transduction proteins with RAS." *EMBO J* 26(13): 3250-9.

Tikk, D., P. Thomas, P. Palaga, J. Hakenberg &U. Leser (2010). "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature." *PLoS Comput Biol* 6: e1000837.

Tillier, E. R. M. &R. L. Charlebois (2009). "The human protein coevolution network." *Genome Research* 19(10): 1861 -1871.

Travers, S. A. A. &M. A. Fares (2007). "Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses." *Molecular Biology and Evolution* 24(4): 1032-1044.

Tuncbag, N., A. Gursoy &O. Keskin (2009). "Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy." *Bioinformatics* 25(12): 1513-20.

Tuncbag, N., A. Gursoy, R. Nussinov &O. Keskin (2011). "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM." *Nat. Protocols* 6(9): 1341-1354.

Venkatesan, K., J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi &M. Vidal (2009). "An empirical framework for binary interactome mapping." *Nat Methods* 6(1): 83-90.

Vlahovicek, K., A. Pintar, L. Parthasarathi, O. Carugo &S. Pongor (2005). "CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures." *Nucleic Acids Res* 33(Web Server issue): W252-4.

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg &M. Vidal (2000). "Protein interaction mapping in C. elegans using proteins involved in vulval development." *Science* 287(5450): 116-22.

Wang, B., P. Chen, D.-S. Huang, J.-j. Li, T.-M. Lok &M. R. Lyu (2006). "Predicting protein interaction sites from residue spatial sequence profile and evolution rate." *FEBS Lett* 580(2): 380-4.

Wass, M. N., G. Fuentes, C. Pons, F. Pazos &A. Valencia (2011). "Towards the prediction of protein interaction partners using physical docking." *Mol Syst Biol* 7: 469.

Williams, S. G. &S. C. Lovell (2009). "The effect of sequence evolution on protein structural divergence." *Molecular Biology and Evolution* 26(5): 1055-1065.

Wu, J., S. Kasif &C. DeLisi (2003). "Identification of functional links between genes using phylogenetic profiles." *Bioinformatics* 19(12): 1524-30.

Xia, J. F., X. M. Zhao, J. Song &D. S. Huang (2010). "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility." *BMC Bioinformatics* 11: 174.

Yanai, I., A. Derti &C. DeLisi (2001). "Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes." *Proc Natl Acad Sci U S A* 98(14): 7940-5.

Yang, Z., Y. Lin, J. Wu, N. Tang, H. Lin &Y. Li (2011). "Ranking support vector machine for multiple kernels output combination in protein-protein interaction extraction from biomedical literature." *Proteomics* 11(19): 3811-7.

Yellaboina, S., A. Tasneem, D. V. Zaykin, B. Raghavachari &R. Jothi (2011). "DOMINE: a comprehensive collection of known and predicted domain-domain interactions." *Nucleic Acids Res* 39(Database issue): D730-5.

Yogurtcu, O. N., S. B. Erdemli, R. Nussinov, M. Turkay &O. Keskin (2008). "Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations." *Biophys J* 94(9): 3475-85.

Yu, C.-Y., L.-C. Chou &D. Chang (2010). "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins." *BMC Bioinformatics* 11(1): 167.

Yuan, Z., J. Zhao &Z.-X. Wang (2003). "Flexibility analysis of enzyme active sites by crystallographic temperature factors." *Protein Eng* 16(2): 109-14.

Zhang, Y., H. Lin, Z. Yang &Y. Li (2011). "Neighborhood hash graph kernel for protein-protein interaction extraction." *J Biomed Inform*.

Zhou, H. X. &Y. Shan (2001). "Prediction of protein interaction sites from sequence profile and residue neighbor list." *Proteins* 44(3): 336-43.

20

Protein-Protein Interaction Networks: Structures, Evolution, and Application to Drug Design

Takeshi Hase[1,2] and Yoshihito Niimura[3]
[1]*Department of Bioinformatics, Graduate School of Biomedical Science, Tokyo Medical and Dental University, Tokyo,*
[2]*The Systems Biology Institute, Tokyo,*
[3]*Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan*

1. Introduction

Since proteins exert their functions through interaction with other proteins rather than in isolation, networks of protein interactions are inevitable for understanding protein functions, disease mechanisms, and discovering novel targets of therapeutic drugs (Hase et al. 2009, Barabasi et al. 2011, Vidal et al. 2011). With the recent influx of genome-wide data of protein interactions, many researchers have studied on the structures and statistics of protein-protein interaction networks (PINs). To discover novel drug target genes, it is informative to understand topological and statistical characteristics of PINs, and how disease and drug target genes are distributed over the networks. Moreover, because those statistical properties of PINs are the results of long-term evolution, analysis of the PIN architecture from the viewpoint of comparative genomics and molecular evolution is of particular importance.

In this chapter, we will first summarize our current knowledge of the statistical properties of PINs. We then argue on possible evolutionary mechanisms generating those properties and review the studies related to drug discovery and diseases as an application of the analyses of PIN structure. Finally, we briefly discuss the possibilities of medical studies as an integration of network and evolutionary biology.

2. Genome-wide data of protein-protein interactions

Genome-wide protein-protein interaction data have been obtained from several organisms, including *Escherichia coli* (Arifuzzaman et al. 2006), *Saccharomyces cerevisiae* (Uetz et al. 2000, Ito et al. 2001, Guldener et al. 2006, Reguly et al. 2006, Yu et al. 2008), *Plasmodium falciparum* (LaCount et al. 2005), *Arabidopsis thaliana* (Arabidopsis Interactome Mapping Consortium 2011), *Caenorhabditis elegance* (Li et al. 2004, Simonis et al. 2009), *Drosophila melanogaster* (Giot et al. 2003), and *Homo sapiens* (Rual et al. 2005, Stelzl et al. 2005). Table 1 summarizes the PIN

datasets that are currently available. These data were mainly obtained by high-throughput experimental techniques such as yeast two-hybrid (Y2H) screens and tandem affinity purification followed by mass spectrometry (APMS) screens (Deane et al. 2002, Parrish et al. 2006, Lavallee-Adam et al. 2011), as well as extensive literature curation by experts.

Species	Number of proteins	Number of interactions	Data type	References
Mycoplasma pneumoniae	410	1,058	APMS	Kuhner et al. (2009)
MRSA 252	608	13,219	APMS	Cherkasov et al. (2011)
Treponema pallidum	726	3,649	Y2H	Titz et al. (2008)
Mesorhizobium loti	1,804	3,121	Y2H	Shimoda et al. (2008)
Escherichia coli	2,448	8,625	APMS	Arifuzzaman et al. (2006)
Campylobacter jejuni	1,301	11,557	Y2H	Parrish et al. (2007)
Yeast	1,647	2,518	Y2H	Yu et al. (2008)
	3,224	11,291	Literature curated	Reguly et al. (2006)
	3,891	7,270	Manually curated	MIPS
	3,278	4,549	Y2H	Ito et al. (2001)
	1,004	957	Y2H	Uetz et al. (2000)
Malaria parasite	1,267	2,726	Y2H	LaCount et al. (2005)
Arabidopsis thaliana	2,661	5,664	Y2H	Arabidopsis Interactome Mapping Consortium (2011)
Worm	2,898	5,240	Y2H	Li et al. (2004)
	2,528	3,864	Y2H	Simonis et al. (2009)
Fly	4,679	4,780	Y2H	Giot et al. (2003)
	2,477	3,546	Y2H	Pacifico et al. (2006)
Human	2,783	6,007	Y2H, Literature curated	Rual et al. (2005)
	1,613	3,101	Y2H	Stelzl et al. (2005)

Table 1. PIN datasets. Y2H, Yeast two-hybrid screens; APMS, tandem affinity purification followed by mass spectrometry screens. "Manually curated" indicates that interactions obtained from high-throughput screens and literatures are manually integrated by experts.

Y2H screens examine an interaction between two proteins, by expressing these genes in yeast nucleus as fusion proteins (Parrish et al. 2006). One protein is fused to a DNA-binding domain of a transcription factor (*e.g.*, Gal4 and LexA), and the other protein is fused to a transcription

activation domain of the transcription factor. When two proteins interact with each other, DNA-binding domain and activation domain are indirectly connected. The activation domain can then interact with the transcription start site of the reporter genes (*e.g.*, LacZ). From the expression of the reporter gene, the interaction between two proteins can be detected. In APMS screens, affinity purification selectively purifies a protein complex that includes a protein of interest (bait protein) (Lavallee-Adam et al. 2011). Then, from the purified complex, mass spectrometry identifies possible interacting partners of the bait protein.

It has been pointed out that genome-wide PIN data identified by high-throughput experiments contains a large number of false positive interactions (Hakes et al. 2008). Y2H screens may detect possible interactions between two proteins that actually reside in different subcellular localizations (Deane et al. 2002). APMS studies identify many false positive interactions caused by inadequate purification (Lavallee-Adam et al. 2011).

Literature-curated PIN datasets are likely to be more reliable, because interactions included in such datasets were obtained from small-scale experiments. However, those data are derived from hypothesis-driven researches focusing on several proteins that are supposed to be biologically important, and thus the datasets can be highly biased (Arabidopsis Interactome Mapping Consortium 2011). Therefore, to study the global structure of PINs, researchers should use several datasets obtained by various methods.

3. Statistical properties of PINs

In PINs, a protein and a physical interaction between two proteins are represented as a node and a link, respectively. A series of studies have revealed that PINs have several interesting properties from the viewpoint of network architecture.

3.1 Scale-freeness

The number of links for a given node is called a degree. The degree distribution $P(k)$, the fraction of nodes with k degrees in a network, has been used to characterize the global structure of a network.

Erdös and Renyi (1960) investigated a random network with N nodes, in which links are attached between each pair of nodes with a uniform probability p. This network contains approximately $pN(N\text{-}1)/2$ randomly placed links. Erdös and Renyi (1960) showed that, in a random network, the distribution $P(k)$ follows the Poisson distribution (Fig 1A, left). Therefore, most nodes have degrees that are nearly equal to the mean degree $<k>$ among all nodes in the network.

On the other hand, the distribution $P(k)$ of various technological, social, and biological networks including PINs is known to follow the power law, *i.e.*, $P(k) \sim k^{-\gamma}$ (Albert et al. 1999; Fig 1A, right). These networks are highly heterogeneous; they have a large number of low-degree nodes and a small but significant number of high-degree nodes that are called hubs. A network following the power law does not have a typical degree characterizing most nodes in the network (*e.g.*, the mean degree $<k>$ in a random network), and thus it is called a "scale-free" network. It was shown that scale-free networks are very robust against random removal of nodes, although selective removal of hubs drastically changes their structures (Jeong et al. 2001, Han et al. 2004).

3.2 Small-worldness

The cluster coefficient of nodes i is defined as $C_i = 2e_i/k_i(k_i\text{-}1)$, where k_i is the degree of node i, and e_i is the number of links among k_i neighbors of node i (Watts & Strogatz 1998) (see Fig. 1B). In other words, e_i is the number of triangles that pass through node i. C_i is equal to one when all neighbors of node i fully interact with one another, while C_i is 0 when there are no links among the neighbors of node i. The mean of the cluster coefficient among all nodes, <C>, reflects the density of triangles ("cliques") within a network.

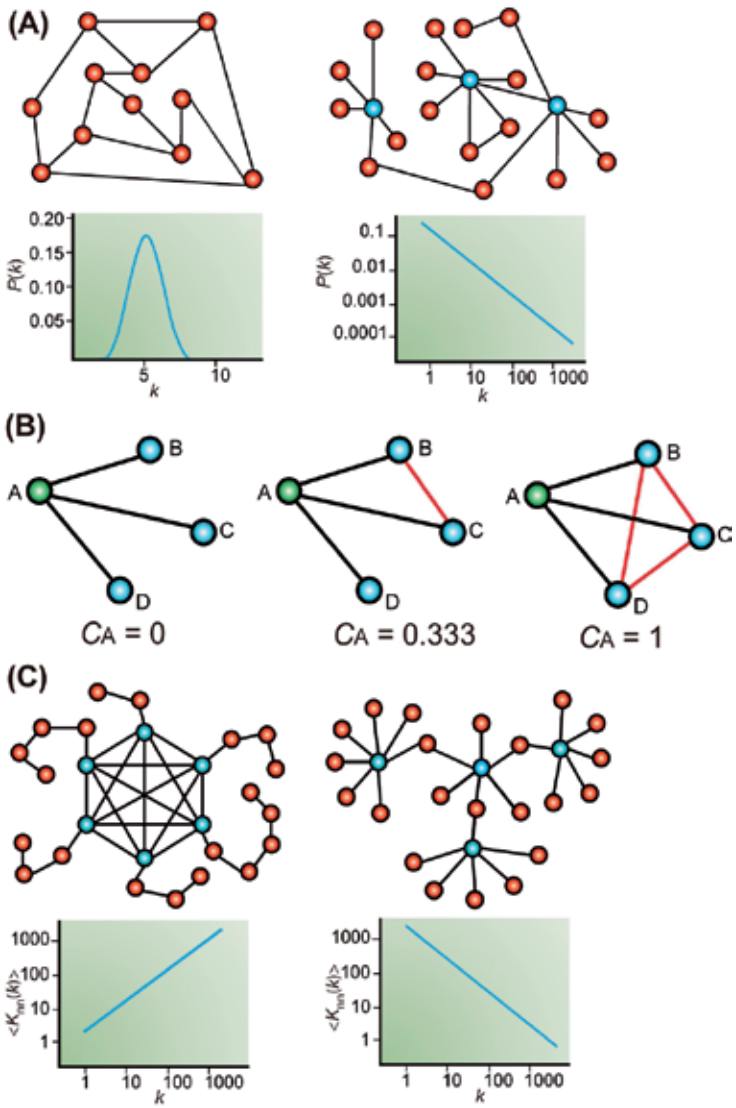


Fig. 1. Measures of a network structure
(A) A random network (left) and a scale-free network (right). The degree distribution $P(k)$ is shown below the networks. **(B)** Cluster coefficient. Red lines represent links among three neighbors of node A. The numbers of links (e_A) among nodes B, C, and D (the neighbors of

node A) in the left, middle, and right networks are 0, 1, and 3, respectively. The cluster coefficient C_A of node A is shown for each network. **(C)** Assortative (left) and disassortative (right) networks. The distribution of $<K_{nn}(k)>$ is shown below the networks. Blue and Red nodes indicates hubs and non-hubs, respectively.

The shortest path length between a pair of nodes is the smallest number of links (distance) that are necessary for travelling from one node to the other (Barabasi & Oltvai 2004). The mean shortest path length among all possible pairs of nodes in a network is denoted by $<L>$. Watts and Strogatz (1998) found that a random network has a much smaller value of $<L>$ compared with a regular lattice. Based on this observation, they defined a "small-world" network as a network that has a value of $<L>$ as small as a random network but is highly clustered like a regular lattice. In a random network, $<L> \sim \log N/\log<k>$, and $<C> = <k>/N$, where N is the number of nodes.

In PINs, the value of $<L>$ is small and the value of $<C>$ is much higher than a random network; therefore, PINs are generally considered to be small-world networks. However, several studies showed that PINs are actually "ultra-small", because $<L>$ is considerably smaller than that in a random network (Chung & Lu 2002, Cohen & Havlin 2003, Hase et al. 2008). In a PIN, proteins are located close to each other, suggesting that perturbations given to a single protein would affect the behaviour of many other proteins and even the entire PIN.

3.3 Assortativity

Another statistic characteristic of a network is the correlation between degrees of nodes that are linked to each other (Callaway et al. 2001, Newman 2002, Costa et al. 2007). Pearson correlation coefficient r of the degrees at both ends of a link is used to evaluate the degree correlation. Networks with $r > 0$ and $r < 0$ are called as assortative and disassortative networks, respectively. In an assortative network, hubs tend to be connected to each other (Fig 1C, left), while in a disassortative network, hubs tend to have links to low-degree nodes (Fig 1C, right).

$<K_{nn}(k)>$, the mean degree among the neighbors of all k-degree nodes ("nn" in $<K_{nn}(k)>$ means "nearest neighbors"), is also used to evaluate the assortativity of a network (Pastor-Satorras et al. 2001, Maslov & Sneppen 2002, Costa et al. 2007, Hase et al. 2008). In an assortative network, $<K_{nn}(k)>$ increases as k increases, while $<K_{nn}(k)>$ in a disassortative network follows decreasing functions of k (Fig 1C). If there are no correlations between degrees of nodes at both ends of a link (*e.g.*, $r = 0$), $<K_{nn}(k)>$ is independent from k and is equal to $<k^2>/<k>$.

It has been shown that the yeast PIN is a disassortative network (Maslov & Sneppen 2002). Therefore, in the yeast PIN, interactions between high- and low-degree nodes are favoured, while those between hubs are suppressed. The biological significance of this structure is unclear. Maslov and Sneppen (2002) proposed that, in the yeast PIN, a hub protein forms a functional module of a cell together with a large number of low-degree neighbors. They then hypothesized that the suppression of links between hubs minimizes unfavourable cross-talks among different functional modules and makes networks robust against perturbations.

If this hypothesis is true, disassortative structure observed in the yeast PIN is under the natural selection, and the disassortativity should be commonly found among PINs in any

organisms. However, by examining PINs from five eukaryote species, Hase et al. (2010) found that the disassortative structure is not a common feature of PINs. The distribution of $<K_{nn}(k)>$ in the PIN can be approximated by $<K_{nn}(k)> \sim k^{-\nu}$, and the value of ν is used to quantify the extent of disassortative structure of a network. Hase et al. (2010) showed that the yeast, worm, fly, and human PINs are disassortative (ν = 0.47, 0.29, 0.35, and 0.26, respectively), while the malaria parasite PIN is not disassortative (ν = 0.02). This observation indicates that the "selectionist view" by Maslov and Sneppen (2002) is not necessary for explaining the disassortative structure of PINs. In section 4, we will see the evolutionary mechanisms generating the difference in assortativity among species.

4. Evolutionary mechanisms generating structures of PINs

To account for the emergence of PIN architecture mentioned above, researchers developed several network growth models and conducted simulation studies using these models. Moreover, statistical properties of PINs were analyzed from the viewpoint of comparative genomics and molecular evolution. In this section, we review evolutionary studies of PINs.

4.1 Preferential attachment and gene duplication

Barabasi and Albert (1999) suggested that the emergence of scale-freeness can be explained by two basic mechanisms: network growth and preferential attachment. The process of network growth adds a new node into a network (red node in Fig 2A). The process of preferential attachment introduces a new link between the new node and each of the other nodes with the probability proportional to the degree of the latter node. For example, the probability that the red node in Fig. 2A gains a new link connected to a blue node is three times higher than that to a black node (Fig 2A). Due to these two processes, a node with a higher degree gains a larger number of links, and thus the degrees of high-degree nodes increase faster than those of low-degree nodes, generating a scale-free network.

In fact, Eisenberg and Levanon (2003) demonstrated that the number of interactions that a protein gained during its evolution is roughly proportional to the degree of the protein by comparing the genomes of *E. coli*, *A. thaliana*, *Schizosaccharomyces pombe*, and *S. cerevisiae*. This observation is consistent with the preferential attachment.

What is the genetic mechanism of network growth and preferential attachment in the evolution of PINs? A plausible mechanism is gene duplication. Let us consider a small PIN containing both high- (node A) and low-degree nodes (node B, C, and D) (Fig 2B, middle). We assume that the number of nodes in a network increases by gene duplication, and a new node has the same interacting partners as the original node. When node B is duplicated, for example, node A acquires a new link and thus the degree of node A increases by one. When node C or node D is duplicated, the same thing happens. On the other hand, if node A is duplicated, each of the degrees of nodes B, C, and D increases by one. Under the assumption that gene duplication occurs randomly with an equal probability for all nodes, the probability that node A acquires a new link is three times higher than the other node does. In general, when we compare a high-degree node (*e.g.*, A) and a low-degree node (*e.g.*, B), a given node (*e.g*, C) is more likely to be a neighbor of a high-degree node than that of a low-degree node. Therefore, a high-degree node gains new links faster than a low-degree node does. For this reason, gene duplication can account for the mechanism of "rich-get-richer".

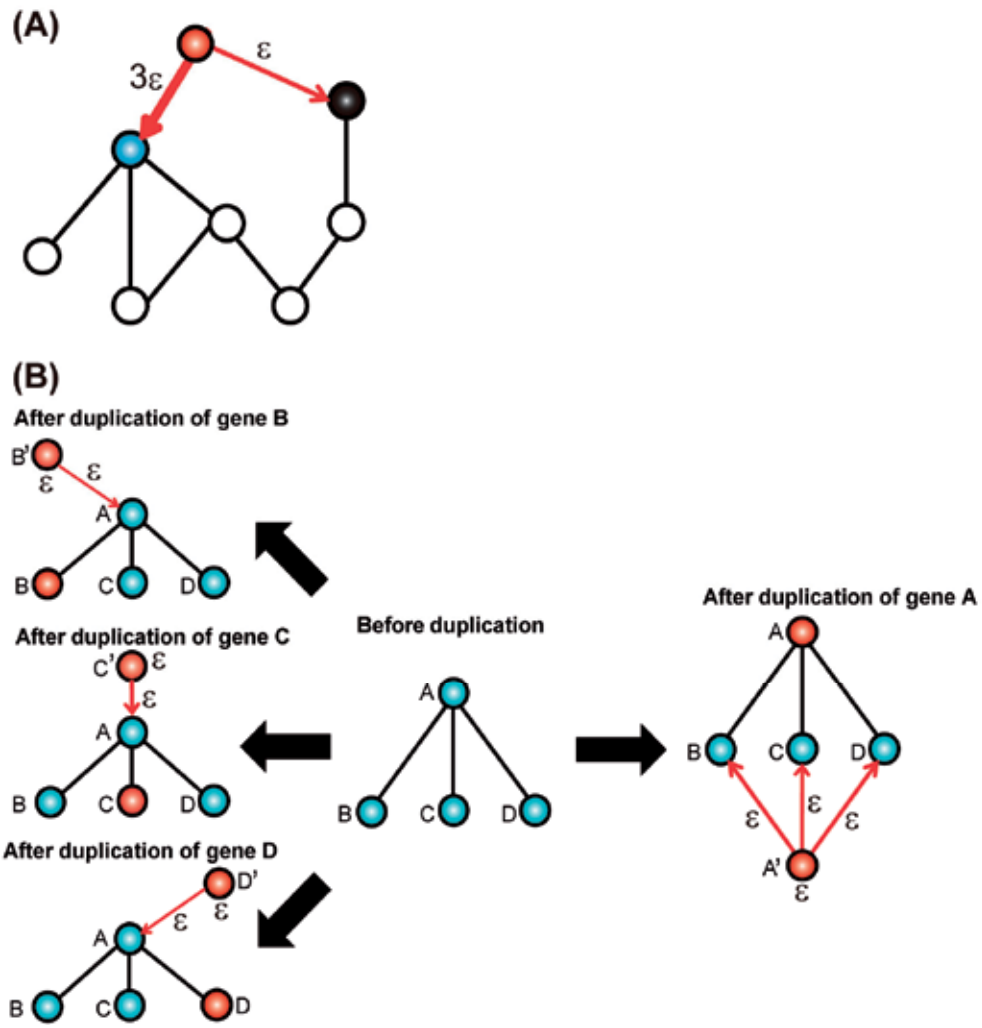


Fig. 2. Network growth by preferential attachment
(A) Preferential attachment. A red node is added to the network. The probability that a new link is attached between red and blue nodes (3ε) is three times higher than that between red and black nodes (ε). **(B)** Network growth with gene duplication. Red nodes represent duplicated nodes. Gene duplication occurs with an equal probability (ε) for all nodes. When node A is duplicated, degrees of nodes B, C, and D increase by one (right), whereas when either node B, C, or D is duplicated, degree of node A increases by one (left).

4.2 Duplication and divergence model

A pair of genes generated by duplication will undergo one of three fates, namely, (i) neofunctionalization, (ii) subfunctionalization, and (iii) nonfunctionalization. After gene duplication, one of the duplicated genes becomes free from selective pressure because of the presence of redundant copies of the gene. Therefore, the gene can tolerate to the accumulation of random mutations and in some cases acquire a novel function (Ohno 1970).

This process is called neofunctionalization. On the other hand, in subfunctionalization process, each of the duplicated genes accumulates mutations, and the functions of the ancestral gene are assigned to the two genes (Force et al. 1999). In nonfunctionalization process, one of the duplicated genes loses its function and becomes a pseudogene due to deleterious mutations. Among the three processes, neofunctionalization and subfunctionalization contribute to the evolution of proteins (Lynch et al. 2000, Blanc et al. 2004, He et al. 2005, Freilich et al. 2006).

In the duplication-divergence model, neofunctionalization and subfunctionalization are modelled as attachment of new links and removal of the links generated by gene duplication, respectively. As for subfunctionalization process, there are two different models, the symmetric divergence and asymmetric divergence. In the former, links are eliminated from both of the duplicated nodes, while in the latter, elimination of links occurs only in one of the two nodes generated by duplication (Fig 3A).

