IntechOpen

Bioinformatics in the Era of Post Genomics and Big Data

Edited by Ibrokhim Y. Abdurakhmonov

BIOINFORMATICS IN THE ERA OF POST GENOMICS AND BIG DATA

Edited by **Ibrokhim Y. Abdurakhmonov**

Bioinformatics in the Era of Post Genomics and Big Data
http://dx.doi.org/10.5772/intechopen.71349
Edited by Ibrokhim Y. Abdurakhmonov

Contributors

Etsuko N. Moriyama, Adam Voshall, Sergio Juárez-Méndez, Vanessa Villegas-Ruíz, Elishai Ezra Tsur, Jing-Doo Wang, Manju Bansal, Aditya Kumar, Satoshi Mizuta, Alain Sewer, Marja Talikka, Florian Martin, Julia Hoeng, Manuel C. Peitsch, Minja Zorc, Jernej Ogorevc, Peter Dovč, Emmanouil Malandrakis, Olga Dadali

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in London, United Kingdom, 2018 by IntechOpen
eBook (PDF) Published by IntechOpen, 2019
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

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

Bioinformatics in the Era of Post Genomics and Big Data
Edited by Ibrokhim Y. Abdurakhmonov

p. cm.

Print ISBN 978-1-78923-268-4

Online ISBN 978-1-78923-269-1

eBook (PDF) ISBN 978-1-83881-563-9

Meet the editor

Ibrokhim Y. Abdurakhmonov received his BS degree (1997) in Biotechnology from the National University of Uzbekistan, MS degree (2001) in Plant Breeding from the Texas A&M University of the USA, PhD degree (2002) in Molecular Genetics, Doctor of Science degree (2009) in Genetics, and full professorship (2011) in Molecular Genetics and Molecular Biotechnology from the Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan. He founded (2012) and is currently leading the Center of Genomics and Bioinformatics of Uzbekistan. He serves as an associate editor and an editorial board member of several international and national journals on plant sciences. He received the Government Award, 2010 Chest Badge "Sign of Uzbekistan," 2010 TWAS Prize, and "ICAC Cotton Researcher of the Year 2013" for his outstanding contribution to cotton genomics and biotechnology. He was elected as the World Academy of Sciences (TWAS) Fellow (2014) on Agricultural Science and as a co-chair/chair of "Comparative Genomics and Bioinformatics" workgroup (2015) of International Cotton Genome Initiative (ICGI). He was elected (2017) as a member and the Vice-president of the Academy of Sciences of Uzbekistan. He was appointed (2017) as a minister of Innovational Development of Uzbekistan.

Contents

Preface

Bioinformatics, being an interdisciplinary scientific field, has evolved significantly from the simple analyses of few DNA/protein sequences to the handling of a large volume of entire genomic sequences, proteomic and metabolomic networks, and complex genetic pathways of biological samples.

Even further, huge advancements were made toward storing, handling, mining, comparing, extracting, clustering and analysis as well as visualization of big macromolecular data using novel computational approaches, machine and deep learning methods, and web-based server tools. There are extensively ongoing world-wide efforts to build the resources for regional hosting, organized and structured access and improving the pre-existing bioinformatics tools to efficiently and meaningfully analyze day-to-day increasing Big Data.

In this book, *Bioinformatics in the Era of Post Genomics and Big Data*, we intended to provide the reader the latest advances of bioinformatics science in the era of post genomics and big data. This is to motivate new generation researchers to efficiently mine this big data and generate meaningful results, enabling "translational bioinformatics."

Toward this goal, here we successfully compiled 9 chapters covering topics such as new generation transcriptome assembly, gene expression analysis, genome-wide association, novel approaches for mining genetic markers and visualization of biological sequences using the latest advance in bioinformatics. Chapters also discussed advances in data modeling and network-based systems using bioinformatics tools.

Although limited to specific topics, chapters do represent interesting aspects of bioinformatics studies of the present time, which should be useful and helpful for scientists, students and readers of life science direction.

I would like to thank all the authors of the book chapters for their valuable contributions. I would also like to thank the IntechOpen book department for giving me the opportunity to work on this book project, and Ms. Maja Bozicevic, IntechOpen's Publishing Process Manager, for her coordination of my book editing process.

Ibrokhim Y. Abdurakhmonov
Center of Genomics and Bioinformatics
Academy of Sciences of Uzbekistan
Tashkent, Uzbekistan

Section 1

Genomics Data Analyses

Chapter 1

Genome-Guided Transcriptomics, DNA-Protein Interactions, and Variant Calling

Emmanouil E. Malandrakis and Olga Dadali

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76842

Abstract

Nowadays, molecular biology has definitely become an interdisciplinary science. Toward the study of the functions and the interactions of the biological molecules, such as nucleic acids and proteins, computer science and engineering, along with chemistry and statistics, are routinely engaged. In molecular biology, techniques and methods are constantly developed, and new techniques emerge. Next-generation sequencing and bioinformatics have become the cornerstones of molecular biology. The developing technologies have led to a decrease of the cost per molecular unit analyzed, but at the cost of computer integration and intensification. Many research methods require a reference nucleic acid sequence. Considering the necessary integration of sequencing data and methodology, combining the "omics" approaches can help to elucidate more complex null hypotheses. Here, data processing basics, with an emphasis to commonly used techniques, are summarized. The knowledge gaps are discussed as well as further prospective for integrating next-generation sequencing data.

Keywords: next-generation sequencing, data analysis, Unix, scripting

1. Introduction

The study of the functions and the interaction of the biological molecules such as nucleic acids and proteins has become a daily laboratory routine. By recently, Sanger sequencing was extensively used to uncover new genomic sequences. Sanger sequencing method was named after Frederick Sanger (1918–2013), the British biochemist who invented it and won the Nobel Prize in Chemistry for the second time (1980). Until now, the method is based on PCR amplification and capillary electrophoresis. Each sequencing reaction generates a ladder of ddNTP-terminated, dye-labeled products, which then are submitted to high-resolution electrophoretic separation within one of 96 or 384 capillaries in one run of a sequencing instrument. The generated fragments are labeled with fluorescent substances and pass the laser, which allows the four different nucleotides to excite and emit different colors of the light spectrum. A camera then captures the colors, and the results are extracted in various formats for further analysis. The analysis of Sanger sequencing data is more or less a straightforward procedure. The sequences can be optically validated and cross-checked through the chromatogram. High-quality reads should not contain ambiguities, and the peaks must be well spaced. On the other hand, poor quality reads have low signal/noise ratio, overlapping peaks and low confidence score. Consequently, a comparison of our sequence can be done with Basic Local Alignment Search Tool (BLAST) by NCBI. BLAST is the cornerstone of sequence analysis, since it facilitates the comparison among amino acid or nucleotide sequences.

Next-generation sequencing (NGS) is an emerging technology with high-throughput outcome. Recently, the rapid development of high-throughput technologies has led to advances in the study of genome function. High-throughput molecular techniques are generally used to study nucleic acids of different species such as DNA, mRNA, lncRNA, etc. Typically, the nucleic acids are fragmented, amplified (or not), and sequenced using various technologies. Although nucleic acid extraction and sequencing are a typical workflow for many laboratories worldwide, integrative software to deal with the analysis workflow in a user-friendly manner is scarce. An intermediate user who is thinking about starting a new NGS project has to set up a Unix-based server with the appropriate software toolkit in order to deal with a huge amount of data. A basic knowledge of R programming language is essential, and the bioconductor project includes an important amount of applications for NGS data processing [1]. Moreover, essential knowledge of scripting languages (Perl, Python) is necessary. In a few words, the data resulting from the sequencer have to be quality checked, filtered, and finally evaluated. This can be achieved on a substantial bioinformatic level.

2. Transcriptomics

Transcriptome sequencing (transcriptomics) enables the characterization of all RNA transcripts for a given organism, including both the coding mRNA and noncoding RNA. For many years, our knowledge on the transcriptome was derived from cloning and sequencing of individual cDNA sequencing. Therefore, it was limited, low-throughput, and partial. However, transcriptomics with next-generation sequencing (NGS) and RNA-Seq is able to increase our knowledge on the dynamic RNA landscape. Compared to the limited capability of Sanger sequencing, a typical RNA-Seq experiment can provide an integrated snapshot of an organism's transcriptome. Normally, regarding RNA-Seq, there are two major experimental setups: de novo assembly of the transcriptome and reference-guided assembly. The former is adequate either when reference genome (or transcriptome) is not available or we want to expand the existing knowledge of an organism's transcriptome. Furthermore, it is mainly utilized in cancer transcriptomics to find fused transcripts or in organisms with trans-splicing. A typical analysis pipeline is presented in **Figure 1**. mRNA sequencing has many advantages over conventional methods. Gene expression can be accurately quantified, and genes with alternative

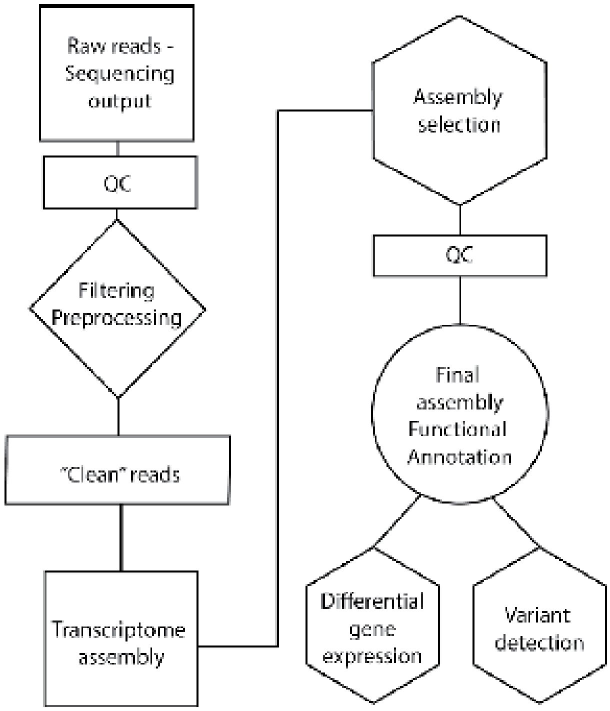


Figure 1. Typical workflow of an RNA-Seq experiment.

splice variants can be identified. Furthermore, reconstructing a transcriptome from short reads is really challenging in terms of computer resources.

To ensure a high-quality transcriptome assembly, the design of the experiment should be carefully designed. If differential expression analysis is planned, biological replication is vital. As a rule of thumb, three to four biological replicates are adequate, but it depends upon the specific experiments. Technical replicates are not essential, but can be used to check for any barcoding effects on results. Usually, technical replicates are highly reproducible (e.g., [2]). In Illumina platforms multiplexing samples are useful for two reasons. Firstly, if a lane fails to produce data, it is still likely that many results from the other lanes can be extracted. Secondly, technical replicates are produced and barcoding effect can be determined.

The very first step upon the receipt of the sequenced data is a secure backup. No one wants to lose precious samples and hundreds of man-hours due to a failure of a hard disk. Data can be stored (and published) in public databases such as NCBI SRA. The reads can be recovered from the database, and a set of metadata is available detailing the experimental conditions.

2.1. Material

A typical flowchart of RNA-Seq experiments usually includes RNA extraction, which must be of high purity and integrity. Most sequencing companies recommend an RNA integrity number (RIN), using Agilent Bioanalyzer 2100 higher than 8. Except for difficult materials (i.e., FFPE samples, fossils), normally preserved samples could easily achieve this score, with standard extraction protocols. Two types of protocols are used for RNA extraction: affinity based (in column) and organic extraction (phenol, chloroform, isoamyl alcohol). The former is compatible with various sample types (animal tissue, plant cells, bacteria, yeast, etc.). Furthermore, DNAse treatment which eliminates contaminating genomic DNA is highly facilitated in a column-based extraction. In this way, excellent RNA purity and integrity are achieved. In addition, automated RNA extraction process is able to reduce working time and at the same time provides opportunity in increasing reproducibility and quality of results [3]. Starting from total RNA, two strategies are available for RNA-Seq: enrichment for mature transcripts using poly(A) tails and depletion of abundant ribosomal sequences. In that way, mature mRNA is abundant in the sample for further processing. Since rRNA represents the 80% of the total RNA and mRNA is 5%, mRNA enrichment is crucial in order to achieve a decent sequencing depth. Sequencing depth is the mean number of times that each nucleotide is sequenced. This stands only for genome, where nucleotides remain relatively stable. For transcriptome, differential expression plus biases in sample processing and sequencing can result in genes with lack of coverage.

2.2. Data quality control and filtering

There are numerous pipelines that check the quality of the data produced by the sequencer. Although millions of reads are typically produced by high-throughput sequencers, simple quality controls are essential in order to be sure that the data could be further processed. If any problems or biases are spotted in the dataset, corrective measures can be taken in most cases. FastQC [4] is a very fast and reliable application that can process different data formats such as *fastq*, compressed *fastq*, *SAM*, and *BAM*. By using simple bash scripts, one could easily analyze multiple datasets at once. FastQC can run in a graphical user interface (GUI) environments even in Unix platforms. An application designed to better group FASTQC result data in whole experiments is FQC [5]. The results are stored in simple *html* files and can be viewed with any web browser available. The output includes simple statistics such as the number of reads, sequence length, etc. A more important statistic for the quality of the available reads is the diagram of the quality score over the nucleotide position in the sequence. In the *fastq* format, each read is tagged with a quality score known as Phred quality score. In general, a Phred quality score of 10 means that there is a possibility of the called base being correct of 90%, 20 is 99%, 30 is 99.9%, etc. As a rule of thumb, bases with score over 20 are considered as bases of good quality.

RNA-Seq reads need further preprocessing before assembly and gene expression analysis. Usually, $5'$ or $3'$ ends present lower-quality or ambiguous sequences. Consequently, these reads are trimmed at both ends. In case the reads have more low-quality or ambiguous nucleotides, they are totally excluded from the analysis. Some good tools for the preprocessing of data include PRINSEQ [6] and Trimmomatic [7]. Although rRNA is routinely removed during library preparation, many sequences are present in raw reads. An efficient tool for rRNA removal is SortMeRNA [8]. SortMeRNA leverages public rRNA databases such as SILVA [9] and Greengenes [10], to identify rRNA sequences. Firstly, developed for metagenomic studies, the software has the ability to extract rRNA sequences in fastq files for further processing. Although designed for single fastq files, two scripts that split and merge back paired-end files, respectively, are provided. Another one critical step during sequence filtering is adapter and primer removal from the dataset. One could combine adapter clipping with quality trimming with Trimmomatic. Cutadapt [11] is definitely a dedicated tool for adapter and PCR primer clipping and removing. The software supports flexible scanning and removal of contaminating sequences.

Consequently, following the previous steps, the data could be used for further processing. In each step, a quality control could be adequate to safeguard the efficiency of process. As a result, the user is able to further process the data or repeat the step, with different settings, until the results are satisfying.

2.3. De novo transcriptome versus genome-guided assembly

De novo transcriptome assembly is a computer-intensive process. Despite the constant increase of available tools, transcriptome assembly from short reads still remains a very challenging process. Probably, the most popular tool for transcriptome assembly is Trinity [12]. Beyond assembly, Trinity incorporates many post-assembly tools which include assembly QC, full-length transcript analysis, abundance, and differential gene expression analysis. Furthermore, protein-coding analysis and functional annotation software are included.

Evaluating a de novo transcriptome assembly is really a hard job. There is a plethora of metrics to assess the accuracy and completeness of a transcriptome assembly. Honaas et al. [13] concluded that a *combination* of metrics can be used in the following order: a number of reads mapping to the assembly; recovery of conserved, widely expressed genes; N_{50} length statistics; and the total number of unigenes. A number of tools are available for this purpose such as BUSCO [14], DETONATE [15], and TransRate [16].

On the other hand, reference-guided transcriptome assembly could be very solid. However, the accuracy of reference-based transcriptome assembly depends on correct read alignment and genetic variants such as alternative splice variants, CNVs, etc. Transcripts are distinguished from the reference genome by Cufflinks [17], and supporting applications in the suite can be used for further analysis.

2.4. Read mapping and counting

The first major data processing step in sequencing studies for species with a reference genome is the mapping of sequencing reads to the reference (genome or transcriptome). Mapping of the reads is defined as the prediction of the loci from which the reads originate. There are many alignment algorithms such as BWA [18] and Bowtie [19] which are unspliced read mappers and TopHat [20] which is a spliced one. The choice of aligner often influences the final results, as different algorithms show various false-positive and/or false-negative rates. There is no single mapper that can align all reads to a reference. This could be due to sequencing errors or polymorphic loci in the reference. Indeed, unmapped reads could be analyzed for identification of such variants. After mapping, a consequential SAM (Sequence Alignment/Map) file result is typically converted to the compressed binary version BAM. This is typically achieved with the *samtools view* command [21].

The differential expression analysis of NGS data includes the counting of the mapped reads on the transcripts. The Trinity suite incorporates various applications for further analysis of transcriptome data. Four tools are employed for read counting, namely, alignment-based tools RSEM [22] and eXpress [23], as well as alignment-free kallisto [24] and salmon [25].

2.5. Differential expression analysis

The output from read counting is a matrix of raw counts that is used as input for R-based software such as DESeq [26], DESeq2 [27], or edgeR [28]. Since each software draws upon different statistical methods, differences in outputs may arise. Furthermore, there are available capabilities of clustering the differentially expressed genes or plotting the results in diagrams. One of the most favorite plots available is the heat map, where the expression of the genes is presented in colors that distinguish the control from the treatment groups. Typically, a palette of red and blue colors is used to indicate up- and downregulation, respectively. Finally, the most important thing on the whole procedure is the discovery of genes (or gene clusters) associated with the biological questions posed.

3. DNA-protein interactions

3.1. Introduction

Proteins bind DNA in order to regulate genome function. Among the proteins that bind DNA, most characteristics are the transcription factors (TFs). Transcription factors regulate transcription by switching on and off genes. They act either alone or synergistically with other proteins as cofactors. Furthermore, groups of TFs function in a coordinated fashion to trigger many fundamental genomic processes (cell division, cell death, development) and periodically in reaction to signals coming from outside the cell.

DNA-protein interactions can be studied by using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) [29]. Accordingly, RNA-protein interactions can be unveiled using cross-linking immunoprecipitation (CLIP), RNA-DNA interactions using CHART and CHiRP, and DNA-DNA interactions (using 3C-based methods, including circularized chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), Hi-C, and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [30]. Further down, we are going to focus on DNA-protein interaction methodology. During this step, our purpose is to capture characteristic read distribution at the chromatin interaction sites and detect significantly enriched regions.

3.2. Material

When conducting an immunoprecipitation experiment, probably the most major consideration is the selection of the antibody. The antibody must be specific and work in chromatin immunoprecipitation. Both monoclonal and polyclonal antibodies are able to work with ChIP, though monoclonal is usually more specific. On the other hand, polyclonal antibodies recognize multiple epitopes on the targets.

This kind of assays starts with cross-linking DNA and protein. During this procedure, the cross-linking substance penetrates intact cells and fixes DNA-protein complexes. The most common stabilizer in ChIP is formaldehyde, and after stabilization chromatin is sheared in fragments commonly 200–600 bp [31].

3.3. Methodology

Following sequencing the dataset has to be aligned on the reference. Typical aligners are used such as Bowtie or BWA, and the corresponding SAM files are converted to their binary analogs (BAM). The alignments can be visualized with the stand-alone genome browser IGV [32]. When uploading an alignment file to the browser, the browser is going to search for the appropriate index file. To create the index file, the BAM file must be sorted according to its chromosomal coordinates. The indexing can be achieved with samtools index.

Another way to visualize these alignments is through the BigWig format which is an indexed binary format. Firstly, BAM is converted into a bedGraph file with BEDTools [33] and then is turned into BigWig using the bedGraphtoBigWig application from the UCSC tools [34]. BEDTools include ready-to-use files for human and mouse genomes. These files can be loaded in genomic viewers such as IGV and zoom in specific genes or chromosomal loci of interest.

MACS analysis [35] was first developed to identify transcription factor-binding sites. MACS empirically models the length of the sequenced ChIP fragments and uses it to improve the spatial resolution of predicted binding sites. MACS can be used for CHIP-Seq data alone or with control sample to increase specificity. In that sense, control sample is highly recommended to distinguish positive binding sites over background noise. Peak files generated from MACS can be uploaded to Ensembl for further analysis. Furthermore, it is advised to look at genes or regulatory elements that are located in proximity with identified regions. PeakAnalyzer [36] is a stand-alone program for the processing of genomic loci, with an emphasis on datasets consisting of ChIP-derived signal peaks. Gene ontology functional annotation can be applied in the closest downstream genes. Finally, it is really interesting to associate motif-binding sites with motifs or sequence patterns. These motifs can be compared to known motifs available in databases such as JASPAR and UniPROBE.

4. Variant calling

4.1. Methodology

Polymorphisms are generally studied in biology, under the prism of various null hypotheses. In population studies, genotype-trait associations, rare diseases and evolutionary biology, and polymorphisms are studied to answer fundamental biological questions. The starting material could be either DNA or RNA, depending on the experimental design. High-quality reads, high coverage, and a thorough bioinformatic pipelines are prerequisites to identify polymorphisms on a solid basis.

Variant callers demand different preprocessing steps before the actual processing of insertions-deletions (InDels), single-nucleotide polymorphisms (SNPs), and structural variants (SVs). Typical steps include duplicate removals and local realignment. Picard tools by Broad Institute include a Java-based set of command-line tool, including *MarkDuplicates.jar* command for the removal of the duplicate reads. The documentation of the software provides an extended walk-through and describes the metrics produced by the software.

Samtools mpileup is one of the options considered for variant calling. This option demands a reference (either genome or transcriptome) in a FASTA file and the BAM file of the aligned reads. The output is in VCF (variant call format) which is converted to its binary analogue (BCF). Samtools scripts are available that can be used to filter for low mapping quality, low coverage, gaps, and similar biases. All these artifacts are known to increase false-positive rates in SNP calls. VCFtools software [37] has the ability to filter, merge, subset, and query VCF files. Furthermore, it is able to produce simple descriptive statistics such as InDel length, transversion/transition ratio, etc. All these results can be visualized either with simple tools such as the *tview* command of the samtools package or with more sophisticated viewers such as IGV [32].

4.2. Annotation

Raw variant calling files contain many false-positive results, which may be due to the sequence quality of the reads, PCR artifacts, or other biases. Annotating these variations may mark these SNVs as less confident. In addition, important mutations could be identified according to the effect they bear on the genome. For these purposes, two potential applications are Annovar [38] and SnpEff [39]. While the latter is able to use directly VCF files, Annovar uses a specific input format that files should be converted to.

Using predefined gtf (general transfer format) models, all SNPs can be classified as synonymous, non-synonymous, loss of function, start loss, stop loss, start gain, start loss, etc. according to their effect on the genome. The 1000 Genomes data as well as the dbSNP can be used to extract data of features for annotated genomes. In case of non-model species, genes should firstly be dully annotated. Finally, any important variants should be spotted using Sanger sequencing for validation.

5. Perspectives

Although thousands of papers have been published, many things have to be done toward integration of NGS data and processing. Most of the work done is not reproducible, and data processing pipelines could not be shared among different experiments. Although guidelines for result validation have been extensively reviewed [40–43], processing parameters should be recorded thoroughly. The complexity of the NGS experiments demands complete description of the parameters through metadata; minimum information about any (x) sequence (MIxS) creates a single-entry point to all minimum information checklists for sequence data [44].

Beside reproducibility issues, one could safely argue that NGS data processing is not actually user-friendly. Expertise in informatics and more specifically in Unix-based systems is essential. In that sense, biologists are able to handle sequencing data in association with computer scientists. Furthermore, for assembly and annotation purposes, intense computing is needed that diverges from personal computers' capabilities. Therefore, small servers to large computing clusters need to be employed for processing. The development of graphical user interface (GUI) software for NGS processing is essential. Furthermore, software suites that include all steps of processing (QC, preprocessing, filtering, assembling, mapping, and differential analysis) combined with machine learning systems could facilitate analysis from beginners or intermediate computer users. In other words, more sophisticated software could propose the user, according to the experiment and the data walk-through to analyze the dataset.

Author details

Emmanouil E. Malandrakis* and Olga Dadali

*Address all correspondence to: emalandrak@uth.gr

University of Thessaly, Volos, Greece

References

[1] Gentleman RC et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology. 2004;**5**(10):R80

[2] Mortazavi A et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;**5**(7):621-628

[3] Tan SC, Yiap BC. DNA, RNA, and protein extraction: The past and the present. Journal of Biomedicine & Biotechnology. 2009;**2009**:574398

[4] Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[5] Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;**33**(19): 3137-3139

[6] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;**27**(6):863-864

[7] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**(15):2114-2120

[8] Kopylova E, Noe L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;**28**(24):3211-3217

[9] Yilmaz P et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research. 2014;**42**(Database issue):D643-D648

[10] DeSantis TZ et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. 2006;**72**(7):5069-5072

[11] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;**17**(1):10-12

[12] Haas BJ et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;**8**(8):1494-1512

[13] Honaas LA et al. Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One. 2016;**11**(1):e0146062

[14] Simao FA et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;**31**(19):3210-3212

[15] Li B et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology. 2014;**15**(12):553

[16] Smith-Unna R et al. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research. 2016;**26**(8):1134-1144

[17] Trapnell C et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013;**31**(1):46-53

[18] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;**25**(14):1754-1760

[19] Langmead B et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;**10**(3):R25

[20] Trapnell C, Pachter L, Salzberg SL. TopHat: Dscovering splice junctions with RNA-Seq. Bioinformatics. 2009;**25**(9):1105-1111

[21] Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; **25**(16):2078-2079

[22] Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;**12**:323

[23] Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods. 2013;**10**(1):71-73

[24] Bray NL et al. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;**34**(5):525-527

[25] Patro R et al. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017;**14**(4):417-419

[26] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;**11**(10):R106

[27] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;**15**(12):550

[28] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;**26**(1):139-140

[29] Solomon MJ, Larsen PL, Varshavsky A. Mapping protein DNA interactions in vivo with formaldehyde – Evidence that histone-H4 is retained on a highly transcribed gene. Cell. 1988;**53**(6):937-947

[30] Sims D et al. Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews. Genetics. 2014;**15**(2):121-132

[31] Park PJ. ChIP-seq: Advantages and challenges of a maturing technology. Nature Reviews. Genetics. 2009;**10**(10):669-680

[32] Robinson JT et al. Integrative genomics viewer. Nature Biotechnology. 2011;**29**(1):24-26

[33] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;**26**(6):841-842

[34] Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Briefings in Bioinformatics. 2013;**14**(2):144-161

[35] Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biology. 2008;**9**(9): R137

[36] Salmon-Divon M et al. PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. BMC Bioinformatics. 2010;**11**:415

[37] Danecek P et al. The variant call format and VCFtools. Bioinformatics. 2011;**27**(15):2156-2158

[38] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;**38**(16):e164

[39] Cingolani P et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;**6**(2):80-92

[40] Roy S et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists. The Journal of Molecular Diagnostics. 2018;**20**(1):4-27

[41] Jennings LJ et al. Guidelines for validation of next-generation sequencing-based oncology panels: A joint consensus recommendation of the association for molecular pathology and college of american pathologists. The Journal of Molecular Diagnostics. 2017;**19**(3):341-365

[42] Kim J et al. Good laboratory standards for clinical next-generation sequencing cancer panel tests. Journal of Pathology and Translational Medicine. 2017;**51**(3):191-204

[43] Endrullat C et al. Standardization and quality management in next-generation sequencing. Applied & Translational Genomics. 2016;**10**:2-9

[44] Yilmaz P et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology. 2011;**29**:415

Next-Generation Transcriptome Assembly: Strategies and Performance Analysis

Adam Voshall and Etsuko N. Moriyama

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73497

Abstract

Accurate and comprehensive transcriptome assemblies lay the foundation for a range of analyses, such as differential gene expression analysis, metabolic pathway reconstruction, novel gene discovery, or metabolic flux analysis. With the arrival of next-generation sequencing technologies, it has become possible to acquire the whole transcriptome data rapidly even from non-model organisms. However, the problem of accurately assembling the transcriptome for any given sample remains extremely challenging, especially in species with a high prevalence of recent gene or genome duplications, those with alternative splicing of transcripts, or those whose genomes are not well studied. In this chapter, we provided a detailed overview of the strategies used for transcriptome assembly. We reviewed the different statistics available for measuring the quality of transcriptome assemblies with the emphasis on the types of errors each statistic does and does not detect. We also reviewed simulation protocols to computationally generate RNAseq data that present biologically realistic problems such as gene expression bias and alternative splicing. Using such simulated RNAseq data, we presented a comparison of the accuracy, strengths, and weaknesses of nine representative transcriptome assemblers including *de novo*, genome-guided, and ensemble methods.

Keywords: RNAseq, transcriptome, assembly, *de novo*, genome-guided, ensemble approach

1. Introduction

Transcriptome assembly from high-throughput sequencing of mRNA (RNAseq) is a powerful tool for detecting variations in gene expression and sequences between conditions, tissues, or strains/species for both model and non-model organisms [1, 2]. However, the ability to accurately perform such analyses is crucially dependent on the quality of the underlying assembly [3]. Especially for the detection of sequence variations, but also for isoform detection and transcript quantification, mis-assembly of genes of interest can increase both the false positive and false negative rates, depending on the nature of the mis-assembly [4]. These problems are exacerbated in non-model organisms where genomic sequences that can be used as the references, if available at all, are sufficiently different than those from the individuals sequenced [5].

Transcripts can be mis-assembled in several ways [6]. Two of the most drastic assembly errors are fragmentation, where a single transcript is assembled as one or more smaller contigs, and chimeras, where a contig is assembled using part or all of more than one transcript. Fragmentation errors tend to result from fluctuations in the read coverage along a transcript, with the breaks in the transcript sequence occurring in regions that have lower coverage. By contrast, chimera errors often occur because of ambiguous overlaps within the reads, coupled with algorithms that choose the longest possible contig represented by the data, or by adjacent genes on the genome being merged. Both of these types of errors can have major impacts especially on gene identification. Small (single or few) nucleotide alterations to the contig sequence also happen as mis-assemblies. Sequence mistakes are often the result of mis-sequenced reads, but can also result from ambiguity for highly similar reads e.g. from heterozygous genes and from duplicated genes. In some cases, these errors can shift the reading frame for the contig, which can have significant impacts on the translated protein sequence. Finally, transcripts can be mis-assembled when alternative transcripts are collapsed into a single contig [6].

In the following sections, we will first review strategies used for transcriptome assembly as well as how their performance can be assessed. We then compare the performance of representative transcriptome assembly methods using a simulated human transcriptome and RNAseq. Finally we discuss a possible strategy to improve transcriptome assembly accuracy.