Wagner (2002) reported that one of the duplicated proteins retain a significantly larger number of interactions than the other. For this reason, several network growth models adopted the asymmetric divergence model (Kim et al. 2002, Wagner 2003, Chung et al. 2003, Ispolatov et al. 2005c). However, "complete" asymmetric divergence in which links are eliminated from only one of the duplicates is unrealistic, and the actual divergence process should be intermediate between symmetric and asymmetric divergence (Hase et al. 2010).

Sole et al. (2002) proposed a model on the basis of neofunctionalization and asymmetric divergence. According to their model, after duplication generates a new node, neofunctionalization process attaches a new link between either of the duplicated nodes and each of the other nodes with a uniform probability θ, and then asymmetric divergence eliminates links to only one of the duplicated nodes with a uniform probability α (Fig 3A). Simulation and analytical studies have demonstrated that this model can generate scale-free networks with a small-world property (Sole et al. 2002, Kim et al. 2002, Pastor-Satorras et al. 2003, Chung et al. 2003, Raval 2003).

However, it has been pointed out that some statistical features of PINs could not be regenerated by the model of Sole et al. (2002). The yeast and fly PINs show a much larger $<C>$ than the networks by Sole et al. with the same number of nodes and links as the actual PINs (Sole et al. 2002, Middendorf et al. 2005, Ispolatov et al. 2005a). To overcome this problem, Vazquez et al. (2003) proposed the heterodimerization (HD) model. In their model, symmetric divergence eliminates links from both of the duplicated nodes with a uniform probability α, and the HD process attaches a new link between two duplicated nodes with another uniform probability β, forming a heterodimer (Fig 3A).

When gene duplication occurs for a self-interacting protein, the duplicated proteins will interact to each other. Therefore, β in Vazquez et al. (2003) represents the probability that a randomly selected protein is self-interacting and the new HD link between two duplicated proteins survives after divergence. Simulation and analytical studies have showed that the HD model could reproduce scale-free networks with a similar $<C>$ to the yeast and fly PINs (Vazquez et al. 2003, Middendorf et al. 2005, Ispolatov et al. 2005a). This is because an HD process creates triangles, and a network containing a large number of triangles shows a high value of $<C>$. A computational study based on machine learning technique showed that the

HD model could best reproduce the fly PIN among seven network growth models (Middendorf et al. 2005).

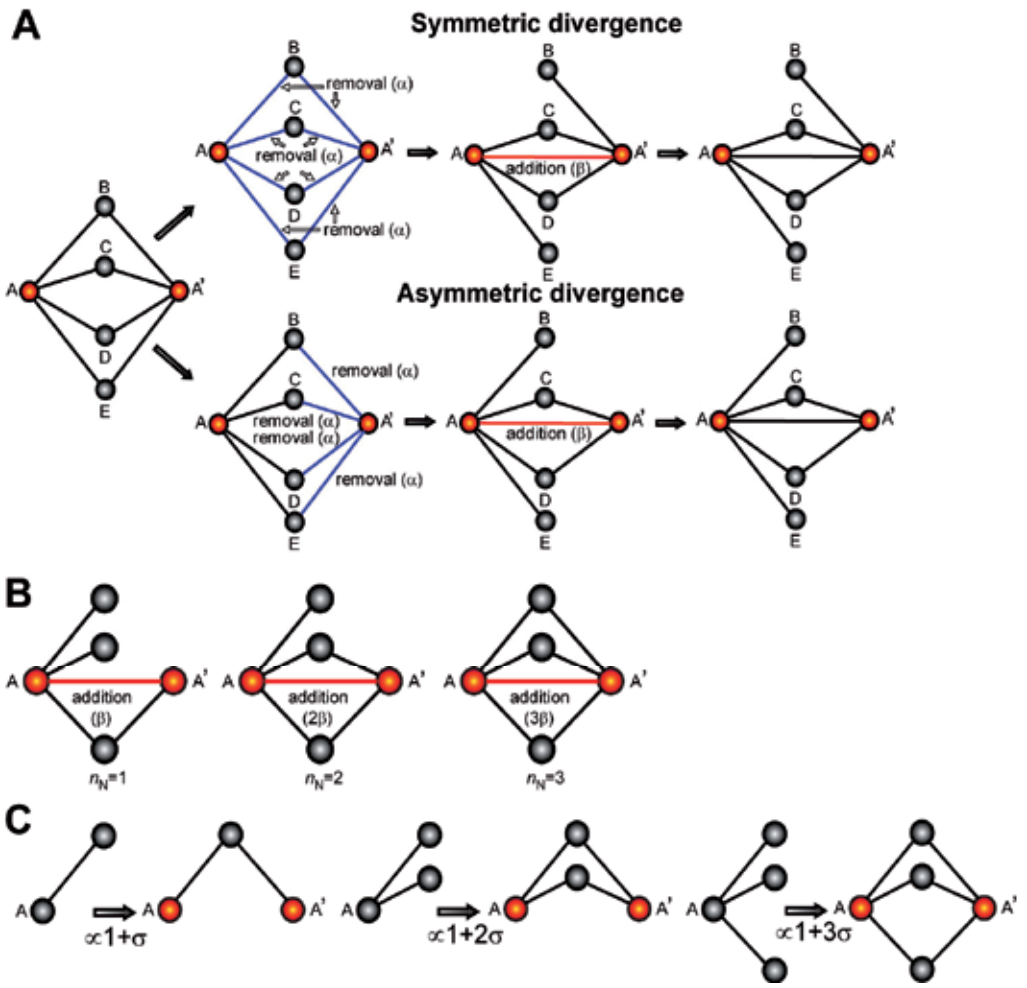


Fig. 3. Network growth models based on gene duplication and divergence
A pair of two red nodes are generated by gene duplication. **(A)** The HD model with asymmetric or symmetric divergence processes. Nodes A and A′ are generated by gene duplication. In the symmetric divergence, each of the links to nodes A and A′ is eliminated with a uniform probability α. On the other hand, in the asymmetric divergence, each of the links to node A′ is eliminated with a uniform probability α. After the divergence process, an HD link (a red line) between two duplicated nodes (nodes A and A′) is attached with a uniform probability β. **(B)** The NHD model. An HD link (red link) is attached between nodes A and A′ with a probability proportional to the number (n_N) of common neighbors shared by these nodes. **(C)** The DDD model. A probability of duplication of a given node is dependent on the degree of the node. If a node has k links, the node is duplicated with the probability proportional to $1 + k\sigma$, where σ is a parameter of the duplicability of a node.

4.3 Non-uniform heterodimerization model

By conducting simulation studies, Hase et al. (2008) showed that, to reproduce the value of <*C*> in the yeast PIN by the HD model, the number of HD links in the networks by the HD model has to be much larger than that in the yeast PIN. Similar observation was made for the fly PIN (Ispolatov et al. 2005a and b). This means that the HD model is insufficient for explaining the evolution of PINs.

As shown in Fig. 3B, when two duplicated nodes share one, two, and three common neighbors, an HD link between them generates one, two, and three new triangles, respectively. The high <*C*> in a PIN indicates that the network contains many triangles. Therefore, if a new HD link is attached more preferentially between duplicated nodes sharing a larger number of common neighbors, the value of <*C*> in a simulation-generated network is expected to become higher. By considering in this way, Hase et al. (2008) proposed the non-uniform heterodimerization (NHD) model in which a new HD link is added between duplicated nodes with a probability proportional to the number of neighbors shared by those nodes (Fig 3B). Simulation studies demonstrated that the NHD model could indeed reproduce both the high value of <*C*> and the small number of HD links in the yeast PIN.

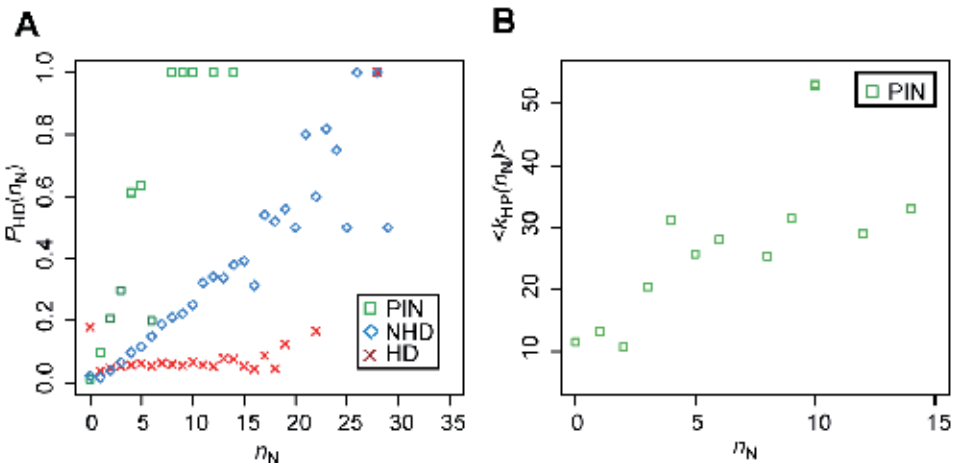


Fig. 4. HD links in the yeast PIN and in the networks by the HD and NHD models (Hase et al. 2008). **(A)** Distribution of $P_{HD}(n_N)$, the probability that an HD link exists between two homologous proteins when they share n_N common neighbors. Green squares, blue diamonds, and red crosses indicate the values for the yeast PIN, the network by the NHD model, and that by the HD model, respectively. **(B)** Distribution of $k_{HP}(n_N)$, the mean degree of proteins that are connected by an HD link and share n_N common neighbors.

In the evolution of PINs, duplication of a self-interacting protein adds an HD link between duplicated proteins. Some HD links were conserved in evolution, while others were eliminated because of occurrence of mutations at interacting sites in these duplicated proteins. In the HD model, the survival rate of HD links is uniform; on the other hand, the NHD model assumes it to be proportional to the number of their common neighbors (Fig. 4A). In the yeast PIN, the probability that two homologous node retain an HD link increases

as the number of neighbors shared by the two nodes increases, which is consistent to the NHD model rather than the HD model (Fig. 4A).

A possible explanation for this observation is as follows. It is expected that, when a given pair of proteins share a large number of common neighbors, the degree of these proteins should be high. In fact, in the yeast PIN, when two homologous proteins are connected by a HD link, there is a positive correlation between the number of common neighbors to the homologues and the mean degree of the two proteins (Fig. 4B). Moreover, several studies showed that high-degree proteins tend to show low evolutionary rate in the yeast PIN (Fraser et al. 2002, 2003, Fraser 2005). Therefore, it is suggested that the survival rates of HD links are also positively correlated with the number of common neighbors shared by the two homologous proteins.

4.4 Degree-dependent duplicability and assortativity

Duplication and divergence models including the NHD model can explain various aspects of the architecture of PINs (Pastor-Satrras et al. 2003, Vazquez 2003, Hase et al. 2008). However, these models cannot explain the differences in overall structures of PINs among species. As mentioned in section 3, the yeast, worm, fly, and human PINs are disassortative, while the malaria parasite PIN is non-disassortative.

A possible evolutionary scenario that can explain the difference in assortativity of PINs among different species is as follows (Hase et al. 2010). Let us consider a disassortative network containing low- and high-degree nodes (*e.g.*, A and C, respectively), in which the low- and high-degree nodes are linked to each other (Fig 5A, middle). Duplication of a low-degree node (*e.g.*, node A) causes the value of ν in the disassortative network to be higher, because the degree of its high-degree neighbor increases (Fig 5A, left). On the other hand, duplication of a high-degree node (*e.g.*, node C) makes the degree of its low-degree neighbors higher, and thus the value of ν decreases (Fig 5A, right). For this reason, duplication of low- and high-degree nodes would make the value of ν in a disassortative network larger and smaller, respectively.

Hase et al. (2010) proposed a novel duplication and divergence model named "degree–dependent duplication (DDD) model", in which duplication of nodes occurs depending on their degree (see Fig 3C). Simulation studies based on the DDD model revealed that preferential duplication of low-degree nodes can successfully reproduce the disassortative structure observed in the yeast, worm, and fly PINs, while preferential duplication of high-degree nodes generate non-disassortative networks similar to the malaria parasite PIN (see Fig 5B and 5C). Moreover, Hase et al. (2010) evaluated the dependency of gene duplicability on their degrees by analyzing orthologous relationships of genes extracted from 55 eukaryotic proteomes. The analyses demonstrated that proteins with a lower degree indeed have higher duplicability in disassortative PINs (the yeast, worm, and fly PINs) (Fig 5D), whereas high-degree proteins tend to have high duplicability in non-disassortative PINs (the malaria parasite PIN) (Fig 5E). Therefore, it is suggested that assortativity of a PIN is related with the gene duplicability dependent on the degrees of genes. If this is the case, disassortative structure of PINs is merely a byproduct of preferential duplication of low-degree proteins, and we do not need to assume any adaptive meaning for this structure, as mentioned in section 3.

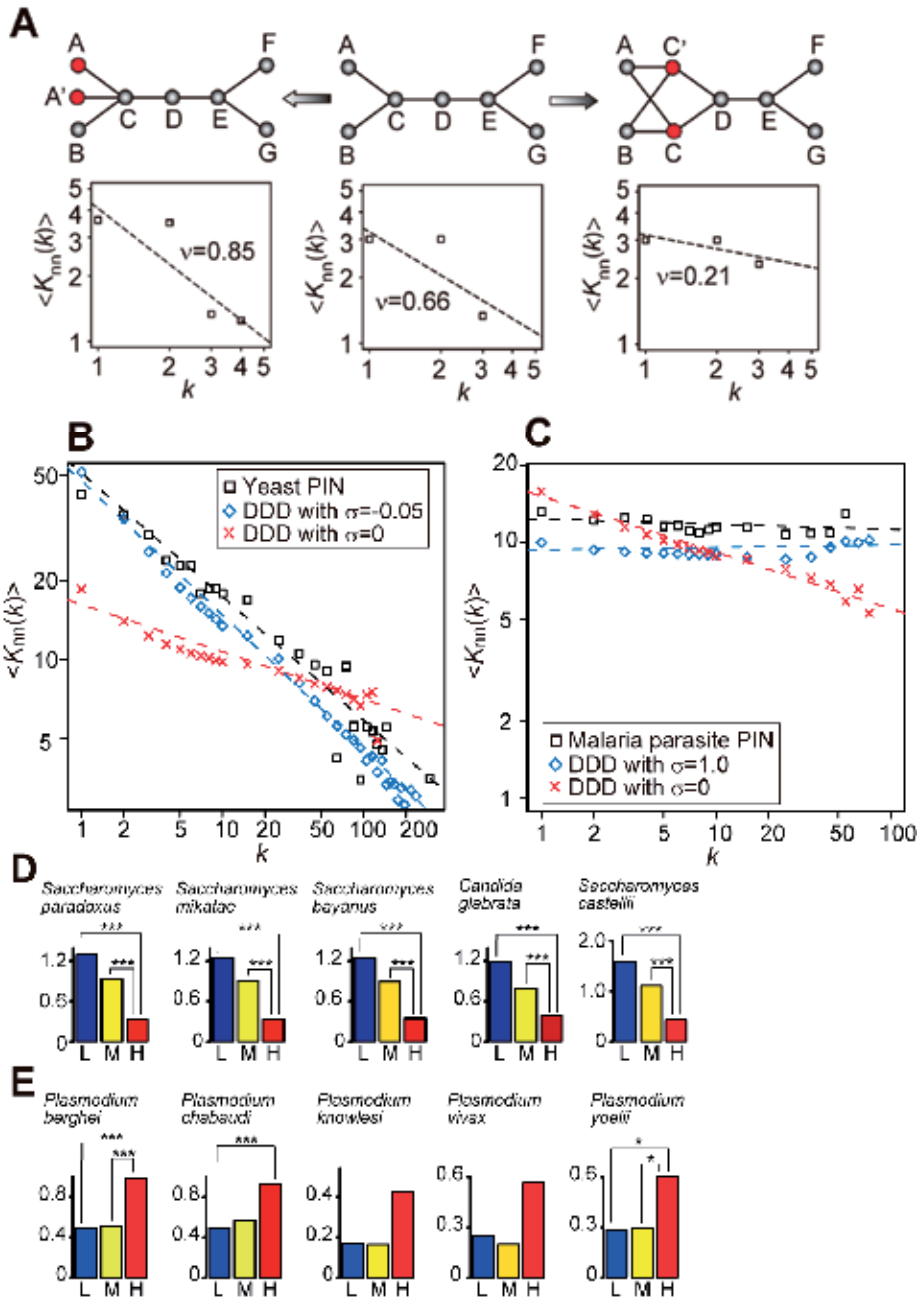


Fig. 5. The DDD model and the extent of assortativity in networks (Hase et al. 2010).
(A) Duplication of a node alters the distribution of $<K_{nn}(k)>$ and the value of ν in a network. A diagram below a network shows the distribution of $<K_{nn}(k)>$ and the value of ν in the network. **(B)** The distribution of $<K_{nn}(k)>$ in the networks generated by the DDD model for yeast. Blue diamonds and red crosses show the results of simulation with σ = -0.05 and 0, respectively (as for σ, see Fig. 3C). Black squares represent $<K_{nn}(k)>$ in the yeast PIN. Dashed

lines in black, blue, and red represent $k^{-0.47}$, $k^{-0.51}$, and $k^{-0.18}$, respectively. **(C)** The distribution of $<K_{nn}(k)>$ in the networks generated by the DDD model for malaria parasite. Blue diamonds and red crosses show the results of simulation with $\sigma = 1.0$ and 0, respectively. Black squares represent $<K_{nn}(k)>$ in the malaria parasite PIN. Dashed lines in black, blue, and red represent $k^{-0.02}$, $k^{0.01}$, and $k^{-0.22}$, respectively. **(D)** and **(E)** indicate correlations between the degree and the duplicability in the yeast and malaria parasite PINs, respectively. Bars in blue, yellow, and red show the mean duplicability among low-, middle-, and high-degree proteins, respectively. A species name above each diagram denotes the species of which genome was compared with *S. cerevisiae* or *P. falciparum*. *, **, and *** represent $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively, by the Wilcoxon rank-sum test with the Bonferroni correction.

5. Structures of PINs and their relationships with disease genes and drug targets

As we have seen above, PINs are characterized by several interesting properties that are different from those of a random network. Therefore, understanding diseases and mechanisms of drug action in the context of PIN architecture may allow us to address some fundamental properties of disease genes and drug target molecules. Indeed, number of disease genes and that of drug targets are very small. Only 10% of the human genes are known to be disease genes (Amberger et al. 2009), and only 435 genes are target genes of therapeutic drugs (Rask-Andersen et al. 2011). Why is the number of drug targets and disease genes so small? Are they distributed randomly over the human PIN? Are there any quantifiable correlations between drug target genes and their statistical properties in the human PIN? To address these questions, drug target and disease genes were mapped onto the human PIN and their statistical properties in the PIN were investigated. Moreover, by using biological networks including the human PIN data, several studies showed that side effects of drugs depend on their statistical features in the network. In this and subsequent sections, we review the application of network analyses to medical researches.

5.1 Statistical properties of disease genes and drug targets in the human PIN

Elimination of a hub protein affects many proteins in a network (Jeong et al. 2001, Yu et al. 2008). Therefore, it was previously hypothesized that genes encoding hub proteins are associated with diseases (Barabasi et al. 2011). Several studies reported that the mean degree among disease genes is in fact significantly higher than that among non-disease genes (Wachi et al. 2005, Jonsson & Bates 2006, Xu & Li 2006).

A human gene is defined to be essential, when knock-out of its orthologous gene causes embryonic and postnatal lethality or sterility in mouse (Liang & Li 2007). Liang & Li (2007) reported that essential genes tend to encode hub proteins in the human PIN.

However, Wachi et al. (2005), Jonsson & Bates (2006), and Xu & Li (2006) took no account for the fact that there are only a small number of disease genes that are also essential (essential disease genes), while vast majority of disease genes are actually non-essential. Because essential disease genes encode hub proteins, the mean degree of disease genes became apparently high in the three studies. In contrast, non-essential disease genes do not show any tendency to encode hub proteins (Goh et al. 2007). Rather, they tend to encode low- and

middle-degree proteins (Feldman et al. 2008). Mutations in high-degree proteins cause dysfunctionality of many neighbor proteins, leading severe impairment of developmental and physiological processes. Individuals having such mutations cannot survive until reproductive years and are likely to be removed from population. For this reason, non-essential disease genes are enriched among low- and middle-degree genes.

Hase et al. (2009) investigated drug target genes to see whether they have specific statistical features in the PIN or not. They found that most drug target genes are middle-degree proteins and some are low-degree, while there are almost no drug targets among high-degree proteins (see Fig 6). The degree distribution is similar to that of disease genes, and, not surprisingly, drug target genes significantly overlap with disease genes (Yao & Rzhetsky 2008). These results indicate that middle-degree proteins are likely to be most advantageous targets for therapeutic drugs.

Oncogenes tend to be high-degree proteins (Jonsson & Bates 2006), and thus they are less likely to be targets for drugs, or one must accept major potential side effects. A possible strategy for designing anti-cancer therapy with less severe side effects is to develop a novel combination of drug compounds that targets several low- and middle-degree proteins, because such combination could generate synergetic effects to cancer cure, and low- or middle-degree targets are expected to induce less severe side effects compared with high-degree targets.

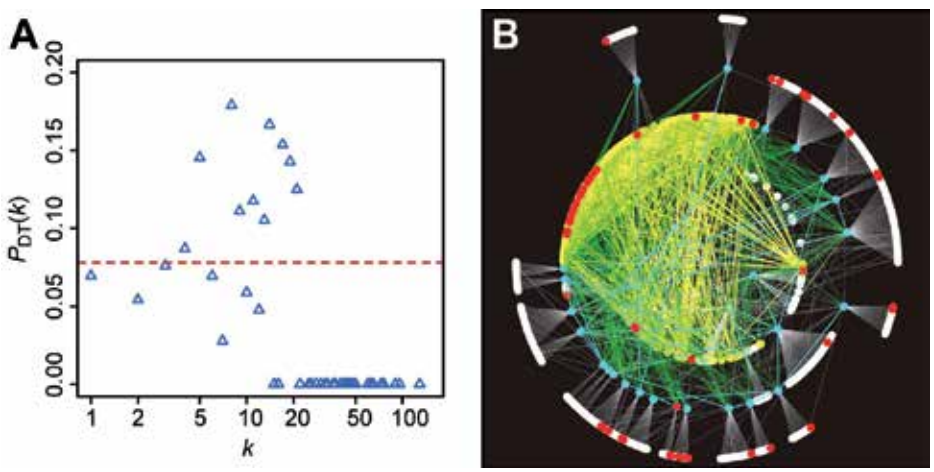


Fig. 6. Degree distribution of drug targets (Hase et al. 2009).
(A) $P_{DT}(k)$ represents the fraction of drug targets to all proteins for the degree of k. The dashed line in red represents the probability that a randomly selected protein is a drug target. **(B)** White, yellow, and blue nodes represent low- (k = 1 – 5), middle- (k = 6 – 30) and high-degree (k > 30) proteins, respectively. Drug targets (red nodes) are mapped on the network. White, yellow, green, and blue links represent interactions between high- and low-degree proteins, those between middle-degree proteins, those between high- and middle-degree proteins, and those between high-degree proteins, respectively. Middle-degree proteins are extensively connected to each other, while links between high-degree proteins are rather suppressed. For clarity, low- and middle-degree proteins that do not have any interactions with high-degree proteins are not shown.

5.2 Predicting candidate drug targets and their side effects based on biological networks

To develop a new drug, it is critical to accurately predict its side effect, because almost 30% of candidate drugs are rejected in clinical stages due to their unexpected toxicity or concerns about drug safety (Kola & Landis 2004, Billingsley 2008). Severe adverse reactions may be found long after the approval of drugs (*e.g.*, Rosiglitazone), and in such cases, those drugs would go out of production (*e.g.*, Rofecoxib) (Moore et al. 2007).

The chemical structures of drugs have been used to predict their adverse side effects and target proteins (Kuhn et al. 2008, Campillos et al. 2008, Yamanishi et al. 2010). Campillos et al. (2008) developed a large-scale database of adverse side effects of drugs. By using the database with information of chemical structure of drugs, they made a similarity metric between two drugs. Under the assumption that drugs with higher similarity in their metric more tend to share the same target proteins, they inferred candidate targets for the drugs.

However, if target proteins of two drugs are close in a molecular network, such drugs may cause similar downstream effects in the network and thus have similar side effects. To understand the molecular mechanisms of drug action and associated adverse effects in greater details, it makes sense to view targets of drugs in the context of biological networks including the genome-wide human interactome (Pache et al. 2008, Zanzoni et al. 2009).

Recently, Brouwers et al. (2011) investigated how side effect similarities of targets depend on their closeness in the human PIN. They found that a certain number of pairs of two drugs without common targets show similar side effects, when they are close in the human PIN. Moreover, Wang et al. (2011) reported that drug side effects are significantly associated with network distances between drug target genes and diseases genes, *i.e.*, targets for failure drugs that make severe adverse side effects are closer to disease genes than targets for approved drugs. Thus, selecting targets that are too close to diseases genes are not always the best strategy (Wang et al. 2011), although the pharmaceutical industry tends to select targets of new drugs that are close with the corresponding disease genes in the biological networks, especially after 1996 (Yildirim et al. 2007).

With recent influx of information of biological networks, especially the human interactome, analyses like Brouwers et al. (2011) or Wang et al. (2011) can be refined and adapted to infer still unknown adverse side effects of drugs and to predict possible target genes. Indeed, by integrating information of the human PIN with similarities between two genes (*e.g.*, GO semantic and sequence similarity) and those between two drugs (*e.g.*, chemical and drug therapeutic similarity), several recent researches attempted to develop a method to predict possible targets for therapeutic drugs (Zhao & Li 2010, Perlman et al. 2011).

6. Possibilities of medical studies with integration of PINs and evolutionary studies

The human PIN is still incomplete and there are many proteins without any information of protein-protein interactions (Venkatesan et al. 2009). Evolutionary information (*e.g.*, evolutionary rate and duplicability) of genes is significantly correlated with their statistical properties in PINs (see sections 2 and 3); therefore, such information can be utilized to complement to the incompleteness of the human PIN.

Rambaldi et al. (2008) reported that most of the cancer genes are singletons and have interactions with many genes. This finding indicates that both gene duplicability and network information are useful for predicting candidate cancer genes. Modification of currently available methods by integrating evolutionary information would improve the accuracy of predicting disease and drug target genes.

Recently, large-scale PINs became available from several prokaryotes, including Methicillin-resistant *Staphylococcus aureus* (MRSA) (Cherkasov et al. 2010), *Treponema pallidum* (Titz et al. 2008), *Campylobacter jejuni* (Parrish et al. 2007), *Mycoplasma pneumonia* (Kuhner et al. 2009), and *Mesorhizobium loti* (Shimoda et al. 2008) (see Table 1). Some of them are pathogenic. By investigating the evolution of their PINs, we may be able to understand the process of acquiring the pathogenicity and developing drug resistance from the viewpoint of network architecture.

Cherkasov et al. (2010) suggested that, in the MRSA PIN, hubs are essential for network stability and may be prospective antimicrobial drug targets. However, almost all known antimicrobial targets have relatively few interactions and hubs have largely been overlooked as drug targets. If hubs in pathogens have no orthologous genes in human and evolve very slowly, by targeting such hubs, we may be able to develop novel antibacterial drugs with high efficacy and small side effects, and without development of resistance to the drugs. With a recent influx of PINs from pathogenic organisms and genomes from various bacterial species, analyses integrating comparative genomics with PINs will become keys to identify still unknown disease mechanisms and novel targets for antibacterial drugs.

7. Conclusion

In this chapter, we describe various aspects of architecture of PINs, such as scale-freeness, small-world properties, and assortativity. Computational studies based on network-growth models and comparative genomics revealed how accumulation of local changes in PINs affects their overall architecture during evolution. We also discussed possible application of PINs and evolutionary studies to medical researches. With expected explosion of OMICs data (*e.g.*, PINs and SNPs from human) in the near future, an integration of networks and genetics will be among the most powerful strategies to elucidate unknown mechanisms of disorders and discover novel targets for efficacious drugs.

8. Acknowledgements

This study was supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, grant 23770271 to YN.

9. References

Albert, R.; Jeong, H. & Barabasi, AL. (1999). Diameter of the World-Wide Web, *Nature*, vol. 401, pp. 130–131.

Amberger, J.; Bocchini, CA.; Scott, AF. & Harmosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM), *Nucleic Acids Research*, vol. 37, pp. D793–D796.

Arabidopsis Interactome Mapping Consortium (2011) Evidence for network evolution in an Arabidopsis interactome map, *Science*, vol. 333, pp. 601–607.

Arifuzzaman, M.; Maeda, M.; Itoh, A.; Nishikata, K.; Takita, C.; Saito, R.; Ara, T.; Nakahigashi, K.; Huang, HC.; Hirai, A.; Tsuzuki, K.; Nakamura, S.; Altaf-Ul-Amin, M.; Oshima, T.; Baba, T.; Yamamoto, N.; Kawamura, T.; Ioka-Nakamichi, T.; Kitagawa, M.; Tomita, M.; Kanaya, S.; Wada, C. & Mori, H. (2007). Large-scale identification of protein-protein interaction of Escherichia coli K-12, *Genome Research*, vol. 16, pp. 686–691.

Barabasi, AL. & Albert, R. (1999). Emergence of scaling in random networks, *Science*, vol. 286, pp. 509–512.

Barabasi, AL. & Oltvai, ZN. (2004). Network biology : understanding the cell's functional organization, *Nature Reviews Genetics*, vol. 5, pp. 101–113.

Barabasi, AL.; Gulbahce, N. & Lascalzo, J. (2011). Network medicine : a network-based approach to human disease, *Nature Reviews Genetics*, vol. 12, pp. 56–68.

Billingsley, ML. (2008). Druggable targets and targeted drugs: enhancing the development of new therapeutics, *Pharmacology* vol. 82, pp. 239–244.

Blanc, G. & Wolfe, KH. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution, *The Plant Cell*, vol. 16, pp. 1679–1691.

Brouwers, L.; Iskar, M.; Zeller, G.; Noort, V. & Bork, P. (2011). Network neighbors of drug targets contribute to drug side effect similarity, *PLoS ONE*, vol. 6, e22187.

Callaway, DS.; Hopcroft, JE.; Kleinberg, JM.; Newman, MEJ. & Strogatz, SH. (2001). Are randomly grown graphs really random?, *Physical Review E*, vol. 64, 041902.

Campillos, M.; Kuhn, M.; Gavin, AC.; Jensen, LJ. & Bork, P. (2008). Drug target identification using side effect similarity, *Science*, vol. 321, pp. 263–266.

Cherkasov, A.; Hsing, M.; Zoraghi, R.; Foster, LJ.; See, RH.; Stoynov, N.; Jiang, J.; Kaur, S.; Lian, T.; Jackson, L.; Gong, H.; Swayze, R.; Amandoron, E.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Santos-Filho, O.; Axerio-Cilies, P.; Byler, K.; McMaster, WR.; Brunham, RC.; Finlay, BB. & Reiner, NE. (2011). Mapping the protein interaction network in Methicillin-resistant Staphylococcus aureus, *Journal of Proteome Research*, vol. 10, pp. 1139–1150.

Chung, F. & Lu, L. (2002). The average distances in random graphs with given expected degrees, *Proceedings of National Academy of Sciences USA*, vol. 99, pp. 15879–15882.

Chung, F.; Lu, L.; Dewey, TG. & Galas DJ. (2003). Duplication models for biological networks, *Journal of computational biology*, vol. 10, pp. 677–687.

Cohen, R. & Havlin, S. (2003). Scale-free networks are ultrasmall, *Physical Review Letters*, vol. 90, 058701.

Costa, LF.; Rodrigues, FA.; Travieso, G. & Boas, RRV. (2007). Characterization of complex networks : A survey of mesurements, *Advances in Physics*, vol. 56, pp. 167–242.

Deane CM.; Salwinski L.; Xenarios I. & Eisenberg D. (2002) Protein interactions, *Molecular & Cellular Proteomics*, vol. 1, pp. 349–356.

Eisenberg, E. & Levanon, EY. (2003). Preferential attachment in the protein network evolution, *Physical Review Letters*, vol. 91, 138701.

Erdös, P. & Renyi, A. (1960). On the evolution of random graphs, *Publication of the Mathematical Institute of the Hungarian Academy of Science*, vol. 5, pp. 17–61.

Feldman, I.; Rzhetsky, A. & Vitkup, D. (2008). Network properties of genes harbouring inherited disease mutations, *Proceedings of National Academy of Sciences USA*, vol. 105, pp. 4323–4328.

Force, A.; Lynch, M.; Pickett, FB.; Amores, A.; Yan, YL. & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations, *Genetics*, vol. 151, pp. 1531–1545.

Fraser, HB.; Hirsh, AE.; Steinmetz, LM.; Scharfe, C. & Feldman, MW. (2002). Evolutionary rate in the protein interaction network, *Science*, vol. 296, pp. 750–752.

Fraser, HB.; Wall, DP. & Hirsh, AE. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions, *BMC Evolutionary Biology*, vol. 3, 11.

Fraser, HB. (2005). Modularity and evolutionary constraint on proteins, *Nature Genetics*, vol. 37, pp. 351–352.

Freilich, S.; Massingham, T.; Blanc, E.; Goldovsky, L. & Thornton, JM. (2006). Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse protein, *Genome Biology*, vol. 7, R89.

Giot, L.; Bader, JS.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, YL.; Ooi, CE.; Godwin, B.; Vitols, E.; Vijayadamodar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, CA.; Finley, RL Jr.; White, KP.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, RA.; McKenna, MP.; Chant, J. & Rothberg, JM. (2003). A protein interaction map of Drosophila melanogaster, *Science*, vol. 302, pp. 1727–1736.

Goh, KI.; Cusick, ME.; Valle, D.; Childs, B.; Vidal, M. & Barabasi, AL. (2007). The human disease network, *Proceedings of National Academy of Sciences USA*, vol. 104, pp. 8685–8690.

Guldener, U.; Munsterkotter, M.; Oesterheld, M.; Pagel, P.; Ruepp, A.; Mewes, HW. & Stumpflen, V. (2006). Mpact: the MIPS protein interaction resource on yeast, *Nucleic Acids Research*, vol. 34, pp. D436–441.

Hakes, L.; Pinney, JW.; Robertson, DL. & Lovell, SC. (2008). Protein-protein interaction networks and biology–what's the connection?, *Nature Biotechnology*, vol. 26, pp. 69–72.

Han, JDJ.; Bertin, N.; Hao, T.; Goldberg, DS.; Berriz, GF.; Zhang, LV.; Dupuy, D.; Walhout, AJM.; Cusick, ME.; Roth, FP. & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, vol. 430, pp. 88–93.

Hase, T.; Niimura, Y.; Kaminuma, T. & Tanaka, H. (2008). Non-uniform survival rate of heterodimerization links in the evolution of the yeast protein-protein interaction network, *PLoS ONE*, vol. 3, e1667.

Hase, T.; Tanaka, H.; Suzuki, Y.; Nakagawa, S. & Kitano, H. (2009). Structure of protein interaction networks and their implications on drug design, *PLoS Computational Biology*, vol. 5, e1000550.

Hase, T.; Niimura, Y. & Tanaka, H. (2010). Difference in gene duplicability may explain the difference in overall structure of protein-protein interaction networks among eukaryotes, *BMC Evolutionary Biology*, vol. 10, 358.

He, X. & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution, *Genetics*, vol. 169, pp. 1157–1164.

Ispolatov, I.; Krapivsky, PL.; Mazo, I. & Yuryev, A. (2005a). Cliques and duplication-divergence network growth, *New Journal of Physics*, vol. 7, 145.

Ispolatov, I.; Yuryev, A.; Mazo, I. & Maslov, S. (2005b). Binding properties and evolution of homodimers in protein-protein interaction networks, *Nucleic Acids Research*, vol. 33, pp. 3629–3635.

Ispolatov, I.; Krapivsky, PL. & Yuryev, R. (2005c). Duplication-divergence model of protein interaction network, *Physical Review E*, vol. 71, 061911.

Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proceeding of National Academy of Sciences USA*, vol. 98, pp. 4569–4574.

Jeong, H.; Mason, SP.; Barabasi, AL. & Oltvai, ZN. (2001). Lethality and centrality in protein interaction networks, *Nature*, vol. 411, pp. 41–42.

Jonsson, PF. & Bates, PA. (2006). Global topological features of cancer proteins in the human interactome, *Bioinformatics*, vol. 22, pp. 2291–2297.

Kim, J.; Krapivsky, PL.; Kahng, B. & Redner, S. (2002). Infinite-order precolation and giant fluctuations in a protein interaction network, *Physical Review E*, vol. 66, 055101.

Kola, I. & Landis, J. (2004). Can the phamaceutical industry reduce attrition rates?, *Nature Reviews Drug Discovery*, vol. 3, pp. 711–715.

Kuhn, M.; Campillos, M.; Gonzalez, P.; Jensen, LJ. & Bork, P. (2008). Large-scale prediction of drug-target relationships, *FEBS Letters*, vol. 582, pp. 1283–1290.

Kuhner, S.; Noort, VV.; Betts, MJ.; Leo-Macias, A.; Batisse, C.; Rode, M.; Yamada, T.; Maier, T.; Bader, S.; Beltran-Alvarez, P.; Castano-Diez, D.; Chen, WH.; Devos, D.; Guell, M.; Norambuena, T.; Racke, I.; Rybin, V.; Schmidt, A.; Yus, E.; Aebersold, R.; Herrmann, R.; Bottcher, B.; Frangakis, AS.; Russell, RB.; Serrano, L.; Bork, P. & Gavin, AC. (2009). Proteome organization in a genome-reduced bacterium, *Science*, vol. 326, pp. 1235–1240.

LaCount, DJ.; Vignali, M.; Chettier, R.; Phansalkar, A.; Bell, R.; Hesselberth, JR.; Schoenfeld, LW.; Ota, I.; Sahasrabudhe, S.; Kurschner, C.; Fields, S. & Hughes, RE. (2005). A protein interaction network of the malaria parasite Plasmodium falciparum, *Nature*, vol. 438, pp. 103–107.

Lavallee-Adam, M.; Cloutier, P.; Coulombe, B. & Blanchette, M. (2011) Modeling contaminants in AP-MS/MS experiments, *Journal of Proteome Research*, vol. 10, pp. 886–895.

Li, S.; Armstrong, CM.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, PO.; Han, JD.; Chesneau, A.; Hao, T.; Goldberg, DS.; Li, N.; Martinez, M.; Rual, JF.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, SL.; Zhang, LV.; Berriz, GF.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, HW.; Elewa, A.; Baumgartner, B.; Rose, DJ.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, SE.; Saxton, WM.; Strome, S.; Heuvel, VDS.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, KC.; Harper, JW.; Cusick, ME.; Roth, FP.; Hill, DE. & Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans, *Science*, vol. 303, pp. 540–543.

Liang, H. & Li, WH. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse, *Trends in Genetics*, vol. 23, pp. 375–378.

Lynch, M. & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization, *Genetics*, vol. 154, pp. 459–473.

Maslov, S. & Sneppen, K. (2002). Specificity and stability in topology of protein networks, *Science*, vol. 296, pp. 910–913.

Middendorf, M.; Ziv, E. & Wiggins, CH. (2005). Inferring network mechanisms : the Drosophila melanogaster protein interaction network, *Proceedings of National Academy of Sciences USA*, vol. 102, pp. 3192–3197.

Moore, TJ.; Cohen, MR. & Furberg, CD. (2007). Serious adverse drug events reported to Food and Drug Administration, 1998-2005, *Archives of Internal Medicine*, vol. 167, pp. 1752–1759.

Newman, MEJ. (2002). Assortative mixing in networks, *Phyical Review Letters*, vol. 89, 208701.

Ohno, S. (1970). *Evolution by gene duplication*, Springer-Verlag, New-York, USA.

Pache, RA.; Zanozoni, A.; Naval, J.; Mas, JM. & Aloy, P. (2008). Towards a molecular characterisation of pathological pathways, *FEBS Letters*, vol. 582, pp. 1259–1265.

Pacifico, S.; Liu, G.; Guest, S.; Parrish, JR.; Fotouhi, F. & Finley, RL. (2006). A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for Drosophila, *BMC Bininformatics*, vol. 7, 195.

Parrish, JR.; Gulyas, KD. & Finley, RL. (2006) . Yeast two-hybrid contributions to interactome mapping, *Current Opinion in Biotechnology*, vol. 17, pp. 387–393.

Parrish, JR.; Yu, J.; Liu, G.; Hines, JA.; Chan, JE.; Mangiola, BA.; Zhang, H.; Pacifico, S.; Fotouhi, F.; DiRita, VJ.; Ideker, T.; Andrew, P. & Finley, RL. (2007). A proteome-wide protein interaction map for Campylobacter jejuni, *Genome Biology*, vol. 8, R130

Pastor-Satorras, R.; Vazquez, A. & Vespignani, A. (2001). Dynamical and correlation properties of the internet, *Physical Review Letters*, vol. 87, 258701.

Pastor-Satorras, R.; Smith, ED. & Sole, RV. (2003). Evolving protein interaction networks through gene duplication, *Journal of Theorerical Biology*, vol. 222, pp. 199–210.

Perlman, L.; Gottlieb, A.; Atias, N.; Ruppin, E. & Sharan, R. (2011). Combining drug and gene simiality measures for drug-target elucidation, *Journal of Computational Biology*, vol. 18, pp. 133–145.

Rambaldi, D.; Giorgi, FM.; Capuani, F.; Ciliberto, A. & Ciccarelli, FD. (2008). Low duplicability and network fragility of cancer genes, *Trends in Genetics*, vol. 24, pp. 427–430.

Rask-Andersen, M.; Almen, MS. & Schioth, HB. (2011). Trends in the exploitation of novel drug targets, *Nature Reviews Drug Discovery*, vol. 10, pp. 579–590.

Raval, A. (2003). Some asymptotic properties of duplication graphs, *Physical Review E*, vol. 68, 066119.

Reguly, T.; Breitkreutz, A.; Boucher, L.; Breitkreutz, BJ.; Hon, GC.; Myers, CL.; Parsons, A.; Friesen, H.; Oughtred, R.; Tong, A.; Stark, C.; Ho, Y.; Botstein, D.; Andrews, B.; Boone, C.; Troyanskya, OG.; Ideker, T.; Dolinski, K.; Batada, NN. & Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae, *Journal of Biology*, vol. 5, 11.

Rual, JF.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, GF.; Gibbons, FD.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem,

M.; Milstein, S.; Rosenberg, J.; Goldberg, DS.; Zhang, LV.; Wong, SL.; Franklin, G.; Li, S.; Albala, JS.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, RS.; Vandenhaute, J.; Zoghbi, HY.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, ME.; Hill, DE.; Roth, FP. & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, vol. 437, pp. 1173–1178.

Shimoda, Y.; Shinpo, S.; Kohara, M.; Nakamura, Y.; Tabata, S. & Sato, S. (2008). A large scale analysis of protein-protein interactions in Nitrogen-fixing Bacterium Mesorhizobium loti, *DNA Research*, vol. 15, pp. 13–23.

Simonis, N.; Rual, JF.; Carvunis, AR.; Tasan, M.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Sahalie, JM.; Venkatesan, K.; Gebreab, F.; Cevik, S.; Klitgord, N.; Fan, C.; Braun, P.; Li, N.; Ayivi-Guedehoussou, N.; Dann, E.; Bertin, N.; Szeto, D.; Dricot, A.; Yildirim, MA.; Lin, C.; Smet, AS.; Kao, HL.; Simon, S.; Smolyar, A.; Ahn, JS.; Tewari, M.; Boxem, M.; Milstein, S.; Yu, H.; Dreze, M.; Vandenhaute, J.; Gunsalus, KC.; Cusick, ME.; Hill, DE.; Tavernier, J.; Roth, FP. & Vidal, M. (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network, *Nature Methods*, vol. 6, pp. 47–54.

Sole, RV.; Pastor-Satorras, R.; Smith, ED. & Kepler, T. (2002). A model of large-scale proteome evolution, *Advances in Complex Systems*, vol. 5, pp. 43–54.

Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, FH.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksöz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H. & Wanker, EE. (2005). A human protein-protein interaction network: a resource for annotating proteome, *Cell*, vol. 122, pp. 957–968.

Titz, B.; Rajagopala, SV.; Goll, J.; Hauser, R.; McKevitt, MT.; Palzkill, T. & Uetz, P. (2008). The binary protein interactome of Treponema pallidum—the syphilis spirochete, *PLoS ONE*, vol. 3, e2292.

Uetz, P.; Giot, L.; Cagney, G.; Mansfield, TA.; Judson, RS.; Knight, JR.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S. & Rothberg, JM. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, vol. 403, pp. 623–627.

Vazquez, A.; Flammini, A.; Maritan, A. & Vespignani, A. (2003). Modeling of protein interaction networks, *Complexus*, vol. 1, pp. 38–44.

Vazquez, A. (2003). Growing networks with local rules : preferential attachment, clustering hierarchy, and degree correlations, *Physical Review E*, vol. 67, 056104.

Venkatesan, K.; Rual, JF.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, KI.; Yildirim, MA.; Simonis, N.; Heinzmann, K.; Gebreab, F.; Sahalie, JM.; Cevik, S.; Simon, C.; Smet, AS.; Dann, E.; Smolyar, A.; Vinayagam, A.; Yu, H.; Szeto, D.; Borick, H.; Dricot, A.; Klitgord, N.; Murray, RR.; Lin, C.; Lalowski, M.; Timm, J.; Rau, K.; Boone, C.; Braun, P.; Cusick, ME.; Roth, FP.; Hill, DE.; Tavernier, J.; Wanker, EE.; Barabasi, AL. & Vidal, M. (2009). An empirical framework for binary interactome mapping, *Nature Methods*, vol. 6, pp. 83–90.

Vidal, M.; Cusick, ME. & Barabasi, AL. (2011). Interactome networks and human disease, *Cell*, vol. 144, pp. 986–998.

Wachi, S.; Yoneda, K. & Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, *Bioinformatics*, vol. 21, pp. 4205–4208.

Wang, J.; Li, ZX.; Qui, CX.; Wang, D. & Cui, QH. (2011). The relationship between rational drug design and drug side effects, *Briefings in Bioinformatics*, (in press).

Wagner, A. (2002). Asymmetric functional divergence of duplicate genes in yeast, *Molecular Biology and Evolution*, vol. 19, pp. 1760–1768.

Wagner, A. (2003). How the global structure of protein interaction networks evolves, *Proceedings of the Royal Society B: Biological Sciences*, vol. 270, pp. 457–466.

Watts, DJ. & Strogatz, SH. (1998). Collective dynamics of 'small-world' networks, *Nature*, vol. 393, pp. 440–442.

Xu, J. & Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network., *Bioinformatics*, vol. 22, pp. 2800–2805.

Yamanishi, Y.; Kotera, M.; Kanehisa, M. & Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework, *Bioinformatics*, vol. 26, pp. i246–254.

Yao, L. & Rzhetsky, A. (2008). Quantitative systems-level determinants of human genes targeted by successful drugs, *Genome Research*, vol. 18, pp. 216–213.

Yildirim, MA.; Goh, KI.; Cusick, ME.; Barabasi, AL. & Vidal, M. (2007). Drug-target network, *Nature Biotechnology*, vol. 25, pp. 1119–1126.

Yu, H.; Braun, P.; Yildirim, MA.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, JF.; Dricot, A.; Vazquez, A.; Murray, RR.; Simon, C.; Tardivo, L.; Tam, S.; Svrzikapa, N.; Fan, C.; de Smet, AS.; Motyl, A.; Hudson, ME.; Park, J.; Xin, X.; Cusick, ME.; Moore, T.; Boone, C.; Snyder, M.; Roth, FP.; Barabási, AL.; Tavernier, J.; Hill, DE. & Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network, *Science*, vol. 322, pp. 104–110.

Zanzoni, A.; Soler-Lopez, M. & Aloy, P. (2009). A network medicine approach to human disease, *FEBS Letters*, vol. 583, pp. 1759–1765.

Zhao, S. & Li, S. (2010). Network-based relating pharmacological and genomic spaces for drug target identification, *PLoS ONE*, vol. 5, e11764.

21

A Survey on Evolutionary Analysis in PPI Networks

Pavol Jancura and Elena Marchiori
Radboud University Nijmegen
The Netherlands

1. Introduction

The analysis and application of the evolutionary information, as measured by means of the conservation of protein sequences, using protein-protein interaction (PPI) networks, has become one of the central research areas in systems biology from the last decade. It provides a promising approach for better understanding the evolution of living systems, for inferring relevant biological information about proteins, and for creating powerful protein interaction and function prediction tools. The aim of this survey is to give a general overview of the relevant literature and advances in the analysis and application of evolution in PPI networks. Due to the broad scope and vast literature on this subject, the present overview will focus on a representative selection of research directions and state-of-the-art methods to be used as a solid knowledge background for guiding the development of new hypothesis and methods aiming at the extraction and exploitation of evolutionary information in PPI networks.

This survey consists of two main parts (see Fig. 1). The first part deals with research works concerning the relation between evolution and the topological structures of a PPI network, in particular trying to discover and assess the evidence of such a relation and its strength at different granularity levels. Specifically, we consider works analysing evolution at the single protein level as well as at the level of a collection of proteins present in a PPI network. The second part of this survey describes works analysing how such evolutionary evidence can be exploited for knowledge discovery, in particular for inferring relevant biological information, such as protein interaction prediction and the discovery of functional modules conserved across multiple species.

The main terms and concepts underlying protein interaction and evolution which are used throughout the survey are summarized in the sequel. In general, a protein-protein interaction can represent different types of relations, such as a true physical bond or a functional interplay between proteins. Here, if not explicitly stated, a PPI represents a physical protein interaction as detected by experimental methods, such as yeast two-hybrid (Y2H) screening, co-immunoprecipitation or tandem affinity purification.