2. Transcriptome assembly strategies

2.1. *De novo* assemblers

De novo assemblers generate contigs based solely on the RNAseq data [7–13]. Most of the *de novo* assemblers rely on de Bruijn graphs generated from kmer decompositions of the reads in the RNAseq data [14]. The reads are subdivided into shorter sequences of a given length *k* (the kmers) and the original sequence is reconstructed by the overlap of these kmer sequences. One major limitation of the de Bruijn graphs is the need for a kmer to start at every position along the original sequence in order for the graph to cover the full sequence [13]. This limitation creates a tradeoff in regard to the length of the kmers. Shorter kmers are more likely to fully cover the original sequence, but are more likely to be ambiguous, with a single kmer corresponding to multiple reads from multiple transcripts. While by using longer kmers such ambiguity can be avoided, those kmers may not cover the entire sequence of some transcripts causing e.g. fragmented assembly. Consequently, each transcript, with its unique combination of expression level (corresponding to the number of reads in the RNAseq data generated from that transcript) and sequence will have a different best kmer length for its assembly [15]. As a result, even using the same *de novo* assembly algorithm, performing two assemblies with different kmer lengths will generate a different set of contigs, inevitably with a varying set of correctly assembled contigs [16].

Examples of popularly used *de novo* assemblers include idba-Tran [9], SOAPdenovo-Trans [8], rnaSPAdes [12], and Trinity [7]. Idba-Tran is unique among these *de novo* assemblers, as it runs individual assemblies across a range of kmer lengths and merges the results to form the final prediction. The remaining assemblers use only the results of a single kmer length. For SOAPdenovo-Trans and Trinity, a kmer length needs to be chosen (default kmer: 23 and 25, respectively), while rnaSPAdes dynamically determines the kmer length to be used based on the read data. While all of these tools use the same fundamental strategies to construct, revise, and parse the de Bruijn graph for the assemblies, each method uses different thresholds and different assumptions to make decisions. These differences lead to different subsets of transcripts being correctly assembled by each method. An example of how these tools produce different sets of contigs is shown in Section 4.2.

2.2. Genome-guided assemblers

Genome-guided assemblers avoid the ambiguity of kmer decompositions used in de Bruijn graphs by mapping the RNAseq data to the reference genome. In order to account of introns, mapping of the reads for genome-guided assembly needs to allow them to be split, where the first part of the read maps to one location (an exon), and the other half maps to a downstream location (another exon). This mapping is done by split-read mappers such as TopHat [17], STAR [18], HISAT [19], or HPG-aligner [20]. Each of these methods maps the reads slightly differently, which may impact the quality of subsequent assembly.

This read mapping greatly reduces the complexity of transcript assembly by clustering the reads based on genomic location rather than relying solely on overlapping sequences within the reads themselves [3]. However, this approach still has some major drawbacks. The most obvious drawback is that genome-guided assemblers require a reference genome, which is not available for all organisms. The quality of the reference genome, if it is available, also impacts the quality of the read mapping and, by extension, the assembly. This impact is particularly noteworthy when genes of interest contain gaps in the genome assembly, preventing the reads necessary to assemble those genes from mapping to part or all of the transcript sequence. Ambiguity occurs also when reads map to multiple places within a genome. How the specific algorithm handles choosing which potential location a read should map to can have a large impact on the final transcripts predicted [6]. This problem is expounded when working with organisms different from the reference, where not all of reads map to the reference without gaps or mismatches.

Examples of popularly used genome-guided assemblers include Bayesembler [21], Cufflinks [22], and StringTie [23]. While each of these methods uses the mapped reads to create a graph representing the splice junctions of the transcripts, how they select which splice junctions are real differs fundamentally. Cufflinks constructs transcripts based on using the fewest number of transcripts to cover the highest percentage of mapped reads. StringTie uses the number of reads that span each splice junction to construct a flow graph, constructing the transcripts based in order of the highest flow. Bayesembler constructs all viable transcripts for each splice junction and uses a Bayesian likelihood estimation based on the read coverage of each potential transcript to determine which combination of transcripts is most likely. Due to these fundamentally different approaches, each of these tools produces different sets of transcripts from the same set of reads. An example of assemblies produced by these methods and how the assembled contigs differ is described in Section 4.3.

2.3. Ensemble approach

While a core set of transcripts are expected to be assembled correctly by many different assemblers, many transcripts will be missed by any individual tool [24] (also see Section 4). Through combining the assemblies produced by multiple methods, ensemble assemblers such as EvidentialGene [25] and Concatenation [26] attempt to address the limitations of individual assemblers, ideally keeping contigs that are more likely to be correctly assembled and discarding the rest. Both of EvidentialGene and Concatenation filter the contigs obtained from multiple assemblers (usually *de novo*) by clustering the contigs based on their sequences, predicting the coding region of the contig, and using features of the overall contig and the coding region to determine the representative sequence for each cluster. EvidentialGene recommends using several different tools across a wide range of kmer lengths. It uses the redundancy from multiple tools generating nearly identical sequences, clusters them, scores the sequences in each cluster based of the features of the sequence (e.g. lengths of the 5′ and 3′ untranslated regions), and returns one representative sequence from each cluster (keeping also some alternative sequences). In contrast, Concatenation recommends using only three assemblers, with one kmer length each. Concatenation merges nucleotide sequences that are identical or perfect subsets, only filters contigs with no predicted coding region.

These approaches greatly reduce the number of contigs by removing redundant and highly similar sequences. However, there is no guarantee that the correct representative sequence is kept for a given cluster or that each cluster represents one unique gene. Because they require multiple assemblies to merge, they also come at a far greater computational cost. An example of how these ensemble assembly strategies perform compared to individual *de novo* and genome-guided methods is shown in Section 4.4.

2.4. Third generation sequencing

All of the methods described so far primarily use short but highly accurate reads from Illumina sequencing for assembly, with or without a reference. With the rise of third-generation sequencing technologies from Pacific Biosciences (PacBio SMRT) and Oxford Nanopore Technologies (ONT MinION), it is becoming possible to sequence entire mRNA molecules as one very long read, though with a high error rate [27]. The ability to sequence the entire mRNA molecule is especially beneficial for detecting alternative splice forms, which remain a challenge for short-read only assembly, and potentially for more accurate transcript quantification if there is no bias in the mRNA molecules sequenced.

While many tools exist to perform genome assemblies using either these long reads alone or by combining long reads and Illumina reads, at present no short read transcriptome assemblers take advantage of long-reads in transcriptome assembly. If these long reads can be sufficiently error-corrected (e.g. [28, 29]), they can be used for a snapshot of the expressed transcriptome, without requiring assembly or external references [30, 31]. Alternatively, after an independent *de novo* assembly of short reads, the long reads can be used to confirm alternative splice forms present in the assembly [32]. The long reads can be also mapped to a reference genome similar to the split-read mapping methods used for genome-guided short-read assemblers discussed above [27, 33–35]. With their accuracy increasing, in the future, long reads can be used more to improve transcriptome assembly quality.

3. Performance metrics used for transcriptome assembly

In this section, we will discuss commonly used metrics to assess the quality of transcriptome assemblies.

3.1. Metrics based on contig count and lengths

The most straightforward assembly metrics are those based on the number and lengths of the sequences produced [36]. The number of sequences can be presented either or both of:

- the number of contigs
- the number of scaffolds

where for contigs no further joining of the sequences is performed after assembly, and for scaffold contigs that have some support for being from the same original sequence are combined together with a certain number of gaps between them.

Several different statistics are available for presenting the lengths of the sequences (either contigs or scaffolds). The most commonly reported metrics are:

- minimum length (bp): the length of the shortest sequence produced
- maximum length (bp): the length of the longest sequence produced
- mean length (bp): the average length of the sequences produced
- median length (bp): the length where half of the sequences are shorter, and half of the sequences are longer
- N50 (bp): a weighted median where the sum of the lengths of all sequences longer than the N50 is at least half of the total length of the assembly
- L50: the smallest number of sequences whose combined length is longer than the N50

Additional metrics similar to N50 (e.g. N90) based on different thresholds are also used.

For genome assemblies where the target number of sequences is known (one circular genome plus any smaller plasmids for prokaryotic organisms and the number of chromosomes for eukaryotic organisms), these metrics provide an estimate for the thoroughness of the assembly

[36]. For instance, in prokaryotic assemblies, the vast majority of the sequence is expected to be in one long sequence, and having many shorter sequences indicates fragmentation of the assembly [15]. In this context, longer sequences (e.g. larger N50) tend to indicate higher quality assemblies. For transcriptome assemblies, however, the length of the assembled contigs varies depending on the lengths of the transcripts being assembled. For the human transcriptome, for example, while the longest transcript (for the gene coding the Titin protein) is over 100 kb, the shortest is only 186 bp, with a median length of 2787 bp [37]. Emphasizing longer contigs also rewards assemblers that over-assemble sequences, either by including additional sequence incorrectly within a gene, or by joining multiple genes together to form chimeric contigs. Therefore, for transcriptome assembly, metrics based on contig lengths do not necessarily reflect its quality.

3.2. Metrics based on coded protein similarity

Rather than focusing on the number or length of the sequences produced by the assembly, performing similarity searches with the assembled sequences can provide an estimate of the quality of the contigs or scaffolds [24, 38]. Typically, the process consists of either similarity searches against well annotated databases (such as the protein datasets of related genomes or targeted orthologs, the BLAST non-redundant protein database [39] or the UniProt/Swiss-Prot database [40]), conserved domain search within the contig sequence that determines the potential function of the gene (such as PFAM or Panther [41, 42]), or a search against a lineage specific conserved single-copy protein database (such as BUSCO [43]). These similarity searches are usually performed on the predicted protein sequences for the contigs (e.g. using GeneMarkS [44]), but can also be performed directly from the assembled nucleotide sequences using BLASTX where translated nucleotide sequences are used to search against a protein database [38]. If the organism being sequenced is closely related to a model organism with a well-defined transcriptome, nearly all of the contigs that are not erroneously assembled and code proteins should have identifiable potential homologs in the database. If a large percentage of the contigs do not have similar proteins identified in the database, there is a high probability that the sequences are incorrectly assembled, regardless of the length of the sequences. By performing similarity searches, over-assemblies or chimera contigs (those covering more than one gene) can be also detected as large gaps in the alignment between the query and the hits. As protein sequence annotations are necessary for most downstream analyses, they also provide a convenient metric without the need for additional, otherwise unnecessary analyses.

Despite these advantages, there are some limitations to using protein-similarity based metrics for assembler performance. First, the more divergent the organism being sequenced is from the sequences in the database searched and the more species-specific genes in the transcriptome, the lower the percentage of contigs with hits will be. This can result in some organisms appearing to have a lower quality assembly solely due to their divergence from those well represented in the databases. By extension, assemblies that recover more transcripts whose coded proteins have few similar sequences in the database will appear worse than assemblies that only recover conserved genes. This limitation can be somewhat mitigated by comparing only genes that are universally single-copy across different species, which are more likely to be conserved and similar enough to be identified. This is the strategy used in BUSCO [43]. However, this comparison at best uses only a subset of the assembled contigs. Second, and more problematic, this metric rewards assemblies that artificially duplicate conserved genes with only small differences in the nucleotide sequence. In the extreme, this can result in several times as many contigs in the assembly than were present in the actual transcriptome, but with nearly all of the contigs coding conserved protein sequences. This is particularly an issue when the analysis depends on identifying the gene copy numbers in the assembly. It also has a large impact on the accuracy of contig quantification and differential expression analyses [45].

3.3. Assembly metrics based on benchmark transcriptomes

The only way to overcome the limitations of the metrics described in the previous sections is to compare the assembly output against a benchmark transcriptome where correct sequences of all transcripts are known. When an RNAseq data generated from a well-established model organism is used for assembly, many of correctly assembled contigs can be identified. However, variability in the transcriptome among e.g. cell types limits the amount of information that can be gained for incorrectly assembled contigs. It is also not possible to determine whether sequences from the reference that are missing from the assembled transcriptome are due to assembly errors, or whether they were not expressed in the library sequenced. Transcriptome sequences may also vary between the individual under study and the reference. Such variations can mask assembly errors that affect the contig sequences. Although this limitation can be mitigated by sequencing an individual that is genetically identical to the reference, it severely limits the types of organisms that can be used for the benchmark.

To comprehensively assess all of the assembly errors, we need to obtain RNAseq data from a transcriptome where all transcript sequences and expression patterns are known. Ideally, such a benchmark transcriptome would be synthetically produced and sequenced using standard protocols. However, currently no such synthetic mRNA library exists. An alternative approach is to simulate the sequencing of a given benchmark transcriptome. There are several tools that can generate simulated reads modeling short Illumina reads [46, 47] and/or long third-generation sequencing reads such as PacBio SMRT and ONT MinION [48, 49]. These tools typically either focus on identifying the statistical distribution of reads across the sequences and errors within the reads, as is the case for RSEM [46], PBSIM [48], and Nanosim [49], or attempt to reconstruct each step of the library preparation and sequencing pipeline, mimicking the errors and biases introduced at each step, as is the case for Flux Simulator [47].

Using simulated RNAseq data with a known transcriptome as a benchmark gives the most detailed and close to true performance metric for assemblies. Specifically, this strategy allows the quantification of each of the following categories:

- correctly assembled sequences (true positives or TPs)
- sequences that are assembled with errors (false positives or FPs)
- sequences in the reference that are missing from the assembly (false negatives or FNs)

“Correctness” and “incorrectness” (or error) can be defined using varying degrees of sequence similarities. Using the strictest threshold, a contig sequence is assembled “correctly” only if the entire nucleotide or coded protein sequence is identical to a reference transcript. All other contigs found in the assembly, including those whose sequences have no similarity in the reference transcriptome (missing contigs), are considered to be assembled “incorrectly” (FPs) regardless of the similarity against the reference sequences.

Note that true negatives (TNs) can be counted only if the assembly experiments are done including reads that are derived from transcripts that are not part of the reference transcriptome (negative transcripts). Using these categories, following assembly metrics can be calculated:

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
- Sensitivity (or recall) = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$
- Precision = $\frac{TP}{TP+FP}$
- F-measure (or F_1 score) = $\frac{2(TP)}{2(TP)+FP+FN}$
- False discovery rate (FDR) = $\frac{FP}{FP+TP}$

Often in an RNAseq simulation, negative transcripts are not included; hence TN cannot be counted. In such cases, we can calculate an alternative metric as the accuracy:

- Accuracy* = $\frac{TP}{TP+FP+FN}$

Despite the added benefits of simulation for measuring the performance of assemblers, these metrics assume that the simulation accurately reflects the nature of real RNAseq data. Differences in the distribution of reads or errors between the simulations and real data can impact the relative performance of the assemblers. Assemblers that perform well on simulated data may perform poorly on real data if those assumptions are not met. Consequently, great care must be taken to ensure that the simulated data captures the features of real data as accurately as possible to best characterize the performance of different assembly strategies.

4. Performance analysis of transcriptome assemblers

In this section, as an example, we compare the performance of transcriptome assemblers using a simulated benchmark transcriptome dataset.

4.1. Benchmark transcriptome and simulated RNAseq

RNAseq datasets were generated by Flux Simulator [47] using the hg38 human genome (available at https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg38) as the reference. The older hg19 human genome (available at http://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19) was also used as an alternate reference genome to assess the impact of using a different reference with genome-guided assemblers. The gene expression profile was generated by Flux Simulator using the standard parameters from the hg38 reference genome and transcriptome model. Approximately 250 million pairs of reads were computationally generated with the given expression model with no PolyA tail. The simulated library construction was fragmented uniformly at random, with an average fragment size of 500 (±180) nucleotides (nt). Because reads overlapping within read pairs can cause problems for some assemblers, fragments shorter than 150 nt were removed. The simulated sequencing was performed using paired-end reads of length of 76 nt using the default error model based on the read quality of Illumina-HiSeq sequencers. Note that only reference transcripts with full coverage of RNAseq data were included in the benchmarking, as transcripts without full coverage cannot be correctly assembled as a single contig. This filtering removed 2700 transcripts expressed in the benchmark transcriptome, leaving 14,040 unique sequences derived from 8557 genes (5309 with no alternative splicing; on average 1.64, ranging up to 13, isoforms per gene).

The read pairs generated by Flux Simulator were quality filtered using Erne-filter version 2.0 [50]. The reads were filtered using ultra-sensitive settings with a minimum average quality of q20 (representing a 99% probability that the nucleotide is correctly reported). The filtering was performed in paired-end mode to ensure that both reads of the pair were either kept or discarded concurrently to keep the pairs together. The remaining reads were normalized using Khmer [51] with a kmer size of 32 and an expected coverage of 50×. The normalization was also performed in paired-end mode to maintain pairs.

4.2. *De novo* assemblies

We compared the performance among four *de novo* transcriptome assemblers: idba-Tran (version 1.1.1) [9], SOAPdenovo-Trans (version 1.03) [8], rnaSPAdes (version 3.11.0) [12], and Trinity (version 2.5.1) [7], using the simulated human RNAseq dataset as described in the previous section. The resulted assemblies were compared against the benchmark transcriptome. As shown in **Table 1**, all of the tools underestimated the number of transcripts present, generating fewer contigs than the number of transcripts expected (14,040). The best performing tool among the four compared was Trinity with the most correct contigs (5782) and the highest correct/incorrect ratio (C/I = 0.84). However, even with Trinity, still only 41% (5782/ 14,040) of transcripts in the benchmark were correctly assembled; the remaining almost 60% of contigs either contained errors in the sequence or were missed entirely. rnaSPAdes assembled the largest number of transcripts (874 more unique transcripts compared to Trinity). The number of unique transcripts generated, 13,513, is also the closest to the expected total number of transcripts (96% of 14,040). However, fewer of those sequences (36%) were correctly assembled, lowering the overall performance across all statistics than Trinity.

Performance statistics for each assembler is given in **Table 2**. Precision is a measure of how likely an assembled contig is to be correct, and recall is a measure of how likely the assembler is to correctly assemble a contig. In these terms, for assemblers with high precision, the contigs produced are more likely to be correct, but the assembly may miss a large number of sequences present in the sample. Conversely, assemblers with high recall values correctly assemble more of

Methods	Total[a]	Unique[a]	Correct[a] (%)[b]	Incorrect[a]	C/I[c]
[Default]					
idba-Tran	11,943	11,941	3504 (24.96)	8437	0.4153
SOAPdenovo-Trans	12,902	11,830	3754 (26.74)	8076	0.4648
rnaSPAdes	15,670	13,513	5014 (35.71)	8499	0.5900
Trinity	14,044	12,639	5782 (41.18)	6857	0.8432
[Pooled][d]					
idba-Tran	170,358	41,849	6391 (45.52)	35,458	0.1802
SOAPdenovo-Trans	297,192	50,504	6059 (43.16)	44,445	0.1363
rnaSPAdes	765,525	113,975	6665 (47.47)	107,310	0.0621
Trinity	89,126	25,045	6452 (45.95)	18,593	0.3470

[a] Number of contigs assembled.
[b] Proportion (%) of transcripts in the benchmark that were correctly assembled.
[c] (Number of correctly assembled contigs)/(number of incorrectly assembled contigs).
[d] Pooled results from using multiple kmers as follows: 15, 19, 23, 27, and 31 for Trinity; 15 kmer values ranging from 15 to 75 in increments of 4 for SOAPdenovo-Trans and rnaSPAdes; 20, 30, 40, 50, and 60 for idba-Tran.

Table 1. Performance of individual *de novo* assemblers on simulated RNAseq library using default parameters or pooled across multiple kmer lengths.

Methods	Precision	Recall	Accuracy*	F_1	FDR
idba-Tran	0.2934	0.2496	0.1559	0.2697	0.7066
SOAPdenovo-Trans	0.3173	0.2674	0.1697	0.2902	0.6827
rnaSPAdes	0.3711	0.3571	0.2225	0.3640	0.6289
Trinity	0.4575	0.4118	0.2767	0.4334	0.5425

Table 2. Performance statistics of individual *de novo* assemblers using default parameters on simulated RNAseq library.

the sequences present in the sample, but may do so at the cost of accumulating a large number of incorrectly assembled contigs. In these statistics, both the modified accuracy score (accuracy*; see Section 3.3) and the F_1 score are a measure of the number of correctly assembled contigs relative to the number of missing and incorrectly assembled contigs. FDR is the proportion of assembled reads that are incorrect. Based on these statistics, Trinity is the best performing *de novo* assembler with the highest precision, recall, accuracy* and F_1 score, and the lowest FDR, followed by rnaSPAdes then SOAPdenovo-Trans. Despite idba-Tran running multiple kmers and merging the results, it performed worst across every metric.

In **Table 1**, the results from pooling (taking the union of) the outputs of multiple runs of each assembler across a range of kmer lengths are also shown. With these pooled assemblies, the proportion of correctly assembled transcripts in the benchmark for Trinity increased from 41 to 46%, and for rnaSPAdes from 36 to 47%. However, the pooling process also accumulated several times more unique incorrect sequences than additional correct sequences recovered.

For Trinity, the C/I decreased from 0.8432 to 0.3470, and for rnaSPAdes this ratio decreased from 0.5900 to 0.0621.

Although the four *de novo* assembly methods use the same core approach, each method assembled a different set of sequences correctly (**Figure 1A**). Only a set of 5331 contigs were correctly assembled by all of the four *de novo* assemblers with at least one kmer length. Additional 813, 567, and 670 contigs were correctly assembled by at least three, at least two, and only one of the assemblers, respectively. In contrast, the vast majority of the incorrectly assembled contigs were produced by only one assembler (**Figure 1B**). For these contigs, 3764 were produced by all four assemblers, while an additional 2692, 7977 and 166,720 were produced by at least three, at least two or only one of the assemblers, respectively.

4.3. Genome-guided assemblies

We next compared the transcriptome assembly performance among three genome-guided assemblers: Bayesembler (version 1.2.0) [21], Cufflinks (version 2.2.1) [22], and StringTie (version 1.0.4) [23]. To demonstrate the impact of using different reference genomes on genome-guided transcriptome assemblies, we used both of the hg38 as well as hg19 genomes as the references. Assembly assessment was done against the hg38 benchmark transcriptome.

Table 3 shows the performance of each of these tools in the two scenarios (RNAseq data and the reference were derived from the same or different genomes). As observed with *de novo* methods, all of these genome-guided methods underestimated the number of transcripts present, even more severely than *de novo* methods. In terms of the number of contigs correctly assembled, StringTie performed slightly better than other two methods. All three methods had comparable percent correct (36–41% with the same reference) and C/I (0.87–0.88 with the same

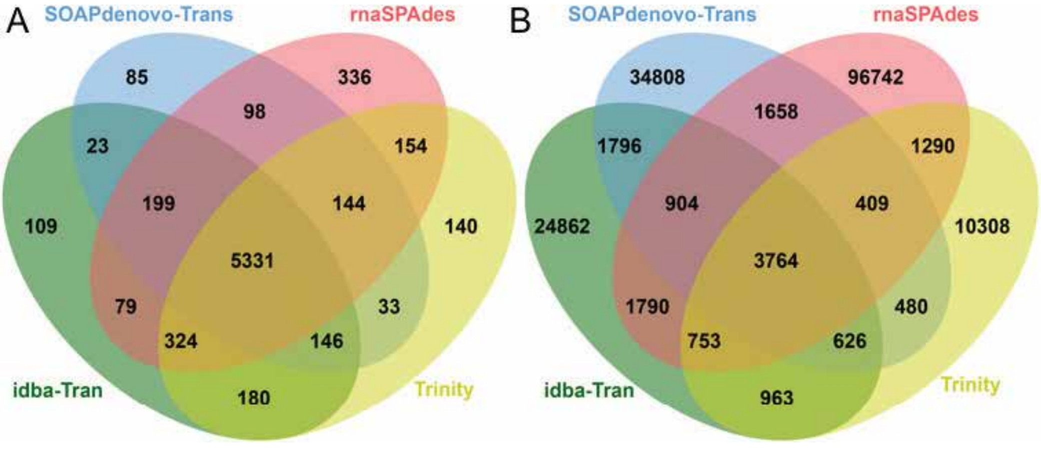


Figure 1. Comparisons of the contigs correctly (A) and incorrectly (B) assembled among four *de novo* assemblers. For each assembler, results from multiple kmers were pooled. Correctly assembled sequences were identified when the protein sequence of the contig matched the protein sequence in the benchmark transcriptome. Incorrectly assembled sequences were identified when the protein sequence of the contig did not exactly match any protein sequence in the benchmark transcriptome.

Methods	Total	Unique	Correct (%)	Incorrect	C/I
[Same reference]					
Bayesembler	12,989	11,482	5327 (37.94)	6155	0.8655
Cufflinks	11,257	10,733	4992 (35.56)	5741	0.8695
StringTie	13,218	12,147	5696 (40.57)	6451	0.8830
[Different reference]					
Bayesembler	8536	7479	3345 (23.82)	4134	0.8091
Cufflinks	7234	6906	3078 (21.92)	3828	0.8041
StringTie	8608	7867	3466 (24.69)	4401	0.7875

Table 3. Performance of individual genome-guided assemblers using default parameters on simulated RNAseq library with both the same and different references genome as the benchmark. See **Table 1** for the description of numbers shown.

reference). While none of the genome-guided assemblers produced as many correctly assembled contigs as the best performing *de novo* assembler (Trinity), proportions of correctly assembled contigs were higher with genome-guided methods (C/I = 0.87–0.88 with the same reference) than with the four *de novo* methods (C/I = 0.41–0.84). When the performance metrics are compared between the best performing *de novo* assembler (Trinity) and genome-guided assembler (StringTie) (**Table 4**), while both methods showed similar accuracy, StringTie (when using the same reference) showed slightly higher precision, accuracy* and F_1 and lower FDR compared to Trinity, but a slightly lower recall. It reflects fewer FPs and FNs produced by StringTie.

As with the *de novo* assemblers, each of these tools correctly assembled a different set of transcripts (**Figure 2A** and **C**). When the assemblies were performed using the same reference as the simulation, all of the genome-guided tools correctly assembled a core set of 4013 transcripts (**Figure 2A**). There were nearly a quarter as many (936) that were unique to only one genome-guided tool. When a different reference was used, the number of sequences correctly assembled by all of the tools dropped to 2546 (**Figure 2C**). Similar to the *de novo* assemblers, most of the

Methods	Precision	Recall	Accuracy*	F_1	FDR
[Same reference]					
Bayesembler	0.4639	0.3794	0.2638	0.4174	0.5361
Cufflinks	0.4651	0.3556	0.2524	0.4030	0.5349
StringTie	0.4689	0.4057	0.2780	0.4350	0.5311
[Different reference]					
Bayesembler	0.4473	0.2382	0.1841	0.3109	0.5527
Cufflinks	0.4457	0.2192	0.1723	0.2939	0.5543
StringTie	0.4406	0.2469	0.1880	0.3164	0.5594

Table 4. Performance statistics of individual genome-guided assemblers using default parameters on simulated RNAseq library with both the same and different references genome as the benchmark.

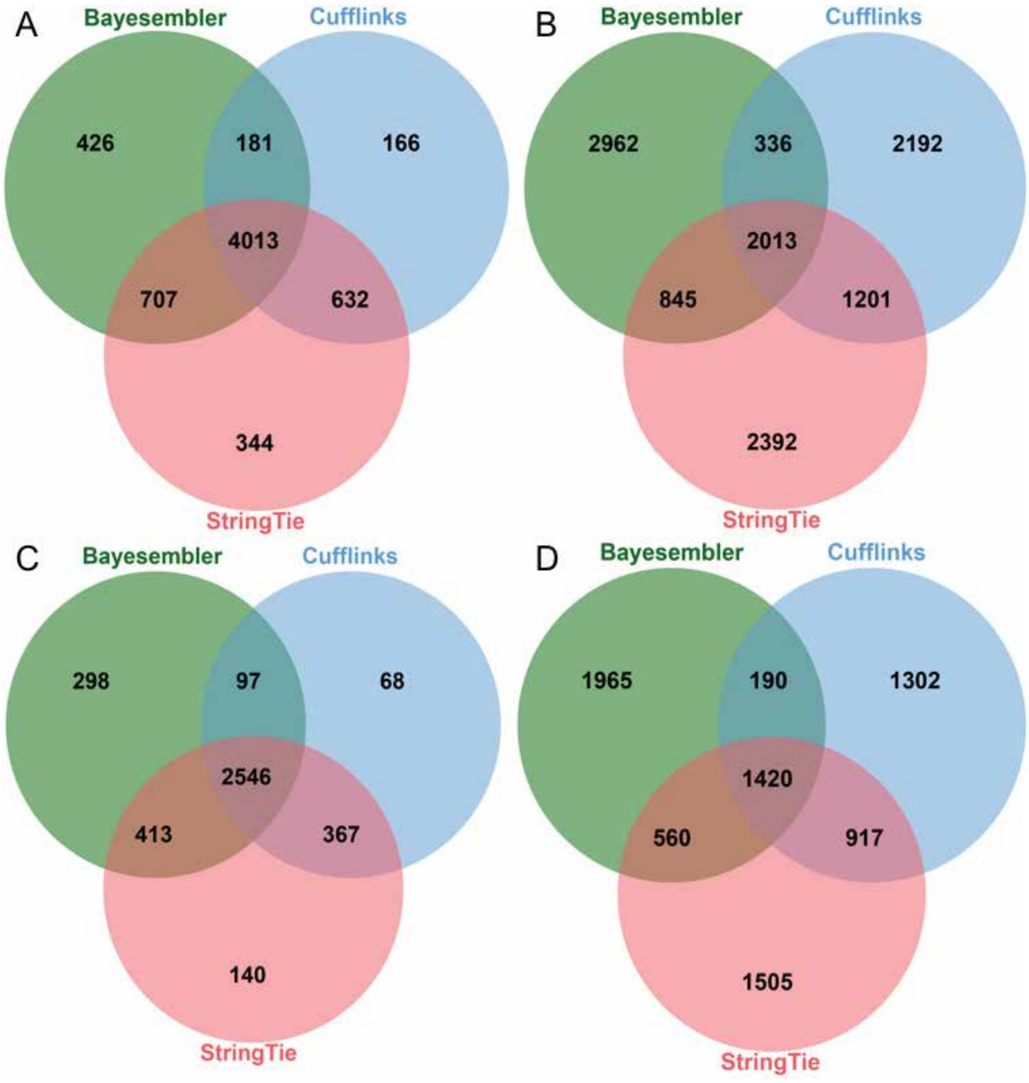


Figure 2. Comparisons of the contigs correctly (A and C) and incorrectly (B and D) assembled among three genome-guided assemblers. Correctly assembled sequences were identified when the protein sequence of the contig matches the protein sequence in the same (A) or different (C) reference genome. Incorrectly assembled sequences were identified when the protein sequence of the contig does not exactly match any protein sequence in the same (B) or different (D) reference genome.

incorrectly assembled contigs produced by each of the genome-guided assemblers were produced by only one assembler regardless of the reference genome used (**Figure 2B** and **D**). For assemblies using the same reference genome, 2013 incorrectly assembled contigs were produced by all of the tools, while an additional 2382 and 7546 were produced by any two or only one tool, respectively (**Figure 2B**). For assemblies using a different reference genome, 1420 incorrectly assembled contigs were produced by all of the tools, while an additional 1667 and 4772 were produced by any two or only one tool, respectively (**Figure 2D**).