Two proteins are called *homologous* if they share high sequence similarity. There are two main types of homologous proteins: *orthologous* and *paralogous*. Here, for simplicity, we consider a protein pair to be orthologous if the proteins of the pair are from different species. We refer to the proteins of an orthologous pair as orthologs. Analogously, a protein pair is considered to

be paralogous if its proteins belong to the same species, in this case their proteins are called paralogs. A general assumption is that the proteins of an orthologous pair originated from a common ancestor, having been separated in evolutionary time only by a speciation event, while paralogous proteins are the product of gene duplication without speciation. The concept of orthology can be directly extended to more than two species, where one can consider clusters of orthologous proteins containing at least one protein of each species.

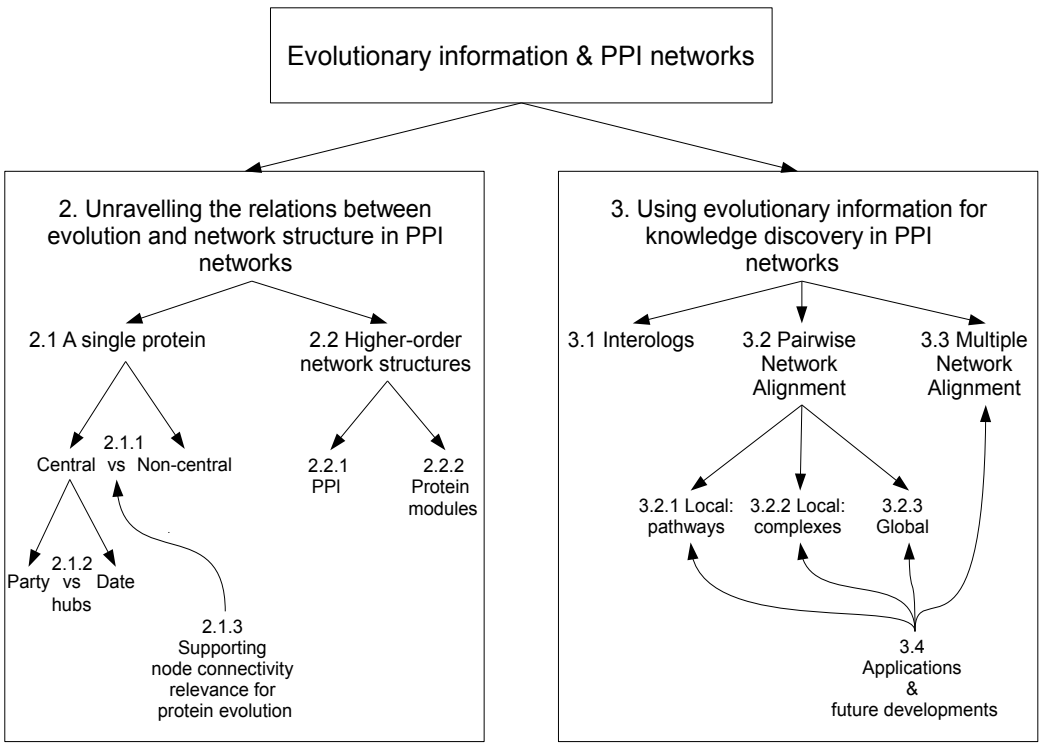


Fig. 1. The structure of the survey.

2. Unravelling the relations between evolution and network structure in PPI networks

We begin with a summary of those studies that involve the analysis of evolutionary information in a single PPI network. One can divide these works into the following two main groups. The first group studies evolutionary conservation with respect to topological properties of a PPI network. The second one primarily investigates the role of evolution with respect to the functional modules present in a PPI network.

The aim of the first group of studies is to describe how the topology of a single PPI network reflects the evolutionary signal present in the proteins it contains. This evolutionary signal is represented by the set of orthologs and it is retrieved with respect to a different species. Specifically, given a PPI network of the species to be investigated and a set of proteins of a

distinct species, those proteins of the network being a part of orthologous pairs or clusters (resulting from a sequence comparison of proteins of the two or multiple species respectively) are considered to be source of the evolutionary or orthology signal in the network. Then, having established the orthology relationship between proteins of the two or multiple species, one can estimate the evolutionary rate or distance of aligned protein sequences (see e.g. Yang & Nielsen, 2000). The higher the rate, the faster is considered the evolution of proteins. Consequently, proteins which evolve slowly are well-conserved and a little or none change to them can be observed throughout the evolution. Other protein evolutionary measures have been also considered, as propensity for gene loss, evolutionary excess retention or protein age (see Table 1).

Type of evolutionary measure	**Evolutionary measure**	**References**
Evolutionary conservation	Evolutionary Rate	e.g. Yang & Nielsen (2000), Wall et al. (2005),
	Propensity for Gene Loss	Krylov et al. (2003)
	Evolutionary Excess Retention	Wuchty (2004)
	Phyletic Retention	Gustafson et al. (2006), Chen & Xu (2005), Fang et al. (2005)
Protein age classification	Time of Origin	Kunin et al. (2004)
	Protein Age Group	Ekman et al. (2006), Kim & Marcotte (2008)

Table 1. Measures of evolutionary signal at protein level

2.1 Relation between a single protein in a PPI network and evolution

Various features of a PPI network topology can be investigated with respect to evolutionary information; the first and simplest ones are measures acting on the single nodes of the network. One can associate with a node different topological measures which estimate the relative relevance of the node within the network, here called *centrality* or *connectivity* of a node.

A basic centrality measure of a node is its degree. The degree of a node is the number of edges containing the node or, in terms of a PPI network, it is the number of proteins with which the protein represented by the node in the network interacts. It has been observed that a protein degree distribution of PPI networks follows a power law and thus PPI networks fall into a class of scale-free networks (see e.g. Jeong et al., 2001). Scale-free networks have a few highly connected nodes, called hubs, and numerous less connected nodes, which mostly interact only with one or two nodes.

2.1.1 Essentiality, centrality and conservation of a protein

As a decade ago large protein physical interaction data were not yet available, researchers mainly focussed on the study of the correlation between importance of a protein function for a living cell (essentiality, dispensability) and its evolutionary conservation rate. The generally accepted premise is that essential genes or proteins should evolve at slower rates

than non-essential ones (see e.g. Kimura, 1983). Although empirical studies have cast doubts on the validity of this hypothesis (see e.g. Hurst & Smith, 1999; Pal et al., 2003; Rocha & Danchin, 2004), in the end the vast majority and late evidences favour the existence of correlation between gene essentiality or dispensability and evolutionary conservation (see e.g. Fang et al., 2005; Fraser et al., 2002; 2003; Hahn & Kern, 2005; Hirsh & Fraser, 2001; 2003; Jordan et al., 2002; Krylov et al., 2003; Ulitsky & Shamir, 2007; Wall et al., 2005; Wang & Zhang, 2009; Waterhouse et al., 2011; Zhang & He, 2005). In particular, as recently stated by Wang & Zhang (2009), the correlation remains weak yet still conveniently sufficient for practical use.

After the growth of protein interaction data, also the correlation between essentiality and centrality, and evolutionary conservation and centrality started to be investigated. At first the *centrality-essentiality relationship* was mostly investigated by examining the degree of a node, proving the existence of the correlation (see e.g. Fraser et al., 2002; 2003; Hahn & Kern, 2005; Jeong et al., 2001; Krylov et al., 2003). However Coulomb et al. (2005) showed no correlation between essentiality and centrality, where centrality was assessed not only by the degree but also by higher order centrality measures, namely average neighbours' degree of a node and clustering coefficient of a node, suggesting that the correlation centrality-essentiality could be an artefact of the dataset. These findings were later supported by Gandhi et al. (2006) who considered a set of PPI networks and also did not observe any significant relationship between a node degree and the essentiality of the corresponding protein. Interestingly, Coulomb et al. (2005) did not test other centrality measures as betweenness and closeness, which showed a higher correlation with essentiality than just the simple degree (Hahn & Kern, 2005). Nevertheless, Batada, Hurst & Tyers (2006) reaffirmed the existence of the correlation between the node degree and essentiality taking into account Coulomb et al.'s concerns. However, Yu et al. (2008) again disputed the correlation using the compilation of Yeast high quality PPI data. Results contradicting this work appeared in two consecutive studies by Park & Kim (2009) and Pang, Sheng & Ma (2010). The first study (Park & Kim, 2009) considered also other centrality measures than just the degree of a node. As a result, the correlation could be successfully revealed, whereas the highest correlation was observed with measures based on betweenness and closeness, similarly to Hahn & Kern (2005). In the other study (Pang, Sheng & Ma, 2010) the newer, updated yeast PPI dataset was used and the correlation between degree of a node and its (protein) essentiality could be detected.

Although, the above works support that there is a connection between topological position of a node and functional importance, it seems one cannot explain this centrality-lethality rule just by the degree distribution (He & Zhang, 2006; Zotenko et al., 2008). This seems to be in accordance with the analysis conducted in (Lin et al., 2007) showing that protein domain complexity is not the single determinant of protein essentiality and that there is a correlation between the number of protein domains and the number of interactions (Schuster-Bockler & Bateman, 2007). In addition, Kafri et al. (2008) showed that highly connected essential proteins tend to have duplicates which can compensate their deletion thus decreasing the deleterious effect of their removal, a phenomenon that could possibly explain the findings that genes with no duplicates are more likely to be essential (Giaever et al., 2002). Therefore higher order topological features appear to be more appropriate for capturing gene essentiality, especially those based on node-betweenness and node-closeness (Hahn & Kern, 2005; Park & Kim, 2009; Yu et al., 2007), which are believed to estimate better the local connectivity or centrality of a

node within the network. Moreover, these features also relate with gene expression (Krylov et al., 2003; Pang, Sheng & Ma, 2010; Yu et al., 2007).

We consider now works that analyse the correlation between evolution and centrality. Also in this case the two main features used to estimate this correlation are the degree of a node and the evolutionary rate. At first, it was hypothesized that proteins with a higher degree should evolve slower (Fraser et al., 2002). A main criticism to this hypothesis was based on the fact that the analysis conducted in (Fraser et al., 2002) did not take into account the presence of a possible bias and of noise in data obtained from high-throughput experiments (Bloom & Adami, 2003; Jordan et al., 2003a;b). Nevertheless Fraser et al. (2003), Fraser & Hirsh (2004) and Lemos et al. (2005) could confirm the existence of such correlation by taking into account these objections. Kim et al. (2007) also confirmed interconnection between centrality, essentiality and conservation and showed that peripheral proteins of the PPI network are under positive selection for species adaptation. Moreover, the link between the connectivity of a node and its evolutionary history was further substantiated by works studying the correlation between node degree and other evolutionary measures such as propensity for gene loss (Krylov et al., 2003), evolutionary excess retention (Wuchty, 2004) and protein age (Ekman et al., 2006; Kunin et al., 2004). However Batada, Hurst & Tyers (2006) again pointed to a lack of evidence for a significant correlation between the evolutionary rate and the connectivity of a node. Moreover, Makino & Gojobori (2006) classified proteins according to two criteria, clustering coefficient of a node and protein's multi-functionality, and showed that multi-functional proteins of sparse parts of yeast PPI network (with a low clustering coefficient) evolve at the slowest rate regardless of the degrees of the connectivity. This suggests that clustering coefficient is a better descriptor of protein evolution within the global network of protein interactions.

A possible explanation for these conflicting results was proposed by Saeed & Deane (2006) who showed that the strength and significance of the correlation between evolution and centrality varies depending upon the type of PPI data used. Also Saeed & Deane (2006) found that more accurate datasets demonstrate stronger correlations between connectivity and evolutionary rate than less accurate datasets. Another reason may be the existence of two distinct types of highly connected nodes, so-called *party* and *date hubs*, which appear to satisfy different evolutionary constraints.

2.1.2 Evolution of party and date hubs

Specifically, Han et al. (2004) observed a bimodal distribution of average Pearson correlation coefficients between the expression profiles of proteins and its interacting partners. This yielded a classification of hubs into party hubs, having similar co-expression profiles with their neighbours, and date hubs, having different co-expression profiles with their neighbours. As a consequence, party hubs tend to interact simultaneously ("permanently") with their partners and to connect proteins within functional modules while date hubs tend to interact with different partners at different time/space ("transiently") and to bridge different modules. Thus, one may also refer to party hubs as *intramodule* and to date hubs as *intermodule* (Fraser, 2005).

Fraser (2005) was the first to investigate the difference in evolution between date and party hubs and found that party hubs are highly evolutionary constrained, whereas date hubs are

more evolutionary labile. This is clearly in accordance with findings of Mintseris & Weng (2005) who argued that residues in the interfaces of permanent protein interactions tend to evolve at a relatively slower rate, allowing them to co-evolve with their interacting partners, in contrast to the plasticity inherent in transient interactions, which leads to an increased rate of substitution for the interface residues and leaves little or no evidence of correlated mutations across the interface. The work of Fraser (2005) was, in addition, later corroborated by Bertin et al. (2007). Examining three dimensional properties of proteins also supported this hypothesis, as multi-interface hubs were found to be more evolutionary conserved and essential as well as more likely to correspond to party hubs (Kim et al., 2006). Defining singlish- and multi-Motif hubs further substantiated these findings, because multi-Motif hubs were found to be more evolutionary conserved, more essential and to correlate with multi-interface hubs (Aragues et al., 2007). In addition, other features as orderness of regions in protein sequences and the solvent accessibility of the amino acid residues was shown to be different between party and date hubs and to contribute in the lowering of the evolutionary rate of party hubs (Kahali et al., 2009). Recently, Mirzarezaee et al. (2010) applied feature selection methods and machine learning techniques to predict party and date hubs based on a set of different biological characteristics including amino acid sequences, domain contents, repeated domains, functional categories, biological processes, cellular compartments, etc.

However, other researchers disputed not only the evolutionary differences between party and date hubs but the existence of hub types as such (Agarwal et al., 2010; Batada, Reguly, Breitkreutz, Boucher, Breitkreutz, Hurst & Tyers, 2006; Batada et al., 2007). Indeed, some datasets do not exhibit clear or robust bimodal distribution of hubs' gene co-expression profiles (Agarwal et al., 2010) and in some cases there is even a complete lack of bimodality (Batada, Reguly, Breitkreutz, Boucher, Breitkreutz, Hurst & Tyers, 2006; Batada et al., 2007). Therefore, Pang, Cheng, Xuan, Sheng & Ma (2010) argue that the average Pearson correlation coefficient is a weak measure of whether a protein acts transiently or permanently with its interacting partners and they propose a new measure, a co-expressed protein-protein interaction degree. This measure estimates the actual number of partners with which a protein can permanently interact. One can interpret it as a degree of 'protein party-ness' and it offers more a continuum-like estimate of the protein's interaction property. This seems to be in accordance with Nooren & Thornton (2003) who suggest that rather a continuum range exists between distinct types of protein interactions and that their stability very much depends on the physiological conditions and environment.

Pang, Cheng, Xuan, Sheng & Ma (2010) firstly corroborated the results of Saeed & Deane (2006) on the correlation variations between connectivity and evolutionary rate of a protein on different datasets and then they showed that the co-expression-dependent node degree correlates significantly with the protein's evolutionary rate irrespectively of the specific dataset used. However, their topological measure is derived by using an external source of experimental data on gene expression. The further investigation on purely topological features of a PPI network which would distinguish transient and permanent interactions, and party and date hubs could bring more insights on how the evolutionary history of a protein is wired in its position within the network of all the protein interactions in an organism. In this perspective, network path-based measures, such as betweenness and closeness, seem to be promising (Yu et al., 2007). All the more, these measures also appear to relate to

protein essentiality (Park & Kim, 2009; Yu et al., 2007) and it could clarify the link between essentiality and evolution as such. Thereafter, they could improve on the prediction of essential genes from the topology of a PPI network in combination with protein evolutionary information, such as phyletic retention (Gustafson et al., 2006), as already corroborated by several application of machine learning techniques for essential gene detection, prioritizing drug targets and determining virulence factors (see e.g. Chen & Xu, 2005; Deng et al., 2011; Doyle et al., 2010; Gustafson et al., 2006; McDermott et al., 2009).

2.1.3 Node connectivity is relevant for protein evolution

Since the factors relevant for protein evolution could be of a multiple character (Wolf et al., 2006), it is interesting to investigate whether protein connectivity plays a central or a more subtle role. In the latter case, the link between protein connectivity and evolution could be the results of spurious correlations due to other underlying biological processes (Bloom & Adami, 2003). In order to address this issue, the contribution of protein connectivity to protein evolutionary conservation has been also studied in an integrated way (Pal et al., 2006) using multidimensional methods such as principal component analysis (PCA) and principal component regression (PCR).

The first successful application of PCA was given by Wolf et al. (2006) on seven genome-related variables. The derived first component reflected a gene's 'importance' and confirmed positive correlation between lethality, expression levels and number of protein-protein interaction which at the same time constrained protein evolution measures. Interestingly, the component also showed that the number of paralogs positively contributes to gene essentiality, which contradicts the finding of Giaever et al. (2002) that non-duplicated genes tend to be essential. However, the study of Drummond et al. (2006) revealed by using PCR only single determinant of protein evolution, namely translational selection, which is almost entirely determined by the gene expression level, protein abundance, and codon bias. Later, Plotkin & Fraser (2007) re-examined the use of PCR method and showed noise in biological data can confound PCRs, leading to spurious conclusions. As a result, when they equalized for different amounts of noise across the predictor variables no single determinant of evolution could be found indicating that a variety of factors-including expression level, gene dispensability, and protein-protein interactions may independently affect evolutionary rates in yeast. This observation was further substantiated by a recent study (Theis et al., 2011) where 16 genomic variables were analysed using Bayesian PCA. The study supports the evidence for the three above-discussed correlations. It also demonstrates how different definitions of paralogs may lead to different conclusions on their effect on essentiality, and thus commenting on Wolf et al.'s conflicting result (Wolf et al., 2006).

2.2 Higher-order structures in a PPI network and evolution

Researchers have also focused on other topological structures of a PPI network than just a node and their relation to evolutionary conservation. With increasing topological complexity we may talk about a single protein-protein interaction (an edge in PPI network), topological motifs, and protein clusters or modules as detected by their interaction density or network traffic.

2.2.1 Evolution and protein-protein interaction

Unlike in the case of a single protein, where various well-established methods for measuring sequence evolution are developed, to the best of our knowledge only a recent attempt has been made in order to estimate the evolutionary rate of protein-protein interaction (Qian et al., 2011). However, this study is limited to a small set of PPIs in yeasts and can not be yet applied for large-scale studies due to the lack of data. Thus, the research has extensively focused on estimating correlated evolution of a protein pair and their functional or physical interaction (Pazos & Valencia, 2008).

It is generally assumed that proteins which co-evolve tend to participate together in a common biological function. This hypothesis is supported by many examples of functionally interacting protein families that co-evolve (see e.g. Galperin & Koonin, 2000; Moyle et al., 1994). Co-evolution of proteins may be assessed at sequence level (*sequence co-evolution*) by correlating evolutionary rates (Clark et al., 2011), or at gene family level (*gene family evolution*) by correlating occurrence vectors (Kensche et al., 2008). An occurrence vector or a phylogenetic profile (phyletic pattern) (Tatusov et al., 1997) is an encoding of protein's (homologue's) presence or absence within a given set of species of interest (Kensche et al., 2008). In general, the methods for correlating protein evolution have been successfully applied to predict a physical or functional interaction between proteins (Clark et al., 2011; Kensche et al., 2008), where sequence co-evolution is powerful in predicting the physical interaction and phylogenetic profiling is a good indicator of functional interplay between proteins in a broader sense. Large-scale co-evolutionary maps have also been constructed and analysed for better understanding the evolution of a species and its link to protein interactions (see e.g. Cordero et al., 2008; Tillier & Charlebois, 2009; Tuller et al., 2009). All these works suggest that the topology of PPIs should reflect the evolutionary processes behind the proteins which formed such network.

The first systematic study of linked genes and their evolutionary rates was done by Williams & Hurst (2000) who showed that the rates of linked genes are more similar than the rates of random pairs of genes. Pazos & Valencia (2001) performed the first successful large-scale prediction of physical PPIs based on sequence co-evolution by correlating phylogenetic trees. Another large-scale study by Kim et al. (2004) on domain structural data of interacting protein families also revealed their high co-evolution but also showed a high diversity in the correlation of rates of each family pair. Specifically, protein families with a greater number of domains were shown to be more likely to co-evolve. However, Hakes et al. (2007) argued that this correlation of evolutionary rates is not responsible for the covariation between functional residues of interacting proteins. Nevertheless, other studies have been able to predict interacting domains from co-evolving residues between domains or proteins (see e.g. Jothi et al., 2006; Yeang & Haussler, 2007) indicating that different organisms use the same 'building blocks' for PPIs and that the functionality of many domain pairs in mediating protein interactions is maintained in evolution (Itzhaki et al., 2006).

Another perspective on co-evolution of interacting partners was given by Mintseris & Weng (2005), who distinguished between transient and obligate interactions. The authors concluded that obligate complexes are likely to co-evolve with their interacting partners, while transient interactions with an increased evolutionary rate show only little evidence for a correlated evolution of the interacting interfaces. This observation was later corroborated by Brown

& Jurisica (2007) who analysed the presence of protein interactions across multiple species via orthology mapping and found that the greater the conservation of a protein interaction is, the higher the enrichment for stable complexes. Beltrao et al. (2009) also observed that stable interactions are more conserved than transient interactions, by studying evolution of interactions involved in phosphoregulation. Finally, Zinman et al. (2011) extracted protein modules from a yeast integrated protein interaction network using various source of PPI evidence, and showed that interactions within modules were much more likely to be conserved than interactions between proteins in different modules.

The preference of conserved protein interactions to be placed in modular parts of a network was also observed by Wuchty et al. (2006) by extending the paradigm of protein's connectivity and its evolutionary conservation to the connectivity of a protein-protein interaction. Specifically, they used the hypergeometric clustering coefficient to estimate the interaction cohesiveness of the PPI's neighbourhood and orthologous excess retention in order to asses the evolutionary conservation of PPIs. They used the same clustering coefficient as that given by the presence of orthologs of interacting proteins in another organism and showed that PPIs with highly clustered environment were accompanied by an elevated propensity for the corresponding proteins to be evolutionary conserved as well as preferably co-expressed (Wuchty et al., 2006). These findings are significant all the more they were shown to be stable under perturbations. This propensity of interacting proteins to be more conserved and prevalent among taxa was later confirmed by Tillier & Charlebois (2009) who used evolutionary distances to estimate the protein's conservation. Yet another perspective on conservation of PPIs was given by Kim & Marcotte (2008) who classified proteins into four groups (from oldest to youngest) according their age and found a unique interaction density pattern between different protein age groups, where the interaction density tends to be dense within the same group and sparse between different age groups.

2.2.2 Evolution and modularity of PPI networks

All the evidences above that PPIs whose proteins are evolutionary correlated tend to form stable complexes and to be embedded in cohesive areas of a network topology support the premise that modularity of PPI networks is maintained by evolutionary pressure (Vespignani, 2003). Indeed, when examining networks solely built from sequence co-evolution, gene context analysis or gene family evolution of completely sequenced genomes, one may observe that these networks exhibit high modularity with clusters corresponding to known functional modules, thus revealing the structure of cellular organization (Cordero et al., 2008; Tuller et al., 2009; von Mering et al., 2003).

Regarding the networks of physically interacting proteins, to the best of our knowledge the first direct evidence that evolution drives the modularity of PPI networks was provided by Wuchty et al. (2003). They looked beyond a single protein pair and studied the more complex patterns of interacting proteins, called topological motifs. In general, they found that, as the number of nodes in a motif and number of links among its constituents increase, a greater and stronger conservation of the proteins could be observed. This was corroborated by Vergassola et al. (2005) who focused on specific instances of motifs known as cliques. Cliques are topological patterns where all protein constituents interact with each other. Vergassola et al. (2005) provided evidence for co-operative co-evolution within cliques of interacting

proteins. Later, Lee et al. (2006) investigated motifs at a higher resolution level, by defining for each motif different motif modes based on functional attributes of interacting proteins: again their findings indicated that motifs modes may very well represent the evolutionary conserved topological units of PPI networks. More recently, Liu et al. (2011) studied network motifs according to the age of their proteins and discovered that the proteins within motifs whose constituents are of the same age class tend to be densely interconnected, to co-evolve and to share the same biological functions. Moreover, these motifs tend to be within protein complexes.

The finding that modularity of PPI networks is constrained by evolution and that conserved interactions are enriched in dense motifs and regions of a PPI network also suggest that protein complexes present in such cohesive areas should be evolutionary driven (Jancura et al., 2012). As putative protein complexes can be extracted from a PPI network by means of clustering techniques, Jancura et al. (2012) detected such protein complexes in the PPI network consisting of only yeast proteins having an ortholog in another organism and compared them with those protein complexes derived either by using the global topology of a yeast PPI network or by using a network induced by randomly selected proteins. The in-depth examination of enriched functions in these three types of protein complexes revealed that evolutionary-driven complexes are functionally well differentiated from other two types of protein complexes found in the same interaction data. As a consequence, new complexes and protein function predictions could be unravelled from PPI data by using a standard clustering approach with the inclusion of evolutionary information. In addition, evolutionary-driven complexes were found to be differentially conserved, in particular some complexes were detected for all distinct set of orthologs as determined by comparison with different species, some exhibited only a subset of proteins identifiable in a complex across all species, and some complexes being found only for one specific set of orthologs. This suggests that presence of evolution in modularity of PPI networks is more versatile and flexible with different degrees of conservation.

The findings of Jancura et al. (2012) seem to conform with related studies that focused on evolutionary cohesiveness of protein functional modules in order to investigate whether a group of proteins which functionally interact, co-evolve more cohesively than a random group of proteins. Either known protein complexes and pathways were analysed (Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004) or putative protein modules usually derived from integrated networks of functional link evidences (Campillos et al., 2006; Zhao et al., 2007; Zinman et al., 2011). A different strategy was employed by Yamada et al. (2006) who at first detected evolutionary modules which were afterwards compared with enzyme connectivity in a metabolic network.

Although the co-evolution of modules is assessed by the presence or absence of modules' constituents across a set of species, there is no standard method to measure the degree to which a module evolves cohesively (Fokkens & Snel, 2009). For instance, Snel & Huynen (2004) used the deviation of the number of modules' orthologs per species from the average number of modules' orthologs per species, whereas Campillos et al. (2006) measured the fraction of joined evolutionary events given the reconstructed, most parsimonious evolutionary scenario of the genes in a module over their phylogenetic profiles.

Despite this measures' diversity, the common conclusion is that the majority of modules evolve flexibly (Campillos et al., 2006; Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004; Yamada et al., 2006). Also, it appears that curated modules evolve more cohesively than modules derived from high throughput interaction data (Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004). Moreover, there is a different enrichment in functions which co-evolve. For example, biochemical pathways, certain metabolic and signalling processes, as well as core functions like transcription and translation, tend to have higher rate of evolutionary cohesiveness (Campillos et al., 2006; Fokkens & Snel, 2009; Zhao et al., 2007). This is also supported by methods which cluster phylogenetic profiles in order to detect biochemical pathways or to predict functional links and thus exploiting the predictive power of phylogenetic methods (Glazko & Mushegian, 2004; Li et al., 2009; Watanabe et al., 2008). These methods show a relatively good performance in characterizing biochemical pathways but seem to have a limited coverage for physically interacting proteins (Watanabe et al., 2008). A dubious result was reported on inter-connectivity of cohesive and flexible modules. Specifically, Fokkens & Snel (2009) demonstrated that components of cohesive modules are less likely to interact with each other than in the case of flexible modules, while two other studies (Campillos et al., 2006; Zinman et al., 2011) suggest cohesive modules to be more highly connected.

It is possible that the above studies underestimated the actual degree of evolutionary cohesiveness present in the modularity of protein interaction networks due to their conservative approach, the limitations in ortholog detection as well as the cohesiveness measures which are restricted to phylogenetic profiles. Nevertheless, they show that, as evolution is a complex process, its presence in modularity of protein interaction networks also exhibits a very complex nature, whose understanding is far from being complete. Evolution itself, indeed, can be expected to be asynchronous and heterotactous along the tree of life.

In general, the interim evidence shows different evolutionary pressure for different types of protein interactions and data. In particular, the slowly evolving interacting partners are enriched in stable, permanent complexes, and functional modules such as biochemical pathways and curated complexes exhibit higher evolutionary cohesiveness than high throughput complexes. It seems that the co-evolutionary degree of modules within PPI networks increases with greater integration of various sources of evidence for proteins to functionally interact (Zinman et al., 2011). Also, not all protein complexes and functional modules need to be co-evolutionary modules (Fokkens & Snel, 2009). There is a continuum from extremely conserved to rapidly changing modules, where those modules found to be co-evolving appear to be enriched in certain, specific functional categories (Campillos et al., 2006). In addition, the degree of conservation and co-evolution of functional modules within interaction networks seem to reflect cellular organization and their spatio-temporal characteristics. For instance, cohesive modules can be classified according to their evolutionary age as ancestral, intermediate and young, where one may observe ancient, ancestral modules to be highly conserved and perform essential, core processes such as information storage and metabolism of amino acids, while young modules are less conserved and responsible for the communication with the environment (Campillos et al., 2006). Therefore one might expect ancestral modules to contain static, obligate interactions as the proteins of essential functions tend to involve multiple domains with slow evolutionary

rates, whereas young modules can be enriched with dynamic, transient interactions with less but fast evolving protein domains to allow adaptation to the environment.

3. Using evolutionary information for knowledge discovery in PPI networks

The tendency of functionally linked or physically interacting proteins and densely interacting motifs to exhibit correlated evolution and/or to be conserved across species is at the core of methods for inferring relevant biological information using PPI networks. Although such biological information can be limited and biased towards specific type of known interactions and protein functions, it allows one to infer new, unknown functions of proteins, to improve the understanding of biological systems, and to guide the discovery of drug-target interaction. In its basic form, the knowledge discovery process is based on the transfer of information involving a single interaction between two organisms, while in its most complex form it involves the identification and transfer of protein complexes across multiple species. In the sequel we summarize concepts and techniques used to achieve these goals, in particular the notions of "interologs" and of multiple PPI networks alignment.

3.1 Predicting protein interaction: Interologs

If two proteins physically interact in one species and they have orthologous counterparts in another species, it is likely that their orthologs interact in that species too. If such conserved interactions exist, they are called *interologs*. This simple method of protein interaction inference was firstly introduced and tested by Walhout et al. (2000) on proteins involved in vulval development of nematode worm, where potential interactions between these proteins were identified based on interactions of their orthologs in other species. Later, Matthews et al. (2001) performed a large-scale analysis of this inference technique using the yeast PPI network as a model and proteins of worm as a target. Although the success rate of detection of inferred interactions by Y2H analysis was between 16%-31%, it represented a 600-1100-fold increase compared to a conventional approach at that time (Matthews et al., 2001).

The interologs-based protein interaction prediction has become one of the standard methods for *in silico* PPI prediction. The method can be easily extended to more PPI data from multiple species. In particular, having two groups of orthologs, where each ortholog group contains proteins from the same N species, and observing an interaction between proteins of these orthologous groups in $(N-1)$ species, the interaction between proteins of the N-th species present in the ortholog groups can be predicted. This multidimensional character of interolog inference has been extensively used to predict and build databases of the whole interactome for various species, either as a stand alone approach or in combination with other *in silico* methods, which often integrate multiple data types including the gene co-expression, co-localization, functional category, the occurrence of orthologs and other genomic context methods. In this way researchers could provide, for instance, the first sketch of human interactome (Lehner & Fraser, 2004), build the interactome of plants (Geisler-Lee et al., 2007; Gu et al., 2011), and improve the understanding of processes in a malarial parasite (Pavithra et al., 2007) or in cancer (Jonsson & Bates, 2006). Also, three, up-to-date, tools have been recently implemented and made available to perform this inference task (Gallone et al., 2011; Michaut et al., 2008; Pedamallu & Posfai, 2010).

Several algorithmic enhancements of the interologs-based approach have been introduced since the first proposal of a systematic use of interolog inference (Matthews et al., 2001). For instance, Yu et al. (2004) have strengthen the definition of ortholog by using a reciprocal best-hit approach and compared it to the original one-way best-hit approach implemented by Matthews et al. (2001). In addition, they required a minimum level for a joint similarity of orthologous sequences in order to perform interolog mapping. Their method yielded a 54% accuracy in contrast to a 30% of the previous method by Matthews et al. (2001).

Other approaches exploited the knowledge on a higher conservation rate of PPIs in dense network motifs. For instance Huang et al. (2007) scored interologs according to the density of the topological pattern containing the respective PPI of the interolog in a model species as determined by the extraction of maximal quasi-cliques from the PPI network of the model species. This score was integrated with scores of other various features used for PPI prediction, such as tissue specificity, sub-cellular localization, interacting domains and cell-cycle stage. The use of multiple types of features was shown to yield more accurate predictions of PPIs in comparison with other interolog-based methods used to build interactome databases. More recently, Jaeger et al. (2010) proposed another interesting method based on two steps. First a set of all candidate interologs is built across the considered species. Next, interologs are assembled into maximal conserved and connected patterns by detecting frequent sub-graphs appearing in the interolog network of the candidate set. Only functionally coherent patterns were used for interolog inference.

The interolog concept was also modified and used in other ways and application domains. In particular, Tirosh & Barkai (2005) proposed a method to assess and increase the confidence of a predicted PPI by examining the co-expression of proteins of its potential interolog in other species. Chen et al. (2007) extended interolog mapping for homologous inference of interacting 3D-domains and they built a database of so-called 3D-interologs (Lo, Chen & Yang, 2010). Chen et al. (2009) used interologs to transfer conserved domain-domain interactions. Recently, Lo, Lin & Yang (2010) combined this interolog domain transfer with the former 3D-interolog detection technique and implemented an integrated tool for searching homologous protein complexes. Finally, Lee et al. (2008) exploited interologs to predict inter-species interactions.

Despite the successful use of interolog inference, a gap was observed between the actual, observed number of conserved interactions and the expected theoretical coverage (Gandhi et al., 2006; Lee et al., 2008). In order to test the reliability of interolog transfer, Mika & Rost (2006) performed a comprehensive validation of the method on several datasets. Their findings suggested that interolog transfers are only accurate at very high levels of sequence identity. In addition, they also compared the interolog transfer within species and across species. In the case of within-species interolog inference a PPI is transferred onto proteins which are sequence similar to the proteins of the considered PPI in the same species. Surprisingly, such paralogous interolog transfers of protein-protein interactions were shown to be significantly more reliable than the orthologous ones. This result was later substantiated by Saeed & Deane (2008), indicating that homology-based interaction prediction methods may yield better results when within-species interolog inference is also considered. In addition, Brown & Jurisica (2007) argued that one also needs to take into account whether all interactions have equal probability of being transferred between organisms. For example, the dynamic components of the interactomes are less likely to be accurately mapped from

distantly related organisms. Moreover, there is apparent bias of interologs to be enriched in stable, permanent complexes (Brown & Jurisica, 2007), which is completely in accordance with findings on the different evolution of transient and permanent interactions. On the other hand, it is likely that the performance of interolog inference could be underestimated since its accuracy is assessed using experimentally tests based on Y2H techniques or high-throughput datasets with a high abundance in Y2H interactions, which were found to be highly enriched in transient and inter-complex connections (Yu et al., 2008).

3.2 Pairwise protein network alignment

Detection and transfer of an interolog between species have motivated the study and exploration of interspecies conservation of protein interactions on a global scale. In particular, instead of focusing on a conserved interaction alone one can compare and align whole interactome maps of distinct species, which mimics the idea behind sequence alignment methods. This approach gave a rise to so-called *network alignment* approach (Sharan & Ideker, 2006).

Using protein network alignment, one can either search for conserved functional network structures such as protein complexes and pathways, or identify functional orthologs across species. As a result this approach should provide a greater evidence and support for protein function and protein interaction prediction for yet uncharacterized or unknown biological processes. Protein network alignment methods can be classified into two main groups:*local network alignments* and *global network alignments*.

As most of the research attention has focused on comparing PPI networks of two different species, here we discuss the successive development of methods for, so-called, *pairwise network alignment*. In sequel we survey local pairwise alignments for detecting evolutionary conserved pathways, local pairwise alignments for detecting conserved protein complexes, and global pairwise network alignment techniques.

3.2.1 Local pairwise network alignment for pathway detection and query tasks

The main goal of local protein network alignment is to detect conserved pathways and protein complexes across species, by searching for local regions of input networks having both high topological similarity between the regions and high sequence similarity between proteins of these regions. The standard approach to this task consists of two main phases: *an alignment phase* and *a searching phase*. In the first phase a merged network representation of compared PPI networks is constructed, called *alignment or orthology graph*. The second phase performs a search for the structures of interest in the orthology graph. Each output result corresponds to a pair or multiplet of complexes or pathways which are evolutionary conserved across the two or more (PPI networks of the) species, respectively.

The first alignment method of whole PPI networks of two species using protein sequence similarity was introduced by Kelley et al. (2003). In this method, called *PathBLAST*, first a many-to-many mapping between proteins of the two species is determined by considering each pair of proteins with a sequence similarity higher than a given threshold as putative orthologs. Next, every orthologous pair is encoded in one alignment node of the new alignment graph and three types of edges (direct, gap and mismatch edge) are identified

between these alignment nodes as follows. The direct edge corresponds to the case when a PPI between proteins of two orthologous pairs exists in the PPI networks of both species. The gap edge represents the case when in one species the respective proteins of alignments nodes are connected indirectly through a common neighbour. Finally, the mismatch edge between alignments nodes is formed if such indirect connection is found between the corresponding proteins in the PPI networks of both species. Gap and mismatch edges are used to describe possible evolutionary variations or account for experimental errors in data (Kelley et al., 2003). In the search phase, the alignment graph is turned into acyclic sub-graphs by random removal of alignment edges, which allows to extract high-scoring paths in linear time by a dynamic programming approach. The score of a path is computed as the sum of log probabilities of true orthology encoded in alignment nodes of the path and of true conserved interactions encoded by alignment edges contained in the path. Interestingly, the method was also applied to align a PPI network with its own copy. In this way they could identify conserved (paralogous) pathways within one species.

The work of Kelley et al. (2003) was followed by other alignment techniques for discovering conserved pathways based on evolutionary conservation. The main drawbacks of *PathBLAST* are that it detects conserved linear pathways in protein interaction data, which is represented as an undirected graph, and it has an exponentially worsening efficiency with the expected increasing length of a pathway to be detected. To circumvent these limitations Pinter et al. (2005) proposed an alignment technique designed explicitly for metabolic networks with directed links between enzymes. The method also handles more complex structures than a simple path, because the scoring of the alignment is based on sub-tree homeomorphism, which can be solved by an efficient deterministic approximation. Another enhancement for the pathway alignment problem was proposed by Wernicke & Rasche (2007) who designed a method that does not impose topological restrictions upon pathways and exploits the biological and local properties of pathways within the network. Another effective approach to metabolic network alignment was developed by Li et al. (2008) which uses an integrative score on compound and enzyme similarities. Pathway alignment has been further extensively investigated and various other techniques have been proposed (see e.g. Cheng et al., 2008; Koyutürk, Kim, Subramaniam, Szpankowski & Grama, 2006; Li et al., 2007).

The evolutionary mapping of *PathBLAST* can also be used to query a known pathway of one species into the PPI network of another species. However, due to limitations and algorithmic constraints of *PathBLAST*, many other methods have been developed with a focussed application of orthologous querying of biological functional complexes, and tools and web-services are available for querying general pathways and other types of protein functional modules across species (see e.g. Bruckner et al., 2009; Dost et al., 2008; Qian et al., 2009; Yang & Sze, 2007).

3.2.2 Local pairwise network alignment for protein complex detection

Another group of methods which followed *PathBLAST* focus on detection of conserved protein complexes across (PPI networks of two or more) species. As these methods compare networks of physical interactions, the identified complexes can be used for interolog prediction as well as for protein function prediction of yet uncharacterized proteins. The detected conserved complexes are either (putative) entire physical complexes or conserved parts of them.

To the best of our knowledge, the first method for detecting conserved complexes using pairwise comparison of PPI networks was introduced by Sharan, Ideker, Kelley, Shamir & Karp (2005) and called *NetworkBLAST*. It can be viewed as a direct extension of *PathBLAST* for the task of complex detection across species. The method employs a comprehensive probabilistic model for conservation of protein complexes and searches for heavy induced sub-graphs in the weighted orthology graph. As the maximal induced sub-graph problem is computationally intractable, *NetworkBLAST* employs a bottom-up greedy heuristic for this task.

Many alignment network techniques which followed *NetworkBLAST* are motivated by the computational intractability issue derived from the problem of a finding maximal common or induced sub-graph in an ortholog graph, and are based on different heuristics. For instance, Koyutürk, Kim, Topkara, Subramaniam, Grama & Szpankowski (2006) partitions the alignment graph into smaller clusters by performing an approximated balanced ratio-cut. In another method by Koyutürk, Kim, Subramaniam, Szpankowski & Grama (2006) the most frequent interaction motifs are extracted from an orthology-contracted graph. Liang et al. (2006) transforms the problem of maximal common sub-graph into the problem of finding all maximal cliques in the graph. Recently, Tian & Samatova (2009) introduced an algorithm based on detection of connected-components of the orthology graph solvable in a very efficient way.

Other researchers propose to restrict the search space to cope with intractability issue of searching phase instead of performing heavy heuristics. For example Li et al. (2007) pre-clusters one PPI network in order to detect candidate complexes which are afterwards aligned to the target species network with an exact integer programming algorithm. Jancura & Marchiori (2010) proposed a pre-processing algorithm based on detection of network hubs for dividing PPI networks, prior to their alignment, into smaller sub-networks containing potential conserved modules. Each possible pair of sub-networks can be later aligned with a state-of-the-art alignment method where the search phase can be performed by means of an exact algorithm, allowing one to perform network comparison in a fully modular fashion and possibly to parallelize the computation. An interesting modular approach was introduced by Narayanan & Karp (2007), where an orthology graph is not constructed but rather networks are compared and split consecutively in several recursive steps until all possible solutions, conserved sub-graphs, are found. Similarly, Gerke et al. (2007) only compares, but does not merge, local hub-centred regions of PPI networks as identified by clustering coefficients and node degrees. The method by Ali & Deane (2009) is again another example of approach where an alignment graph is not explicitly constructed; there interspecies protein similarities are considered as new edges in such a way that species PPI networks and similarity edges between them are encoded into a single global meta-graph which can be searched by standard clustering techniques.

There are also alignment methods which try to incorporate or use other types of information than just the one based on sequence similarity and interaction conservation. For instance, Guo & Hartemink (2009) exploited the findings on co-evolving interacting domains which mediate PPIs and, instead of using putatively homologous proteins for alignment, compares PPI networks across species according to conserved domains of protein-protein interactions. Ali & Deane (2009) propose a functionally guided alignment of PPI networks, where a scoring function incorporates not only sequence and topological similarity of aligned proteins but also

their gene co-expression characteristics and coherence of functional annotations. Thus, the method can be seen as detecting functional modules shared across species rather than strictly evolutionary modules. Finally, Berg & Lässig (2006) developed a generalized alignment Bayesian method applicable to different biological networks.

Despite various pairwise alignment techniques have been introduced, only a few of them embody an evolutionary model of PPI networks in the scoring scheme of an alignment. Notably, Koyutürk, Kim, Topkara, Subramaniam, Grama & Szpankowski (2006) were the first to introduce a method that builds the orthology graph following the duplication/divergence model based on gene duplications. Another interesting method was proposed by Hirsh & Sharan (2007) who extended the probabilistic score of *NetworkBLAST* to asses the likelihood that two complexes originated from an ancestral complex in the common ancestor of the two species being compared under the evolutionary pressure of duplication and link dynamics events.

3.2.3 Global pairwise network alignment

In contrast to local network alignment, which uses many-to-many homologous mapping between proteins of distinct species to detect local conserved regions of a high topological similarity in the respective PPI networks, global protein network alignment uses this mapping to define an unique, globally optimal mapping across whole topologies of PPI networks (Singh et al., 2007), even if it were locally suboptimal in some regions of the networks. In the most strict form of this unique mapping each node in one input network is either matched to one node in the other input network or has no match in the other network. Thus the goal of global protein network alignment is to define functional orthologs across species, as the solution offers a way to resolve the ambiguity of orthology detection with the use of species interactome map. Naturally, as a by-product the global alignment can also identify conserved complexes or pathways.

To the best of our knowledge, the first method performing explicitly global alignment on pair of networks, called *IsoRank*, was introduced by Singh et al. (2007). Similarly to the local network alignment problem, the global network alignment problem is in general computationally intractable. As a consequence, *IsoRank* employs an approximation using an eigenvalue framework in a manner analogous to Google's PageRank algorithm.

Several advancements have naturally followed the introduction of *IsoRank*. For instance, Evans et al. (2008) proposed an asymmetric network matching algorithm based on a network simulation method called quantitative simulation, where a similarity score of a protein pair is iteratively updated by the similarity scores of their neighbours and vice versa until a unique global optimum is found. Other researchers focused more on formulating global alignment as combinatorial optimization problems. For instance Zaslavskiy et al. (2009) redefined the problem of global alignment as a standard graph matching problem and investigated methods using ideas and approaches from state-of-the-art graph matching techniques. Klau (2009) formalized global network alignment as an integer linear programming problem, where a near-optimal solution with a quality guarantee is found by solving a Lagrangian relaxation of the original optimization formulation. Recently, Chindelevitch et al. (2010) proposed a method where the global alignment is encoded as bipartite matching and applied a very efficient local optimization heuristic used for the well-known Travelling Salesman Problem.

3.3 Multiple protein network alignment

The methods on network alignment discussed so far perform alignment of two PPI networks of distinct species. The next natural extension is aligning more than two PPI networks, that is multiple network alignment. A first attempt to perform multiple local network alignment using three species was done by Sharan, Suthram, Kelley, Kuhn, McCuine, Uetz, Sittler, Karp & Ideker (2005), which exploited the scoring model of *NetworkBLAST*. However, the method scales exponentially with the number of input species and consequently it is ineffective for large scale comparisons.

Apart from the scalability problem, there are also other issues related to the problem of aligning more than two species. For instance, the putative orthologous mapping of certain proteins does not need to span across all species, meaning that proteins may be conserved only for a particular subset of species. This "orthology decay" is more evident when a large number of increasingly distant species are considered in the alignment. As a result, functional modules, such as pathways and complexes, can have a different degree of conservation, with some modules being strictly conserved across all species and some other modules being conserved only for a particular clade. Thus, a good alignment method should allow one to search for conserved modules at different degree of conservation. However, such requirement also increases the complexity of searching and consequently one may need to prune the number of all possible species combinations in alignment.

To the best of our knowledge, the first method capable of an efficient comparison of multiple PPI networks, called *Graemlin*, was introduced by Flannick et al. (2006). The alignment model of the method allows one to perform local as well as global alignment and is also applicable for querying tasks of particular biological modules of interest across PPI networks. It employs a rather involved scoring scheme which allows one to search for conserved pathways as well as for conserved complexes. It also outputs modules with a different conservation degree. *Graemlin* progressively aligns the closest pair of PPI networks according the species distance measured using a phylogenetic tree, until the last pair on the root of the tree is compared, corresponding to the most conserved parts of the aligned networks. The main disadvantage of this approach is that it involves to estimate many parameters. Recently, a supervised, automated parameter learner was proposed to lessen the burden of parameter tuning (Flannick et al., 2009).

Another phylogeny-guided local network alignment was proposed by Kalaev et al. (2008). Although the method uses the same probabilistic scoring for conserved complex as *NetworkBLAST*, it avoids its exponential scalability by redefining the alignment model such that it does not construct the merged representation of aligned networks but represents them as separate layers interconnected via orthologous mapping. Then a seed, that is, a group of putatively orthologous proteins spanning across all species, is selected using the species phylogeny and greedily expanded by adding other proteins being orthologous to each other in all respective species in order to maximize the alignment conservation score. The proposed method, however, identifies only protein complexes conserved across all species and does not detect complexes conserved only for a certain subset of species.

Notably, the functionally guided network alignment method of Ali & Deane (2009), previously mentioned as one of the methods for pairwise alignment, was also shown to perform efficiently local alignment of multiple networks.

All these multiple local network alignments do not reconstruct a plausible evolutionary history of PPI networks based on a model of evolution, although they might be phylogeny-aware. Motivated by this observation, Dutkowski & Tiuryn (2007) introduced a new multiple local network alignment method, called *CAPPI*, which from the given PPI networks of distinct species aims to reconstruct an ancient PPI network of the common ancestor. The method uses a Bayesian inference framework based on a duplication and divergence model of network evolution which mimics the processes by which most protein interactions are formed. After the reconstruction step, the ancestral network is decomposed into connected components which correspond to the ancestral modules of protein interactions and are projected back to the original networks to obtain the actual conserved network residues. Although the demonstrated application of the method was restricted to orthologous groups spanning across all species (Dutkowski & Tiuryn, 2007), to the best of our knowledge *CAPPI* is the only model-based approach for large-scale ancestral network reconstruction.

Among the multiple alignment methods above mentioned, only *Graemlin* was shown to perform a global multiple network alignment, yet it relies on a involved parameter estimation step and phylogeny-guided approximation. Recently Liao et al. (2009) developed another global alignment technique which is fully unsupervised and phylogeny-free. The method, called *IsoRankN*, is built on the *IsoRank* algorithm mentioned above (Singh et al., 2007) and its extension to the multiple global network alignment (Singh et al., 2008a). At first *IsoRankN* scores topological and sequence similarity matching between putatively orthologous proteins of each pair of input networks using *IsoRank*. Then, a maximum k-partite graph matching problem is formulated on the induced graph of pairwise alignment scores (Singh et al., 2008a) and the exact solution is approximated by a spectral graph partitioning algorithm. *IsoRankN* also effectively identifies one-to-one orthologous mappings for all subset of species and appears to out-perform *Graemlin* in terms of coverage and quality of functional enrichments.