4.4. Comparison of *de novo* and genome-guided assemblers

While the overall statistics are comparable between the best *de novo* assemblies and the genome-guided assemblies using the same reference genome, these tools produced different sets of contigs. The overlap of correctly assembled contigs between the assemblers from *de novo* with pooled kmers lengths and the three genome-guided assemblers are shown in **Figure 3A**. All of the *de novo* assemblers and at least one genome-guided assembler correctly assembled 4605 contigs. An additional 629 were assembled by at least three *de novo* and at least one genome-guided assembler and 427 assembled by at least two *de novo* and at least one genome-guided assembler. Conversely, 3861 contigs were correctly assembled by all of the three genome-guided assemblers and at least one *de novo* assembler, with 1338 assembled by at least two genome-guided assemblers and at least one *de novo* assembler (**Figure 3B**). Additionally, these tools produced only 602 correctly assembled contigs that were not predicted by any *de novo* assembly, while 1514 sequences were correctly assembled by at least one *de novo* assembly, but no genome-guided assemblies.

As with the individual assemblies, fewer incorrectly assembled contigs were produced by all of the tools, and most are assembler specific (**Figure 3C** and **D**). In particular, only 1387 incorrectly assembled contigs were produced by all of the *de novo* assemblers and at least one genome-guided assembler (**Figure 3C**), and only 1593 contigs were produced all of the genome-guided assemblers and at least one *de novo* assembler (**Figure 3D**). In contrast, 4823 incorrectly assemblers were produced by at least one genome-guided assembler but no *de novo* assemblers, and 176,397 incorrectly assembled contigs were produced by at least one *de novo* assembler but no genome-guided assemblers.

Overall, these results suggest that genome-guided assemblies provide relatively few correctly assembled contigs relative to performing multiple *de novo* assemblies, even when using the same reference genome. However, they produce far fewer incorrectly assembled contigs than the pooled *de novo* assemblies. If the correctly assembled contigs produced by each of the *de novo* assemblies can be retained while filtering out the incorrectly assembled contigs, *de novo* assemblies can outperform all of the genome-guided assemblies. This result forms the motivation of ensemble assembly strategies, discussed in the next section.

4.5. Ensemble assemblies

We compared the two ensemble transcriptome assembly methods, EvidentialGene (version 2017.03.09) [25] and Concatenation (version 1) [26] using the simulated RNAseq data. The strategies for these assemblies followed the recommendations by each method. For EvidentialGene, the pooled results from all of the four *de novo* assemblies performed across the full range of kmer lengths (described in Section 4.2) were used. For Concatenation, the results of a single assembly each from idba-Tran (using kmer length of 50), rnaSPAdes (with default kmer selection), and Trinity (with default kmer length) were used. These assemblers were chosen to match the assemblies used in [26], substituting the commercial CLC Assembly Cell (https://www.qiagenbioinformatics.com/products/clc-assembly-cell/) with freely available rnaSPAdes.

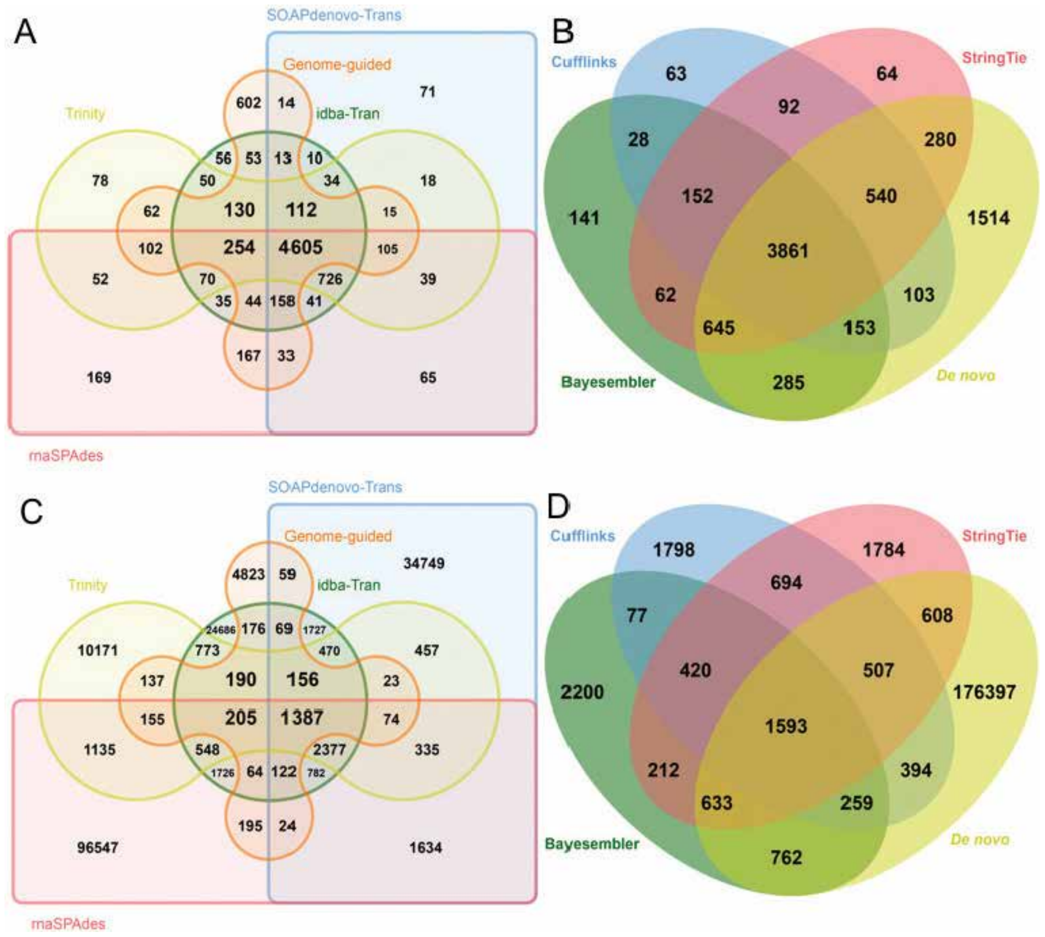


Figure 3. Comparisons of the results among *de novo* and genome-guided transcriptome assemblers. For each *de novo* assembler, results from multiple kmers were pooled. Correctly (A) and incorrectly (C) assembled sequences for each *de novo* assembler are compared with the combined results from genome-guided assemblers. Correctly (B) and incorrectly (D) assembled sequences for each genome-guided assembler are compared with the combined results from *de novo* assemblers.

In addition to the two ensemble methods, we also included three “consensus” approaches taking the consensus of the pooled *de novo* methods. These consensus assemblies involve keeping all of the unique protein sequences produced by any two, three and four tools (named Consensus 2, Consensus 3 and Consensus 4, respectively). Note that Consensus 4 is a subset of Consensus 3, and Consensus 3 is a subset of Consensus 2.

The performance of these ensemble strategies is shown in **Table 5**. Both of EvidentialGene and Concatenation resulted in an over-estimation in the number of transcripts present. Interestingly, while Concatenation produced a larger total number of transcripts (19,767) than EvidentialGene (19,177), ~2300 of those sequences were redundant, leading to fewer unique sequences (17,497 by

Concatenation). Additionally, Concatenation both kept more of the correctly assembled contigs from the individual *de novo* assemblies, and removed more of the incorrectly assembled contigs than EvidentialGene. These differences lead Concatenation to outperform EvidentialGene across every statistic (**Table 6**). The performance of the consensus approach varied based on the number of assemblers required.

Consensus 2 produced the most correctly assembled contigs of any method (6711), but at the cost of more incorrectly assembled contigs than Concatenation (14,433). However, both Consensus 3 and Consensus 4 kept the majority of the correctly assembled contigs while reducing the number of incorrectly assembled contigs by roughly half or three quarters, respectively. Consensus 4 had the highest precision (0.5861) and lowest FDR (0.4139) of any method. However, the additional reduction in the number of correctly assembled contigs lead to Consensus 3 having slightly higher accuracy* (0.2998) and F_1 score (0.4613).

In **Figure 4** all individual methods (both *de novo* and genome-guided) as well as ensemble methods are compared. Concatenation performed more poorly than Trinity despite the Trinity assembly forming part of the ensemble. In contrast, Consensus 3 kept more correctly assembled contigs than any individual assembly, with fewer incorrectly assembled than any approach except Consensus 4. This test highlights the weakness of ensemble assembly strategies to retain the incorrect version of a transcript, even if the correct version of the transcript exists in the individual assemblies. More robust methods, such as the consensus approaches we presented here, are needed to reliably improve over individual assemblies.

Methods	Total	Unique	Correct (%)	Incorrect	C/I
EvidentialGene	19,177	19,175	2267 (16.15)	16,908	0.1341
Concatenation	19,767	17,497	4697 (33.45)	12,800	0.3670
Consensus 2	21,444	21,444	6711 (47.80)	14,433	0.4650
Consensus 3	12,600	12,600	6144 (43.76)	6456	0.9517
Consensus 4	9095	9095	5331 (37.97)	3764	1.416

Table 5. Performance of individual genome-guided assemblers using default parameters on simulated RNAseq library with both the same and different references genome as the benchmark transcriptome. See **Table 1** for the description of numbers shown.

Methods	Precision	Recall	Accuracy*	F_1	FDR
EvidentialGene	0.1182	0.1615	0.0733	0.1365	0.8818
Concatenation	0.2684	0.3345	0.1750	0.2979	0.7316
Consensus 2	0.3174	0.4780	0.2357	0.3815	0.6826
Consensus 3	0.4876	0.4376	0.2998	0.4613	0.5124
Consensus 4	0.5861	0.3797	0.2994	0.4609	0.4139

Table 6. Performance statistics of ensemble assembly strategies using *de novo* assemblies on simulated RNAseq library.

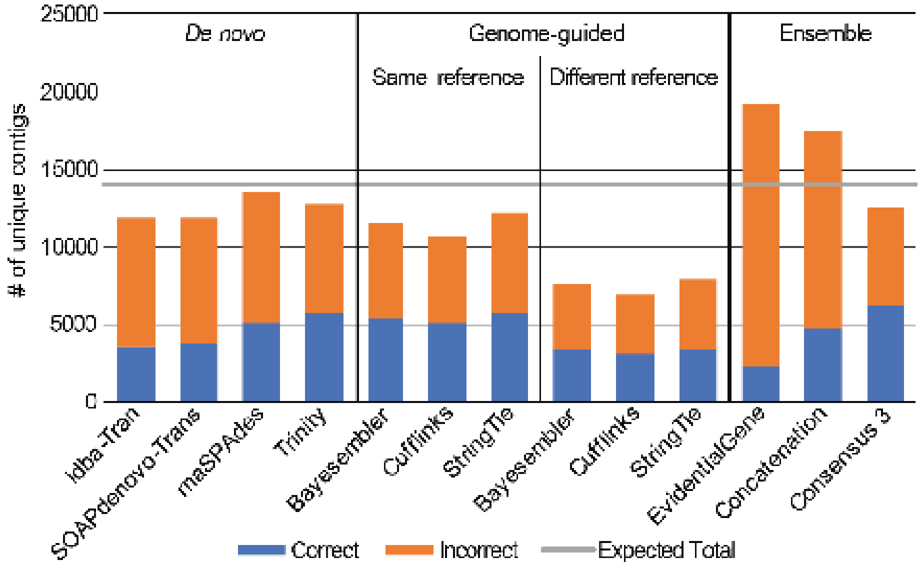


Figure 4. Performance comparison among all assemblers including *de novo*, genome-guided, and ensemble strategies. Simulated RNAseq data were used for testing, and the default parameters were used for each assembler. See **Tables 1**, **3**, and **5** for the actual numbers. The expected number of contigs is 14,040.

5. Conclusions

Transcriptome assembly can be approached from multiple different strategies. Historically, these approaches have revolved around assembling short but highly accurate Illumina reads with or without an existing genome assembly as a reference, referred to as genome-guided or *de novo* assemblers, respectively. All of the widely used *de novo* assemblers decompose the short reads into smaller kmers and use de Bruijn graphs built on these kmers to attempt to reconstruct the original transcripts. Due to the limitations of the de Bruijn graphs, this approach presents a trade-off between the uniqueness of the longer kmers and increased coverage of the shorter kmers. As a result, different kmer lengths can produce drastically different graphs, leading to large differences in the final assemblies.

Genome-guided assemblers avoid the limitations of the de Bruijn graphs by mapping the reads to the reference genome. This mapping, however, introduces its own limitations and trade-offs. Reads that are ambiguous between splice forms in the same genomic locations or across multiple genomic locations create similar challenges to the de Bruijn graphs. These ambiguities are compounded when the mapping must take into account mismatches due to sequencing errors as well as biological variations.

The limitations of the individual tools can potentially be overcome by combining multiple different assemblies in ensemble. As each tool and set of parameters results in a different set of correctly assembled contigs, accurately selecting these correctly assembled contigs without selecting any redundant incorrectly assembled contigs would leverage the strengths of each methods without the weaknesses of any. However, currently available ensemble strategies cannot guarantee that the correct sequence is chosen, leading to ensemble assemblies that are less accurate than individual assemblies. As the selection criteria for ensemble methods improve, such as with the "Consensus" approach shown here, these methods can also leverage new assembly approaches that can better handle certain subsets of transcripts (e.g. alternative splice forms) that may have other weaknesses that prevent them from being competitive as a general transcript assembly tool.

Overall, as our results demonstrated, transcriptome assemblers can still be improved, regardless of the approach used. While the genome-guided assemblers generally perform best when the assembly is performed against the same reference sequence that the RNAseq data was generated from, this is not always possible. When these sequences differ, the genome-guided assemblers may have lower accuracy than the *de novo* assemblers. While ensemble assembly strategies can potentially improve on accuracy over individual assemblies, it is also possible that they instead reduce the accuracy. Improving the performance of these tools, whether individual assemblers, ensemble strategies, or combined with long-read sequencing, will improve not only the accuracy of the reconstructed transcriptome but also the accuracy of downstream analyses, such as sequence annotation, quantification, and differential expression.

Acknowledgements

This work was supported by a grant from the National Science Foundation to ENM (Award #: 1339385).

Author details

Adam Voshall and Etsuko N. Moriyama*

*Address all correspondence to: emoriyama2@unl.edu

School of Biological Sciences and Center for Plant Science Innovation, University of
Nebraska-Lincoln, Lincoln, NE, USA

References

[1] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews. Genetics. 2009;**10**:57-63. DOI: 10.1038/nrg2484

[2] Ozsolak F, Milos PM. RNA sequencing: Advances, challenges and opportunities. Nature Reviews. Genetics. 2011;**12**:87-98. DOI: 10.1038/nrg2934

[3] Huang X, Chen XG, Armbruster PA. Comparative performance of transcriptome assembly methods for non-model organisms. BMC Genomics. 2016;**17**:523. DOI: 10.1186/s12864-016-2923-8

[4] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;**17**:13. DOI: 10.1186/s13059-016-0881-8

[5] Simonis M, Atanur SS, Linsen S, Guryev V, Ruzius FP, Game L, Lansu N, de Bruijn E, van Heesch S, Jones SJ, Pravenec M, Aitman TJ, Cuppen E. Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biology. 2012;**13**:r31. DOI: 10.1186/gb-2012-13-4-r31

[6] Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research. 2016;**26**:1134-1144. DOI: 10.1101/gr.196469.115

[7] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;**29**:644-652. DOI: 10.1038/nbt.1883

[8] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J. SOAPdenovo-trans: *De novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;**30**:1660-1666. DOI: 10.1093/bioinformatics/btu077

[9] Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: A more robust *de novo* de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;**29**:i326-i334. DOI: 10.1093/bioinformatics/btt219

[10] Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research. 2008;**18**:821-829. DOI: 10.1101/gr.074492.107

[11] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;**28**:1086-1092. DOI: 10.1093/bioinformatics/bts094

[12] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012;**19**:455-477. DOI: 10.1089/cmb.2012.0021

[13] Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T, Suhai S. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research. 2004;**14**:1147-1159. DOI: 10.1101/gr.1917404

[14] Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Reviews. Genetics. 2011;**12**:671-682. DOI: 10.1038/nrg3068

[15] Koren S, Treangen TJ, Hill CM, Pop M, Phillippy AM. Automated ensemble assembly and validation of microbial genomes. BMC Bioinformatics. 2014;**15**:126. DOI: 10.1186/1471-2105-15-126

[16] Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL. An ensemble strategy that significantly improves *de novo* assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Research. 2015;**43**:e46. DOI: 10.1093/nar/gkv002

[17] Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;**25**:1105-1111. DOI: 10.1093/bioinformatics/btp120

[18] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;**29**:15-21. DOI: 10.1093/bioinformatics/bts635

[19] Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature Methods. 2015;**12**:357-360. DOI: 10.1038/nmeth.3317

[20] Medina I, Tarraga J, Martinez H, Barrachina S, Castillo MI, Paschall J, Salavert-Torres J, Blanquer-Espert I, Hernandez-Garcia V, Quintana-Orti ES, Dopazo J. Highly sensitive and ultrafast read mapping for RNA-seq analysis. DNA Research. 2016;**23**:93-100. DOI: 10.1093/dnares/dsv039

[21] Maretty L, Sibbesen JA, Krogh A. Bayesian transcriptome assembly. Genome Biology. 2014;**15**:501. DOI: 10.1186/s13059-014-0501-4

[22] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010;**28**(5):511. DOI: 10.1038/nbt.1621

[23] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;**33**:290-295. DOI: 10.1038/nbt.3122

[24] Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. PLoS One. 2014;**9**:e91776. DOI: 10.1371/journal.pone.0091776

[25] Gilbert D. Gene-omes built from mRNA seq not genome DNA. 7th Annual Arthropod Genomics Symposium Notre Dame. 2013

[26] Cerveau N, Jackson DJ. Combining independent *de novo* assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics. 2016;**17**:525. DOI: 10.1186/s12859-016-1406-x

[27] Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy AS. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications. 2016;**7**:11706. DOI: 10.1038/ncomms11706

[28] Salmela L, Walve R, Rivals E, Ukkonen E. Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics. 2017;**33**:799-806. DOI: 10.1093/bioinformatics/btw321

[29] Salmela L, Rivals E. LoRDEC: Accurate and efficient long read error correction. Bioinformatics. 2014;**30**:3506-3514. DOI: 10.1093/bioinformatics/btu538

[30] Hargreaves AD, Mulley JF. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ. 2015;**3**:e1441. DOI: 10.7717/peerj.1441

[31] Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. Gigascience. 2017;**6**:1-13. DOI: 10.1093/gigascience/gix086

[32] Mei W, Liu S, Schnable JC, Yeh CT, Springer NM, Schnable PS, Barbazuk WB. A comprehensive analysis of alternative splicing in paleopolyploid maize. Frontiers in Plant Science. 2017;**8**:694. DOI: 10.3389/fpls.2017.00694

[33] Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nature Biotechnology. 2013;**31**:1009-1014. DOI: 10.1038/nbt.2705

[34] Minoche AE, Dohm JC, Schneider J, Holtgrawe D, Viehover P, Montfort M, Sorensen TR, Weisshaar B, Himmelbauer H. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biology. 2015;**16**:184. DOI: 10.1186/s13059-015-0729-7

[35] Zhang SJ, Wang C, Yan S, Fu A, Luan X, Li Y, Sunny Shen Q, Zhong X, Chen JY, Wang X, Chin-Ming Tan B, He A, Li CY. Isoform evolution in primates through independent combination of alternative RNA processing events. Molecular Biology and Evolution. 2017;**34**:2453-2468. DOI: 10.1093/molbev/msx212

[36] Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, et al. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. Genome Research. 2011;**21**:2224-2241. DOI: 10.1101/gr.126599.111

[37] Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: A tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database: The Journal of Biological Databases and Curation. 2016;**2016**: Article number: baw153. DOI: 10.1093/database/baw153

[38] O'Neil ST, Emrich SJ. Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics. 2013;**14**:465. DOI: 10.1186/1471-2164-14-465

[39] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;**215**:403-410. DOI: 10.1016/S0022-2836(05)80360-2

[40] The UniProt Consortium. UniProt: The universal protein knowledgebase. Nucleic Acids Research. 2017;**45**:D158-D169. DOI: 10.1093/nar/gkw1099

[41] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: The protein families database. Nucleic Acids Research. 2014;**42**:D222-D230. DOI: 10.1093/nar/gkt1223

[42] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: A library of protein families and subfamilies indexed by function. Genome Research. 2003;**13**:2129-2141. DOI: 10.1101/gr.772403

[43] Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution. 2018;**35**:543-548. DOI: 10.1093/molbev/msx319

[44] Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Research. 2001;**29**:2607-2618

[45] Wang S, Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. 2017;**33**: 327-333. DOI: 10.1093/bioinformatics/btw625

[46] Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;**12**:323. DOI: 10.1186/1471-2105-12-323

[47] Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Research. 2012;**40**:10073-10083. DOI: 10.1093/nar/gks666

[48] Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator—Toward accurate genome assembly. Bioinformatics. 2013;**29**:119-121. DOI: 10.1093/bioinformatics/bts649

[49] Yang C, Chu J, Warren RL, Birol I. NanoSim: Nanopore sequence read simulator based on statistical characterization. Gigascience. 2017;**6**:1-6. DOI: 10.1093/gigascience/gix010

[50] Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;**8**:e85024. DOI: 10.1371/journal.pone.0085024

[51] Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, Fenton J, Fenzl T, Fish J, Garcia-Gutierrez L, Garland P, Gluck J, Gonzalez I, Guermond S, Guo J, Gupta A, et al. The khmer software package: Enabling efficient nucleotide sequence analysis. F1000Research. 2015;**4**:900. DOI: 10.12688/f1000research.6924.1

Chapter 3

Modulation of Gene Expression by Gene Architecture and Promoter Structure

Aditya Kumar and Manju Bansal

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76051

Abstract

Regulation of gene expression is achieved by the presence of *cis* regulatory elements; these signatures are interspersed in the noncoding region and also situated in the coding region of the genome. These elements orchestrate the gene expression process by regulating the different steps involved in the flow of genetic information. Transcription (DNA to RNA) and translation (RNA to Protein) are controlled at different levels by different regulatory elements present in the genome. Current chapter describes the structural and functional elements present in the coding and noncoding region of the genome. Further we discuss role of regulatory elements in regulation of gene expression in prokaryotes and eukaryotes. Finally, we also discuss DNA structural properties of regulatory regions and their role in gene expression. Identification and characterization of *cis* regulatory elements would be useful to engineer the regulation of gene expression.

Keywords: DNA structural properties, gene architecture, gene expression, promoter structure

1. Introduction

Genome, the blue print of life, is essentially comprised of coding (genes) and noncoding (regulatory regions and other repetitive sequences) DNA. Genetic information is embedded in the form of coding regions (genes) that encode proteins. This flow of information from gene to proteins is a multistep pathway *viz*. transcription that is synthesis of RNA from the DNA and continues with the translation which is protein synthesis from RNA. Control of this flow of information is crucial for fate of the cell and this phenomenon is known as the regulation of the gene expression. The function of the cell is determined by the amount and type of the

RNA and protein molecules that is achieved by the regulation of the gene expression. There are various steps involved in this flow of information process such as chromatin domain organization, transcription (initiation, elongation and termination), post-transcriptional modification, RNA export (exclusive for eukaryotes), translation and mRNA degradation. Among all these different regulated stages of gene expression transcription initiation is the most utilized point of regulation. Transcription event is coupled with the translation process in the case of prokaryotes due to availability of ribosomes in the same compartment (due to lack of nucleus). However, transcription process is far more complicated in case of eukaryotes due to involvement of additional steps that are RNA splicing and RNA export. These additional steps provide extra stages for the regulation of gene expression process in eukaryotes.

Regulation of gene expression is achieved by harnessing the regulatory elements, located in the noncoding as well as coding regions of the genome. Current chapter focuses on the different structural and functional elements present in the coding regions (genes) and noncoding regions (regulatory regions), which are utilized by the cell to regulate the gene expression process.

2. Gene architecture

2.1. Noncoding elements of the genes

Genes are the repositories for primary information content of inheritance in genome and their expression determines the phenotypes, which in turn decides future of the cell in multicellular organisms. Functioning of gene products *viz*. mRNA (messenger RNA) and ncRNA (noncoding RNA) is modulated by complex gene regulatory networks. Eukaryotic genomes are mostly comprised of compositional properties (repetitive sequences, codon usage bias, mutational information, etc.) and functional signals (TATA box, Inr-element, cap signal, Kozak sequence, GT-AG splicing sites, etc.) [1]. Processing of the transcript is an important phase in the gene expression process, which also provides additional level of regulation in eukaryotes. Transcription and translation events are coupled in prokaryotes due to the availability of ribosomes to the mRNA while transcript undergoes several levels of processing in nucleus and finally processed transcripts are exported to the cytoplasm for translation in eukaryotes. Complexity in the gene structure results into the phenotypic diversity and this complexity arises from the occurrence and arrangement of the noncoding elements interspersing the coding region. Gene expression diversification is achieved by the presence of trailer sequences known as untranslated region (UTR) and intervening noncoding sequences known as introns [2]. These elements exert several direct and indirect functions.

2.2. Untranslated regions

UTRs are the trailer sequences located at the 5′ and 3′ end of the coding region which are the part of the transcribed mRNA but remain untranslated. Presence of alternative promoters or more than one transcription start site result into multiple 5′ UTRs which in turn controls the gene expression in several ways [3–6]. G quadruplex or G4 structure is a predominant secondary structure situated in the guanine rich 5′ UTRs which in turn hinders the translation process [7–9]. Highly and constitutively expressed genes are associated with short and poor in guanine base 5′ UTR in order to facilitate the translation process [10]. Sequential unwinding of natural stem loop structures located in the 5′ UTR in some mRNA is found to be associated with efficient translation [11–13].

IRES (internal ribosome entry sites), located usually upstream of the initiation codon (in the 5′ UTR) are responsible for the translation initiation in a cap independent mechanism by recruiting ribosome near the initiation site [5, 14–16]. The IRES mediated translational regulation occurs under certain stress conditions such as cellular stress, nutritional stress, mitotic stress etc. [17–19]. Conserved upstream open reading frames (uORFs) located in the 5′ UTR are also found to regulate protein translation, which are followed by main start codon (AUG) in the downstream [20–22]. Antibiotic resistance in the pathogenic bacteria is also found to be associated with uORFs mediated regulation [23]. In a recent study, fusion of uORF in the upstream of the auto-activated immune receptor gene developed the resistance to the plant diseases in Arabidopsis and rice [24].

Apart from these regulatory regions located in the 5′ UTR, the 3′ UTR is also rich in regulatory sequences located at the end of the coding sequence or gene. The conserved motif/s associated with 3′ UTR play crucial roles in gene expression at the posttranscriptional level. The 3′ UTR perform various regulatory functions, which are providing stability to the mRNA by polyadenylation, transcript cleavage, serve the binding site for microRNAs etc. Different isoforms of mRNA are derived from the alternative splicing and polyadenylation with alternative 3′ UTR. The varying expression levels and spatiotemporal localization for the same protein is achieved by differing 3′ UTR sequence in human [25–27]. AU rich elements (AREs) which are 50 to 150 nucleotide long and associated with multiple copies of pentanucleotide AUUUA regulate gene expression by stabilizing the mRNA [28, 29]. The abundance of AREs in the 3′ UTR of wide range of gene families indicates significant role in the gene regulation process [30]. MicroRNA response elements (MREs) are mostly located in the 3′ UTR region where single stranded 22 nucleotide long microRNA binds to regulate the expression of mRNA [31]. Poly(A) tail is stretch of adenosine (around 250 nucleotide) attached at the 3′ end of the RNA by adenylation process. The poly(A) binding proteins (PABP), specific class of regulatory proteins (nuclear and cytoplasmic) binds to the poly(A) tail and perform different regulatory functions like stability of mRNA, export and decay of the mRNA. These proteins play vital role in gene regulation [32–35].

2.3. Intronic regions

An intron is a noncoding DNA sequence that is transcribed but not translated; it is removed during the processing of pre-mRNA (precursor mRNA) to final mature RNA also known as RNA splicing. There are four different types of introns based on different splicing mechanisms. Spliceosomal introns are the foremost discovered and well characterized introns, which are excised by spliceosome, a ribonucleoprotein complex [36, 37]. Group I introns, widely present in mRNA, rRNA and tRNA of variety of organisms including algae, fungi, lower eukaryotes and few bacteria [38–42]. Similarly, group II introns are large autocatalytic ribozymes widely present in the mitochondria, chloroplast, plants, fungi, yeast and many bacteria, play major role in genome evolution [43–46]. The tRNA introns widely present in all domains of life are exceptionally different as enzymes are involved in the removal of intron and in the joining of the two halves [47–49]. Gene regulation is modulated to a great extent by count or number, length and position of the introns and they have several direct and indirect biological functions [50]. Multiple protein isoforms of the same gene are derived from the regulated alternative splicing process in eukaryotes [51–54]. Introns modulate gene expression either by the presence of transcriptional regulatory elements or by intron mediated enhancements [55–57]. Introns also regulate the gene expression by mediating the chromatin assembly (chromatin structure modulation) and controlling the mRNA export [58–61]. Apart from these direct biological functions, introns also exert indirect influence, for example position and length of the intron in the gene have potential role in the regulation of the expression level of the transcript [62–65].

3. Promoter structure

3.1. Different promoter elements

Promoters are stretch of genomic sequences where assembly of transcription machinery (RNAP and other accessory proteins) takes place prior to initiation of transcription. Although prokaryotic and eukaryotic polymerase shares functional similarity, promoter architecture differs in complexity [66]. Single type of RNA polymerase along with the specific σ factors recognizes promoter elements in prokaryotes [67]. Where −10 and −35 elements located in the upstream of the transcription start sites (TSSs) are recognized by different domains specific σ factors while UP element, an AT rich sequence situated from −40 to −60 is recognized by CTDs of α subunit of RNAP (**Figure 1**). An extension of extended −10 element, −15 element (TGnT) has been also proposed as new element situated from −15 to −12. It has been found that −15 element determines the overall promoter strength by complementing the weak −10 element.

On the other hand, complexity of promoter architecture in eukaryotes increases from yeast to mammals. Different types of RNA polymerases (normally three) are responsible for the generation of variety of RNA such as ribosomal RNA, messenger RNA (mRNA) and tRNA. As in case of bacterial RNAP, archaeal RNA polymerase and eukaryotic RNA Pol II (responsible for transcribing mRNA) also require specific factors and promoter elements to initiate transcription at specific sites in the genome. Eukaryotic promoters can be broadly classified in to three categories such as core, distal and proximal. The core promoter (approximately 50 nucleotide

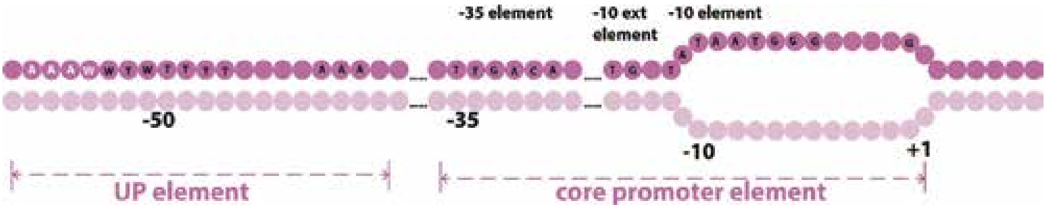


Figure 1. Prokaryotic promoter structure: UP element and core promoter elements (−35, −10 extended and −10 element).

sequence) is a platform where assembly of RNA polymerase and associated general transcription factors (GTFs), collectively referred as pre- initiation complex (PIC) takes place [68, 69]. Various promoter elements (**Table 1**) in the vicinity of the transcription start site; upstream and downstream regions are recognized by Pol II and other factors, such as TATA box, are recognized by TATA-binding protein (TBP), the B recognition element (BRE) by TFIIB and other elements by TBP-associated factors (TAFs) [70] (**Figure 2**). Apart from these, core promoter regions also consist of Inr element and may also contain downstream elements like downstream promoter element (DPE), motif ten element (MTE) (in humans) [71].

Proximal promoters are located in the immediate upstream (up to a few hundred base pairs) of core promoter, are comprised of GC box, CAAT box, *cis*-regulatory modules (CRM) etc. CpG islands are stretch of short DNA sequences, which are rich in GC content located in the upstream of house keeping and other regulated gene promoters [72, 73]. Proximal promoters mostly work as tethering element for distal promoters instead of acting as direct activators. On the other hand, distal promoters work from long distance. Enhancers, silencers and insulators are present in the distal promoter regions. Enhancers, also known as "promoters of promoter" mainly control specificity of gene expression by deploying unique enhancers in deferent cell types [74]. Multiple enhancers associated with single gene and single enhancer activating multiple genes provides additional level of diversity in phenotypes. In contrast to other regulators, enhancers exert their effects over tens of kilobases of DNA [75, 76]. Silencers are sequence specific elements where negative transcription factors bind to down regulate the gene expression [77]. Insulators are also referred to as boundary elements which block the effect of transcriptional activity of neighboring genes [77, 78].

3.2. Promoter structure and nucleosome dynamics

The locations and strengths of transcription factor and RNAP binding sites, also known as *cis*-regulatory elements and list of all nucleosome-binding sites are collectively defining the promoter structure. Nucleosomes are not only involved in the packaging of DNA but also bring order to the eukaryotic genome by regulating replication and transcription [79, 80]. Nucleosomes provide the first line of defense to avoid the unwanted transcription initiation.

Name	Location (relative to TSS at +1)	Associated factor/s
BRE^{u}	Upstream of the TSS	TFIIB
TATA box	−30/−31 to −23/−24	TBP
BRE^{d}	Downstream of the TATA box	TFIIB
Inr	−2 to +4/+5	TAF1 & TAF2
DCE	+6 to +11, +16 to +21, +30 to +34	TAF1
MTE	+18 to +29	TAF6 &TAF9
DPE	+28 to +33	TAF6 &TAF9

Table 1. List of core promoter elements and factors associated with them [72, 73].

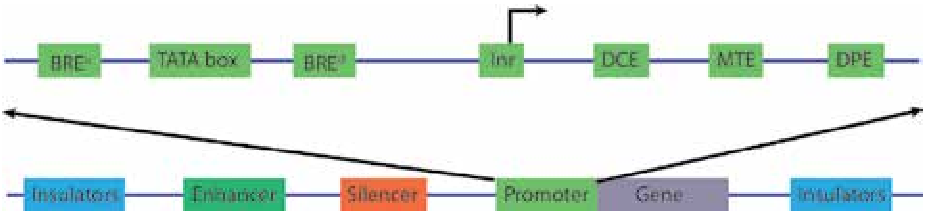


Figure 2. The different types of gene regulatory elements in eukaryotes.

Gene promoters involved in active transcription require accessibility to the DNA by RNAP machinery and associated factors, which is facilitated by nucleosome free region (NFR) or nucleosome-depleted region (NDR) [81, 82]. Nucleosome positioning is the probability of finding nucleosome at given genomic location relative to the surrounding locations while nucleosome occupancy refers to the average number nucleosomes present at the given genomic location in a given population of cells [83, 84]. Cellular gene expression is the final outcome of nucleosome dynamics, which itself depends on a complex interplay between nucleosome positioning and occupancy [85–87].

3.3. DNA structural properties of promoter regions

DNA sequence not only determines the distinct or base specific interactions but also determines the overall conformational shape, which is recognized by different proteins in case of non-base specific interactions [88]. The higher DNA binding specificity is achieved by combing different readout mechanisms by DNA binding proteins, with DNA shape playing an important role in gene regulation and genome organization [89]. The DNA sequence dependent structural properties can be roughly divided in to two categories, conformational and physiochemical [90]. Conformational properties represent the static DNA structure, which are influenced by geometry of base pair steps described by translational (shift, slide and rise) and rotational (tilt, roll and twist) parameters [91]. These also determine variation in the major and minor groove dimensions, which are crucial for DNA protein interactions. The physiochemical properties refer to the dynamic DNA structural properties such as persistence length, stress induced duplex destabilization, DNA duplex stability, protein induced bendability and intrinsic curvature etc.

Structural properties of given DNA sequence can be calculated using different di, tri tetra nucleotide models reported in experimental as well as theoretical studies. These models provide property values (lookup tables) for different oligonucleotides and using these values and appropriate length (sliding window), a given DNA sequence can be converted in to a series of numerical values referred to as a structural property profile. These profiles of given DNA sequence show variation in the structural property over the different regions of the sequence (**Figure 3**). An average structural property profile is calculated by taking mean of the feature value over all positions by aligning the different sequences [92]. DNA structural features such as low stability, protein induced bendability and intrinsic curvature are consistently observed in the prokaryotic and eukaryotic promoters (**Figure 4**) [93–96]. Promoter regions of different categories of transcripts (primary, internal, antisense and noncoding RNA) present in prokaryotic transcriptome show distinctly different DNA

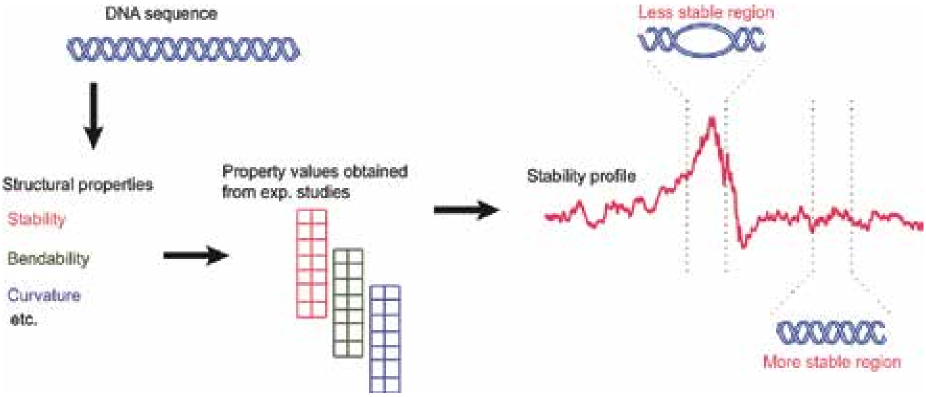


Figure 3. Schematic illustration showing DNA structural properties profile (example shown for DNA duplex stability) using values for di, tri, tetra nucleotide etc. obtained from experimental studies. Stability profile shows variation depending on the DNA sequence.

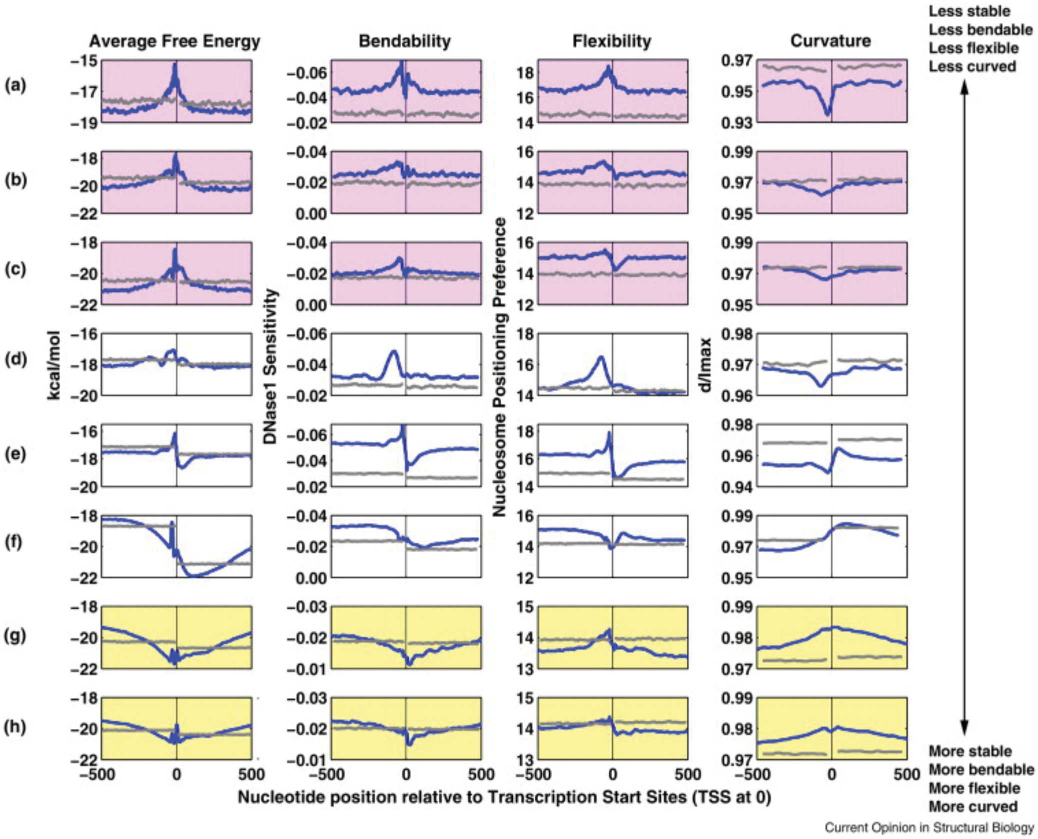


Figure 4. DNA sequence dependent structural properties of the promoter regions (−500 to +500 with respect to transcription start site at 0 position). Profiles of four structural properties (DNA duplex stability, DNase I sensitivity, Nucleosome Positioning Preference and intrinsic curvature) are shown of eight model systems: (a) *H. pylori*, (b) *E. coli*, (c) *K. pneumoniae*, (d) *S. cerevisiae*, (e) *C. elegans*, (f) rice, (g) mouse and (h) human. Figure taken (with permission) from Bansal et al., [96].

structural features [97]. Moreover, promoter regions of orthologous genes show conserved DNA structural properties in prokaryotes and plants [98–100]. These findings suggest that the DNA structural properties of promoter regions are conserved across the various classes of organisms.

4. Modulation of gene expression

The activity of RNAP and RNAPII in prokaryotes and eukaryotes respectively is tightly regulated to ensure proper level of gene expression. Transcription factors (TFs), proteins that bind to specific regulatory sequences (*cis*-regulatory elements or CRE) are the key regulators of transcription [101]. The complex gene regulation in eukaryotes is a consequence of the large number of transcription factors available and localization of *cis*-regulatory elements.

4.1. Gene expression noise and its regulation

A variation in the copy number of mRNA or protein molecules for a given gene in cell is referred as gene expression noise. It is largely under the control of regulatory DNA since it is linked with the promoter structure. TATA box with variable strength, transcription factor binding sites count, strength and their position in the promoter and nucleosome binding sites in the regulatory region have enormous effect on gene expression noise in eukaryotes [102]. Though transcriptional regulation is quite well understood at molecular level, very little is known about gene expression noise in the case of prokaryotes. Transcription factors and inducer molecules play a major role in gene regulation process. Additionally, genome condensation assisted by nucleoid associated proteins and DNA supercoiling also play a vital role in gene regulation in bacteria. Gene expression noise is essential for achieving phenotypic heterogeneity and it has been found to be universal in nature.

4.2. DNA structural properties and their role in gene expression

Nucleosome organization in the genome has been found to be closely associated with the gene expression and its variability [82, 84, 85]. Genes with dissimilar expression levels tend to have sequences with different structural features in order to attain the required nucleosome organization [103–105]. Plasticity of gene expression, also known as gene expression variability is crucial for cell survival, is closely linked with the DNA structural properties of promoter region in *S. cerevisiae*. Promoters of genes with low plasticity (less responsive) are less stable, less bendable and lower nucleosome occupancy compared to the promoters of genes with high plasticity (high responsive) [106, 107]. A recent study in six different prokaryotes with variable genomic GC content (ranging from 39–58%) shows good correlation between DNA structural properties of promoter regions and gene expression. It has been found that promoter regions associated with high gene expression are less stable, less bendable and more curved as compare to the genes associated with low gene expression as seen from **Figure 5**. Intrinsic curvature was found to be most significant property which is distinctly present in the promoter regions associated with high gene expression as compared to those with low

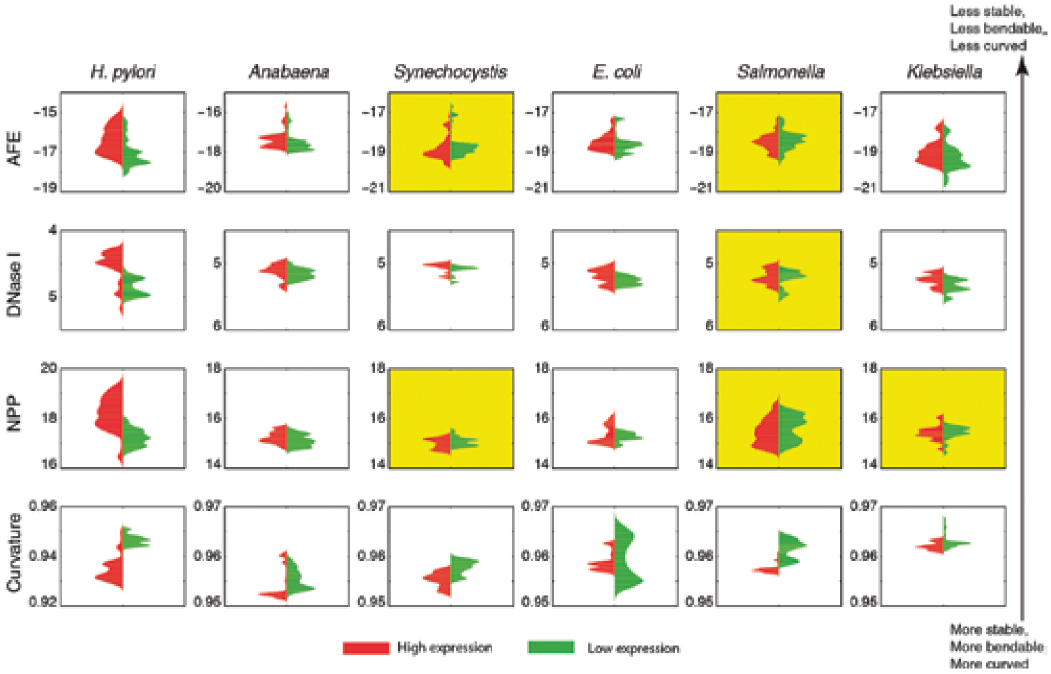


Figure 5. Violin plot of four DNA structural property values in the promoter regions (-100 to 0 nucleotide with respect to TSS at 0) associated with high and low gene expression in six different prokaryotes. The x-axis shows the probability density while y-axis represents the DNA structural feature value. Plots with yellow background indicate the cases which failed to reject the null hypothesis using two sample KS test at the level of significance of P = 0.01 (Figure taken from Kumar & Bansal, [97]).

gene expression across all organisms [97]. Hence estimation and characterization of DNA structural features of promoter regions could be very informative in analyzing the expression of associated gene.

5. Conclusions

The growing plethora of genomic information in the form of whole genome sequences requires its annotation to make sense of it. Mere delineation of coding sequences (gene identification) is not enough to get complete understanding of functional genomics since regulation of gene expression orchestrates the fate of cells. Gene expression regulation depends on different regulatory elements localized in the noncoding and coding region of the genome. Identification and characterization of these regulatory elements is the next level of challenge in the genome annotation process. Studies on DNA structural features of the regulatory regions show quite promising results toward achieving this goal. Moreover, DNA structural properties based characterization of regulatory regions is more sensitive and precise as compared to sequence-based approaches and most importantly it is universal in nature, applicable to all domains of life. Accumulating evidence shows a close relationship between gene expression and structural properties of promoter DNA; furthermore, this information can be used to engineer the regulatory sequences to modulate gene expression.

Acknowledgements

MB is Senior Scientist, Indian National Science Academy (INSA) and the recipient of J. C. Bose National Fellowship of Department of Science and Technology (DST), India.

Conflict of interest

None declared.

Author details

Aditya Kumar[1] and Manju Bansal[2*]

*Address all correspondence to: mb@iisc.ac.in

1 Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam, India

2 Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka, India

References

[1] Huang Y, Chen SY, Deng F. Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. Computational and Structural Biotechnology Journal. 2016;**14**:298-303. DOI: 10.1016/j.csbj.2016.07.002

[2] Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cellular and Molecular Life Sciences. 2012;**69**(21):3613-3634. DOI: 10.1007/s00018-012-0990-9

[3] Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szynkiewicz M, Speirs V, et al. Differential regulation of oestrogen receptor beta isoforms by 5′ untranslated regions in cancer. Journal of Cellular and Molecular Medicine. 2010;**14**(8):2172-2184. DOI: 10.1111/j.1582-4934.2009.00867.x

[4] Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences of the United States of America. 2009;**106**(18):7507-7512. DOI: 10.1073/pnas.0810916106

[5] Leppek K, Das R, Barna M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nature Reviews. Molecular Cell Biology. 2018;**19**(3):158-174. DOI: 10.1038/nrm.2017.103

[6] Srivastava AK, Lu Y, Zinta G, Lang Z, Zhu JK. UTR-dependent control of gene expression in plants. Trends in Plant Science. 2018;**23**(3):248-259. DOI: 10.1016/j.tplants.2017.11.003

[7] Bugaut A, Balasubramanian S. 5'-UTR RNA G-quadruplexes: Translation regulation and targeting. Nucleic Acids Research. 2012;**40**(11):4727-4741. DOI: 10.1093/nar/gks068

[8] Serikawa T, Spanos C, von Hacht A, Budisa N, Rappsilber J, Kurreck J. Comprehensive identification of proteins binding to RNA G-quadruplex motifs in the 5' UTR of tumor-associated mRNAs. Biochimie. 2018;**144**:169-184. DOI: 10.1016/j.biochi.2017.11.003

[9] Cammas A, Dubrac A, Morel B, Lamaa A, Touriol C, Teulade-Fichou MP, Prats H, Millevoi S. Stabilization of the G-quadruplex at the VEGF IRES represses cap-independent translation. RNA Biology. 2015;**12**(3):320-329. DOI: 10.1080/15476286.2015.1017236

[10] Pickering BM, Willis AE. The implications of structured 5′ untranslated regions on translation and disease. Seminars in Cell & Developmental Biology. 2005;**16**(1):39-47. DOI: 10.1016/j.semcdb.2004.11.006

[11] Dmitriev SE, Andreev DE, Terenin IM, Olovnikov IA, Prassolov VS, Merrick WC, Shatsky IN. Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5′ untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. Molecular and Cellular Biology. 2007;**27**(13):4685-4697. DOI: 10.1128/MCB.02138-06

[12] Kapp LD, Lorsch JR. The molecular mechanics of eukaryotic translation. Annual Review of Biochemistry. 2004;**73**:657-704. DOI: 10.1146/annurev.biochem.73.030403.080419

[13] Jackowiak P, Hojka-Osinska A, Gasiorek K, Stelmaszczuk M, Gudanis D, Gdaniec Z, Figlerowicz M. Effects of G-quadruplex topology on translational inhibition by tRNA fragments in mammalian and plant systems in vitro. The International Journal of Biochemistry & Cell Biology. 2017;**92**:148-154. DOI: 10.1016/j.biocel.2017.10.001

[14] Yamamoto H, Unbehaun A, Spahn CMT. Ribosomal chamber music: Toward an understanding of IRES mechanisms. Trends in Biochemical Sciences. 2017;**42**(8):655-668. DOI: 10.1016/j.tibs.2017.06.002

[15] Johnson AG, Grosely R, Petrov AN, Puglisi JD. Dynamics of IRES-mediated translation. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2017;**372**(1716). DOI: 10.1098/rstb.2016.0177

[16] Terenin IM, Smirnova VV, Andreev DE, Dmitriev SE, Shatsky IN. A researcher's guide to the galaxy of IRESs. Cellular and Molecular Life Sciences. 2017;**74**(8):1431-1455. DOI: 10.1007/s00018-016-2409-5

[17] Komar AA, Hatzoglou M. Internal ribosome entry sites in cellular mRNAs: Mystery of their existence. The Journal of Biological Chemistry. 2005;**280**(25):23425-23428. DOI: 10.1074/jbc.R400041200

[18] Nevins TA, Harder ZM, Korneluk RG, Holcik M. Distinct regulation of internal ribosome entry site-mediated translation following cellular stress is mediated by apoptotic

fragments of eIF4G translation initiation factor family members eIF4GI and p97/DAP5/NAT1. The Journal of Biological Chemistry. 2003;**278**(6):3572-3579. DOI: 10.1074/jbc.M206781200

[19] Sajjanar B, Deb R, Raina SK, Pawar S, Brahmane MP, Nirmale AV, Kurade NP, Manjunathareddy GB, Bal SK, Singh NP. Untranslated regions (UTRs) orchestrate translation reprogramming in cellular stress responses. Journal of Thermal Biology. 2017;**65**:69-75. DOI: 10.1016/j.jtherbio.2017.02.006

[20] Wethmar K, Smink JJ, Leutz A. Upstream open reading frames: Molecular switches in (patho)physiology. BioEssays. 2010;**32**(10):885-893. DOI: 10.1002/bies.201000037

[21] Barbosa C, Peixeiro I, Romao L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genetics. 2013;**9**(8):e1003529. DOI: 10.1371/journal.pgen.1003529

[22] Wethmar K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdisciplinary Reviews. RNA. 2014;**5**(6):765-778. DOI: 10.1002/wrna.1245

[23] Dar D, Sorek R. Regulation of antibiotic-resistance by non-coding RNAs in bacteria. Current Opinion in Microbiology. 2017;**36**:111-117. DOI: 10.1016/j.mib.2017.02.005

[24] Xu G, Yuan M, Ai C, Liu L, Zhuang E, Karapetyan S, Wang S, Dong X. uORF-mediated translation allows engineered plant disease resistance without fitness costs. Nature. 2017;**545**(7655):491-494. DOI: 10.1038/nature22372

[25] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science. 2008;**320**(5883):1643-1647. DOI: 10.1126/science.1155390

[26] Matoulkova E, Michalova E, Vojtesek B, Hrstka R. The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biology. 2012;**9**(5):563-576. DOI: 10.4161/rna.20231

[27] Dickson AM, Wilusz J. Polyadenylation: Alternative lifestyles of the A-rich (and famous?). The EMBO Journal. 2010;**29**(9):1473-1474. DOI: 10.1038/emboj.2010.67

[28] Eberle AB, Stalder L, Mathys H, Orozco RZ, Muhlemann O. Posttranscriptional gene regulation by spatial rearrangement of the 3′ untranslated region. PLoS Biology. 2008;**6**(4):e92. DOI: 10.1371/journal.pbio.0060092

[29] Chen CY, Shyu AB. AU-rich elements: Characterization and importance in mRNA degradation. Trends in Biochemical Sciences. 1995;**20**(11):465-470

[30] von Roretz C, Di Marco S, Mazroui R, Gallouzi IE. Turnover of AU-rich-containing mRNAs during stress: A matter of survival. Wiley Interdisciplinary Reviews. RNA. 2011;**2**(3):336-347. DOI: 10.1002/wrna.55

[31] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'xxUTR evolution. Cell. 2005;**123**(6):1133-1146. DOI: 10.1016/j.cell.2005.11.023

[32] Gorgoni B, Richardson WA, Burgess HM, Anderson RC, Wilkie GS, Gautier P, Martins JP, Brook M, Sheets MD, Gray NK. Poly(A)-binding proteins are functionally distinct and have essential roles during vertebrate development. Proceedings of the National Academy of Sciences of the United States of America. 2011;**108**(19):7844-7849. DOI: 10.1073/pnas.1017664108

[33] Wilkie GS, Gautier P, Lawson D, Gray NK. Embryonic poly(A)-binding protein stimulates translation in germ cells. Molecular and Cellular Biology. 2005;**25**(5):2060-2071. DOI: 10.1128/MCB.25.5.2060-2071.2005

[34] Smith RW, Blee TK, Gray NK. Poly(A)-binding proteins are required for diverse biological processes in metazoans. Biochemical Society Transactions. 2014;**42**(4):1229-1237. DOI: 10.1042/BST20140111

[35] Grenier St-Sauveur V, Soucek S, Corbett AH, Bachand F. Poly(A) tail-mediated gene regulation by opposing roles of Nab2 and Pab2 nuclear poly(A)-binding proteins in pre-mRNA decay. Molecular and Cellular Biology. 2013;**33**(23):4718-4731. DOI: 10.1128/MCB.00887-13

[36] Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. Cold Spring Harbor Perspectives in Biology. 2014;**14**(6). DOI: 10.1101/cshperspect.a016071

[37] Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biology Direct. 2012;**7**:11. DOI: 10.1186/1745-6150-7-11

[38] Haugen P, Simon DM, Bhattacharya D. The natural history of group I introns. Trends in Genetics. 2005;**21**(2):111-119. DOI: 10.1016/j.tig.2004.12.007

[39] Cech TR. Self-splicing of group I introns. Annual Review of Biochemistry. 1990;**59**(1): 543-568. DOI: 10.1146/annurev.bi.59.070190.002551

[40] Nielsen H, Johansen SD. Group I introns: Moving in new directions. RNA Biology. 2009;**6**(4):375-383

[41] Hedberg A, Johansen SD. Nuclear group I introns in self-splicing and beyond. Mobile DNA. 2013;**4**(1):17. DOI: 10.1186/1759-8753-4-17

[42] Raghavan R, Minnick MF. Group I introns and inteins: Disparate origins but convergent parasitic strategies. Journal of Bacteriology. 2009;**191**(20):6193-6202. DOI: 10.1128/JB.00675-09

[43] Pyle AM. Group II intron self-splicing. Annual Review of Biophysics. 2016;**45**:183-205. DOI: 10.1146/annurev-biophys-062215-011149

[44] Zhao C, Pyle AM. Structural insights into the mechanism of group II intron splicing. Trends in Biochemical Sciences. 2017;**42**(6):470-482. DOI: 10.1016/j.tibs.2017.03.007

[45] Zhao C, Pyle AM. The group II intron maturase: A reverse transcriptase and splicing factor go hand in hand. Current Opinion in Structural Biology. 2017;**47**:30-39. DOI: 10.1016/j.sbi.2017.05.002

[46] Novikova O, Belfort M. Mobile group II introns as ancestral eukaryotic elements. Trends in Genetics. 2017;**33**(11):773-783. DOI: 10.1016/j.tig.2017.07.009

[47] Fujishima K, Kanai A. tRNA gene diversity in the three domains of life. Frontiers in Genetics. 2014;**5**:142. DOI: 10.3389/fgene.2014.00142

[48] Yoshihisa T. Handling tRNA introns, archaeal way and eukaryotic way. Frontiers in Genetics. 2014;**5**:213. DOI: 10.3389/fgene.2014.00213

[49] Lopes RR, Kessler AC, Polycarpo C, Alfonzo JD. Cutting, dicing, healing and sealing: The molecular surgery of tRNA. Wiley Interdisciplinary Reviews. RNA. 2015;**6**(3):337-349. DOI: 10.1002/wrna.1279

[50] Jo BS, Choi SS. Introns: The functional benefits of introns in genomes. Genomics & Informatics. 2015;**13**(4):112-118. DOI: 10.5808/GI.2015.13.4.112

[51] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 2008;**40**(12):1413-1415. DOI: 10.1038/ng.259

[52] Mastrangelo AM, Marone D, Laido G, De Leonardis AM, De Vita P. Alternative splicing: Enhancing ability to cope with stress via transcriptome plasticity. Plant Science. 2012;**185-186**:40-49. DOI: 10.1016/j.plantsci.2011.09.006

[53] Barbazuk WB, Fu Y, McGinnis KM. Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Research. 2008;**18**(9):1381-1392. DOI: 10.1101/gr.053678.106

[54] Zhang C, Yang H, Yang H. Evolutionary character of alternative splicing in plants. Bioinformatics and Biology Insights. 2015;**9**(Suppl 1):47-52. DOI: 10.4137/BBI.S33716

[55] Shaul O. How introns enhance gene expression. The International Journal of Biochemistry & Cell Biology. 2017;**91**(Pt B):145-155. DOI: 10.1016/j.biocel.2017.06.016

[56] Rose AB. Intron-mediated regulation of gene expression. Current Topics in Microbiology and Immunology. 2008;**326**:277-290

[57] Gallegos JE, Rose AB. The enduring mystery of intron-mediated enhancement. Plant Science. 2015;**237**:8-15. DOI: 10.1016/j.plantsci.2015.04.017

[58] Brown SJ, Stoilov P, Xing Y. Chromatin and epigenetic regulation of pre-mRNA processing. Human Molecular Genetics. 2012;**21**(R1):R90-R96. DOI: 10.1093/hmg/dds353

[59] Valencia P, Dias AP, Reed R. Splicing promotes rapid and efficient mRNA export in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America. 2008;**105**(9):3386-3391. DOI: 10.1073/pnas.0800250105

[60] Carrillo Oesterreich F, Bieberstein N, Neugebauer KM. Pause locally, splice globally. Trends in Cell Biology. 2011;**21**(6):328-335. DOI: 10.1016/j.tcb.2011.03.002

[61] Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nature Structural & Molecular Biology. 2009;**16**(9):990-995. DOI: 10.1038/nsmb.1659

[62] Bradnam KR, Korf I. Longer first introns are a general property of eukaryotic gene structure. PLoS One. 2008;**3**(8):e3093. DOI: 10.1371/journal.pone.0003093

[63] Carmel L, Rogozin IB, Wolf YI, Koonin EV. Evolutionarily conserved genes preferentially accumulate introns. Genome Research. 2007;**17**(7):1045-1050. DOI: 10.1101/gr.5978207

[64] Park SG, Hannenhalli S, Choi SS. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics. 2014;**15**:526. DOI: 10.1186/1471-2164-15-526

[65] Zhang Q, Li H, Zhao XQ, Xue H, Zheng Y, Meng H, Jia Y, Bo SL. The evolution mechanism of intron length. Genomics. 2016;**108**(2):47-55. DOI: 10.1016/j.ygeno.2016.07.004

[66] Decker KB, Hinton DM. Transcription regulation at the core: Similarities among bacterial, archaeal, and eukaryotic RNA polymerases. Annual Review of Microbiology. 2013;**67**:113-139. DOI: 10.1146/annurev-micro-092412-155756

[67] Browning DF, Busby SJ. The regulation of bacterial transcription initiation. Nature Reviews. Microbiology. 2004;**2**(1):57-65. DOI: 10.1038/nrmicro787