3.4 Applications and future developments

Local and global alignment methods have been successfully applied to study evolution of species and to discover relevant biological knowledge. For example, Suthram et al. (2005) applied the network alignment of Sharan, Suthram, Kelley, Kuhn, McCuine, Uetz, Sittler, Karp & Ideker (2005) to examine the degree of conservation between the Plasmodium protein network and other model organisms, such as yeast, nematode worm, fruit fly and the bacterial pathogen Helicobacter pylori. They investigated whether the divergence of Plasmodium at the sequence level is reflected in the configuration of its protein network. Indeed, the alignments showed very little conservation suggesting that the patterns of protein interaction in Plasmodium, like its genome sequence, set it apart from other species (Suthram et al., 2005).

Another application of local network alignment was performed by Tan et al. (2007) who combined transcriptional regulatory interactions with protein-protein interactions and identified co-regulated complexes between yeast and fly revealing different conservation of their regulators. This finding advocates that PPI networks may evolve more slowly than transcriptional interaction networks. In addition, Schwartz et al. (2009) and Dutkowski & Tiuryn (2009) used conserved complexes detected by network alignments for protein interaction prediction in a manner similar to the interologs transfer approach and demonstrated their usefulness. In particular, Schwartz et al. (2009) provided a

comprehensive experimental design which includes PPI prediction using network alignment, and demonstrated how effectively it reduces the cost of interactome mapping.

Furthermore, Bandyopadhyay et al. (2006) presented the first systematic identification of functional orthologs based on protein network comparison. They used the pairwise local alignment model of Kelley et al. (2003) to construct the orthology graph and then they resolved ambiguity of orthology mapping by fitting a logistic function previously trained on a known set of functional orthologs. In contrast, Singh et al. (2008b) predicted functional orthologs in unsupervised manner by using explicitly a global multiple network alignment method.

Finally, Kolar et al. (2008) performed a cross-species analysis of two herpes-viruses using the generalized Bayesian network alignment of Berg & Lässig (2006). Interestingly, the performed alignment employs in its probabilistic scoring system evolutionary rates of sequences and thus it goes beyond the narrow use of orthologous mapping as done in all other alignment techniques. The method predicted meaningful functional associations that could not be obtained from sequence or interaction data alone.

Despite the recent progress and increasing number of network alignment tools, their further development remains an ongoing research issue, in particular for multiple network comparison. Only a few methods perform the scoring of alignment according to evolutionary models and there is only one of them which fully reconstructs network evolutionary history. This clearly is in contrast with the numerous techniques for the reconstruction of evolutionary history of gene families. Also, actual alignment methods do not distinguish among diverse types of interactions, specifically between transient and permanent interactions. For example, the prior knowledge on different evolutionary behaviour of these types of physical interactions could be incorporated into a scoring scheme of alignment construction.

In addition, all but one network comparison methods just rely on the straightforward use of putative orthologous mapping as identified by sequence comparison or available in orthologous databases, but they do not employ evolutionary measures, such as evolutionary distances or retentions, which can be derived from the corresponding sequence alignments. These measures assess the level of evolutionary conservation and they could potentially improve the performance of network alignments.

Mostly all current applications of network alignments have worked with networks of physical interactome. However, the power of network alignment for functional annotation and other system biology applications could be explored when one performs comparison of more general, functional interaction networks. One may expect that such alignment could reveal a higher number of conserved modules as the interspecies conservation of modularity across protein networks increases with combined, integrated evidence for a pair of proteins to be functionally linked. Finally, all available methods here considered focused on conservation of modules but not on the more general concept of module evolutionary cohesiveness or co-evolution. The evolutionary cohesiveness can be assessed especially for the case of multiple alignments. Indeed, all conserved modules are inherently very cohesive, however not all evolutionary modules need to exhibit the correlated conservation at a level as expected by actual multiple network alignments. Protein functional modules differ in the degree of conservation and also in the degree of cohesiveness.

4. References

Agarwal, S., Deane, C. M., Porter, M. A. & Jones, N. S. (2010). Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks, *PLoS Comput Biol* 6(6): e1000817.

Ali, W. & Deane, C. M. (2009). Functionally guided alignment of protein interaction networks for module detection, *Bioinformatics* 25(23): 3166–3173.

Aragues, R., Sali, A., Bonet, J., Marti-Renom, M. A. & Oliva, B. (2007). Characterization of protein hubs by inferring interacting motifs from protein interactions, *PLoS Comput Biol* 3(9): e178.

Bandyopadhyay, S., Sharan, R. & Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison, *Genome Research* 16(3): 428–435.

Batada, N. N., Hurst, L. D. & Tyers, M. (2006). Evolutionary and physiological importance of hub proteins, *PLoS Comput Biol* 2(7): e88.

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hurst, L. D. & Tyers, M. (2006). Stratus not altocumulus: A new view of the yeast protein interaction network, *PLoS Biol* 4(10): e317.

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hurst, L. D. & Tyers, M. (2007). Still stratus not altocumulus: Further evidence against the date/party hub distinction, *PLoS Biol* 5(6): e154.

Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L. & Krogan, N. J. (2009). Evolution of phosphoregulation: Comparison of phosphorylation patterns across yeast species, *PLoS Biol* 7(6): e1000134.

Berg, J. & Lässig, M. (2006). Cross-species analysis of biological networks by Bayesian alignment, *Proceedings of the National Academy of Sciences* 103(29): 10967–10972.

Bertin, N., Simonis, N., Dupuy, D., Cusick, M. E., Han, J.-D. J., Fraser, H. B., Roth, F. P. & Vidal, M. (2007). Confirmation of organized modularity in the yeast interactome, *PLoS Biol* 5(6): e153.

Bloom, J. & Adami, C. (2003). Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets, *BMC Evolutionary Biology* 3(1): 21.

Brown, K. & Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks, *Genome Biology* 8(5): R95.

Bruckner, S., Hüffner, F., Karp, R. M., Shamir, R. & Sharan, R. (2009). Torque: topology-free querying of protein interaction networks., *Nucleic Acids Research* 37(Web Server issue): W106–108.

Campillos, M., von Mering, C., Jensen, L. J. & Bork, P. (2006). Identification and analysis of evolutionarily cohesive functional modules in protein networks, *Genome Research* 16(3): 374–382.

Chen, C.-C., Lin, C.-Y., Lo, Y.-S. & Yang, J.-M. (2009). Ppisearch: a web server for searching homologous protein-protein interactions across multiple species, *Nucleic Acids Research* 37(suppl 2): W369–W375.

Chen, Y.-C., Lo, Y.-S., Hsu, W.-C. & Yang, J.-M. (2007). 3d-partner: a web server to infer interacting partners and binding models, *Nucleic Acids Research* 35(suppl 2): W561–W567.

Chen, Y. & Xu, D. (2005). Understanding protein dispensability through machine-learning analysis of high-throughput data, *Bioinformatics* 21(5): 575–581.

Cheng, Q., Berman, P., Harrison, R. & Zelikovsky, A. (2008). Fast alignments of metabolic networks, *BIBM '08: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE Computer Society, Washington, DC, USA, pp. 147–152.

Chindelevitch, L., Liao, C.-S. & Berger, B. (2010). Local optimization for global alignment of protein interaction networks, *Pacific Symposium on Biocomputing* 15: 123–132.

Clark, G. W., Dar, V.-u.-N., Bezginov, A., Yang, J. M., Charlebois, R. L. & Tillier, E. R. M. (2011). Using coevolution to predict protein-protein interactions, *in* G. Cagney, A. Emili & J. M. Walker (eds), *Network Biology*, Vol. 781 of *Methods in Molecular Biology*, Humana Press, pp. 237–256.

Cordero, O. X., Snel, B. & Hogeweg, P. (2008). Coevolution of gene families in prokaryotes, *Genome Research* 18(3): 462–468.

Coulomb, S., Bauer, M., Bernard, D. & Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks, *Proceedings of the Royal Society B: Biological Sciences* 272(1573): 1721–1725.

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A. A., Hassett, D. J. & Lu, L. J. (2011). Investigating the predictability of essential genes across distantly related organisms using an integrative approach, *Nucleic Acids Research* 39(3): 795–807.

Dost, B., Shlomi, T., Gupta, N., Ruppin, E., Bafna, V. & Sharan, R. (2008). Qnet: A tool for querying protein interaction networks, *Journal of Computational Biology* 15(7): 913–925.

Doyle, M., Gasser, R., Woodcroft, B., Hall, R. & Ralph, S. (2010). Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes, *BMC Genomics* 11(1): 222.

Drummond, D. A., Raval, A. & Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution, *Molecular Biology and Evolution* 23(2): 327–337.

Dutkowski, J. & Tiuryn, J. (2007). Identification of functional modules from conserved ancestral protein-protein interactions, *Bioinformatics* 23(13): i149–158.

Dutkowski, J. & Tiuryn, J. (2009). Phylogeny-guided interaction mapping in seven eukaryotes, *BMC Bioinformatics* 10(1): 393.

Ekman, D., Light, S., Björklund, A. K. & Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae?, *Genome Biology* 7(6): R45.

Evans, P., Sandler, T. & Ungar, L. (2008). Protein-protein interaction network alignment by quantitative simulation, *BIBM '08: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE Computer Society, Washington, DC, USA, pp. 325–328.

Fang, G., Rocha, E. & Danchin, A. (2005). How essential are nonessential genes?, *Molecular Biology and Evolution* 22(11): 2147–2156.

Flannick, J., Novak, A., Do, C. B., Srinivasan, B. S. & Batzoglou, S. (2009). Automatic parameter learning for multiple local network alignment, *Journal of Computational Biology* 16(8): 1001–1022.

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H. & Batzoglou, S. (2006). Graemlin: General and robust alignment of multiple large interaction networks, *Genome Res.* 16(9): 1169–1181.

Fokkens, L. & Snel, B. (2009). Cohesive versus flexible evolution of functional modules in eukaryotes, *PLoS Comput Biol* 5(1): e1000276.

Fraser, H. B. (2005). Modularity and evolutionary constraint on proteins, *Nat Genet* 37(4): 351 – 352.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network, *Science* 296(5568): 750–752.

Fraser, H. & Hirsh, A. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level, *BMC Evolutionary Biology* 4(1): 13.

Fraser, H., Wall, D. & Hirsh, A. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions, *BMC Evolutionary Biology* 3(1): 11.

Gallone, G., Simpson, T. I., Armstrong, J. D. & Jarman, A. (2011). Bio::homology::interologwalk - a perl module to build putative protein-protein interaction networks through interolog mapping, *BMC Bioinformatics* 12(1): 289.

Galperin, M. Y. & Koonin, E. V. (2000). Who's your neighbor? new computational approaches for functional genomics, *Nat Biotech* 18(6): 609–613.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S. & Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets, *Nat Genet* 38(3): 285 – 293.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N. J., Millar, A. H. & Geisler, M. (2007). A predicted interactome for arabidopsis, *Plant Physiology* 145(2): 317–329.

Gerke, M., Bornberg-Bauer, E., Jiang, X. & Fuellen, G. (2007). Finding common protein interaction patterns across organisms, *Evolutionary bioinformatics online* 2: 45–52.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W. & Johnston, M. (2002). Functional profiling of the saccharomyces cerevisiae genome, *Nature* 418: 387–391.

Glazko, G. & Mushegian, A. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns, *Genome Biology* 5(5): R32.

Gu, H., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. (2011). Prin: a predicted rice interactome network, *BMC Bioinformatics* 12(1): 161.

Guo, X. & Hartemink, A. J. (2009). Domain-oriented edge-based alignment of protein interaction networks, *Bioinformatics* 25(12): i240–1246.

Gustafson, A., Snitkin, E., Parker, S., DeLisi, C. & Kasif, S. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis, *BMC Genomics* 7(1): 265.

Hahn, M. W. & Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Molecular Biology and Evolution* 22(4): 803–806.

Hakes, L., Lovell, S. C., Oliver, S. G. & Robertson, D. L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution, *Proceedings of the National Academy of Sciences* 104(19): 7999–8004.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature* 430: 88–93.

He, X. & Zhang, J. (2006). Why do hubs tend to be essential in protein networks?, *PLoS Genet* 2(6): e88.

Hirsh, A. E. & Fraser, H. B. (2001). Protein dispensability and rate of evolution, *Nature* 411: 1046–1049.

Hirsh, A. E. & Fraser, H. B. (2003). Genomic function (communication arising): Rate of evolution and gene dispensability, *Nature* 421(6922): 497–498.

Hirsh, E. & Sharan, R. (2007). Identification of conserved protein complexes based on a model of protein network evolution, *Bioinformatics* 23(2): e170–176.

Huang, T.-W., Lin, C.-Y. & Kao, C.-Y. (2007). Reconstruction of human protein interolog network using evolutionary conserved network, *BMC Bioinformatics* 8(1): 152.

Hurst, L. D. & Smith, N. G. (1999). Do essential genes evolve slowly?, *Current biology* 9: 747–750.

Itzhaki, Z., Akiva, E., Altuvia, Y. & Margalit, H. (2006). Evolutionary conservation of domain-domain interactions, *Genome Biology* 7(12): R125.

Jaeger, S., Sers, C. & Leser, U. (2010). Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction, *BMC Genomics* 11(1): 717.

Jancura, P. & Marchiori, E. (2010). Dividing protein interaction networks for modular network comparative analysis, *Pattern Recognition Letters* 31(14): 2083 – 2096.

Jancura, P., Mavridou, E., Carrillo-De Santa Pau, E. & Marchiori, E. (2012). A methodology for detecting the orthology signal in a ppi network at a functional complex level, *BMC Bioinformatics* 13(Suppl 1). In press.

Jeong, H., Mason, S. P., Barabasi, A.-L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks, *Nature* 411: 41–42.

Jonsson, P. F. & Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome, *Bioinformatics* 22(18): 2291–2297.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, *Genome Research* 12(6): 962–968.

Jordan, I. K., Wolf, Y. & Koonin, E. (2003a). Correction: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly, *BMC Evolutionary Biology* 3(1): 5.

Jordan, I. K., Wolf, Y. & Koonin, E. (2003b). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly, *BMC Evolutionary Biology* 3(1): 1.

Jothi, R., Cherukuri, P. F., Tasneem, A. & Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions, *Journal of Molecular Biology* 362(4): 861 – 875.

Kafri, R., Dahan, O., Levy, J. & Pilpel, Y. (2008). Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy, *Proceedings of the National Academy of Sciences* 105(4): 1243–1248.

Kahali, B., Ahmad, S. & Ghosh, T. C. (2009). Exploring the evolutionary rate differences of party hub and date hub proteins in saccharomyces cerevisiae protein-protein interaction network, *Gene* 429(1-2): 18 – 22.

Kalaev, M., Bafna, V. & Sharan, R. (2008). Fast and accurate alignment of multiple protein networks, *Research in Computational Molecular Biology*, pp. 246–256.

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. & Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment, *Proceedings of the National Academy of Science* 100: 11394–11399.

Kensche, P. R., van Noort, V., Dutilh, B. E. & Huynen, M. A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution, *Journal of The Royal Society Interface* 5(19): 151–170.

Kim, P. M., Korbel, J. O. & Gerstein, M. B. (2007). Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context, *Proceedings of the National Academy of Sciences* 104(51): 20274–20279.

Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights, *Science* 314(5807): 1938–1941.

Kim, W. K., Bolser, D. M. & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (psimap), *Bioinformatics* 20(7): 1138–1150.

Kim, W. K. & Marcotte, E. M. (2008). Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence, *PLoS Comput Biol* 4(11): e1000232.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press.

Klau, G. (2009). A new graph-based method for pairwise global network alignment, *BMC Bioinformatics* 10(Suppl 1): S59.

Kolar, M., Lassig, M. & Berg, J. (2008). From protein interactions to functional annotation: graph alignment in herpes, *BMC Systems Biology* 2(1): 90.

Koyutürk, M., Kim, Y., Subramaniam, S., Szpankowski, W. & Grama, A. (2006). Detecting conserved interaction patterns in biological networks, *Journal of Computational Biology* 13(7): 1299–1322.

Koyutürk, M., Kim, Y., Topkara, U., Subramaniam, S., Grama, A. & Szpankowski, W. (2006). Pairwise alignment of protein interaction networks, *Journal of Computional Biology* 13(2): 182–199.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution, *Genome Research* 13(10): 2229–2235.

Kunin, V., Pereira-Leal, J. B. & Ouzounis, C. A. (2004). Functional evolution of the yeast protein interaction network, *Molecular Biology and Evolution* 21(7): 1171–1176.

Lee, S.-A., Chan, C.-h., Tsai, C.-H., Lai, J.-M., Wang, F.-S., Kao, C.-Y. & Huang, C.-Y. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions, *BMC Bioinformatics* 9(Suppl 12): S11.

Lee, W.-P., Jeng, B.-C., Pai, T.-W., Tsai, C.-P., Yu, C.-Y. & Tzou, W.-S. (2006). Differential evolutionary conservation of motif modes in the yeast protein interaction network, *BMC Genomics* 7(1): 89.

Lehner, B. & Fraser, A. (2004). A first-draft human protein-interaction map, *Genome Biology* 5(9): R63.

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D. & Hartl, D. L. (2005). Evolution of proteins and gene expression levels are coupled in drosophila and are independently associated with mrna abundance, protein length, and number of protein-protein interactions, *Molecular Biology and Evolution* 22(5): 1345–1354.

Li, H., Kristensen, D. M., Coleman, M. K. & Mushegian, A. (2009). Detection of biochemical pathways by probabilistic matching of phyletic vectors, *PLoS ONE* 4(4): e5326.

Li, Y., de Ridder, D., de Groot, M. & Reinders, M. (2008). Metabolic pathway alignment between species using a comprehensive and flexible similarity measure, *BMC Systems Biology* 2(1): 111.

Li, Z., Zhang, S., Wang, Y., Zhang, X.-S. & Chen, L. (2007). Alignment of molecular networks by integer quadratic programming, *Bioinformatics* 23(13): 1631–1639.

Liang, Z., Xu, M., Teng, M. & Niu, L. (2006). Comparison of protein interaction networks reveals species conservation and divergence, *BMC Bioinformatics* 7(1): 457.

Liao, C.-S., Lu, K., Baym, M., Singh, R. & Berger, B. (2009). IsoRankN: spectral methods for global alignment of multiple protein networks, *Bioinformatics* 25(12): i253–258.

Lin, Y.-S., Hwang, J.-K. & Li, W.-H. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome, *Gene* 387(1-2): 109 – 117.

Liu, Z., Liu, Q., Sun, H., Hou, L., Guo, H., Zhu, Y., Li, D. & He, F. (2011). Evidence for the additions of clustered interacting nodes during the evolution of protein interaction networks from network motifs, *BMC Evolutionary Biology* 11(1): 133.

Lo, Y.-S., Chen, Y.-C. & Yang, J.-M. (2010). 3d-interologs: an evolution database of physical protein- protein interactions across multiple genomes, *BMC Genomics* 11(Suppl 3): S7.

Lo, Y.-S., Lin, C.-Y. & Yang, J.-M. (2010). Pcfamily: a web server for searching homologous protein complexes, *Nucleic Acids Research* 38(suppl 2): W516–W522.

Makino, T. & Gojobori, T. (2006). The evolutionary rate of a protein is influenced by features of the interacting partners, *Molecular Biology and Evolution* 23(4): 784–789.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. & Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs", *Genome Research* 11(12): 2120–2126.

McDermott, J. E., Taylor, R. C., Yoon, H. & Heffron, F. (2009). Bottlenecks and hubs in inferred networks are important for virulence in salmonella typhimurium, *Journal of Computational Biology* 16: 169–180.

Michaut, M., Kerrien, S., Montecchi-Palazzi, L., Chauvat, F., Cassier-Chauvat, C., Aude, J.-C., Legrain, P. & Hermjakob, H. (2008). Interoporc: automated inference of highly conserved protein interaction networks, *Bioinformatics* 24(14): 1625–1631.

Mika, S. & Rost, B. (2006). Protein-protein interactions more conserved within species than across species, *PLoS Comput Biol* 2(7): e79.

Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions, *Proceedings of the National Academy of Sciences* 102(31): 10930–10935.

Mirzarezaee, M., Araabi, B. & Sadeghi, M. (2010). Features analysis for identification of date and party hubs in protein interaction network of saccharomyces cerevisiae, *BMC Systems Biology* 4(1): 172.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y. & Wang, X. (1994). Co-evolution of ligand-receptor pairs, *Nature* 368(6468): 251–255.

Narayanan, M. & Karp, R. M. (2007). Comparing protein interaction networks via a graph match-and-split algorithm, *Journal of Computational Biology* 14(7): 892–907.

Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions, *EMBO J* 22(14): 3486–3492.

Pal, C., Papp, B. & Hurst, L. D. (2003). Genomic function (communication arising): Rate of evolution and gene dispensability, *Nature* 421(6922): 496–497.

Pal, C., Papp, B. & Lercher, M. J. (2006). An integrated view of protein evolution, *Nat Rev Genet* 7: 337–348.

Pang, K., Cheng, C., Xuan, Z., Sheng, H. & Ma, X. (2010). Understanding protein evolutionary rate by integrating gene co-expression with protein interactions, *BMC Systems Biology* 4(1): 179.

Pang, K., Sheng, H. & Ma, X. (2010). Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network, *Biochemical and Biophysical Research Communications* 401(1): 112 – 116.

Park, K. & Kim, D. (2009). Localized network centrality and essentiality in the yeast-protein interaction network, *PROTEOMICS* 9(22): 5143–5154.

Pavithra, S. R., Kumar, R. & Tatu, U. (2007). Systems analysis of chaperone networks in the malarial parasite plasmodium falciparum, *PLoS Comput Biol* 3(9): e168.

Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering* 14(9): 609–614.

Pazos, F. & Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions, *EMBO J* 27(20): 2648–2655.

Pedamallu, C. S. & Posfai, J. (2010). Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information, *Source Code for Biology and Medicine* 5(1): 8.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. (2005). Alignment of metabolic pathways, *Bioinformatics* 21(16): 3401–3408.

Plotkin, J. B. & Fraser, H. B. (2007). Assessing the determinants of evolutionary rates in the presence of noise, *Molecular Biology and Evolution* 24(5): 1113–1121.

Qian, W., He, X., Chan, E., Xu, H. & Zhang, J. (2011). Measuring the evolutionary rate of protein-protein interaction, *Proceedings of the National Academy of Sciences* 108(21): 8725–8730.

Qian, X., Sze, S.-H. & Yoon, B.-J. (2009). Querying Pathways in Protein Interaction Networks Based on Hidden Markov Models, *Journal of Computational Biology* 16(2): 145–157.

Rocha, E. P. C. & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins, *Molecular Biology and Evolution* 21(1): 108–116.

Saeed, R. & Deane, C. (2006). Protein protein interactions, evolutionary rate, abundance and age, *BMC Bioinformatics* 7(1): 128.

Saeed, R. & Deane, C. (2008). An assessment of the uses of homologous interactions, *Bioinformatics* 24(5): 689–695.

Schuster-Bockler, B. & Bateman, A. (2007). Reuse of structural domain-domain interactions in protein networks, *BMC Bioinformatics* 8(1): 259.

Schwartz, A. S., Yu, J., Gardenour, K. R., Finley Jr, R. L. & Ideker, T. (2009). Cost-effective strategies for completing the interactome, *Nat Meth* 6(1): 55–61.

Seidl, M. & Schultz, J. (2009). Evolutionary flexibility of protein complexes, *BMC Evolutionary Biology* 9(1): 155.

Sharan, R. & Ideker, T. (2006). Modeling cellular machinery through biological network comparison, *Nature Biotechnology* 24(4): 427–433.

Sharan, R., Ideker, T., Kelley, B. P., Shamir, R. & Karp, R. M. (2005). Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data, *Journal of Computional Biology* 12(6): 835–846.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. & Ideker, T. (2005). From the Cover: Conserved patterns of protein interaction in multiple species, *Proceedings of the National Academy of Sciences* 102(6): 1974–1979.

Singh, R., Xu, J. & Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology, *Research in Computational Molecular Biology* pp. 16–31.

Singh, R., Xu, J. & Berger, B. (2008a). Global alignment of multiple protein interaction networks, *Pacific Symposium on Biocomputing* 13: 303–314.

Singh, R., Xu, J. & Berger, B. (2008b). Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proceedings of the National Academy of Sciences* 105(35): 12763–12768.

Snel, B. & Huynen, M. A. (2004). Quantifying modularity in the evolution of biomolecular systems, *Genome Research* 14(3): 391–397.

Suthram, S., Sittler, T. & Ideker, T. (2005). The plasmodium protein network diverges from those of other eukaryotes, *Nature* 438(7064): 108–112.

Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. (2007). Transcriptional regulation of protein complexes within and across species, *Proceedings of the National Academy of Sciences* 104(4): 1283–1288.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families, *Science* 278(5338): 631–637.

Theis, F. J., Latif, N., Wong, P. & Frishman, D. (2011). Complex principal component and correlation structure of 16 yeast genomic variables, *Molecular Biology and Evolution* 28(9): 2501–2512.

Tian, W. & Samatova, N. F. (2009). Pairwise alignment of interaction networks by fast identification of maximal conserved patterns, *Pacific Symposium on Biocomputing* 14: 99–110.

Tillier, E. R. & Charlebois, R. L. (2009). The human protein coevolution network, *Genome Research* 19(10): 1861–1871.

Tirosh, I. & Barkai, N. (2005). Computational verification of protein-protein interactions by orthologous co-expression, *BMC Bioinformatics* 6(1): 40.