[68] Kadonaga JT. Perspectives on the RNA polymerase II core promoter. Wiley Interdisciplinary Reviews: Developmental Biology. 2012;**1**(1):40-51. DOI: 10.1002/wdev.21

[69] Gupta K, Sari-Ak D, Haffke M, Trowitzsch S, Berger I. Zooming in on transcription Preinitiation. Journal of Molecular Biology. 2016;**428**(12):2581-2591. DOI: 10.1016/j.jmb.2016.04.003

[70] Roy AL, Singer DS. Core promoters in transcription: Old problem, new insights. Trends in Biochemical Sciences. 2015;**40**(3):165-171. DOI: 10.1016/j.tibs.2015.01.007

[71] Kadonaga JT. The DPE, a core promoter element for transcription by RNA polymerase II. Experimental & Molecular Medicine. 2002;**34**(4):259-264. DOI: 10.1038/emm.2002.36

[72] Vilar JM, Saiz L. DNA looping in gene regulation: From the assembly of macromolecular complexes to the control of transcriptional noise. Current Opinion in Genetics & Development. 2005;**15**(2):136-144. DOI: 10.1016/j.gde.2005.02.005

[73] Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011;**25**(10):1010-1022. DOI: 10.1101/gad.2037511

[74] Kim TK, Shiekhattar R. Architectural and functional commonalities between enhancers and promoters. Cell. 2015;**162**(5):948-959. DOI: 10.1016/j.cell.2015.08.008

[75] Krivega I, Dean A. Enhancer and promoter interactions-long distance calls. Current Opinion in Genetics & Development. 2012;**22**(2):79-85. DOI: 10.1016/j.gde.2011.11.001

[76] Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nature Reviews. Genetics. 2013;**14**(4):288-295. DOI: 10.1038/nrg3458

[77] Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics. 2006;**7**:29-59. DOI: 10.1146/annurev.genom.7.080505.115623

[78] Ali T, Renkawitz R, Bartkuhn M. Insulators and domains of gene expression. Current Opinion in Genetics & Development. 2016;**37**:17-26. DOI: 10.1016/j.gde.2015.11.009

[79] Iyer VR. Nucleosome positioning: Bringing order to the eukaryotic genome. Trends in Cell Biology. 2012;**22**(5):250-256. DOI: 10.1016/j.tcb.2012.02.004

[80] Ballare C, Zaurin R, Vicent GP, Beato M. More help than hindrance: Nucleosomes aid transcriptional regulation. Nucleus. 2013;**4**(3):189-194. DOI: 10.4161/nucl.25108

[81] Rando OJ, Ahmad K. Rules and regulation in the primary structure of chromatin. Current Opinion in Cell Biology. 2007;**19**(3):250-256. DOI: 10.1016/j.ceb.2007.04.006

[82] Jiang C, Pugh BF. Nucleosome positioning and gene regulation: Advances through genomics. Nature Reviews. Genetics. 2009;**10**(3):161-172. DOI: 10.1038/nrg2522

[83] Struhl K, Segal E. Determinants of nucleosome positioning. Nature Structural & Molecular Biology. 2013;**20**(3):267-273. DOI: 10.1038/nsmb.2506

[84] Teif VB. Nucleosome positioning: Resources and tools online. Briefings in Bioinformatics. 2016;**17**(5):745-757. DOI: 10.1093/bib/bbv086

[85] Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nature Reviews. Molecular Cell Biology. 2017;**18**(9):548-562. DOI: 10.1038/nrm.2017.47

[86] Liu MJ, Seddon AE, Tsai ZT, Major IT, Floer M, Howe GA, Shiu SH. Determinants of nucleosome positioning and their influence on plant gene expression. Genome Research. 2015;**25**(8):1182-1195. DOI: 10.1101/gr.188680.114

[87] Shukla A, Bhargava P. Regulation of tRNA gene transcription by the chromatin structure and nucleosome dynamics. Biochimica et Biophysica Acta. 2017. DOI: 10.1016/j.bbagrm.2017.11.008

[88] Harteis S, Schneider S. Making the bend: DNA tertiary structure and protein-DNA interactions. International Journal of Molecular Sciences. 2014;**15**(7):12335-12363. DOI: 10.3390/ijms150712335

[89] Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA recognition. Annual Review of Biochemistry. 2010;**79**:233-269. DOI: 10.1146/annurev-biochem-060408-091030

[90] Meysman P, Marchal K, Engelen K. DNA structural properties in the classification of genomic transcription regulation elements. Bioinformatics and Biology Insights. 2012;**6**:155-168. DOI: 10.4137/BBI.S9426

[91] Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, Harvey SC, Heinemann U, Lu XJ, Neidle S, Shakked Z, et al. A standard reference frame for the description of nucleic acid base-pair geometry. Journal of Molecular Biology. 2001;**313**(1):229-237. DOI: 10.1006/jmbi.2001.4987

[92] Yella VR, Kumar A, Bansal M. DNA structure and promoter engineering. In: Singh V, Dhar PK, editors. Systems and Synthetic Biology. Dordrecht: Springer Netherlands; 2015. pp. 241-254

[93] Kanhere A, Bansal M. Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. Nucleic Acids Research. 2005;**33**(10):3165-3175. DOI: 10.1093/nar/gki627

[94] Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW. A DNA structural atlas for Escherichia coli. Journal of Molecular Biology. 2000;**299**(4):907-930. DOI: 10.1006/jmbi.2000.3787

[95] Rangannan V, Bansal M. High-quality annotation of promoter regions for 913 bacterial genomes. Bioinformatics. 2010;**26**(24):3043-3050. DOI: 10.1093/bioinformatics/btq577

[96] Bansal M, Kumar A, Yella VR. Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. Current Opinion in Structural Biology. 2014;**25**:77-85. DOI: 10.1016/j.sbi.2014.01.007

[97] Kumar A, Bansal M. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression. DNA Research. 2017;**24**(1):25-35. DOI: 10.1093/dnares/dsw045

[98] Kumar A, Bansal M. Characterization of structural and free energy properties of promoters associated with primary and operon TSS in *Helicobacter pylori* genome and their orthologs. Journal of Biosciences. 2012;**37**(3):423-431

[99] Kumar A, Manivelan V, Bansal M. Structural features of DNA are conserved in the promoter region of orthologous genes across different strains of *Helicobacter pylori*. FEMS Microbiology Letters. Letters. 2016;**363**(18). DOI: 10.1093/femsle/fnw207

[100] Morey C, Mookherjee S, Rajasekaran G, Bansal M. DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes. Plant Physiology. 2011;**156**(3):1300-1315. DOI: 10.1104/pp.110.167809

[101] Wittkopp PJ, Kalay G. Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews. Genetics. 2011;**13**(1):59-69. DOI: 10.1038/nrg3095

[102] Sanchez A, Choubey S, Kondev J. Regulation of noise in gene expression. Annual Review of Biophysics. 2013;**42**:469-491. DOI: 10.1146/annurev-biophys-083012-130401

[103] Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research. 2008;**18**(7):1073-1083. DOI: 10.1101/gr.078261.108

[104] Choi JK, Kim YJ. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nature Genetics. 2009;**41**(4):498-503. DOI: 10.1038/ng.319

[105] Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biology. 2010;**8**(7):e1000414. DOI: 10.1371/journal.pbio.1000414

[106] Yella VR, Bansal M. DNA structural features and architecture of promoter regions play a role in gene responsiveness of *S. cerevisiae*. Journal of Bioinformatics and Computational Biology. 2013;**11**(6):1343001. DOI: 10.1142/S0219720013430014

[107] Yella VR, Kumar A, Bansal M. Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. Scientific Reports. 2018;**8**(1):4520. DOI: 10.1038/s41598-018-22129-8

Chapter 4

Alternative RNA Splicing: New Approaches for Molecular Marker Discovery in Cancer

Vanessa Villegas-Ruíz and Sergio Juárez-Méndez

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74415

Abstract

In cancer, several alterations driving cell transformation including: imbalances DNA, changes in gene expression as well as protein diversity. The transcriptional regulation is finely driving and controlled by a large number of molecules, including: SR proteins, hnRNPS, RNA, DNA, histones methylation, among others. However, in cancer the regulation is altered. It is little kwon regulations causing alternative splicing in healthy and human diseases. The alternative splicing plays an important role in the generation of diversity of transcripts its proteins resulting. The aberrant transcripts variants expressed in cancer have shown a great potential as biomarkers or therapeutics targets. In this manuscript, we showed the basic in alternative splicing and a simple method using available data for detection alternative transcripts expressed in the tree most common human cancer.

Keywords: alternative splicing, breast cancer, prostate cancer, gene expression, molecular markers

1. Introduction

The molecular biology of cancer is not completely understood. The human transcriptome is an important molecule that to be used as molecular marker, because the RNA is fractionated in coding and non-coding and the functions, locations and structure are very variables. However, in cancer is little known complexity of the transcriptome. In this chapter, we focused in the showed the landscape of the post-transcriptional modifications (RNA splicing), data mining and identification of alternative splicing of available microarray data. We think that splicing and alternative splicing is machinery that is high modified in cancer and the changes in the disease could play a very important role in the diagnosis and prognosis.

2. RNA splicing

The gene expression is orchestrated by means of great interaction of molecules including: SR proteins, hnRNPS, RNA, DNA, histones methylation, among others. The RNA is a fundamental molecule for the life. Recent studies have shown that the human transcriptome is fractionated in coding and non-coding RNA. Interestingly, the coding RNA is representing for only 2% of the human transcriptome and the remains is non-coding RNA, suggesting large versatility to generate protein diversity. The pre-RNA is matured by several events include, addition of a poly (A) tail in the 3′, 5′m7G cap in endings and RNA splicing; those modifications conferring RNA stability, transport efficiency to the cytoplasm, among others.

The RNA splicing involves several steps and includes specific signals that delimited intronic and exonic sequences (splice site, SS), and sequences that help to exon skipping such as: intronic splicing enhancer (ISE), intronic splicing silencer (ISS), exonic splicing enhancer (ESE) and exonic splicing silencer (ESS) [1]. In addition, five small nuclear ribonocleoproteins (snRNP; U1, U2, U4, U5 y U6) and more than 150 additional co-factors that contributing to splicing [2, 3].

Basically, four signals that include: branch point, polypyrimidine tract, splice site 5′ and splice site 3′ [4–6]. Moreover, sequential steps that confer topological changes between RNA and snRNPs forming E complex, A complex (ATP dependent), B complex and finally C complex or spliceosome, which is the catalytic complex [7]. Additionally, the RNA could be subjected to alternative exon skipping by means of alternative splicing AS. The AS is processed using the basic machinery of splicing, and SR and hnRNPs plays an essential role for alternative exon skipping.

The coding RNA is represented by ~25,000 genes, however, more than 300,000 transcripts have been reported [8, 9]. The difference between genes and transcripts is probability by alternative splicing (AS) regulation. Actually, we know that more than 80% of RNA coding are subjected to AS, promoting a great diversity of mRNA and consequently proteins. For example, in Drosophila melanogaster, the gene Down Syndrome Cell Adhesion Molecule (Dscam) could generate more than 38,000 different mRNAs by means of alternative splicing [10]. These findings showed the importance of AS for the biology of the cell.

On the other hand, the long non-coding RNAs (lncRNAs) also could be subject to alternative splicing. However, the diversity of lncRNAs transcripts has been poorly studied. It is thought that AS in lncRNAs could be implicated in several regulatory processes, mainly mediated RNA-RNA, RNA-DNA and RNA-Proteins interaction. All possibility interaction, probability could increase the complex regulatory process.

We will focus in coding RNA. The coding RNA only is representing for ~2% of total RNA. It is known that in eukaryotic cell there are more events of AS than another organism [11, 12]. Among AS patters we found alternative promoters, exon skipping, intron retention, mutually exclusive exons, exon scrambling, Alternative 5′ splice site, alternative 3′splice site, alternative polyadenylation [4]. Interestingly, the proteins product of AS could change their native functions. In several human diseases, the AS contributes to diverse cellular process including: cell proliferation, migration, adhesion, metastasis, among others [9, 13–16]. The transcripts subject to AS and their product could be used as molecular markers and therapeutics targets, because only it is expressed in the disease or its expression is increase [11, 17–19].

3. Alternative splicing and diseases

The AS is regulated by large number of proteins/non-coding RNAs/DNA and large complex network interactions among them provide the perfect capacity of cell regulation. In addition, the posttranscriptional regulation is orchestrated so finely that the cells have capacity to response rapidly before a stimulus and the cell adjust their proteome. Additionally, the cell is exposed daily to several toxics agents, UV radiation, promoting vulnerability to mutations and misregulation. Particularly, the mutations plays an important role in aberrant AS that cause diseases especially neuromuscular, neurodegenerative and multifactorial diseases as cancer [20].

Three sequences are extremely important for RNA processing and mature, the 5′, 3′ splice site; 5′, 3′ introns end and the branch point sequence, which is usually located at ~40 upstream of 3′splice site, because contain the specific sequences of recognition by spliceosome for precise exon joining [21]. However, mutations in those sites disrupt the correct spliceosome assembly. Approximately 10% of genetic diseases are cause by point mutations that disrupt the interaction between RNA and spliceosome [22, 23].

The class mutation and locations in the genome, contributing to different variants of AS such as: exon skipping, exon retention, alternative 5′ and 3′, among others. The severity of the disease could be represented by intensity of expression of the mutate gene, for example: In spinal muscular atrophy (SMA) the SMN2 gene has C → T change in the exon 7, this change promotes an exon skipping (SMNΔ7) and their expression is proportionally mayor ~80% than ~ 20% in the healthy. Other case is in Duchenne muscular dystrophy (DND), in the dystrophin gene there is a substitution of T → A in the exon 31, promoting this exon skipping. In cystic fibrosis, the exclusion of the exon 9 in CFTR modifying the severity of the disease. In the Peutz-Jeghers syndrome the alternative transcript of LBK1 is expressed as consequence of change IVS2 + 1A > G [24].

4. Alternative splicing and cancer

In cancer, several alterations are involved to cell transformation, recently studies have showed that the AS plays an important role in cancer development, because change the transcriptomics and consequently the proteome; contributing to cell transformation [19, 25]. However, there are few studies focused on the identification of transcripts variants in cancer. Computational studies in cancer derived of expression sequence tags has showed that the AS in cancer was slightly lower in tumors than normal tissues [26]. The question is what is the difference between AS in cancer tissues and normal tissues? The aberrant transcripts expressed in cancer have shown a great potential as biomarkers or therapeutics targets. In breast cancer, CD44 gene can to transcribe seven alternate transcripts; the transcript variants five and seven have been involved in diverse pathologies, but the transcript six only is expressed in metastatic cancer and tumorigenic cell lines. These finding suggesting that an alternative transcript six of CD44 could play role in metastasis process [27, 28]. The BRCA1 has been involved in diverse types of cancer, in breast malignancies the mutation c.591C > T is implicated in skipping of exon 18 in BRCA1 transcript, the mutation constitute an important prognostic factor in familiar breast and ovarian cancer [29]. In gastric cancer, the KIT gene has a deletion of ~40 nucleotides, this cleavage promotes aberrant AS and loss of functional protein resulting [30].

In the healthy cell a several proteins are key for DNA repair, transcript regulators, among others. The BCL protein is very important in programed cell death. However, the cancer cell up regulates the expression and AS of BCL-xL, promoting the expression of long protein involved in anti-apoptotic process. In contrast, with short protein BCL-xS is involved in the apoptosis [31]. In ovarian cancer was found a new alternative transcript of p53 (TP53INP2) its expression is strongly associated to migration and cell invasion [32] its expression is associated to adverse prognosis [33]. The leukemia is the most frequency malignance in childhood, this neoplasia is not the exception also has been identifying AS in several transcripts including: CCAR1 promote the complex Par-4/THAP1 y Notch3 [34] and confer unfavorable prognosis as well as hMLH1Delta6 [35]. The Ikaros is a suppressor tumor gene, the variant IK11 is associated to proliferation and anti-apoptotic process [36].

5. Alternative splice transcript/proteins as molecular markers and therapeutic targets

The great challenge in cancer is the identification of the molecular markers and therapeutic targets. The proteins and transcripts products of AS are a magnify molecules because open some new opportunities in cancer. The aberrant AS is a consequence of malignant transformation, the mutations and gene expression modulation promote the expression of new molecules that confers advantage to cancer cell, such as: cell proliferation, migration, invasion, evading programed death, among others. In this context, the identification of molecules expressed in cancer could be a best molecular marker as well as treatment targets, because only are expressed in pathological tissue. There is a little information about of AS profiles in cancer, nevertheless, some molecules have been used such as molecular markers. The CD44 isoforms be predictive to anti CD44 treatment in many types of cancer [37]. The androgen receptor AR-V7 has been used as a predictive marker [38], patients who expressed V7 isoforms are resistant to therapy using enzalutamide and abiraterone [39]. The isoforms of SLC39A14 are used to detection of non- invasive colorectal cancer and the isoform is specific of the colon and rectum [40, 41]. The new transcript variant of VNN1 also be specific of cancer colon cancer and is used detection by their specificity [42].

The prospect for treatment of cancer is based on antibodies specifics for isoform expressed exclusively in the disease. However, there are other strategies that also could be used with RNA target, such as: using stable antisense RNAs, this approach could be used in different types of RNAs (coding and non-coding RNA) inclusively pre-RNA. The interference RNA is other strategy used in the elimination of aberrant expressed transcripts or even splicing variant [43].

6. Alternative splicing methods for detection

The mRNA splice is easily visualized using several tools for molecular biology; the most used is the RT-PCR. The implications for AS detection using the PCR, is based on primers

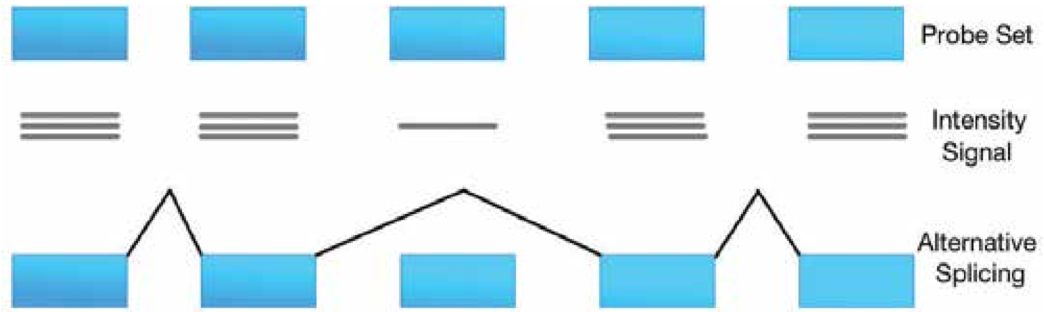


Figure 1. Representative probe set signal in microarray and alternative splicing detection. The figure showed in the top the probe set, the markers that inspect exon level expression. In the middle part depicted signal intensity in microarray hybridization. On the bottom the alternative splicing by low signal intensity is shown.

design. Usually the primers are flanking the exon skipping; however, this method not could detect novel splice transcripts. The in situ hybridization is other method that is used for AS detection as well as PCR this method no could detect novel splice sites. In the last 20 years has been developed massive method for detection gene expression. These tools have provided quickly gene expression profiles diseases-associated. Nowadays, gene expression microarrays and next generation sequencing are being used for detection novel molecules expressed in diverse diseases, include alternative mRNA spliced. Actually, the microarray gene expression (MGE) can measure exon expression, in this context, the low expression or suppression in particular probe set could be indicating AS **Figure 1**. Up to day, there are 25,252 assays performed with Affymetrix Human Exon 1.0 ST; 39,836 assays using Affymetrix GeneChip Human Gene 1.0 ST; and the most recently version 2422 assays with Affymetrix GeneChip Human Gene 2.0 ST, the experiments were performed between 8/7/07 and 8/8/16, 8/12/08 and 12/1/17 and 8/1/13 to 12/19/17, respectively each version array. Moreover, the microarrays data are available for data mining provides extraordinary information about profiles in human diseases including cancer. Additionally, the microarray analysis can be driving to explorer the AS.

7. Alternative splicing in the most common cancer types

The most common type of cancer is the breast cancer with more than 255,000 new cases expected in the United States in 2017, followed lung and prostate cancer according to National Cancer Institute. The question is Which are the transcripts alternatively spliced between normal and cancerous tissues? The major difficulty has been to determine whether the splicing changes detected in cancer are pathogenic [26]. Then we showed different analysis using high density microarrays to identify AS in three models of cancer. We performed data mining of Affymetrix microarrays. The data were download of ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) [44] or Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) [45]. The microarray analysis was performed using Partek Genomics Suite v7.17 according to previous reports [46].

8. Breast cancer

We performed analysis using the data set E-GEOD-81838 available in ArrayExpress web page or GEO-GSE81838 available in GEO, the data that we used was published by Lehmann et al. [47]. The data set was established by 10 breast tumors and 10 stromal cells. Our analysis showed 605 differential and alternatively spliced transcripts, the top 10 overexpress and suppressed are showed in the **Table 1**. We showed the most significant over and down expressed. The DTL transcript have 15 exons, is over expressed in tumor cell and DTL showed a potential alternative cap site **Figure 2** The FGF7 transcript have four exons, the heat map showed two apparent alternative site; cap and polyadenylation **Figure 3**.

Gene Symbol	RefSeq	p-value	Fold-Change
DTL	NM_001286229	1.45E-46	4.97616
ESRP1	NM_001034915	3.48E-77	3.86032
HOOK1	NM_015888	1.36E-71	3.69894
TTK	NM_001166691	3.31E-42	3.6733
GRHL1	NM_198182	4.13E-44	3.5722
ASPM	NM_001206846	7.69E-67	3.53763
DLGAP5	NM_001146015	9.97E-41	3.51451
OCLN	NM_001205254	1.02E-19	3.4842
ELF5	NM_001243080	1.35E-12	3.41892
TDRD5	NM_001199085	1.07E-42	3.33299
INHBA	NM_002192	6.07E-12	−2.99379
TSHZ2	NM_001193421	1.53E-05	−3.01074
FGF10	NM_001142556	1.27E-13	−3.01537
PDZRN4	NM_001164595	4.88E-21	−3.0344
NEXN	NM_001172309	6.90E-53	−3.03644
CXCL12	NM_000609	7.58E-23	−3.06879
IGF1	NM_000618	7.66E-16	−3.46073
COL8A1	NM_001850	6.71E-20	−3.50649
FGF7	NM_002009	3.49E-19	−4.07332
FGF7P2	OTTHUMT00000157659	6.90E-12	−4.13171

Table 1. Main genes whit potential alternative splicing in breast cancer.

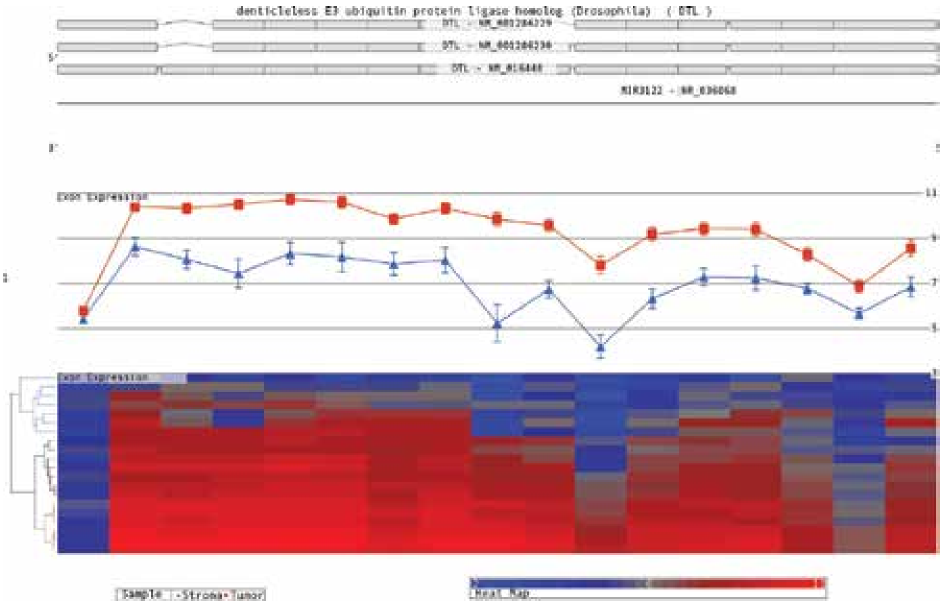


Figure 2. Differential exon expression of DTL gene. The figure showed in the top tree transcripts variants reported. The middle part sowed the level expression, the line red indicates tumor samples and blue indicate stroma samples. The heat map showed exon level expression on the far left, the exon is supressed suggesting an alternative splicing.

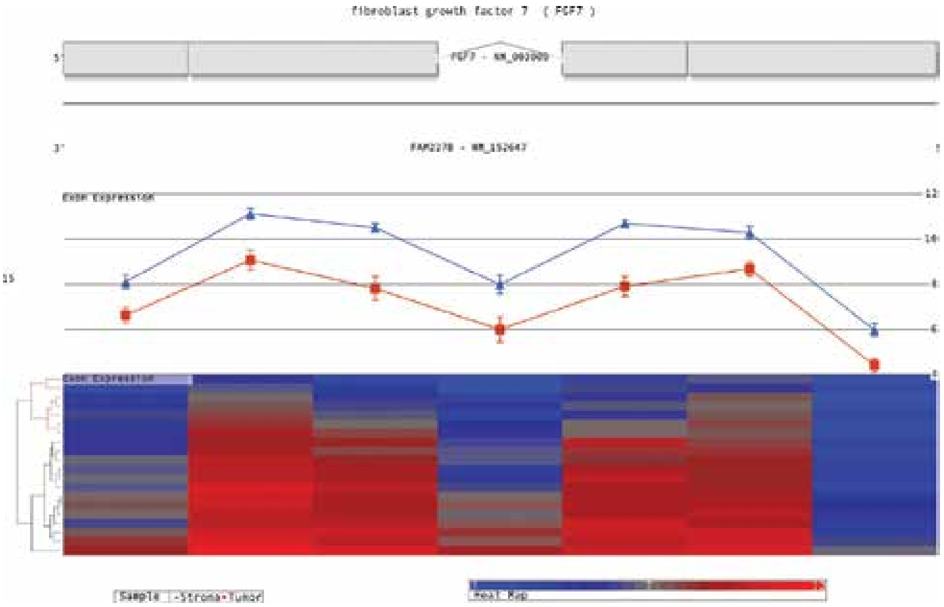


Figure 3. Differential exon expression of FGF7 gene. The figure showed in the top one transcript variant reported, in the middle indicates level expression; the blue line indicates stroma samples and the line red indicates tumor samples. The heat map showed exon level expression on the far left and right, the exons are supressed suggesting that there are two new potential alternative transcripts.

9. Lung cancer

The analysis in lung cancer was performed using data set E-GEOD-30979 available in ArrayExpress web page or GEO-GSE30979, the data was published by Leithner et al. [48]. The model was hypoxic-based in lung cancer. Our analysis revealed 101 transcripts expressed differentially also could have potentially alternative splicing, in the **Table 2** we showed the top 10 over and suppressed transcript identifies in this analysis. One of the most significant AS transcripts was LOX, this transcript has six exons. Our results showed a potential alternative site in cap **Figure 4**. The CEACAM6 was the supressed in hypoxic condition, also apparently showed an alternative cap site **Figure 5**.

Gene symbol	RefSeq	p-value	Fold-Change
MROH9	NM_001163629	8.67E-26	2.95561
LOX	NM_001178102	1.53E-21	2.66476
CLGN	NM_001130675	4.13E-18	2.49403
MME	NM_000902	1.73E-34	2.47655
DDIT3	NM_001195053	9.74E-10	2.46406
NUCB2	NM_005013	5.49E-34	2.41954
FICD	NM_007076	8.59E-07	2.36754
DNAJB9	NM_012328	6.35E-15	2.35193
GBE1	NM_000158	2.04E-60	2.24392
ADM	NM_001124	2.06E-08	2.20975
BPIFA1	NM_001243193	2.45E-05	−2.91878
IGKC	AF113887	9.63E-15	−2.93589
TOP2A	NM_001067	3.08E-75	−3.05223
SFTPB	NM_000542	2.40E-09	−3.0864
HP	NM_001126102	0.00045976	−3.12945
PI15	NM_015886	8.33E-14	−3.13972
IGKV3OR2–268	OTTHUMT00000330418	0.000711524	−3.52251
IGKV2D-30	OTTHUMT00000323285	0.00240377	−3.55542
CEACAM6	NM_002483	6.87E-11	−3.68686
CEACAM5	NM_001291484	2.69E-11	−3.8858

Table 2. Main genes whit potential alternative splicing in lung cancer.

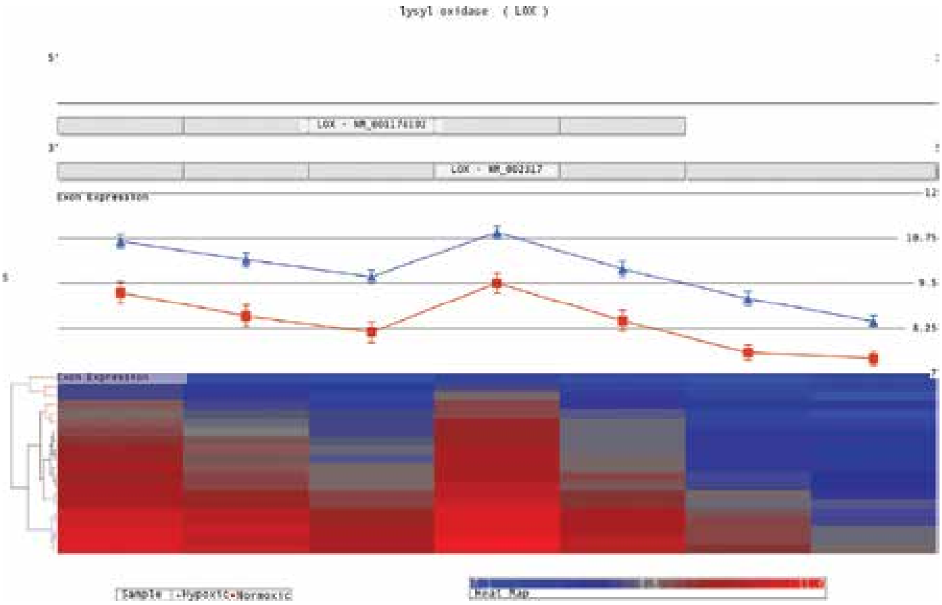


Figure 4. Differential exon expression of LOX gene. The figure showed in the top two alternative transcripts reported, the middle part the blue line indicates hypoxic model and the red line indicates normoxic model. The heat map showed exon level expressions on the far right two probe set are supressed, both markers inspection one exon. Our results could indicate the expression is the LOX NM_001178102 transcript variant.