Tuller, T., Kupiec, M. & Ruppin, E. (2009). Co-evolutionary networks of genes and cellular processes across fungal species, *Genome Biology* 10(5): R48.

Ulitsky, I. & Shamir, R. (2007). Pathway redundancy and protein essentiality revealed in the saccharomyces cerevisiae interaction networks, *Mol Syst Biol* 3: 1–7.

Vergassola, M., Vespignani, A. & Dujon, B. (2005). Cooperative evolution in protein complexes of yeast from comparative analyses of its interaction network, *PROTEOMICS* 5(12): 3116–3119.

Vespignani, A. (2003). Evolution thinks modular, *Nature Genetics* 35(2): 118–119.

von Mering, C., Zdobnov, E. M., Tsoka, S., Ciccarelli, F. D., Pereira-Leal, J. B., Ouzounis, C. A. & Bork, P. (2003). Genome evolution reveals biochemical networks and functional modules, *Proceedings of the National Academy of Sciences* 100(26): 15428–15433.

Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N. & Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development, *Science* 287(5450): 116–122.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B. & Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution, *Proceedings of the National Academy of Sciences* 102(15): 5483–5488.

Wang, Z. & Zhang, J. (2009). Why is the correlation between gene importance and gene evolutionary rate so weak?, *PLoS Genet* 5(1): e1000329.

Watanabe, R., Morett, E. & Vallejo, E. (2008). Inferring modules of functionally interacting proteins using the bond energy algorithm, *BMC Bioinformatics* 9(1): 285.

Waterhouse, R. M., Zdobnov, E. M. & Kriventseva, E. V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi, *Genome Biology and Evolution* 3: 75–86.

Wernicke, S. & Rasche, F. (2007). Simple and fast alignment of metabolic pathways by exploiting local diversity, *Bioinformatics* 23(15): 1978–1985.

Williams, E. J. B. & Hurst, L. D. (2000). The proteins of linked genes evolve at similar rates, *Nature* 407(6806): 900–903.

Wolf, Y. I., Carmel, L. & Koonin, E. V. (2006). Unifying measures of gene function and evolution, *Proceedings of the Royal Society B: Biological Sciences* 273(1593): 1507–1515.

Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network, *Genome Research* 14(7): 1310–1314.

Wuchty, S., Barabasi, A.-L. & Ferdig, M. (2006). Stable evolutionary signal in a yeast protein interaction network, *BMC Evolutionary Biology* 6(1): 8.

Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network, *Nature Genetics* 35(2): 176–179.

Yamada, T., Kanehisa, M. & Goto, S. (2006). Extraction of phylogenetic network modules from the metabolic network, *BMC Bioinformatics* 7(1): 130.

Yang, Q. & Sze, S.-H. (2007). Path matching and graph matching in biological networks, *Journal of Computational Biology* 14(1): 56–67.

Yang, Z. & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, *Molecular Biology and Evolution* 17(1): 32–43.

Yeang, C.-H. & Haussler, D. (2007). Detecting coevolution in and among protein domains, *PLoS Comput Biol* 3(11): e211.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T.,

Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E. & Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network, *Science* 322(5898): 104–110.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput Biol* 3(4): e59.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M. & Gerstein, M. (2004). Annotation transfer between genomes: Protein-protein interologs and protein-dna regulogs, *Genome Research* 14(6): 1107–1118.

Zaslavskiy, M., Bach, F. & Vert, J.-P. (2009). Global alignment of protein-protein interaction networks by graph matching methods, *Bioinformatics* 25(12): i259–1267.

Zhang, J. & He, X. (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution, *Molecular Biology and Evolution* 22(4): 1147–1155.

Zhao, J., Ding, G.-H., Tao, L., Yu, H., Yu, Z.-H., Luo, J.-H., Cao, Z.-W. & Li, Y.-X. (2007). Modular co-evolution of metabolic networks, *BMC Bioinformatics* 8(1): 311.

Zinman, G., Zhong, S. & Bar-Joseph, Z. (2011). Biological interaction networks are conserved at the module level, *BMC Systems Biology* 5(1): 134.

Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality, *PLoS Comput Biol* 4(8): e1000140.

22

Scalable, Integrative Analysis and Visualization of Protein Interactions

David Otasek[1], Chiara Pastrello[2,1] and Igor Jurisica[1,3]
[1]*Ontario Cancer Institute the Campbell Family Institute for Cancer Research, University Health Network, Toronto, Ontario,*
[2]*CRO Aviano, National Cancer Institute, Aviano,*
[3]*Department of Computer Science and Medical Biophysics, University of Toronto, Toronto,*
[1,3]*Canada*
[2]*Italy*

1. Introduction

Biology offers a diversity of problems, leading to many computational biology workflows, including tasks where network visualization is helpful to interpret and analyse data. High-throughput screening techniques generate large amounts of data useful for the comprehension of the biological mechanisms underlying different diseases. The need for agile tools to handle such data and analyse it correctly has become continuously more evident.

Individual network visualization systems differ greatly in terms of the features and standards they support, and consequently the analyses they enable. Importantly, users have a broad range of skills and expectations, ranging from biology to computational biology. As a result, network visualization tools must satisfy diverse requirements and thus offer different user interfaces and features. In this role, they are also fundamental in helping scientists in different fields integrate their knowledge and their data in an interdisciplinary approach to research.

The number of '-omics' disciplines that use high-throughput techniques and that can benefit from a network approach are increasing. The diverse data that can be represented as a graph includes physical protein-protein interactions (PPIs), metabolic networks (Swainston et al., 2011), genetic co-expression (Helaers et al., 2011), gene regulatory networks (Longabaugh, 2012), microRNA-target (Shirdel et al., 2011) and drug-target associations (Morrow et al., 2010). In this chapter we focus on physical PPIs.

Proteins are key players in virtually all biological events that take place within and between cells and often accomplish their function as part of large molecular machines, whose action is coordinated through intricate regulatory networks of transient PPIs. The understanding of the interrelationships between molecules is the basis for an understanding of the behaviour of biological systems (Stein et al., 2011).

The analysis of the full proteome is possible with techniques such as mass spectrometry and protein microarray, which can be integrated with targeted approaches such as yeast-2-hybrid screen, immune precipitation and affinity purification. So far, PPI discovery methods are not accurate enough to be used alone, but the combination of different techniques can help to build an accurate interactome map (Remmerie et al. 2011). Still, this kind of analysis can only indicate that two proteins interact but does not reveal the molecular details or the mechanism of binding captured in high resolution three-dimensional (3D) structures, in which individual residue contacts are resolved and the interaction interfaces characterized. Moreover, they do not capture transient interactions and post translational modifications (PTMs) that can be addressed by techniques such as immobilized metal affinity chromatography (IMAC) mass spectrometry for protein phosphorylation analysis.

It becomes evident that the analysis of protein interactions is already a huge field with a plethora of data coming from different sources that can be improved by computational techniques and integrative network visualization and analysis. It is even more interesting to integrate PPI data with protein-target interaction data to have a wider view of the environmental context that influences network operations.

In this context, a pathway-centric analysis can help to elucidate the role and the importance of proteins in the context of the cell environment, specifically when the pathways can be related to the process/disease being studied. However, it is mandatory to be aware of the limits of this analysis, due to the cross-talk among pathways: a singular protein, in fact, can be associated or interact with multiple pathways so none of the pathways can be considered a single actor but rather a piece of a bigger puzzle (Kreeger & Lauffenburger, 2010).

Another intriguing aspect that protein-target interactions can describe is the relationship between protein exogenous molecules like drugs or toxins (Yu, 2011). The analysis of networks generated from drug-target and protein-target interactions can highlight different molecules that can be responsible of the response or resistance to a certain drug as well as alternative drugs that can target disease specific proteins.

2. Network visualization tools

There are dozens of applications available for the visualization of biological networks, each with its own focus, work-flow and tools (Pavlopoulos et al., 2008; Gehlenborg et al., 2010). We will describe some of the most common features and workflows involved in using these applications, with brief discussion of NAViGaTOR (Brown et al., 2009; McGuffin and Jurisica, 2009; Djebbari et al., 2011), Cytoscape (Smoot et al., 2011), VANTED (Björn et al., 2006) and VisANT (Mellor et al., 2004), four popular multi-platform biological network visualization applications.

2.1 Biological networks as annotated graphs

The most basic mathematical structure common to all of these applications is the graph, a collection of objects connected by links, referred to as nodes and edges. These objects are abstractions of real-world biological entities, where nodes could represent proteins, genes, molecules, drugs, etc. and edges could represent physical protein-protein, metabolic, or genetic interactions, microRNA to target associations, correlation, similarity relationships,

etc. Edges can be directed or undirected, weighted or not. In a case like gene regulation, Gene A may regulate Gene B, but the relationship may not be symmetric, meaning Gene B does not regulate Gene A. These models of biological networks have differing levels of support across various applications. An application may only support a small subset of node and edge types in order to specialize on one particular model, such as VisANT, which integrates many specialized tools for tasks such as Gene Ontology (GO) annotation, name resolution and online searches. Other applications may be more open-ended to provide support for as many models as possible, such as NAViGaTOR, Cytoscape and VANTED. The advantage of such a model is versatility, but it comes at the cost of having to manually define the nature of each node or edge via annotation.

To populate a graph within an application, the application must support one or more input formats. Often, the most basic level of input is either plain text or spreadsheet files such as Excel XLS format. For more graph-specific data, such as layout, GML can be used. To support more complex and structured biological data, several community standards exist: PSI-MI, BioPAX, and SBML.

Adding new nodes and edges to an existing graph can generally be done manually or by adding additional interactions from a supported database or file format. Some applications may have a workspace that supports concurrent, multiple graphs, which can then be combined or compared in various ways. Cytoscape and NAViGaTOR both support this type of workspace.

Once the graph is loaded within an application, a researcher may wish to add additional annotations, such as gene or protein expression, experimental confidence measures or Gene Ontology (The Gene Ontology Consortium, 2000) to their graph objects. Data from in-house sources must generally conform to the application used; generally, this is in the form of spreadsheets or text data with varying degrees of format flexibility. The researcher can also call upon more specialized data from public databases, such as UniProt, Entrez, KEGG or Genbank, either through the import of files or from direct access to the database through the application or a plug-in.

The amount of biological networks available to the researcher is ever expanding, and the size of the networks involved in many types of analysis is in order of thousands of nodes and edges. For example, the yeast interactome comprises 23,918 interactions according to DIP and 152,877 known and predicted interactions in I2D, the Interologous Interaction Database (http://ophid.utoronto.ca/i2d), an integrated database of PPIs from curated databases, experimental sources and predicted interactions (Niu et al., 2010; Brown and Jurisica, 2007; Brown and Jurisica 2005). While the researcher may only be interested in a small portion of the network in question, the scalability of an individual application and its analysis methods to networks of such size can be a considerable advantage.

2.2 Network visualization

Part of the challenge of visualizing a network is the laying out of the graph in a comprehensible manner. For smaller graphs, manual editing of node positions may be sufficient. With the aforementioned instances of graphs in the order of thousands of interactions, more robust tools are available with which to lay out a graph. Automated

graph layout algorithms, such as the force-directed and hierarchical, make the process easier, but often produce messy, uninterpretable graphs. Manual control over the placement of nodes and specialized tools for doing so are often necessary, from simple movement of single nodes to alignments in circles and lines to manipulate groups of nodes.

Algorithms for graph analysis are generally included in each application. Here, the number and type of analyses available are wildly variable. Algorithms can be used to find important graph properties, such as node degree, centrality, shortest paths, cliques and clusters. In addition, diverse biology-specific algorithms exist such as GeneMANIA (Montojo et al., 2010). Some applications may be designed specifically for one type of analysis while others contain a variety of analysis methods and in some cases allow for the addition of third party methods through plug-ins (NAViGaTOR, Cytoscape) or scripting languages (VisANT).

How an application chooses to visualize a graph is also variable. Nodes can be represented as anything from basic geometric shapes with variable size, color and transparency to application specific or user supplied bit-mapped images (Cytoscape, VANTED) or even other data visualizations such as bar charts (VANTED, VisANT). Edges can be straight, curved, displayed with various dot or dash schemes and can have variable widths, colors and transparencies. To make certain attributes readily visible, it is also possible in some instances to map an attribute to a visual property, such as color or size. All four of our example applications have different implementations of such mapping; the utility of a specific implementation is dependent upon the needs and competencies of the individual researcher.

Once the graph satisfies the requirements envisioned by the researcher, its state must be stored or exported. Proprietary formats are generally the norm for most programs, as visualization and data are often application specific and must be stored for later editing. Export formats often take the form of community standards (PSI-MI, BioPAX) and graphical exports. Graphical export is generally the final stage before publication. Usually, this can be done in bitmap (JPEG, TIFF, PNG, etc.) or vector (SVG, PDF, etc) formats, the latter being preferable for publication, as it can be resized and manipulated without loss of quality.

2.3 NAViGaTOR

NAViGaTOR (Network Analysis, Visualization and Graphing Toronto; http://ophid.utoronto.ca/navigator) is a network and graph visualization application with an emphasis on large graphs with integrated data (Brown et al., 2009). Data can be imported using diverse formats, ranging from community standards such as PSI-MI XML (Kerrien et al., 2007), BioPAX (Demir et al., 2010) or GML (Himsolt, 1996), to user-defined text files. Though the application is geared towards protein-protein interactions, the graph implementation within NAViGaTOR is not PPI specific, and can be used to model many types of real world or theoretical objects. Nodes and edges can have data associated with them, from simple numeric or text data to structured XML. Once imported, graphs can be combined from within a multi-graph workspace using combinations of cut, copy and paste operations. Additional data for the annotation of existing graphs can be imported using compatible files or online resources, such as I2D, cPath, or the one of the many online databases implementing the PSICQUIC web service.

Graphs generated by the above methods can quickly increase in size to thousands of nodes and edges. NAViGaTOR was designed with networks of this size in mind. While graphs this

size do create a demand for both memory and processing power to render, layout and navigate, the conservation of important paths and data is important to end-user analysis, particularly since most graphs of interest are subsets of a much larger interaction networks. NAViGaTOR approaches the problem of limited computing resources through the combination of a powerful OpenGL rendering engine through JOGL, and a suite of efficient layout, search and analysis tools. The JOGL rendering system gives the application access to the graphic processing power of the OpenGL compliant hardware of most graphics cards, allowing the application to use the CPU for more intensive graph operations.

NAViGaTOR supports several layout algorithms tailored for large graphs, including GRIP (Graph Drawing with Intelligent Placement) and several variants of the force directed algorithm. These algorithms come in both single and multi-threaded modes to take advantage of computers with multi-core CPUs.

When the structure and data contained within a graph are sufficient, the user can then interact with the graph, identifying significant nodes, edges or subsets of the graph using a variety of searches, spreadsheet tables and algorithms. Online or file supported databases can also be used to indicate known pathways and complexes within the data.

Users can highlight interesting structures within a graph with a variety of methods. Nodes and edges can be assigned visual properties to differentiate them from each other. Nodes can be given different colors, sizes, and highlighting styles. Edges can be given different colors, widths and styles and have the option to be rendered as user adjustable curves. Transparency can be used on both nodes and edges to either increase or decrease the visibility of graph objects.

The user can save the file in native NAViGaTOR format, GML, PSI-MI or delimited plain text. In addition, for presentation or publication purposes, the graph can be exported to one of several graphical formats, including JPEG, PNG, TIFF, SVG and PDF.

3. Iterative expansion of a protein interaction network

The increasing amount of data that can be collected from high-throughput analyses is accelerating research in the field of molecular biology; however, data of this type is also challenging due to its size. It can be used either for knowledge-based targeted analyses, meaning to improve the understanding of the role of an important well-known player in a specific field of interest (for example of BRCA1 in breast cancer), or unbiased analyses to understand the processes involved in a specific behaviour without a priori knowledge (for example, which genes/proteins are responsible for the poor survival of patients with pancreatic cancer?)

For our example, we have a list of potential interactors for a hypothetical protein of interest, PRO1, generated by computational PPI prediction. Also at our disposal are two meta analyses efforts, specifying the number of ovarian or prostate cancer related studies found in which the gene and its interactors were significantly deregulated. All other data will be collected from publicly available resources, including a PPI database, and a catalogue of drugs and their gene targets.

For our example, we will start with our experimental data in a tabular format. Data such as this can be obtained from any number of sources, from high-throughput experiments to

computational predictions. In our case, we have 21,302 predicted PPIs. Our analysis has produced a confidence metric associated with each interaction, ranging from 0 to 1.0. This confidence metric can be used to reduce the number of interactions we are dealing with to a more manageable size by removing lower confidence interactions. Our cut-off for high confidence will be 0.892, a value determined by cross validation. This leaves us with only 39 interactions, a far more manageable number for the next analysis steps. More complex filtering can be done through a simple spreadsheet application, such as Excel, or with a mathematical application such as R or Matlab.

At this point, we translate this data into a pair-wise table of PPIs, and import this table into NAViGaTOR. While NAViGaTOR supports several formats for loading interactions, we have chosen the tab-delimited format to facilitate easy translation from our original data. Other interaction data sets can be imported using community standard file formats, such as BioPAX, GML, PSI-MI XML and PSI-MITAB. Though these formats are harder to construct, they can contain more structured data, and facilitate easier data interchange among diverse programs and databases.

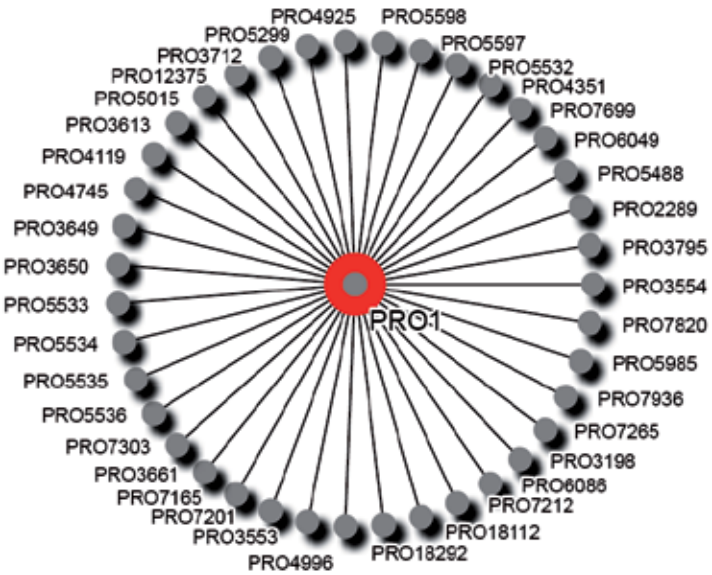


Fig. 1. Example graph containing hypothetical protein PRO1, with interactors loaded from experimental data. Tabular view of the data is available as a supplemental material (http:// http://www.cs.utoronto.ca/~juris/data/intech12/).

Loading our pair-wise data, we get a very basic view (Figure 1). The visualization of this network at this stage is a spoke diagram with PRO1 in the center, and offers little information to the researcher that could not have been seen through a simple spreadsheet. We already have data regarding 39 interactions in the form of the confidence metric imported from our initial study. This can be mapped to one or more visual attributes using NAViGaTORs filter framework. In this case, we can make the highest confidence interactions more visible by applying a filter to map confidence to both edge width and transparency (Figure 2).

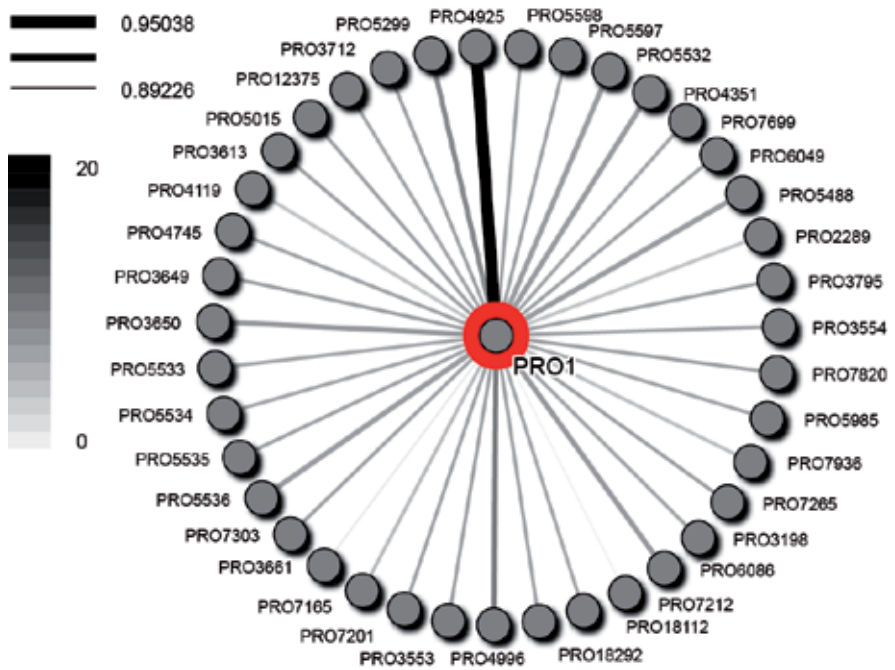


Fig. 2. Example graph with experimental interaction confidence mapped to edge width and transparency.

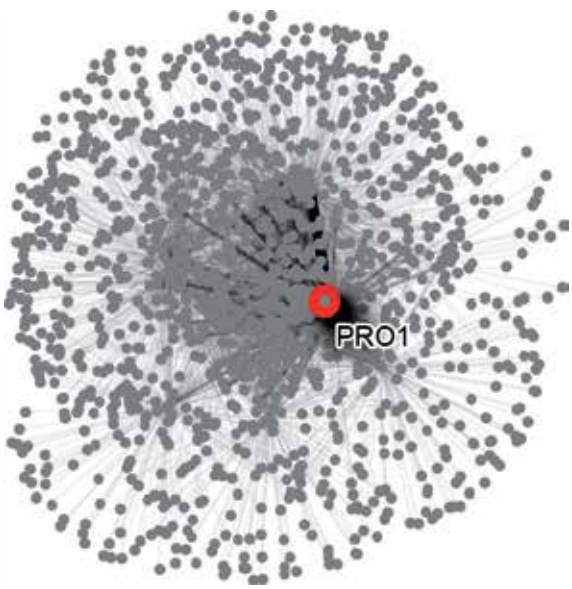


Fig. 3. Example graph enriched with interactions loaded from I2D, and laid out using the GRID algorithm.

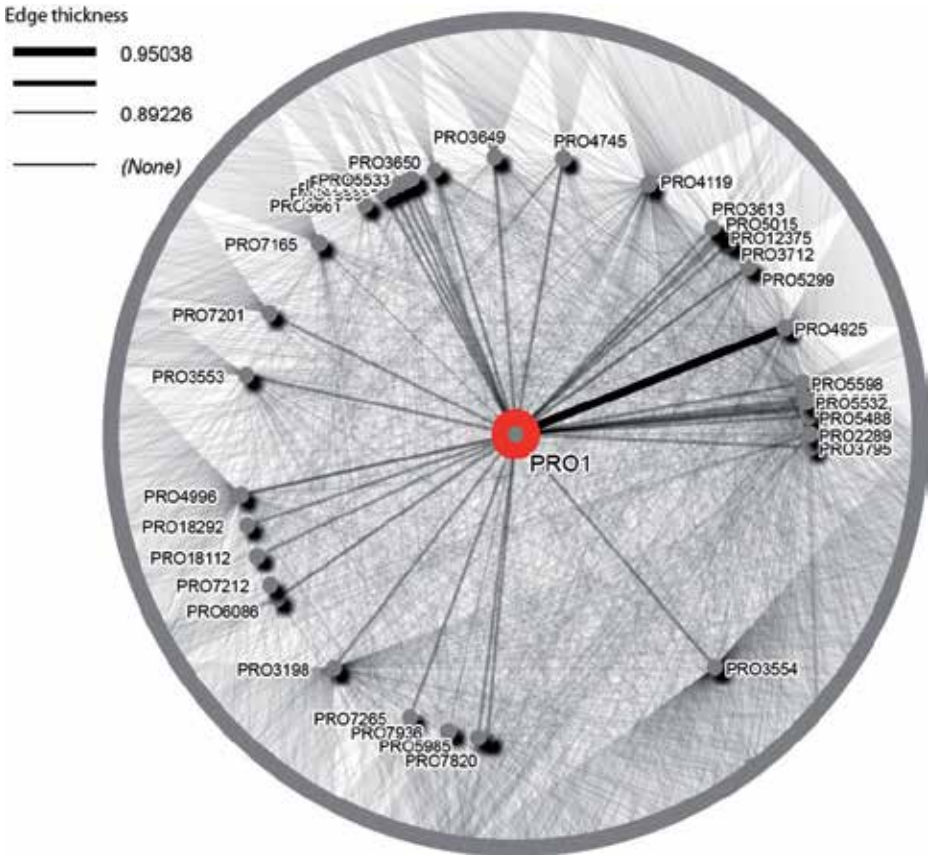


Fig. 4. Example graph laid out hierarchically, with PRO1 in a central position.