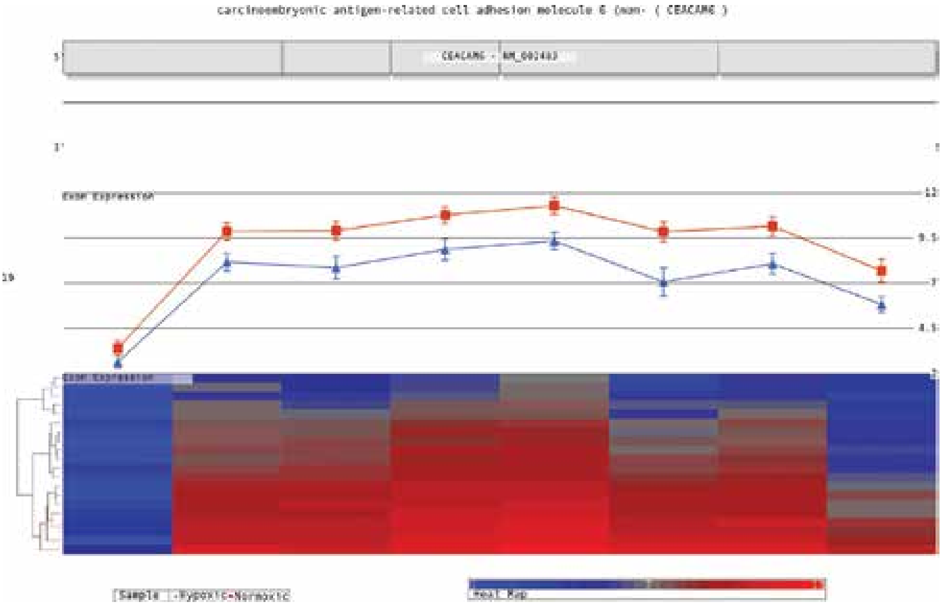


Figure 5. Differential exon expression of CEACAM6 gene. The figure showed in the top one transcripts, in the middle parte the blue line indicates hypoxic model and the red line indicates normoxic model. The heat map showed exon level expression, on the far left one marker is supressed indicating a potential fractioned exon, consequently alternative cap site.

10. Prostate cancer

The prostate cancer is one of the most common malignance in the worldwide. For this chapter we performed data mining using the data set E-GEOD-66852 available in ArrayExpress web page or GEO-GSE66852, the data was published by Nouri et al. [49]. Our results showed 777 transcripts that have significant differential exon expression, the most significant over and down expressed are shown in the **Table 3**. Our results showed the over expression in the CCDC80 transcript also showed an alternative spliced site in the exon six **Figure 6**. The down regulate transcript was DLGAP5, this transcript showed two potential sites of splicing; in the exon four and eight **Figure 7**.

Gene symbol	RefSeq	p-value	Fold-Change
CCDC80	NM_199511	2.98E-52	12.3906
PLA2G2A	NM_000300	1.48E-25	9.88943
PCDH11X	NM_001168360	6.22E-45	8.56495
RIMS1	NM_001168407	2.78E-104	7.77511
SI	NM_001041	1.17E-118	7.63493
IGFBP3	NM_000598	5.86E-35	7.56274
NLGN1	NM_014932	2.60E-22	7.29711
PCDH11X	NM_001168360	3.23E-36	6.91453
LRRN1	NM_020873	4.82E-16	6.45031
EPB41L4A	NM_022140	8.22E-55	6.22688
SHCBP1	NM_024745	1.26E-42	−14.8909
KIF20A	NM_005733	4.31E-69	−15.7441
HMMR	NM_001142556	5.84E-65	−16.2617
FAM111B	NM_001142703	3.52E-18	−17.0768
MELK	NM_001256685	3.51E-64	−17.2272
HIST1H3I	NM_003533	1.58E-10	−19.3335
TOP2A	NM_001067	5.06E-126	−23.1938
PBK	NM_001278945	4.11E-38	−24.005
NCAPG	NM_022346	2.24E-64	−24.1474
DLGAP5	NM_001146015	4.89E-73	−32.3098

Table 3. Main genes whit potential alternative splicing in prostate cancer.

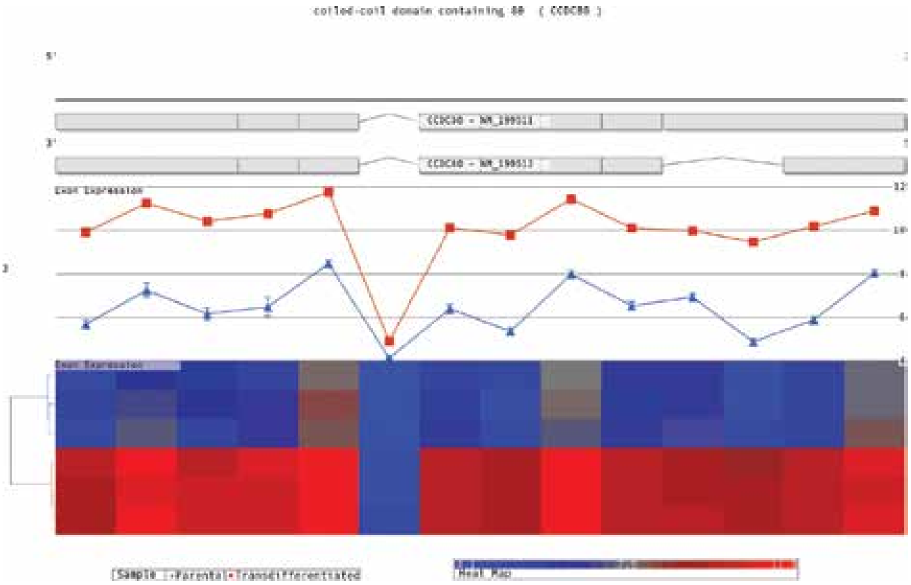


Figure 6. Differential exon expression of CCDC80 gene. The figure showed in the top two alternative transcripts, the middle part the blue line indicates parental cells model and the red line indicates transdifferentiated cells model. The heat map showed exon level expression, on the middle transcript one marker is supressed indicated by blue color in transdifferentiated model.

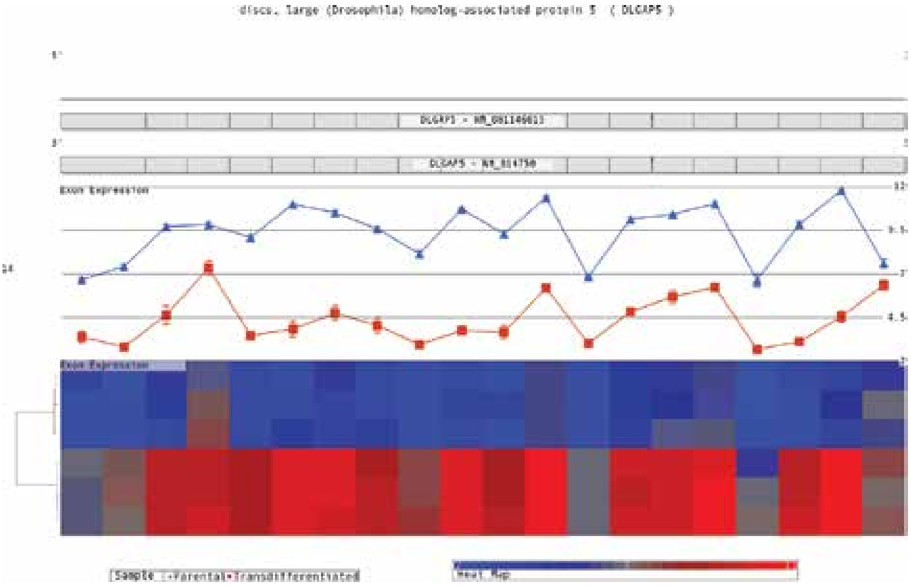


Figure 7. Differential exon expression of DLGAP5 gene. The figure showed in the top two alternative transcripts. The middle the blue line indicates parental cells model and the red line indicates transdifferentiated cells. The heat map showed exon level expression on the right side two markers were supressed in the parental model. Our results suggest two additional transcript variants non-reported are expressed in this model.

11. Conclusions

The alternative splicing is an important transcriptional mechanism that promote protein diversity. In cancer, several alterations in AS has been reported. In this chapter, we showed the generalities of alternative splicing process, the implications of AS in human diseases. The potential use of alternative transcript expressed in cancer as molecular markers and therapeutic targets. Finally, a simple method for identification of alterative transcripts expressed in three models of cancer using available dataset of Affymetrix.

Acknowledgements

This work was supported by SEP-CONACyT Basic Science grant 243233, Sectorial Fund for Research in Health and Social Security grant 272633 and Fiscal Resources, INP 2017.

Competing interests

The authors declare that they have no competing interests.

Author details

Vanessa Villegas-Ruíz and Sergio Juárez-Méndez*

*Address all correspondence to: ser.mend@gmail.com

Experimental Oncology Laboratory, Research Department, National Institute of Pediatrics, Mexico City, Mexico

References

[1] Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: Towards a cellular code. Nature Reviews. Molecular Cell Biology. 2005;**6**(5):386-398. DOI: 10.1038/nrm1645. Epub 2005/06/16;PubMed PMID: 15956978

[2] Kim E, Goren A, Ast G. Alternative splicing: Current perspectives. BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology. 2008;**30**(1):38-47. DOI: 10.1002/bies.20692. Epub 2007/12/18 PubMed PMID: 18081010

[3] Fu XD, Ares M Jr. Context-dependent control of alternative splicing by RNA-binding proteins. Nature Reviews Genetics. 2014;**15**(10):689-701. Epub 2014/08/13. DOI: 10.1038/nrg3778.PubMed PMID: 25112293; PubMed Central PMCID: PMCPMC4440546

[4] Chen J, Weiss WA. Alternative splicing in cancer: Implications for biology and therapy. Oncogene. 2015;**34**(1):1-14. Epub 2014/01/21. DOI: 10.1038/onc.2013.570. PubMed PMID: 24441040

[5] Chen M, Manley JL. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. Nature Reviews. Molecular Cell Biology. 2009;**10**(11):741-754. Epub 2009/09/24. DOI: 10.1038/nrm2777. PubMed PMID: 19773805; PubMed Central PMCID: PMCPMC2958924.

[6] Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. Alternative splicing: A pivotal step between eukaryotic transcription and translation. Nature Reviews. Molecular Cell Biology. 2013;**14**(3):153-165. Epub 2013/02/07. DOI: 10.1038/nrm3525. PubMed PMID: 23385723

[7] Li Q, Lee JA, Black DL. Neuronal regulation of alternative pre-mRNA splicing. Nature Reviews. Neuroscience. 2007;**8**(11):819-831. Epub 2007/09/27. DOI: 10.1038/nrn2237. PubMed PMID: 17895907

[8] Carninci P. Constructing the landscape of the mammalian transcriptome. The Journal of Experimental Biology. 2007;**210**(Pt 9):1497-1506. DOI: 10.1242/jeb.000406. PubMed PMID: 17449815

[9] Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. Genome Research. 1999;**9**(12):1288-1293. PubMed PMID: 10613851; PubMed Central PMCID: PMC310997

[10] Celotto AM, Graveley BR. Alternative splicing of the drosophila Dscam pre-mRNA is both temporally and spatially regulated. Genetics. 2001;**159**(2):599-608. Epub 2001/10/19. PubMed PMID: 11606537; PubMed Central PMCID: PMCPMC1461822

[11] Modrek B, Lee C. A genomic view of alternative splicing. Nature Genetics. 2002;**30**(1):13-19. DOI: 10.1038/ng0102-13. PubMed PMID: 11753382

[12] Graveley BR. Alternative splicing: Increasing diversity in the proteomic world. Trends in Genetics: TIG. 2001;**17**(2):100-107. PubMed PMID: 11173120

[13] Modrek B, Resch A, Grasso C, Lee C. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Research. 2001;**29**(13):2850-2859. PubMed PMID: 11433032; PubMed Central PMCID: PMC55780

[14] Kan Z, Rouchka EC, Gish WR, States DJ. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Research. 2001;**11**(5):889-900. DOI: 10.1101/gr.155001. PubMed PMID: 11337482; PubMed Central PMCID: PMC311065

[15] Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nature Reviews. Genetics. 2002;**3**(4):285-298. DOI: 10.1038/nrg775. PubMed PMID: 11967553

[16] Pick M, Flores-Flores C, Soreq H. From brain to blood: Alternative splicing evidence for the cholinergic basis of mammalian stress responses. Annals of the New York Academy of Sciences. 2004;**1018**:85-98. DOI: 10.1196/annals.1296.010. PubMed PMID: 15240356

[17] Stoilov P, Meshorer E, Gencheva M, Glick D, Soreq H, Stamm S. Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA and Cell Biology. 2002;**21**(11):803-818. DOI: 10.1089/104454902320908450. PubMed PMID: 12489991

[18] Hastings ML, Krainer AR. Pre-mRNA splicing in the new millennium. Current Opinion in Cell Biology. 2001;**13**(3):302-309. PubMed PMID: 11343900

[19] Nissim-Rafinia M, Kerem B. Splicing regulation as a potential genetic modifier. Trends in Genetics: TIG. 2002;**18**(3):123-127. PubMed PMID: 11858835

[20] Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell. 2009;**136**(4):777-793. DOI: 10.1016/j.cell.2009.02.011. Epub 2009/02/26. PubMed PMID: 19239895; PubMed Central PMCID: PMCPMC2866189

[21] Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. Proceedings of the National Academy of Sciences of the United States of America. 2001;**98**(20):11193-11198. Epub 2001/09/27. DOI: 10.1073/pnas.201407298. PubMed PMID: 11572975; PubMed Central PMCID: PMCPMC58706

[22] Cartegni L, Krainer AR. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nature Genetics. 2002;**30**(4):377-384. DOI: 10.1038/ng854. Epub 2002/04/02. PubMed PMID: 11925564

[23] Wang GS, Cooper TA. Splicing in disease: Disruption of the splicing code and the decoding machinery. Nature Reviews. Genetics. 2007;**8**(10):749-761. DOI: 10.1038/nrg2164. Epub 2007/08/30. PubMed PMID: 17726481

[24] Hastings ML, Resta N, Traum D, Stella A, Guanti G, Krainer AR. An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. Nature Structural & Molecular Biology. 2005;**12**(1):54-59. DOI: 10.1038/nsmb873. PubMed PMID: 15608654

[25] Philips AV, Cooper TA. RNA processing and human disease. Cellular and Molecular Life Sciences: CMLS. 2000;**57**(2):235-249. PubMed PMID: 10766020

[26] Kim E, Goren A, Ast G. Insights into the connection between cancer and alternative splicing. Trends in Genetics: TIG. 2008;**24**(1):7-10. Epub 2007/12/07. DOI: 10.1016/j.tig.2007.10.001. PubMed PMID: 18054115

[27] Gunthert U, Hofmann M, Rudy W, Reber S, Zoller M, Haussmann I, et al. A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. Cell. 1991;**65**(1):13-24. PubMed PMID: 1707342

[28] Sneath RJ, Mangham DC. The normal structure and function of CD44 and its role in neoplasia. Molecular Pathology : MP. 1998;**51**(4):191-200. PubMed PMID: 9893744; PubMed Central PMCID: PMC395635

[29] Dosil V, Tosar A, Canadas C, Perez-Segura P, Diaz-Rubio E, Caldes T, et al. Alternative splicing and molecular characterization of splice site variants: BRCA1 c.591C>T as a case study. Clinical Chemistry. 2010;**56**(1):53-61. DOI: 10.1373/clinchem.2009.132274. PubMed PMID: 19892845

[30] Chen LL, Sabripour M, Wu EF, Prieto VG, Fuller GN, Frazier ML. A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. Oncogene. 2005;**24**(26):4271-4280. DOI: 10.1038/sj.onc.1208587. PubMed PMID: 15824741

[31] Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA, et al. Bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell. 1993;**74**(4):597-608. PubMed PMID: 8358789

[32] Moran-Jones K, Grindlay J, Jones M, Smith R, Norman JC. hnRNP A2 regulates alternative mRNA splicing of TP53INP2 to control invasive cell migration. Cancer Research. 2009;**69**(24):9219-9227. DOI: 10.1158/0008-5472.CAN-09-1852. PubMed PMID: 19934309

[33] Hofstetter G, Berger A, Fiegl H, Slade N, Zoric A, Holzer B, et al. Alternative splicing of p53 and p73: The novel p53 splice variant p53delta is an independent prognostic marker in ovarian cancer. Oncogene. 2010;**29**(13):1997-2004. DOI: 10.1038/onc.2009.482. PubMed PMID: 20101229

[34] Lu C, Li JY, Ge Z, Zhang L, Zhou GP. Par-4/THAP1 complex and Notch3 competitively regulated pre-mRNA splicing of CCAR1 and affected inversely the survival of T-cell acute lymphoblastic leukemia cells. Oncogene. 2013;**32**(50):5602-5613. DOI: 10.1038/onc.2013.349. PubMed PMID: 23975424; PubMed Central PMCID: PMC3898485

[35] Peasland A, Matheson E, Hall A, Irving J. Alternative splicing of hMLH1 in childhood acute lymphoblastic leukaemia and characterisation of the variably expressed Delta9/10 isoform as a dominant negative species. Leukemia Research. 2010;**34**(3):322-327. DOI: 10.1016/j.leukres.2009.08.015. PubMed PMID: 19767099

[36] Capece D, Zazzeroni F, Mancarelli MM, Verzella D, Fischietti M, Di Tommaso A, et al. A novel, non-canonical splice variant of the Ikaros gene is aberrantly expressed in B-cell lymphoproliferative disorders. PLoS One. 2013;**8**(7):e68080. DOI: 10.1371/journal.pone.0068080. PubMed PMID: 23874502; PubMed Central PMCID: PMC3706598

[37] Birzele F, Voss E, Nopora A, Honold K, Heil F, Lohmann S, et al. CD44 isoform status predicts response to treatment with anti-CD44 antibody in cancer patients. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. 2015;**21**(12):2753-2762. Epub 2015/03/13. DOI: 10.1158/1078-0432.CCR-14-2141. PubMed PMID: 25762343

[38] Lu J, Van der Steen T, Tindall DJ. Are androgen receptor variants a substitute for the full-length receptor? Nature Reviews. Urology. 2015;**12**(3):137-144. Epub 2015/02/11. DOI: 10.1038/nrurol.2015.13. PubMed PMID: 25666893

[39] Antonarakis ES, Lu C, Wang H, Luber B, Nakazawa M, Roeser JC, et al. AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. The New England Journal of Medicine. 2014;**371**(11):1028-1038. Epub 2014/09/04. DOI: 10.1056/NEJMoa1315815. PubMed PMID: 25184630; PubMed Central PMCID: PMCPMC4201502

[40] Sveen A, Bakken AC, Agesen TH, Lind GE, Nesbakken A, Nordgard O, et al. The exon-level biomarker SLC39A14 has organ-confined cancer-specificity in colorectal cancer.

International Journal of Cancer. 2012;**131**(6):1479-1485. Epub 2011/12/17. DOI: 10.1002/ijc.27399. PubMed PMID: 22173985

[41] Thorsen K, Mansilla F, Schepeler T, Oster B, Rasmussen MH, Dyrskjot L, et al. Alternative splicing of SLC39A14 in colorectal cancer is regulated by the Wnt pathway. Molecular & Cellular Proteomics. 2011;**10**(1):M110 002998. Epub 2010/10/13. DOI: 10.1074/mcp.M110.002998. PubMed PMID: 20938052; PubMed Central PMCID: PMCPMC3013455

[42] Lovf M, Nome T, Bruun J, Eknaes M, Bakken AC, Mpindi JP, et al. A novel transcript, VNN1-AB, as a biomarker for colorectal cancer. International Journal of Cancer. 2014;**135**(9):2077-2084. Epub 2014/04/02. DOI: 10.1002/ijc.28855. PubMed PMID: 24687856

[43] Grimm D, Kay MA. Therapeutic application of RNAi: Is mRNA targeting finally ready for prime time? The Journal of Clinical Investigation. 2007;**117**(12):3633-3641. Epub 2007/12/07. DOI: 10.1172/JCI34129. PubMed PMID: 18060021; PubMed Central PMCID: PMCPMC2096424

[44] ArrayExpress. Available from: https://www.ebi.ac.uk/arrayexpress/ [Accessed: 2017-10-29]; 2017

[45] Gene Expression Omnibus. Available from: https://www.ncbi.nlm.nih.gov/geo/ [Accessed: 2017-10-29]; 2017

[46] Juarez-Mendez S, Zentella-Dehesa A, Villegas-Ruiz V, Perez-Gonzalez OA, Salcedo M, Lopez-Romero R, et al. Splice variants of zinc finger protein 695 mRNA associated to ovarian cancer. Journal of Ovarian Research. 2013;**6**(1):61. DOI: 10.1186/1757-2215-6-61. PubMed PMID: 24007497; PubMed Central PMCID: PMC3847372

[47] Lehmann BD, Jovanovic B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of triple-negative breast cancer molecular subtypes: Implications for Neoadjuvant chemotherapy selection. PLoS One. 2016;**11**(6):e0157368. Epub 2016/06/17. DOI: 10.1371/journal.pone.0157368. PubMed PMID: 27310713; PubMed Central PMCID: PMCPMC4911051

[48] Leithner K, Wohlkoenig C, Stacher E, Lindenmann J, Hofmann NA, Galle B, et al. Hypoxia increases membrane metallo-endopeptidase expression in a novel lung cancer ex vivo model – Role of tumor stroma cells. BMC Cancer. 2014;**14**:40. Epub 2014/01/28. DOI: 10.1186/1471-2407-14-40. PubMed PMID: 24460801; PubMed Central PMCID: PMCPMC3905926

[49] Nouri M, Caradec J, Lubik AA, Li N, Hollier BG, Takhar M, et al. Therapy-induced developmental reprogramming of prostate cancer cells and acquired therapy resistance. Oncotarget. 2017;**8**(12):18949-18967. Epub 2017/02/02. DOI: 10.18632/oncotarget.14850 PubMed PMID: 28145883; PubMed Central PMCID: PMCPMC5386661

Chapter 5

A Novel Approach to Mine for Genetic Markers via Comparing Class Frequency Distributions of Maximal Repeats Extracted from Tagged Whole Genomic Sequences

Jing-Doo Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75113

Abstract

The cost to extract one new biomarker within genomic sequences is very huge. This chapter adopts a scalable approach, developed previously and based on MapReduce programming model, to extract maximal repeats from a huge amount of tagged whole genomic sequences and meanwhile computing the similarities of sequences within the same class and the differences among the other classes, where the types of classes are derived from those tags. The work can be extended to any kind of genomic sequential data if one can have the organisms into several disjoint classes according to one specific phenotype, and then collect the whole genomes of those organisms. Those patterns, for example, biomarkers, if exist in only one class, with distinctive class frequency distribution can provide hints to biologists to dig out the relationship between that phenotype and those genomic patterns. It is expected that this approach may provide a novel direction in the research of biomarker extraction via whole genomic sequence comparison in the era of post genomics.

Keywords: biomarker, comparative genomics, class frequency distribution, maximal repeat, MapReduce programming

1. Introduction

It is very attractive and challenging to discover markers [1] from genomic sequences and then to use these markers for genetic tests [2] to diagnose diseases and for personalized medicine to adverse drug responses [3, 4]. Nowadays, genome-wide association studies (GWASs) [5] have already examined single-nucleotide polymorphisms (SNPs) across human genomes to identify specific SNPs related to some diseases, for example, diabetes, heart abnormalities, Parkinson disease, and Crohn disease [6]. Furthermore, GWAS is also used to predict cancer [7] and to influence human intelligence [8].

Most of GWASs are achieved with SNP arrays [9]. The "Illumina" [10] uses the "whole-genome genotyping" to interrogate SNPs across the entire genome to obtain the most comprehensive view of genomic variation; the Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers which includes more than 906,600 single-nucleotide polymorphisms (SNPs) [11]. The majority of these SNPs are designed to investigate the coding regions of genes in genomic sequences. However, some of the non-coding regions, once being mistaken as "junk DNA," are believed to contain functions to regulate gene transcription and to account for the genetic differences between individuals [12]. Although, on the one hand, the number of SNPs on one chip may be several hundreds of thousands, on the other hand, its coverage is still not enough [13] to figure out the relationship between genotypes and phenotypes in humans as given in the database of "dbGaP" [14].

As the era of post genomics with Next-Generation Sequencing (NGS) is coming, it is expected that the cost of genomic sequencing is decreasing and the availability of complete whole genomes of individual creatures is becoming popular. After using NGS for DNA sequencing [15], as shown on the right side in **Figure 1**, for example, one creature, for example, a virus, is supposed to contain three chromosomes with eight genotypes. On the other side of **Figure 1**, there are three phenotypes, for example, "Drug Resistance" "Envelope," and "Contents," inspected and detected by three domain experts, respectively. Under the assumption that these three phenotypes are totally dominated by those eight genotypes, represented as different icons, without considering the epigenetics [16], as shown in **Figure 2**, it is difficult for biologists in wet laboratory to analyze aimlessly the relationships among these phenotypes and those

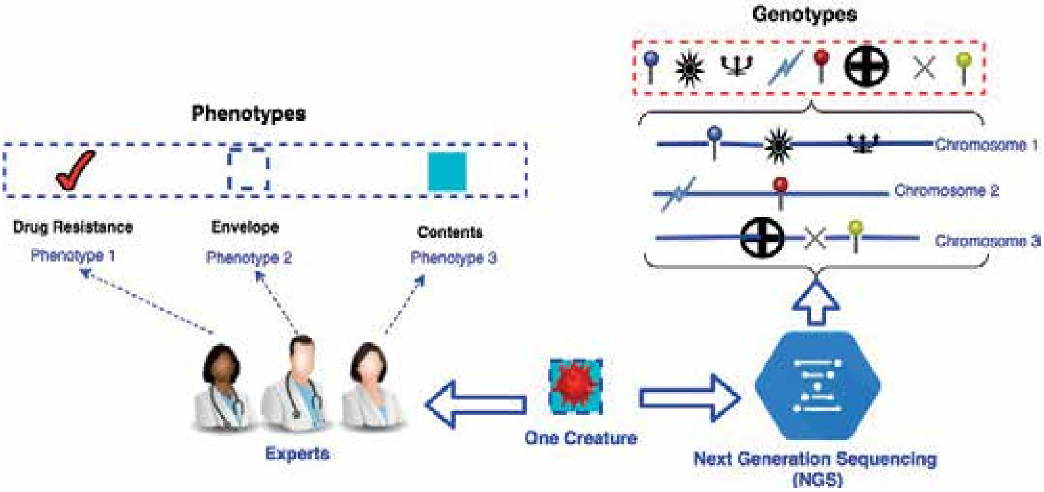


Figure 1. An example of one creature with three phenotypes and eight genotypes.

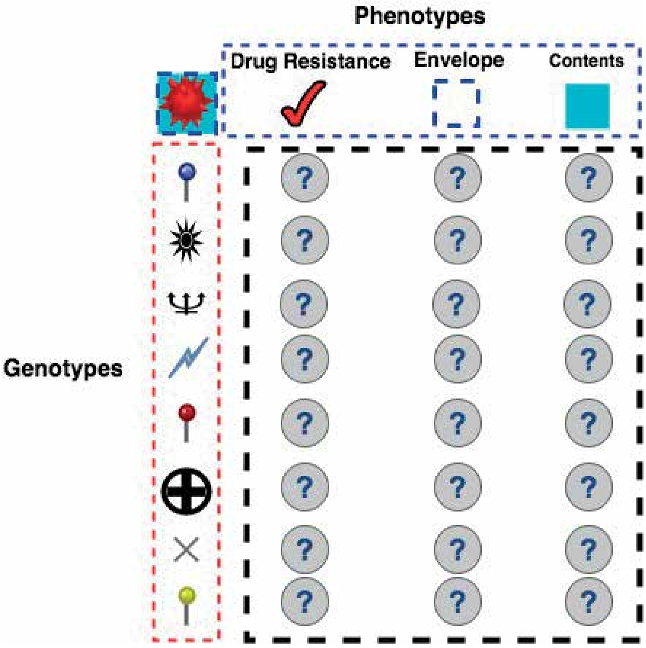


Figure 2. Example: how to identify the relationships among genotypes and phenotypes as described in **Figure 1**.

genotypes without further bioinformatics information or techniques such as comparative genomics [17].

With more and more complete whole genomes of distinctive creatures being available and popular in the coming days, it is very interesting and desired to extract common significant subsequences as candidate genomic markers as genotypes via comparing these creatures' whole DNA sequences according to the classes (or types) of their phenotypes observed and specified by domain experts. **Figure 3** shows the conceptual diagram of the corresponding classes for each of these three phenotypes given in **Figure 1**. With precise observations or experiments (phenotypes), biologists or experts can divide these creatures with complete whole genomes into disjoint classes if possible. Then, it is highly expected for biologists that some distinctive patterns (genotypes) hidden within their DNA sequences can be extracted as the candidates of class markers (phenotypes) if the frequency distributions of these patterns among classes are extremely biased, or some patterns are just in one class solely and appear in all instances belonging to that class ideally. To achieve the earlier-mentioned goal, one needs to extract repeats and to compute class frequency distributions of these repeats from a huge amount of tagged genomic sequences, where the types of classes are derived from the tags.

Due to the availability of genomic sequences in National Center for Biotechnology Information (NCBI) [18], The Cancer Genome Altas (TCGA) [19], it is interesting to have class frequency distribution of maximal repeats from these tagged genomic sequences for mining the biomarker or specific patterns. As the age of Next-Generation Sequencing (NGS) is going to be introduced for the project "Cancer Moonshot" in the National Cancer Institute [20], it is very attractive to identify specific biomarkers from these genomic sequences with tags, such as cancer types or distinctive genotypes. **Figure 4** gives the conceptual diagram of how to reduce the gap between phenotypes and genotypes by using the phenotypes as classes to identify those subsequences that appear in unique class only as biomarkers.

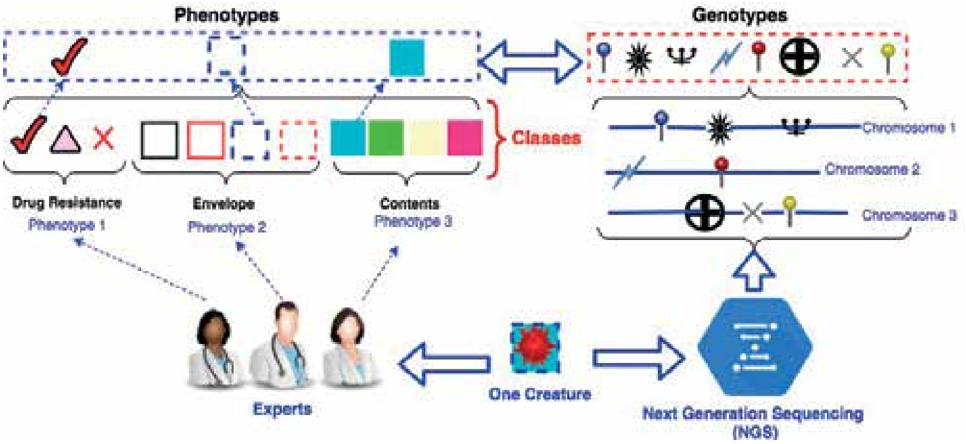


Figure 3. Mining the relationship of phenotypes and genotypes via classes comparison.