This is better, but still not that much more informative. One way of enriching our isolated data is by viewing it in the context of known and predicted interactions. I2D, the Interologous Interaction Database (http://ophid.utoronto.ca/i2d; (Brown et al., 2005, Brown et al., 2007)), will be our source for these interactions. NAViGaTOR offers an I2D plug-in, which enables the researcher to easily add interactions to the existing graph. NAViGaTOR also has the PSICQUIC search plug-in, which supports the searching of databases that implement the PSICQUIC interface (Aranda et al., 2011). To further support the openness and versatility of PPI integration, NAViGaTOR can import additional interactions from the same file formats listed above. If a database does not support any of these formats, finding or building a representation of the database in tab-delimited format may be an option as well. Our interaction search returns 1,367 nodes and 3,192 edges (Figure 3).

At this point, the graph has become more complex, and the force-directed layout is not helpful in interpreting it. Several options exist at this point for manually laying out objects in the graph. The user can select 'fix' nodes within the graph and either move them manually (which would be very labor intensive and inflexible) or lay them out with an array of tools such as linear, circular, arc or radial layout. We will use the radial layout method, starting with PRO1 as our central node and extending to a depth of 2. This gives us a hierarchical arrangement of

nodes starting with PRO1 in the centre, with its immediate interactors arranged circularly around it, and their interactors in turn arranged around them (Figure 4).

3.1 Ambiguity of protein names

When combining data from different sources, the users' choice of protein nomenclature becomes extremely important. Although a researcher knows which genes or proteins they are referring to, queries to a database require additional levels of specificity to resolve ambiguities in entity names.

For example, DLC1 has the following SwissProt identifiers: Q96QB1, Q9Y238, P63167, Q7Z5R8, Q45XF9, Q86UC6. However, names in literature could be ambiguous and confusing, potentially resulting in incorrect interpretation and analyses:

- **DLC1** (ARHGAP7) (KIAA1723) (STARD12) [**Rho GTPase-activating protein 7** (Rho-type GTPase-activating protein 7) (Deleted in liver cancer 1 protein) (Dlc-1) (StAR-related lipid transfer protein 12) (START domain-containing protein 12) (StARD12) (HP protein)]
- **DLEC1** (DLC1) [**Deleted in lung and esophageal cancer protein 1** (Deleted in lung cancer protein 1) (DLC-1)]
- **DYNLL1** (DLC1) (DNCL1) (DNCLC1) (HDLC1) **[Dynein light chain 1, cytoplasmic** (Dynein light chain LC8-type 1) (8 kDa dynein light chain) (DLC8) (Protein inhibitor of neuronal nitric oxide synthase) (PIN)]
- **DLC1** [**Deleted in liver cancer 1** variant 2 (Fragment)]
- **DLC1** [**DLC1** protein]

Similarly, many papers refer to SHC – but details about which variant and which species are frequently "hidden" in the supplemental information (http://www.cs.utoronto.ca/~juris/data/intech12/). Yet, there are at least four variants in mouse and human. Sometimes, a radical change in nomenclature is required, such as in case of Caspases (Alnemri et al., 1986). Systematic analysis led to redefying various ICE, MACH, MCH genes into Caspase1-10 (Alnemri et al., 1986).

There are many different standards of referring to genes and proteins: UniProt (http://www.uniprot.org) (Jain et al., 2009), Ensembl (http://www.ensembl.org) (Flicek et al., 2012), EBI IPI (http://www.ebi.ac.uk) (Kersey et al., 2004), Gene Cards (http://www.genecards.org) (Safran et al., 2010), NCBI Gene (http://www.ncbi.nlm.nih.gov) (Maglott et al., 2010) are just a few examples of databases that attempt to systematically characterize and describe genes and proteins. Each database has its own focus and strengths, and different interaction or annotation databases may choose any one of these standards to organize their data. In this example, and in many other case uses of NAViGaTOR, the user may have to import data from one or more databases that use different nomenclatures. To facilitate the use of multiple nomenclatures, NAViGaTOR can store multiple IDs per node as a text feature, allowing alternative keys for node identification. When combining data from two or more databases using different formats, the user must translate between these different nomenclatures. This must be done very carefully and methodically, as this additional translation step often effects the data returned. For example, UniProt stores mappings from its own accession IDs to Ensembl

Gene IDs, and Ensembl stores mappings from its own IDs to UniProt. However, respectively, they return 55,639 unique UniProt accession IDs for 20,995 unique Ensembl gene IDs and 21,735 unique Ensembl gene IDs for 63,370 unique UniProt accession IDs. The mapping is clearly different depending on which method is used. There is no definitive mapping available in situations such as these: it is up to the individual researcher to choose and document the translations used to amalgamate their data in a fashion that is replicable. Bearing this in mind during the earlier stages of experiment design will make this process much easier and less prone to confusion or ambiguity.

3.2 Associating data with an existing graph

Though better organized, we still have in excess of 1,000 nodes and 3,000 interactions, and to better identify nodes and edges that represent novel research material, we must associate more data with those objects.

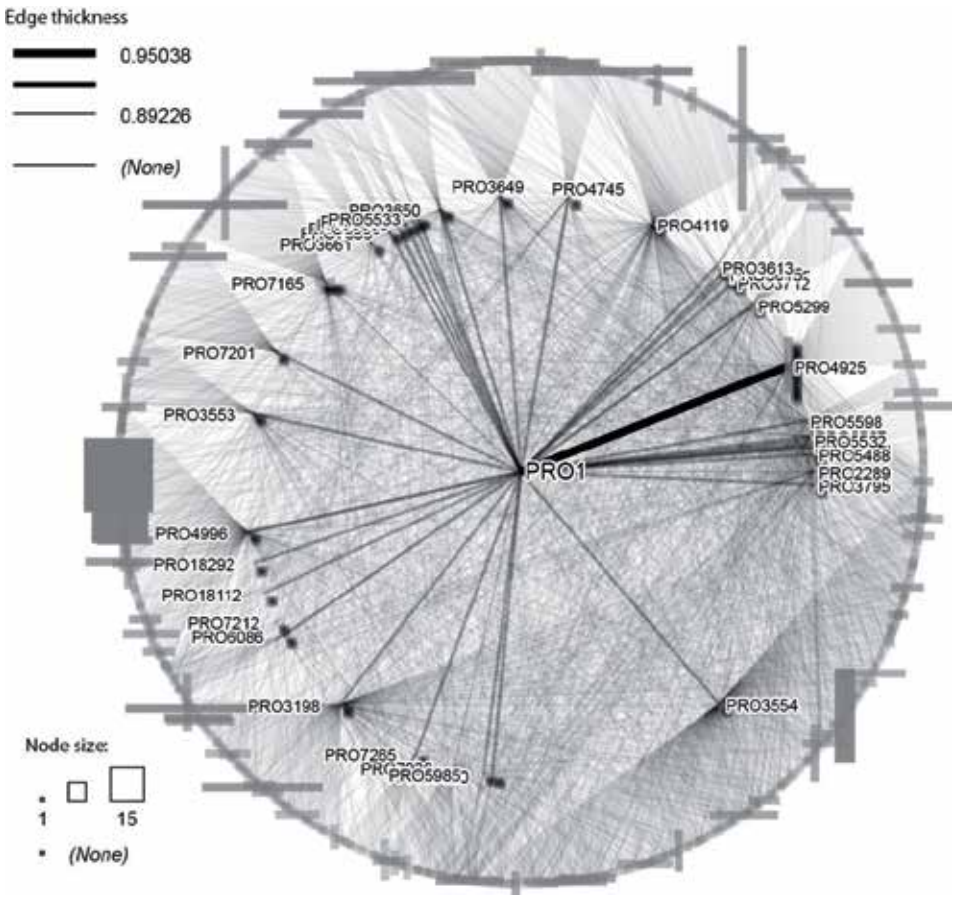


Fig. 5. Example graph with numbers of referencing studies in ovarian and prostate mapped to node width and height.

We can for example integrate PPIs with the gene expression results obtained from our literature studies. Each file contains several values associated with each gene, specifying the number of studies in which the gene was down-regulated, up-regulated and a total representing both (Figure 5). We will also generate a third file representing the total studies in which the gene was found to have been significantly deregulated, which simply sums the totals for the previous two files. Similarly to the opening of the initial experiment, NAViGaTOR requires a unique identifier column to be specified. In this case, because we are only concerned with data to be associated with nodes, the program only requires a single Node ID column. This process is the same for the prostate, ovarian cancer and generated data sets. To visualize this data, we will add another filter, this time mapping the total number of significantly deregulated studies in ovarian cancer to node width, and the total number of significantly deregulated studies in prostate cancer to its height. It is immediately evident which nodes have already been described to be up/down regulated in either one or both types of cancer. This can be useful to parallel the information already known from one cancer to the other. In addition, we can map the generated total of studies to node transparency, making genes with less disease evidence less obtrusive.

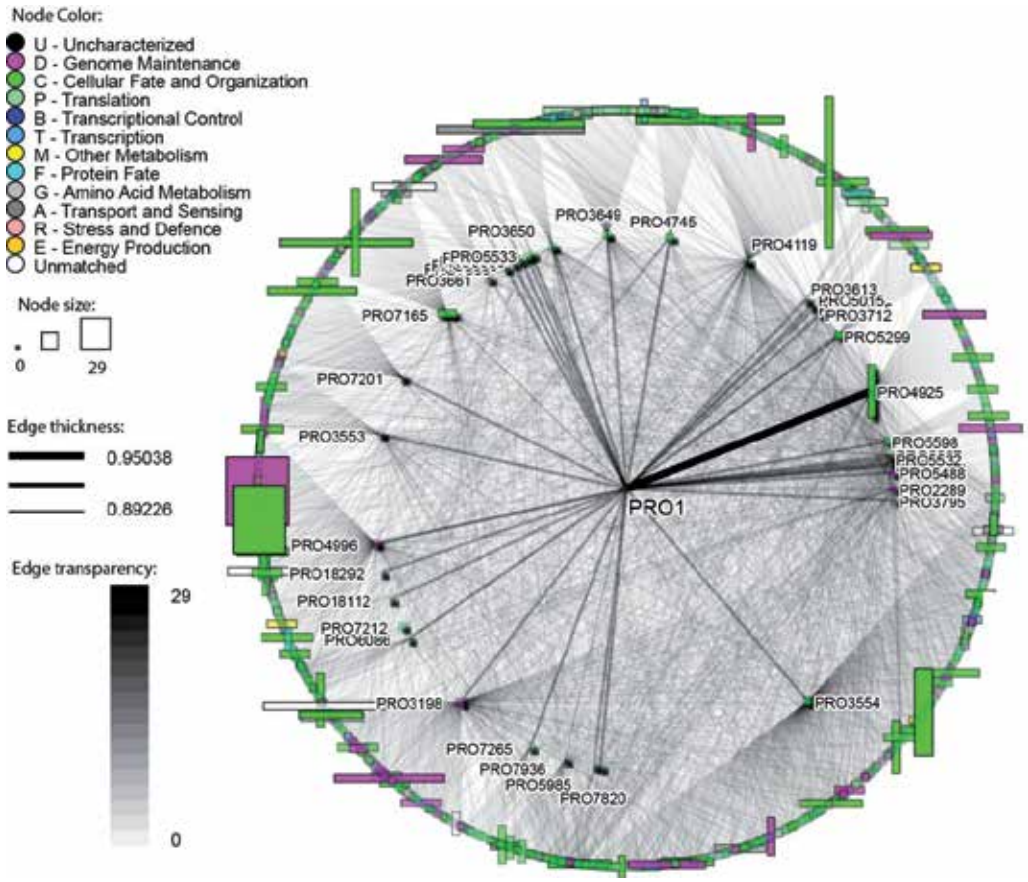


Fig. 6. Example graph with GO Annotation mapped to a color scheme.

We can also import structured data, in the form of GO attributes, retrieved from the I2D plug-in(Figure 6). We can view this data per individual node in the Node side panel, revealing the list of individual GO attributes and their descriptions. To get a graph-wide view of these attributes, we will add a filter to map the GO data to one of several categories, each with its own colour. The same result can be obtained by applying GO terms or other attributes, like pathways to which the node belongs, retrieved from other sources to the nodes as features and editing the filter in the desired way.

3.3 Importing drug-protein interactions

Finally, we will import a list of drugs and their gene targets as additional interactions. This expands our network to 2,707 nodes and 5,257 edges (Figure 7). Through a combination of manual layout and radial layout tools, we arrange the drugs in a circle around PRO1, its interactors, and their interactors from I2D. The edges connecting drugs to proteins are coloured blue to differentiate them from PPIs. To see the impact of individual drugs to this network, we map their degree to node size and transparency. Thus, large nodes represent drugs that target many of the proteins in the network. The top six of these drugs are labelled for convenience. Analogously, some proteins have a high degree of blue edges and connect to small nodes, such as ProX. These drugs show strong specificity to ProX. The initial data will be available in ASCII tab-delimited format and the final figure in NAViGaTOR 2 XML file at http://www.cs.utoronto.ca/~juris/data/intech12/.

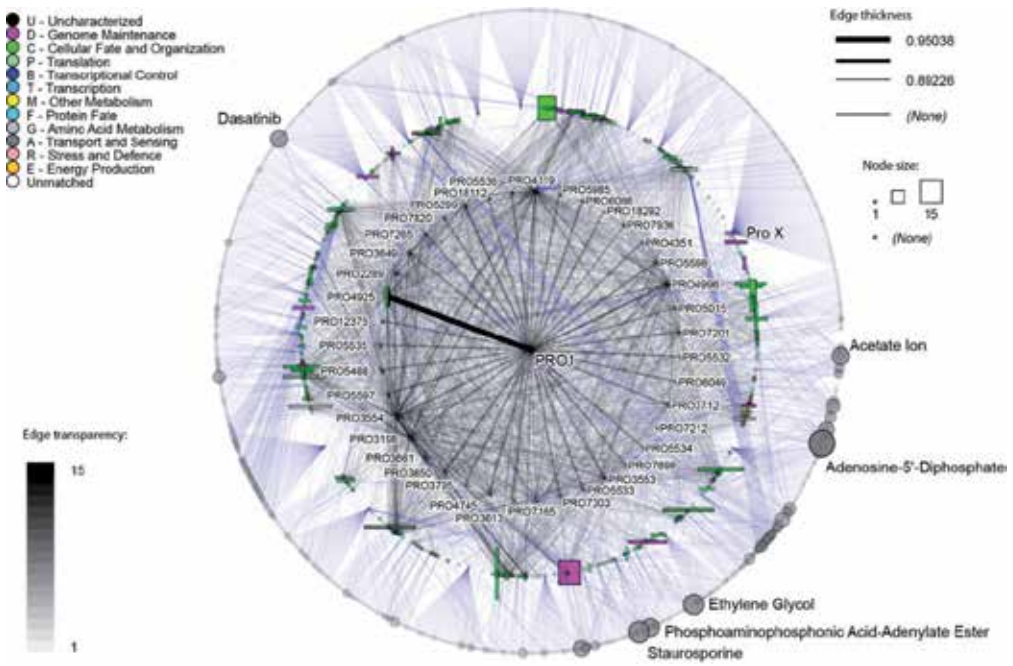


Fig. 7. Final graph, with drug interactions included and the size of nodes representing drugs derived from number of interactions within the graph. NAViGaTOR 2 XML file for the final figure is available in supplemental material (http://www.cs.utoronto.ca/~juris/data/intech12/).

4. Conclusions

Integrated databases and resources are only useful when they can be effectively accessed, navigated and analyzed. Several biological network visualization tools are currently available, providing a diverse range of approaches and algorithms. While many existing visualization tools are effective and widely used, there are several critical areas where these applications require improvement. Scalability is essential to visualize the tens of thousands of known PPIs, which is a challenge for current layout algorithms and software. Biological graph drawing software must also be able to handle richly annotated data, including genomic and proteomic profiles, pathways, Gene Ontology annotations and data in PSI-MI and BioPAX formats, in addition to the vast quantity of microarray and proteomic data that is available.

Individual tools need a good balance of performance and useful features. The features that are needed for each use are highly dependent on the available data and the workflow. As in any creative activity, a tool may enable new workflows by providing novel features, but the tool may also lack certain important features, or offer features that are not needed. There is no single solution that satisfies all of these requirements at the present time, and as data and workflows change over time, network visualization tools must also evolve.

As the data grow more complex, the performance of layout algorithms will need to improve, and new options of differentiating multiple attributes will be required. As certain workflows become more main-stream, they may be turned into *analysis patterns* and implemented as plug-ins. Standardizing file formats, APIs and plug-ins will further intertwine existing tools, enabling their easier integration and specialization.

With new data and advances in computational biology, user tasks are modified, which must be reflected by types of algorithms that support analyses and the user interfaces that effectively enable them. New graph theory algorithms for faster and biologically meaningful network layouts and algorithms for network structure analysis will need to be integrated into network visualization tools. Importantly, none of these algorithms would make a broad difference unless a user interface appropriate for biologists is available (Viau et al., 2010).

5. Acknowledgment

This research was funded in part by Ontario Research Fund (GL2-01-030), Canada Foundation for Innovation (CFI #12301 and CFI #203383), and the Ontario Ministry of Health and Long Term Care. The views expressed do not necessarily reflect those of the OMOHLTC. CP was funded in part by Friuli Exchange Program. IJ is supported in part by the Canada Research Chair Program.

The authors would like to thank Max Kotlyar, Dan Strumpf, Fiona Broackes-Carter and the entire Jurisica lab for useful comments and discussions.

6. References

Alnemri, Emad, S., David J Livingston, Donald W Nicholson, Guy Salvesen, Nancy A Thornberry, Winnie W Wong, Junying Yuan (1986). Human ICE/CED-3 protease nomenclature, *Cell*, 87(2):171.

Aranda, B.,Blankenburg, H.,Kerrien, S.,Brinkman, F.S.L., Ceol, A., Chautard, E., Dana, J.M.,De Las Rivas, J., Dumousseau, M.,Galeota, E., Gaulton, A., Goll, J., Hancock, R.E.W., Isserlin, R., Jimenez, R.C., Kerssemakers, J., Khadake, J., Lynn, D.J., Michaut, M.,O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G.D., Cesareni, G., Donaldson, I.M., Eisenberg, D., Kleywegt, G.J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M.,Hermjakob, H. (2011). PSICQUIC and PSISCORE: Accessing and scoring molecular interactions. *Nature Methods*, 8(7): 28-529.

Björn H. Junker, Christian Klukas and Falk Schreiber (2006). VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109

Brown, K.R., and Jurisica, I. (2005). Online Predicted Human Interaction Database. *Bioinformatics*, 21(9):2076-82.

Brown, K.R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 8(5):R95.

Brown, K.R., Otasek D, Ali M, McGuffin, M.J., Xie W, Devani B, van Toch I.L., and Jurisica I. (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto, *Bioinformatics,* 25(24): 3327-3329.

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*. 28(9):935-42.

Djebbari, A., Ali, M., Otasek, D., Kotlyar. M., Fortney, K., Wong, S., Hrvojic, A. and Jurisica, I. (2011). NAViGaTOR: Scalable and Interactive Navigation and Analysis of Large Graphs. *Internet Mathematics*, 7(4):314-347.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM. (2012). Ensembl. *Nucleic Acids Res*. [Epub ahead of print]

Gehlenborg N., O'Donoghue S.I., Baliga N.S., Goesmann A., Hibbs M.A., Kitano H., Kohlbacher O., Neuweger H., Schneider R., Tenenbaum D., Gavin A.C. (2003). Visualization of omics data for systems biology. *Nat Methods,* 7(3 Suppl):S56-68.

Helaers R, Bareke E, De Meulder B, Pierre M, Depiereux S, Habra N, Depiereux E. (2011). gViz, a novel tool for the visualization of co-expression networks. *BMC Res Notes*. 4(1):452.

Himsolt, M. (1996). GML: A portable Graph File Format. Syntax. Retrieved from http://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/brandenburg/projekte/gml/gml-technical-report.pdf

Himsolt, M. (1996). GML: A portable Graph File Format. Syntax. Retrieved from http://www.fim.uni-passau.de/fileadmin/files/lehrstuhl/brandenburg/projekte/gml/gml-technical-report.pdf

Hu Z., Hung J.H., Wang Y., Chang Y.C., Huang C.L., Huyck M., DeLisi C. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res,* 37, W115–W121.

Hu, Z., Mellor, J., Wu, J. and DeLisi, C. (2004). VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5, 17.

Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B.E., Martin M.J., McGarvey P., Gasteiger E. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, 10:136.

Junker B.H., Klukas C. & Schreiber F. (2006). VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics,* 7(109).

Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D, Tyers M, Salama JJ , Moore S, Ceol A, Chatr-Aryamontri A, Oesterheld M, Stümpflen V, Salwinski L, Nerothin J, Cerami E, Cusick ME, Vidal M, Gilson M, Armstrong J, Woollard P, Hogue C, Eisenberg D, Cesareni G, Apweiler R, Hermjakob H (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol*. 5, 44

Kersey P. J., Duarte J., Williams A., Karavidopoulou Y., Birney E., Apweiler R. (2004). The International Protein Index: An integrated database for proteomics experiments. *Proteomics* 4(7): 1985-1988

Kreeger P.K., Lauffenburger D.A. (2010). Cancer systems biology: a network modeling perspective*, Carcinogenesis*, 31(1):2-8.

Longabaugh WJ. (2012). BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. *Methods Mol Biol*. 786:359-94.

Maglott D, Ostell J, Pruitt KD, Tatusova T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 39(Database issue):D52-7.

McGuffin, M, and Jurisica, I. (2009). Interaction techniques for selecting and manipulating subgraphs in network visualizations. *IEEE Trans Vis Comput Graph*, 15 (6): 937-944.

Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD (2010). GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, 26: 22

Morrow JK, Tian L, Zhang S. (2010). Molecular networks in drug discovery. *Crit Rev Biomed Eng*. 38(2):143-56.

Niu, Y., Otasek, D., Jurisica, I. (2011). Evaluation of linguistic features useful in extraction of interactions from PubMed; Application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, 26(1): 111-9.

Pavlopoulos G.A., Wegener A.L., Schneider R. (2008). A survey of visualization tools for biological network analysis, *BioData Min,* 1(12).

Remmerie N., De Vijlder T., Laukens K., Dang T.H., Lemière F., Mertens I., Valkenborg D., Blust R., Witters E. (2011). Next generation functional proteomics in non-model plants: A survey on techniques and applications for the analysis of protein complexes and post-translational modifications. *Phytochemistry*, 72(10):1192-218.

Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A and Lancet D. (2010). GeneCards Version 3: the human gene integrator *Database*; doi: 10.1093/database/baq020

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13:2498–2504.

Shirdel EA, Xie W, Mak TW, Jurisica I. (2011) NAViGaTing the micronome--using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs. *PLoS One*. 6(2):e17429.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 1;27(3):431-2.

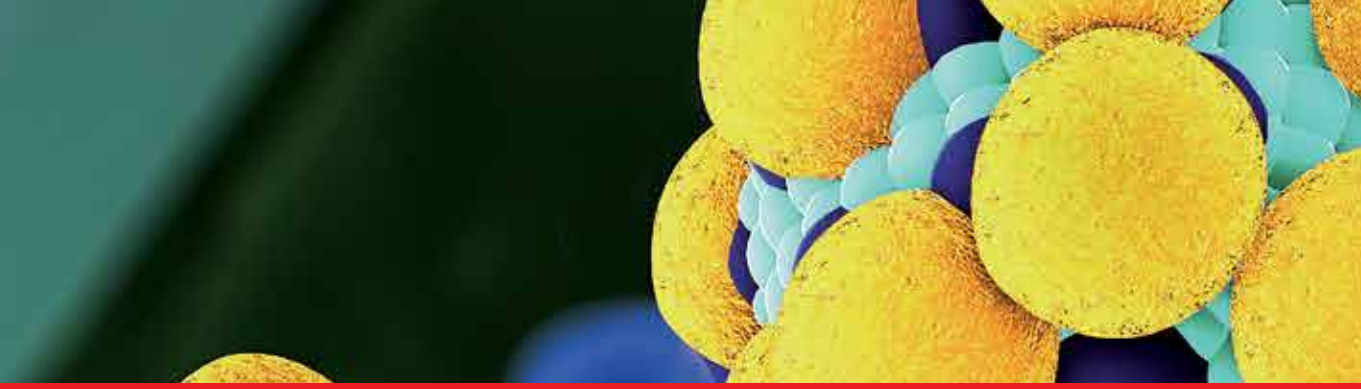
Stein A., Mosca R., Aloy P. (2011). Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr Opin Struct Biol*, 21(2):200-8.

Swainston N, Smallbone K, Mendes P, Kell D, Paton N. (2011). The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform*. 8(2):186.

The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet*. 25(1):25-9.

Viau, C., McGuffin, M J., Chiricota, Y., and Jurisica, I. (2010). The FlowVizMenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration. *IEEE Trans Vis Comput Graph*, 16(6):1100-8.

Yu L.R. (2011). Pharmacoproteomics and toxicoproteomics: The field of dreams. *J Proteomics*, 74(12):2549-53.



Edited by Weibo Cai and Hao Hong

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

ISBN 978-953-51-4312-3

Photo by Ugreen / iStock

IntechOpen