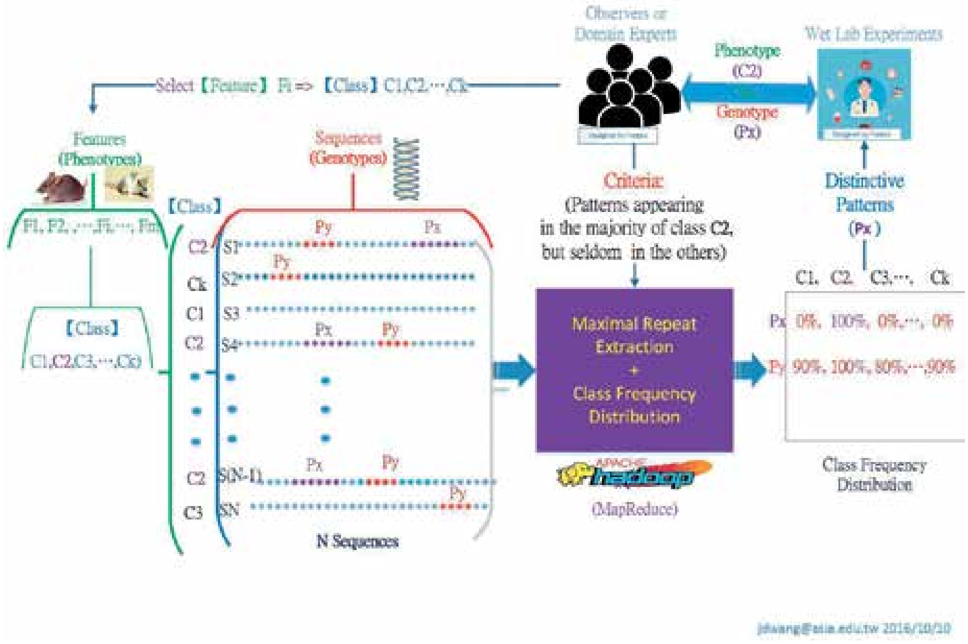


Figure 4. The conceptual diagram of reducing the gap between phenotypes and genotypes.

The remainder of this chapter is organized as follows. Section 2 gives the review of potential applications with class frequency distributions of maximal repeats. Section 3 shows the scalable approach to extract maximal repeat from tagged sequential data. Section 4 describes the most recent work [21] that compute co-occurrences of DNA maximal repeat patterns appearing in both humans and viruses. Section 5 concludes and discusses on future works.

2. Potential applications with class frequency distribution of maximal repeats extracted from tagged sequential data

The previous work in [22] was a scalable approach based on Hadoop MapReduce programming model to overcome the computational bottleneck of using single computer with external memory [23, 24]. Furthermore, it had been applied for a USA patent (US-2017-0255634-A1) [25] whose publication data is as "Sep. 7, 2017" [25]. Recently, in these 2 years, many novel and potential applications, derived from that work, were launched in diverse fields successfully, due to its scalability being able to handle a huge amount of sequential data. There were many experiments in diverse applications with a huge amount of tagged sequential data, such as textual data for trend analysis [26–28], genomic sequences for biomarker identification [21, 29, 30], time-stamped gantry sequences for significant travel time intervals [31] and, most recently, the sequences of product traceability for quality control [32].

3. Methods

The scalable approach of maximal repeat extraction adopted in this chapter is based on Hadoop MapReduce programming model, and the details can be found in [22]. To illustrate the concept of the earlier approach clearly, as shown in **Figure 5**, there are 20 creatures generated manually. Each of them is with three phenotypes, "Drug Resistance," "Envelope," and "Contents," as given in **Figure 3**, and all of its chromosomes are concatenated into one line which may contain genotypes including motifs, domains, or unknown DNA segments that are represented as icons for simplicity. Even though with the conceptual diagram as shown in **Figure 5**, it is still very difficult for users to catch the hidden connection (or relationship) among these three phenotypes and those icons (genotypes) at first glance, let alone each of these icons (genotypes) presents one continuous subsequence whose length is not fixed and its location is unknown within chromosomes.

To reveal the possible mapping of phenotype "Drug Resistance," for example, to genotypes on purpose, **Figure 6** presents the rearrangement in the order of these 20 chromosomes which may contain icons as hidden or unknown DNA segments. The mapping of different types of phenotype "Drug Resistance" to the corresponding genotypes (icons) can be observed. Similarly, one can have the mapping of different types of phenotype "Envelope"

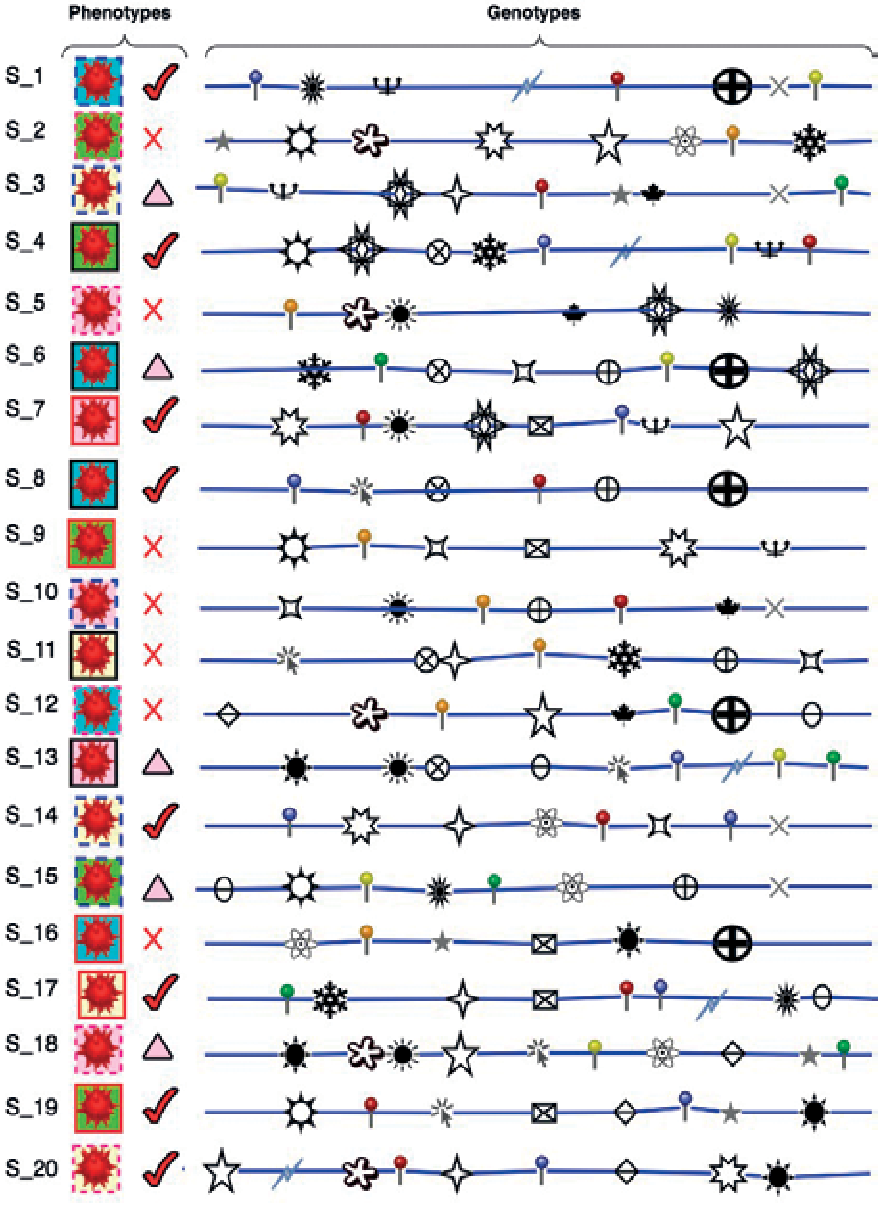


Figure 5. Each of 20 creatures is with three kinds of phenotypes as given in **Figure 3** and all of its chromosomes are concatenated as one line containing several icons as motif, domain, or unknown patterns.

and "Contents" to the corresponding genotypes (icons). Due to the page limitation, the corresponding mapping of figures for "Envelope" and "Contents" are given in the supplements. Focusing on the repeats whose class frequency distributions are biased, as shown in

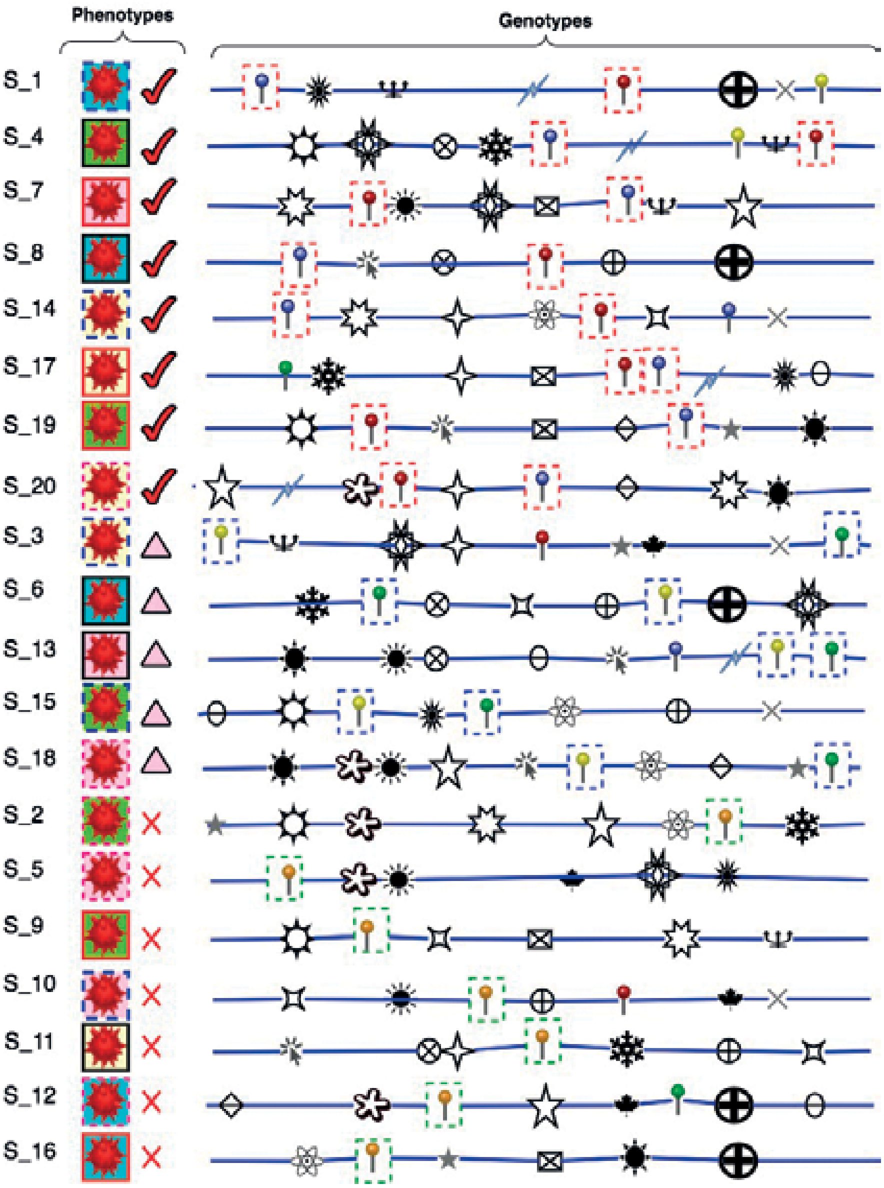


Figure 6. The mapping of different types of phenotype "Drug Resistance" to the corresponding genotypes (icons).

Figure 7, one can estimate these repeats as candidate class markers which can be the clues for further experiments of analyzing the mapping of phenotypes and genotypes derived from 20 creatures in **Figure 5**.

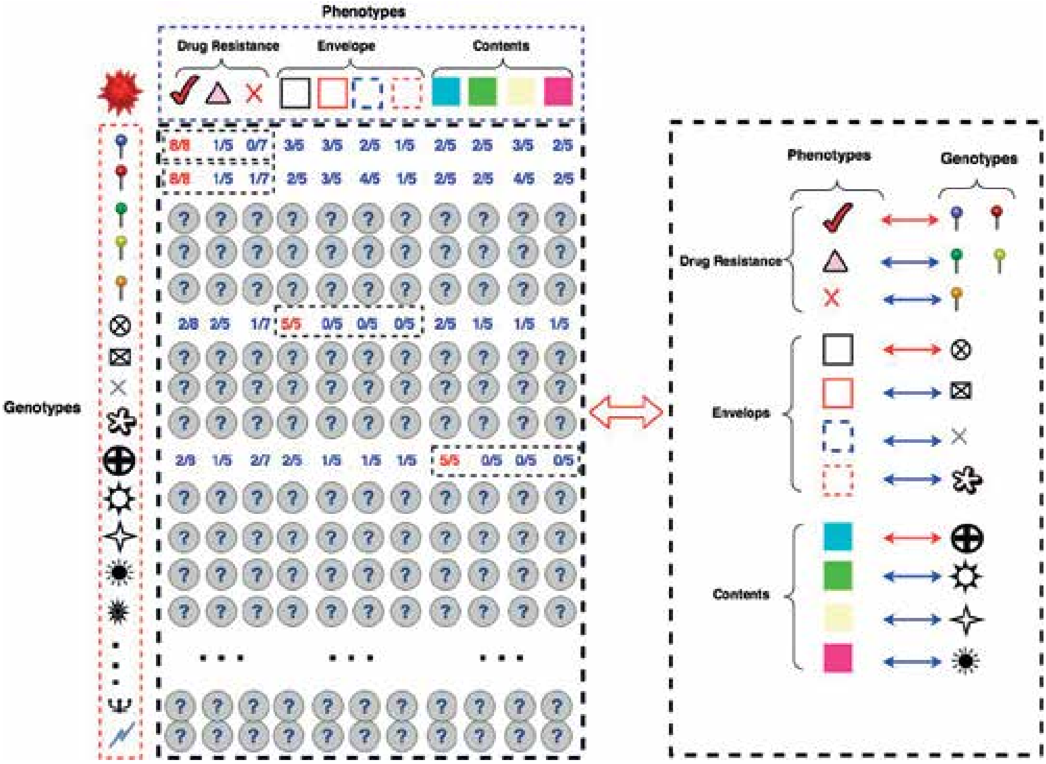


Figure 7. The mapping of phenotypes and genotypes derived from 20 creatures in **Figure 5**.

4. Case study: mining for the co-occurrences of DNA maximal repeat patterns in both human and viruses

There were three studies with a huge amount of genomic sequences [21, 29, 30] based on the scalable approach of maximal repeat extraction with class frequency distribution mentioned in this chapter. This chapter only describes the most recent work [21] that the co-occurrences of DNA maximal repeat patterns appearing in both humans and viruses are extracted via a scalable approach that is based on Hadoop distributed computing [22]; that work aimed to mine for specific DNA patterns within human genomes via observing class frequency distribution of DNA maximal repeats extracted from the whole genomic DNA sequences of humans and 559 virus genuses. The detail in [21] is described for reference in the following.

4.1. Genome resources

In [21], Wang et al. extracted significant DNA sequences appearing in both the genomes of humans and viruses. In this study, the taxonomic level of viruses is "genus" and is selected as the classes (tags) for further experiments. Experimental resources included the complete whole genomes of humans (GRCh38.p7 Primary Assembly) downloaded from the NCBI FTP [33] and that of 559 virus genuses, including 2712 viruses that had genus name and were selected from the total of 4388 viruses download from in NCBI FTP [34] on January 14, 2017. **Table 1** shows the partial statistics of 560 classes, including 559 virus genuses and the humans as "C248." Note that each of the 24 human chromosomes is estimated as one individual instance for observing the frequency distribution among human chromosomes. This chapter, for

Class ID	Human and virus genuses	No of Instances
C1	Alfamovirus	1
C2	Allexivirus	6
C3	Allolevivirus	3
C4	Alpha3microvirus	2
C5	Alphabaculovirus	40
C6	Alphacarmotetravirus	1
C7	Alphabaculovirus	7
…	…	…
C247	Human mastadenovirus E	1
C248	HumanGenomes_23_Assembled	24
C249	Hunnivirus	1
C250	Hypovirus	4
C478	Sobemovirus	15
C479	Solendovirus	1
C480	Soymovirus	4
…	…	…
C553	Xipapillomavirus	1
C554	Xp10virus	5
C555	Yatapoxvirus	2
C556	Yatapoxvirus	3
C557	Zeavirus	1
C558	Zetapapillomavirus	1
C559	Zetatorqueviurs	1
C560	primate papillomaviruses	1

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 1. The partial statistics of 559 virus genuses and human genomes (C248).

simplicity, only takes the positive-strand DNA sequences of humans and viruses for further experiments.

4.2. Computational time and environment

To show the scalability of this approach from a practical view of point, as shown in **Figure 8**, the computational platform was the Hadoop cluster with eight computing nodes, two name (master) nodes, and six data (slave) nodes; **Table 2** showed the specifications of hardware and software of one computing node; the computational time was about 37.5 h when the maximum length of maximal repeat patterns was limited to 500 bp (base pair).

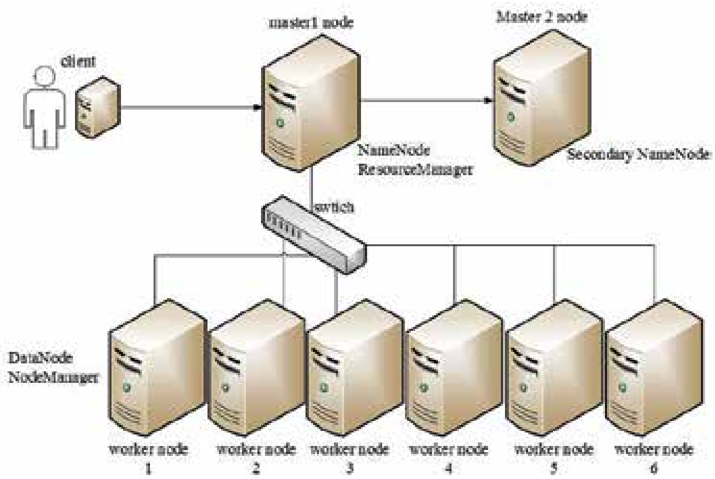


Figure 8. The conceptual diagram of a Hadoop cluster with two name (master) nodes, and six data (worker) nodes; "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [41].

Hardware	CPU	Intel® Xeon® Processor E5-2630 v3 (8 cores)
	RAM	128 GB (16GB*8, ECC/REG DDR4 2133)
	Hard Disk	6 TB (SATA3 2 TB*3, 7200 rpm 3.5 inch)
	Network Card	Intel Ethernet X540 10GBASE-T RJ45 DualPort *4
Software	OS	CentOS 6.7
	Hadoop	Hadoop 2.6 ("Cloudera Express 5.4.5")

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 2. The hardware and software of one computing node in Hadoop cluster.

Length	Virus (only)	Human (only)	Human and virus
5	426	127	1341
6	245	102	4234
7	84	48	16,454
8	29	26	65,556
9	5	11	262,154
10	1	9	1,048,579
11	956	4093	4,189,216
12	95,386	1,198,404	15,310,125
13	547,437	23,069,913	34,360,563
14	788,030	110,159,534	42,567,207
15	547,766	273,869,697	36,497,761
16	305,641	322,333,237	22,317,495
17	206,969	209,993,387	10,170,128
18	86,585	103,569,439	3,920,359
19	47,417	48,474,700	1,407,005
20	66,719	25,284,157	493,326
21	25,068	15,882,880	175,934
22	18,507	11,902,168	67,700
23	39,947	9,921,624	29,793
24	14,802	8,649,670	14,795
25	12,227	7,794,361	8749
…	…	…	…
98	165	107,159	15
99	710	102,830	15
100	707	99,579	13
101	1607	96,326	13
102	608	93,630	12
103	638	92,129	11
…	…	…	…
460	19	1933	1
461	27	2000	1
462	22	1812	1
463	26	1936	1
464	23	1993	
465	19	1817	
…	…	…	…

Length	Virus (only)	Human (only)	Human and virus
495	7	1542	
496	16	1564	
497	27	1408	
498	14	1451	
499	16	1494	
500	22	1542	

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 3. The partial of frequency distribution of DNA maximal repeats (length 5–500 bp).

4.3. The length distribution of DNA maximal repeats in both the genomes of human and 559 virus genuses

Comparing the maximal repeats that appear only in virus (Virus only), only in humans (Human only) or in both human and virus (Human and virus), **Table 3** shows the partial frequency distribution of maximal repeats whose lengths are from 5 to 500 bp. It is observed that the majority of those maximal repeats whose length range from 7 to 11 almost belong to the "Human and Virus." Note that there may exist extra nucleic acid codes, for example, "N," within these DNA sequences such that the number of maximal repeat (length = 5) appearing in both humans and viruses in **Table 3** is $1,341$ and that is great than 4^5 (= 1024).

4.4. The longest DNA maximal repeat (length = 463 bp) appearing in both the genomes of human and 559 virus genuses

Table 3 shows the length of the longest maximal repeat extracted in both the genomes of humans and selected viruses of 559 virus genuses is 463 bp. In [21], the result of blasting two sequences, "*Homo sapiens* chromosome 5" (NC_000005.10) and "Human herpesvirus 6B" (NC_000898.1), as shown in **Table 4**, that longest repeat appears 109 times within human chromosome 5 and two times within virus "Human herpesvirus 6B." To further inspect the longest maximal repeat, as show in **Figure 9**, one can find that the longest one is a tandem repeat (TAACCC) and appears within virus "Human herpesvirus 6B" at two intervals, the front (8249–8711 bp) and tail (161570–162,032 bp), that are located within the regions of direct repeats (DR) [35]. **Figure 10** gives one of two longest patterns aligned within "Human herpesvirus 6B" (8249–8711 bp) in **Figure 9**.

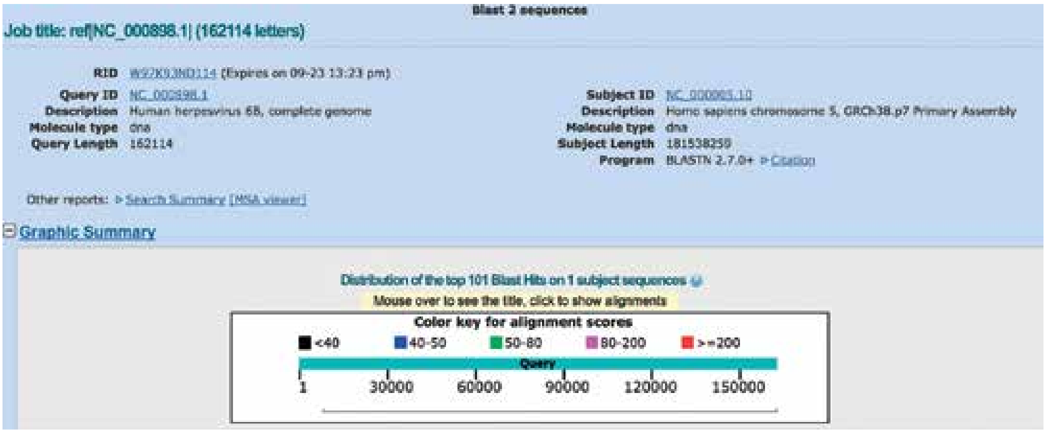
4.5. The statistics of DNA maximal repeat patterns (length = 100 bp) appearing in both human and 559 virus genuses

Table 5, for example, shows the statistics of 13 DNA maximal repeat patterns (length = 100 bp) appearing in both human and 559 virus genuses. It is observed that the three repeats as the

Maximal repeat patterns	DF	TF	Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 Primary assembly)	Viruses
ctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccc taaccctaaccctaaccctaaccctaacc ctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccc taaccctaaccctaaccctaaccctaacc Ctaaccctaaccctaaccctaaccctaac Cctaaccctaaccctaaccctaaccctaa Ccctaaccctaaccctaaccctaacccta Accctaaccctaaccctaaccctaaccc	2	111	463	(C248#1#109) (C442#1#2)	(TAACCC)n	5	Human herpesvirus 6B

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 4. The longest DNA maximal repeat patterns (Length = 463 bp) appearing in both humans and viruses.



Blast 2 sequences

Job title: ref|NC_000898.1| (162114 letters)

RID W97K93ND114 (Expires on 09-23 13:23 pm)

Query ID NC_000898.1
Description Human herpesvirus 6B, complete genome
Molecule type dna
Query Length 162114

Subject ID NC_000005.10
Description Homo sapiens chromosome 5, GRCh38.p7 Primary Assembly
Molecule type dna
Subject Length 181538259
Program BLASTN 2.7.0+ ▷ Citation

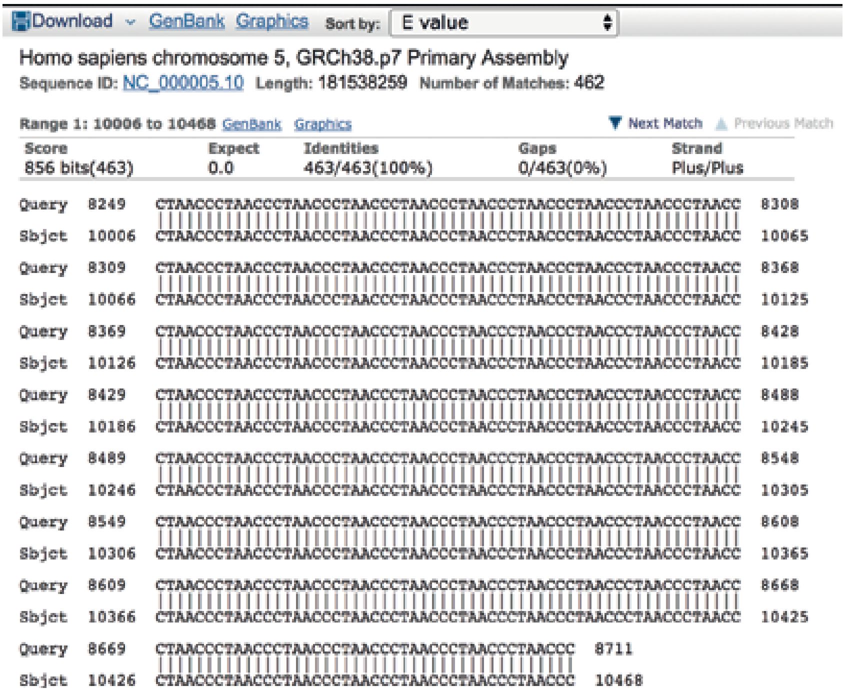
Other reports: ▷ Search Summary [MSA viewer]

⊟ Graphic Summary

Distribution of the top 101 Blast Hits on 1 subject sequences

Mouse over to see the title, click to show alignments

Figure 9. BLAST: "*Homo sapiens* chromosome 5" versus "human herpesvirus 6B"; "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [21].



Homo sapiens chromosome 5, GRCh38.p7 Primary Assembly

Sequence ID: NC_000005.10 Length: 181538259 Number of Matches: 462

Range 1: 10006 to 10468 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
856 bits(463)	0.0	463/463(100%)	0/463(0%)	Plus/Plus

```
Query  8249   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8308
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10006  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10065

Query  8309   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8368
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10066  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10125

Query  8369   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8428
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10126  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10185

Query  8429   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8488
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10186  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10245

Query  8489   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8548
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10246  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10305

Query  8549   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8608
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10306  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10365

Query  8609   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  8668
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  10366  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC  10425

Query  8669   CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC  8711
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct  10426  CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC  10468
```

Figure 10. One of two aligned patterns (8249–8711 bp) in **Figure 9** "Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017" [41].

1st, the 6th, and the11th, for example, have the similar regular expression as "(AACCCT)n", "(CTAACC)n," and "(TAACCC)n", respectively, and all of them appear in human chromosomes "1," "5," "10" and "12"; all of these three repeats appear in these viruses, "Cyprinid herpesvirus 1," "Falconid herpesvirus 1," "Gallid herpesvirus 2," "Human herpesvirus 6A," "Human herpesvirus 6B," and "Equid herpesvirus 3." It is very interesting to investigate the relationship between these human chromosomes and those viruses for further research. On the other hand, from the biological viewpoint, furthermore, (AACCCT)n, (CCCTAA)n, and (CTAACC)n may comprise the same maximal repeat pattern with different repeat frame; (GGGTTA)n, and (AGGGTT)n can also comprise the same maximal repeat pattern in complementary sequence. Moreover, the (GGGTTA)n is expected to be targeted by cisplatin [36].

4.6. Phenotypes: "Group I(dsDNS)" in Baltimore virus classification

It is observed that all of these viruses in **Table 5** belong to the "Group I(dsDNS)" of Baltimore classification [37], as shown in **Table 6**, and most of them are from the family "Herpesviridae" and order "Herpesvirales." Indeed, it is very interesting and attractive to have all the viruses compared with human whole genome and to inspect these co-occurrences of repeats for virus prevention from the genomic point of view in the future.

		Total		Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 primary assembly)	Viruses
	Maximal Repeat Patterns	DF	TF				C248	C5,C14,C149,C284,C305,C357, C442C541
1	aaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacc	10	591	100	(Cl 49# 1 # 18) (C248#4#299) (C305#2#41) (C442#2#208) (C541#l#25)	(AACCCT)n	1, 5, 10, 12	Cyprinid herpesvirus 1, Falconid herpesvirus 1, Gallid herpesvirus 2, Human herpesvirus 6A, Human herpesvirus 6B, Equid herpesvirus 3
2	aatagaatagaatagaatagaatagaatagaatagaatagaatagaataga atagaatagaatagaatagaatagaatagaatagaatagaatagaatag	3	42	100	(C248#1#18) (C284#1#11) (C357#1#13)	(AATAG)n	X	Rabbit fibroma virus, Taterapox virus,
3	agggttagggttagggttagggttagggttagggttagggttagggttagg gttagggttagggttagggttagggttagggttagggttagggttaggg	9	188	100	(C248#7#152) (C305#2#36)	(AGGGTT)n	2,12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2
4	atatatatatatatatatatatatatatatatatatatatatatatatatatatatata tatatatatatatatatatatatatatatatatatatatat	6	533	100	(C14#1#34) (C248#5#499)	(AT)n	2, 3, 7, 19, X	Gryllus bimaculatus nudivirus
5	ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccct aaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccc	3	4	100	(C248#1#1) (C305#1#2) (C541#1#1)	(CCCTAA)n	5	Meleleagrid herpesvirus 1, Equid herpesvirus 3
6	ctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa ccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaa	11	591	100	(C149#1#18) (C248#4#298) (C305#2#41) (C442#3#210) (C541#l#24)	(CCCTAA)n	1, 5,10,12	Cyprinid herpesvirus 1, Falconid herpesvirus 1, Gallid herpesvirus 2, Human herpesvirus 6A, Human herpesvirus 6B, Human herpesvirus 7, Equid herpesvirus 3
7	gagagagagagagagagagagagagagagagagagagagagagag agagagagagagagagagagagagagagagagagagagagagaga gagagagaga	3	55	100	(C149#1#32) (C248#2#23)	(GA)n	6, 11	Cyprinid herpesvirus 3
8	ggggttagggttagggttagggttagggttagggttagggttagggttagg gttagggttagggttagggttagggtagggttagggttagggttaggg	3	3	100	(C248#2#2) (C305#1#1)	G(GGGTTA)n	13, 18	Meleleagrid herpesvirus 1
9	gggttagggttagggttagggttagggttagggttagggttagggttaggg ttagggttagggttagggttagggttagggttagggttagggttagggt	9	188	100	(C248#7#152) (C305#2#36)	(GGGTTA)n	2, 12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2

		Total		Length	Class frequency distribution (ClassID#DF#TF)	Regular expression	Human chromosome (GRCh38.p7 primary assembly)	Viruses
10	gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtg tgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtg	5	78	100	(C248#4#38) (C5#1#40)	(GT)n	2,10, 16, 19	Orgyia pseudotsugata MNPV
11	taaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac	10	588	100	(C149#1#16) (C248#4#298) (C305#2#41) (C442#2#208) (C541#1#25)	(TAACCC)n	1, 5,10,12	Cyprinid herpesvirus 1, Gallid herpesvirus 2, Falconid herpesvirus 1 Human herpesvirus 6A, Human herpesvirus 6B, Rquid herpesvirus 3
12	tgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt	5	76	100	(C248#4#35) (C5#1#41)	(GT)n	2,10, 16, 19	Orgyia pseudotsugata MNPV
13	ttagggttagggttagggttagggttagggttagggttagggttagggttag ggttagggttagggttagggttagggttagggttagggttagggttag	9	186	100	(C248#7#150) (C305#2#36)	(AGGGTT)n	2,12,13,18,22, X, Y	Falconid herpesvirus 1, Gallid herpesvirus 2

Table 5. The statistics of 13 DNA maximal repeat patterns (length = 100 bp) appearing in both humans and viruses.

Viruses	Class ID	The International Committee on Taxonomy of Viruses (ICTV)			Baltimore classification
		Genus	Family	Order	
Orgvia pseudotsugata MNPV	C5	Alphabaculovirus	Baculoviridae	N	Group I(dsDNA)
Gryllus bimaculatus nudiviras	C14	Alphanudivirus	Nudiviridae	N	Group I(dsDNA)
Cyprinid herpesvirus 1	C149	Cyprinivirus	Alloherpesviridae	Herpesvirales	Group I (dsDNA)
Rabbit fibroma virys	C284	Leporipoxvirus	Poxviridae	N	Group I (dsDNA)
Falconid herpesvirus 1	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Gallid herpesvirus 2	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Meleagrid herpesvirus 1	C305	Mardivirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Taterapox virus	C357	Orthopoxvirus	Poxviridae	N	Group I(dsDNA)
Human herpesvirus 6A	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Human herpesvirus 6B	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Human herpesvirus 7	C442	Roseolovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)
Equid herpesvirus 3	C541	Varicellovirus	Herpesviridae	Herpesvirales	Group I(dsDNA)

"Reproduced with permission from International Conference on BioInformatics and BioEngineering (BIBE); published by IEEE, 2017." [21].

Table 6. The taxonomy of 12 viruses selected in **Table 5**.

5. Conclusions and future works

Except considering the phenotypes that result from the epigenetics [38], it is believed that some of the phenotypes of creatures (or organisms) are determined by their genotypes as they are born in the beginning. This chapter proposes a novel approach to mine for genetic markers via comparing class frequency distributions of maximal repeats extracted from tagged genomic sequences of creatures, where the classes are derived from the tags given by domain experts. Once domain experts can divide the creatures into disjoint classes as precisely as possible according to their features (phenotypes), then they can adopt the scalable approach developed in [22] to extract the maximal repeats and compute class frequency distributions of these repeats via comparing the whole genomic sequences of these creatures. The repeats or the combination of some repeats that are with extremely biased class frequency distribution can be seen as class markers (genotypes) and can provide clues to biologists to analyze the relationship among these class markers (genotypes) and their corresponding features (phenotypes).

Due to the availability of cloud computing with flexible infrastructure, nowadays, it becomes possible to compute class frequency distributions of maximal repeats from a huge amount of tagged whole genomic sequences of many creatures across species via the scalable maximal repeat extraction approach [22] with Hadoop MapReduce programming model. The function mentioned in this chapter is somewhat like "Archimedes' Law of the Lever," as shown in

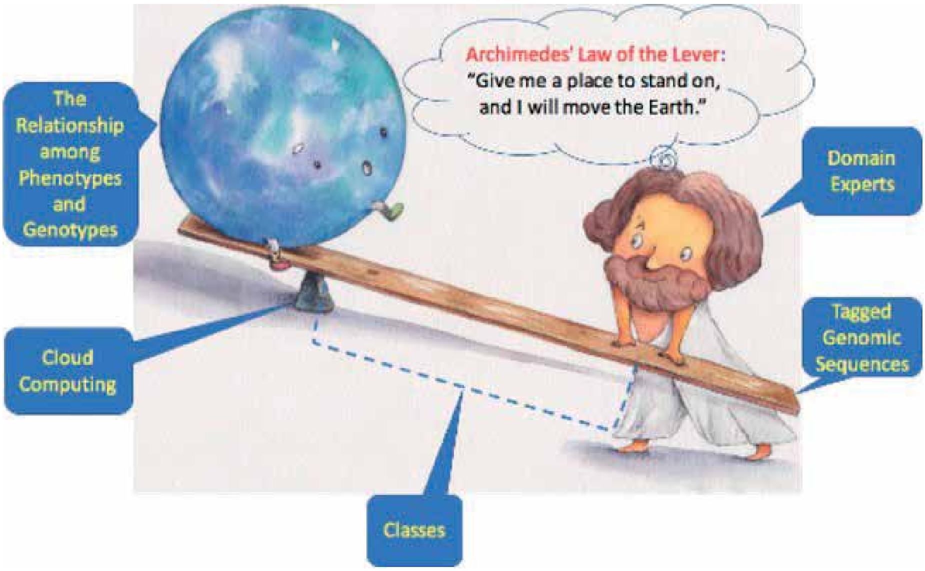

Archimedes' Law of the Lever:
"Give me a place to stand on, and I will move the Earth."
The Relationship among Phenotypes and Genotypes
Domain Experts
Tagged Genomic Sequences
Cloud Computing
Classes

Figure 11. "Mining for biomarkers via observing class frequency distributions of maximal repeats from tagged genomic sequences" is somewhat like "Archimedes' Law of the Lever".

Figure 11, the Archimedes, an outstanding ancient Greek scientist, said that "Give me a place to stand on, and I will move the Earth." With scalable computing power and enough tagged genomic sequences, in other words, a domain expert can figure out the relationship among phenotypes and genotypes if the classes are properly and precisely defined. It is desired to have further cooperation with domain experts, especially who have collected the whole genomes of diverse organisms and desire to find or identify the relationship between genomic signatures and the features they concern in the future.

From a practical point of view, it is inconvenient for general users to have experiments of maximal repeat extraction by themselves in the beginning because there are a lot of preprocessing works and need considerable hardware infrastructure to support such a big-data computing. Furthermore, it might be a bottleneck or nightmare for general users, for example, biologists, to implement Hadoop MapReduce programming as described in [22]. Therefore, it is highly desired if maximal repeat extraction can be provided in public cloud services, such as Amazon Elastic Container Service (AWS ECS) [39], Google Cloud Platform [40], and Azure Container Service (AKS) [41]. It is highly expected that one will develop novel comparative genome with tagged genomic sequences and bring users with novel cloud services of computing class frequency distribution of maximal repeats in the future.

Acknowledgements

This study is supported by Ministry of Science and Technology, Taiwan, under project *MOST 106-2632-E-468-002*. I thank Prof. Jeffrey J.P. Tsai who provided computational environments and finical support. I also thank Prof. Yi-Chun Wang for collecting the human chromosomes and inspecting experimental results about viruses and humans. I also thank Prof. Rouh-Mei Hu who helped in explaining the SNPs and the relationship between genotypes and phenotypes; Prof. Jan-Gowth Chang and Charles C.N. Wang had valuable discussions; I thank Jazz Wang for providing valuable Hadoop programming discussions. Finally, I thank Ling-Yu Ji for offering her drawing to show the conceptual diagram of "Archimedes' Law of the Lever."

Author details

Jing-Doo Wang[1,2]*

*Address all correspondence to: jdwang@asia.edu.tw

1 Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan

2 Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

References

[1] Azuaje F. Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine. Wiley; 2011

[2] Novelli G, Ciccacci C, Borgiani P, Amati MP, Abadie E. Genetic tests and genomic biomarkers: Regulation, qualification and validation. Clinical Cases in Mineral and Bone Metabolism. 2008;**5**(2):149154

[3] Glauser TA. Biomarkers for antiepileptic drug response. Biomarkers in Medicine. 2011; **5**(5):635641

[4] Sun W et al. Common genetic polymorphisms influence blood biomarker measurements in COPD. PLoS Genetics. 2016;**12**(8):e1006011

[5] What are genome-wide association studies? https://ghr.nlm.nih.gov/primer/genomicresearch/gwastudies.

[6] Genome-wide association studies. https://www.yourgenome.org/stories/genome-wide--association-studies.

[7] Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: Current insights and future perspectives. Nature Reviews Cancer. 2017;**17**:692704

[8] Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E, Taskesen E, Hammerschlag AR, Okbay A, Zabaneh D, Amin N, Breen G, Cesarini D, Chabris CF, Iacono WG, Arfan Ikram M, Johannesson M, Koellinger P, Lee JJ, Magnusson PKE,

McGue M, Miller MB, Ollier WER, Payton A, Pendleton N, Plomin R, Rietveld CA, Tiemeier H, van Duijn CM, Posthuma D. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nature Genetics. 2017;**49**:1107. EP –, 05

[9] Louhelainen J. SNP arrays. Microarrays. 2016;**5**(4):27

[10] Illumina genotyping solutions. https://www.illumina.com/techniques/popular-applications/genotyping.html.

[11] Genome-Wide Human SNP Array 6.0. https://www.thermofisher.com/order/catalog/product/901182

[12] Clark DP, Pazdernik NJ. Chapter e9 - genomics and systems biology. In: Clark DP, Pazdernik NJ, editors. Molecular Biology. 2nd ed. Boston: Academic Press; 2013. p. e110, e117

[13] Ha N-T, Freytag S, Bickeboeller H. Coverage and efficiency in current snp chips. European Journal of Human Genetics. 2014;**22**:11241130

[14] The database of Genotypes and Phenotypes (dbGaP). https://www.ncbi.nlm.nih.gov/gap.

[15] Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;**107**(1):1-8

[16] Han Y, He X. Integrating epigenomics into the understanding of biomedical insight. Bioinformatics and Biology Insights. 2016;**10**(267289)

[17] Brown JR. Comparative Genomics: Basic and Applied Research. CRC Press; 2007

[18] NCBI Whole Genomes FTP Site. ftp://ftp.ncbi.nih.gov/genomes.

[19] The Cancer Genome Altas (TCGA). https://cancergenome.nih.gov/

[20] Cancer Moonshot. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

[21] Wang J-D, Wang Y-C, Hu R-M, Tsai J. Extracting the co-occurrences of dna maximal repeats in both human and viruses. In: The 17th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2017); 2017

[22] Wang J-D. Extracting significant pattern histories from timestamped texts using mapreduce. The Journal of Supercomputing. 2016:1-25

[23] Wang J-D. An external memory approach to compute the statistics of maximal repeats across classes from whole genome sequences. In: 2005 National Computer Symposium, Taiwan, R.O.C. p. BIC1–2, 2005

[24] Wang J-D. External memory approach to compute the maximal repeats across classes from DNA sequences. Asian Journal of Health and Information Sciences. 2006;**1**(2):276-295

[25] Wang C-T. Method for extracting maximal repeat patterns and computing frequency distribution tables, Sep 2017. US Patent App. 15/208,994

[26] Wang J-D. A novel approach to compute pattern history for trend analysis. In: The 8th International Conference on Fuzzy Systems and Knowledge Discovery; 2011. pp. 1796-1800

[27] Wang J-D, Heri W. Extracting retrospective patterns from time-stamped texts according to variable query time interval. In: The International Multi-Conference on Engineering and Technology Innovation 2015 (IMETI2015); 2015

[28] Wang J-D, Jiang A-K, Chen J-C. Shape query for pattern history in PubMed literatures via Haar wavelet. International Journal of Advanced Information Technologies. 2015;**9**(6):67-76

[29] Chan W-L, Wang J-D, Chang J-G, Tsai J. Genome-wide functional identification of maximal consensus patterns derived from multiple species pirnas. In: The 16th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2016); 2016

[30] Wang J-D, Chan W-L, Wang CCN, Chang J-G, Tsai JJP. Mining distinctive DNA patterns from the upstream of human coding and non-coding genes via class frequency distribution. In 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2016); 2016

[31] Wang J-D, Hwang M-C. A novel approach to extract significant patterns of travel time intervals of vehicles from freeway gantry timestamp sequences. Applied Sciences. 2017;**7**(9)

[32] Wang J-D. A novel approach to improve quality control by comparing the tagged sequences of product traceability. In: The 3rd International Conference on Inventions; 2017

[33] NCBI Whole Genomes FTP Site *Homo Sapiens* Assembled Chromosomes. ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/seq/

[34] NCBI Whole Genomes FTP Site Virus Whole Genomes. ftp://ftp.ncbi.nih.gov/genomes/Viruses/all.gbk.tar.gz

[35] Dominguez G, Dambaugh TR, Stamey FR, Dewhurst S, Inoue N, Pellett PE. Human Herpesvirus 6B genome sequence: Coding content and comparison with human Herpesvirus 6A. Journal of Virology. 1999;**73**(10):8040-8052

[36] Nguyen HTQ, Galea AM, Murray V. The interaction of cisplatin with a human telomeric DNA sequence containing seventeen tandem repeats. Bioorganic & Medicinal Chemistry Letters. 2013;**23**(4):1041-1045

[37] Baltimore D. Animal Virology. Number 4. Elsevier Science; 1976

[38] Felsenfeld G. A brief history of epigenetics. Cold Spring Harbor Perspectives in Biology. 2014;**6**(1)

[39] Amazon Elastic Container Service (AWS ECS). https://aws.amazon.com/tw/documentation/ecs/

[40] Google Cloud Platform : CONTAINER ENGINE. https://cloud.google.com/container-engine/

[41] Introduction to Azure Container Service (AKS). https://docs.microsoft.com/en-us/azure/aks/intro-kubernetes

Chapter 6

Mining for Structural Variations in Next-Generation Sequencing Data

Minja Zorc, Jernej Ogorevc and Peter Dovč

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76568

Abstract

Genomic structural variations (SVs) are genetic alterations that result in duplications, insertions, deletions, inversions, and translocations of segments of DNA covering 50 or more base pairs. By changing the organization of DNA, SVs can contribute to phenotypic variation or cause pathological consequences as neurobehavioral disorders, autoimmune diseases, obesity, and cancers. SVs were first examined using classic cytogenetic methods, revealing changes down to 3 Mb. Later techniques for SV detection were based on array comparative genome hybridization (aCGH) and single-nucleotide polymorphism (SNP) arrays. Next-generation sequencing (NGS) approaches enabled precise characterization of breakpoints of SVs of various types and sizes at a genome-wide scale. Dissecting SVs from NGS presents substantial challenge due to the relatively short sequence reads and the large volume of the data. Benign variants and reference errors in the genome present another dimension of problem complexity. Even though a wide range of tools is available, the usage of SV callers in routine molecular diagnostic is still limited. SV detection algorithms relay on different properties of the underlying data and vary in accuracy and sensitivity; therefore, SV detection process usually utilizes multiple variant callers. This chapter summarizes strengths and limitations of different tools in effective NGS SV calling.

Keywords: bioinformatics, genome organization, next-generation sequencing, structural variation, variant calling

1. Introduction

First, efforts in exploring genetic variations were focused on single-nucleotide polymorphisms (SNPs) which were initially considered the main source of genetic and phenotypic human variation [1], while larger variations were thought to be rare events. However, in 2004 two studies [2, 3] revealed an unexpectedly large amount of large-scale variations (several kb to hundreds of kb) in the human genome. The evidence for the prevalence of structural variants (SVs), such as deletions, duplications, and inversions, began to accumulate. By changing the organization of the DNA, SVs can contribute to the phenotypic differences among healthy individuals or cause severe phenotypic consequences. SVs are involved in a wide range of diseases and conditions, such as autism spectrum disorders [4–6], schizophrenia [7], Crohn's disease [8], rheumatoid arthritis [9], lupus erythematosus [10], psoriasis [11], obesity [12], and cancers [13, 14]. Among the different classes of genetic variations, SVs have remained the most challenging to detect and characterize. SVs were examined since the identification of chromosomal abnormalities using classic cytogenetic methods, revealing changes down to 3 Mb. Later techniques for SVs detection are based on array comparative genome hybridization (aCGH) and single-nucleotide polymorphism arrays. Next-generation sequencing (NGS) has enabled methods for precise definition of breakpoints of SVs of different sizes and types. Characterization of SVs from high-throughput sequencing data presents complex task due to the volume of the data and short sequence reads.

2. Structural variations

Genomic structural variations (SVs) are genetic alterations that result in amplifications, losses, inversions, and translocations of segments of DNA greater than 50 bp. SVs are a normal part of genomic variation but can also cause disorders. Standard detection methods include chromosome banding, fluorescent in situ hybridization (FISH), and array comparative genome hybridization (aCGH) that is very useful to detect copy number variations (CNVs) but cannot detect copy-neutral SVs (inversions, balanced translocations) [15]. Recent methods include employment of NGS to identify SVs, which are not detectable by cytogenetic methods.

Chromosomal rearrangements can occur on a single chromosome (interchromosomal SVs) or can involve exchange of genomic DNA between chromosomes (intrachromosomal SVs). Intrachromosomal SVs are a product of one or more double-strand breaks, which may result in deletions, inversions, and duplications. Deletions and duplications are copy number variations and are easily detected by employing NGS data (read coverage method), whereas inversions are copy number-neutral. Intrachromosomal translocation is the exchange of genetic material between two non-homologous chromosomes. In a reciprocal translocation, two broken-off pieces of two non-homologous chromosomes are exchanged, usually producing two balanced derivative chromosomes. Unless breakpoints disrupt important developmental genes, balanced translocations do not affect phenotype [15]. However, during gamete formation such chromosomes may segregate in unbalanced manner or unbalanced translocations may occur de novo and lead to monosomy and trisomy of different chromosome segments [16], which account for approximately 1% of developmental delay and intellectual disability cases in human [17–19]. Robertsonian translocations are a type of SVs resulting from chromosome end breaks near centromeric regions of two acrocentric chromosomes and their reciprocal exchange, which results in one large metacentric chromosome and one very small chromosome that is usually lost without phenotype effect. In case three or more chromosomal breakpoints are involved, we speak of complex chromosome rearrangements, which may result in balanced or unbalanced state [20].

3. Next-generation sequencing

The first commercially available next-generation sequencing platform was released in 2005 [21]. The technology has been continuously upgraded and has fundamentally changed the field of genetics studies. Next-generation sequencing (NGS), also known as high-throughput sequencing, parallelizes the sequencing process and produces millions of short reads (50–400 bp each) in a single experimental run. It has contributed to rapid progress in single-nucleotide polymorphisms detection. Due to the nature of the NGS short-read sequences, the category of longer variants remained poorly characterized. Variants in range 10–100 kb are small for detection by cytogenetic methods [22] but too large for reliable detection with short-read sequencing. SVs affect more bases than single-nucleotide polymorphisms [23] and present an important class of genetic variation. Moreover, many SVs have been shown to play relevant roles in phenotypic variability and disease [24].

3.1. NGS data analysis pipeline

Once the samples are sequenced, the NGS data analysis becomes the task in bioinformatics field. The computational analysis and interpretation of the data generated remains one of the major bottlenecks in NGS projects. The basic steps for analyzing NGS data are quality assessment, reads alignment (mapping) to a reference sequence, and variant identification. The second stage of analysis comprises variant analysis, visualization, and interpretation of the variants in relation to phenotypes. Commercial packages such as CLCBio Genomic Workbench, CASAVA, and SeqNext often provide all-in-one solutions, while academic pipelines typically consist of sequential tools for specific steps in the analysis.

The output from the sequencing machines are reads, which are usually stored in text-based FASTQ files. The data obtained from NGS are compromised by sequence artifacts, including read errors, poor-quality reads, and primer contamination [25]. To avoid erroneous conclusions, the artifacts should be removed. A number of bioinformatics tools for sequencing quality assessment, such as FastQC, FASTX-Toolkit, PRINSEQ [26], TagDust [27], and NGS QC Toolkit [28] are designed. Next step in NGS data analysis is alignment of short reads to corresponding positions on a reference sequence. A variety of algorithms have been developed for this task. Representative read mappers are Bowtie2 [29], BWA [30], and Novoalign. The typical output from the read mapper is BAM file which contains information about qualities and positions of aligned sequences. Variant analysis consists of genotyping, variant calling, annotation, and prioritization. Genomic variants, such as SNPs and short-scale insertions and deletions are identified by variant callers. Widely used tools for variant calling are Genome Analysis Tool Kit HaplotypeCaller (GATK-HC) [14], Samtools mpileup [31], Freebayes and Torrent Variant Caller (TVC). Variant callers take in a BAM file and return a list of variants. To annotate

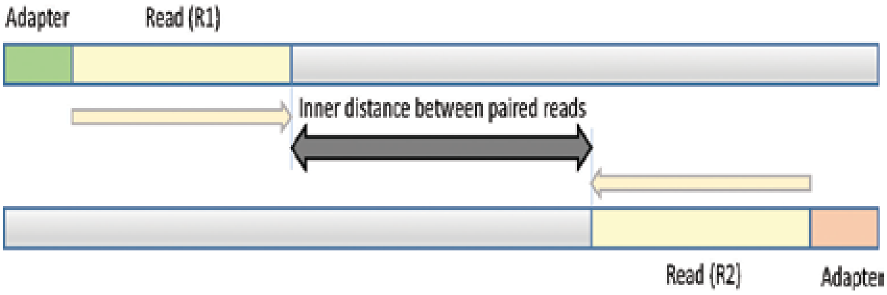


Figure 1. Paired-end sequencing; the inner distance between paired reads (R1 and R2) is known.

variants, SnpEff [32], VariantAnnotator from the GATK [33], and ANOVAR [34] tools are used. To systematically filter, evaluate, and prioritize thousands of variants VAAST 2.0 [35], VarSifer [36], KGGseq [37], and commercial software Ingenuity Variant Analysis are available.

3.2. Single-read and paired-end sequencing

Initially, NGS technologies produced extremely short reads (25–36 bp), sequenced from only one end of the DNA (single-read sequencing) [38]. As technology developed, read lengths consistently increased and sequencers have been improved to sequence both ends of a fragment with or without a non-sequenced stretch in between (paired-end sequencing). This not only has the benefit from doubling the number of reads but also improves accuracy and offers additional information for structural variants detection.

The reads obtained from paired-end sequencing (R1 and R2) come from the same fragment of DNA. The length of the fragment is usually longer than the length of reads (R1 + R2), so there is a gap between them (**Figure 1**). Although the sequence of the fragment between reads is not known, the knowledge that R1 and R2 are next to each other on the known distance and have opposite orientation is useful.

4. Overview of the structural variation detection algorithms

Using NGS technologies, large volume of sequence data at an unprecedented speed and constantly reducing cost is produced. Consequently, the computational tools for analysis of massive amounts of genomic data are in demand. There is a growing awareness that structural variations represent a significant contribution to genotypic and phenotypic diversity [39]. However, the accurate detection of structural variants from NGS is a daunting task [40]. A number of algorithms have been proposed to address the issue of structural variants calling from NGS data [41]. SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. The algorithms follow one or a combination of strategies, which could be classified into categories: (1) read depth (RD), (2) paired-end (PE), (3) split reads (SR), and (4) de novo assembly (AS). The most suitable method for SV detection depends on the size and variant type as well as characteristics of the sequencing data [42]. SV detection process usually utilizes multiple variant callers.

4.1. Algorithms based on read depth

Read depth (RD) algorithms are able to identify CNVs. RD-based algorithms can accurately predict absolute copy-numbers [43] but are unable to detect copy-number neutral variants such as inversions and balanced translocations. The breakpoint identification resolution is low and depends on the sequence coverage.

RD algorithms divide the reference sequence in intervals and calculate the number of reads aligned within them. The read depth per interval should follow a normal distribution centered at the average read depth for the entire reference sequence. When the read depth of contiguous intervals significantly differs from the average observed, the CNV is detected (**Figure 2**). Deleted regions show reduced read depth when compared to entire reference sequence (**Figure 3**).

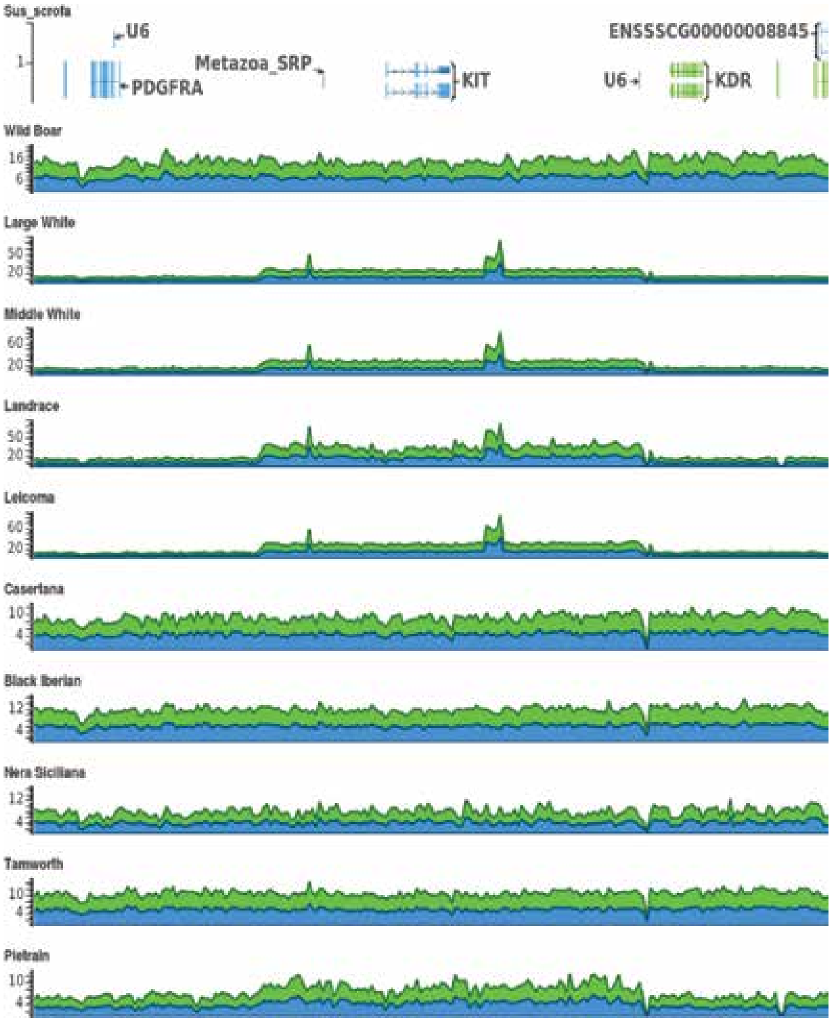


Figure 2. An example of CNV including gene *KIT* with flanking regions in four pig genomes. The read coverage is higher in the region of the CNV. The figure was made using Golden Helix GenomeBrowse.

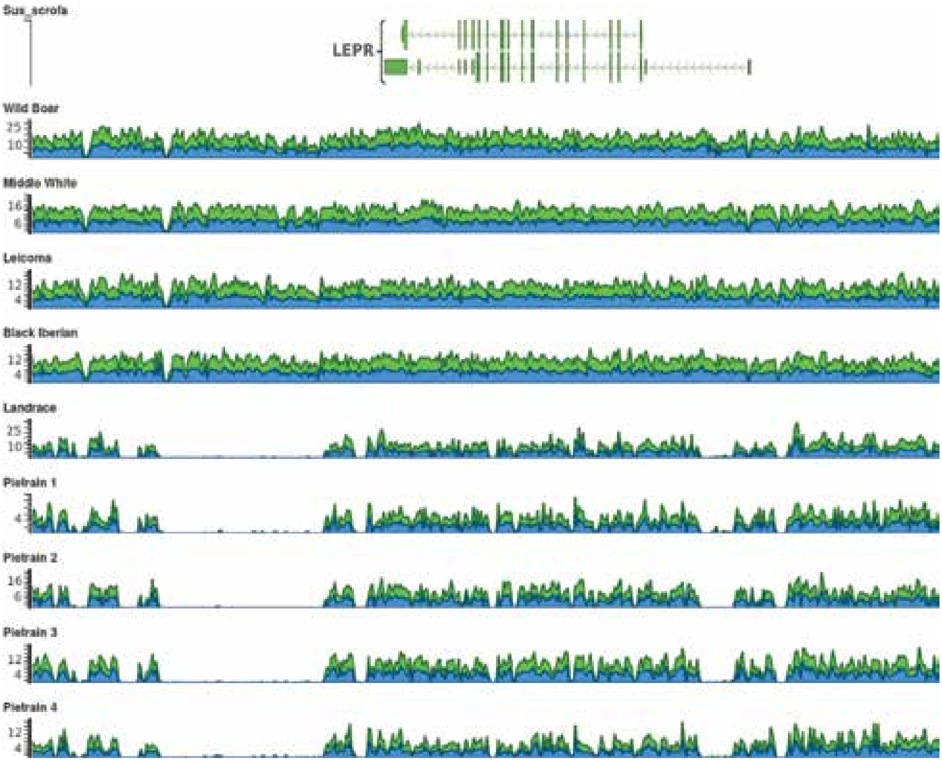


Figure 3. An example of deletion within upstream and downstream regions of *LEPR* locus in five pig genomes. The read coverage is low in the region of deletion. The figure was made using Golden Helix GenomeBrowse.

4.2. Paired-end approaches

Paired-end sequencing data allow detection of many types of SVs. Paired-end (PE) SV calling approaches detect deviations from expected library insert size (donor reads map at inconsistent distances). When a pair of reads does not overlap with any SV, the distance between them is the same as the size of the library insert and reads have correct orientation (concordant pairs). When the read pair overlaps a SV, the mapping distance of paired reads differs from the library insert size and their orientation may be inverted. Discordantly mapped paired-reads can be (1) further apart than expected, (2) closer together than expected, (3) in inverted orientation, (4) in incorrect order, (5) on different chromosomes. Clusters of read pairs aligned to the same genomic regions with the distance shorter than expected can be explained by insertion in the sequenced samples (donor). Larger distances between reads than expected can be explained by deletion in the sample (donor) (**Figure 4**). The resolution of the breakpoints detected by this approach depends on the library's insert size and on the read coverage. Insertions larger than the library insert size cannot be detected.

4.3. Algorithms based on split-reads

Split-read (SR) algorithm can detect SVs with a single base-pair resolution. Split reads contain the breakpoint of the structural variant. Their alignments to the reference genome are split

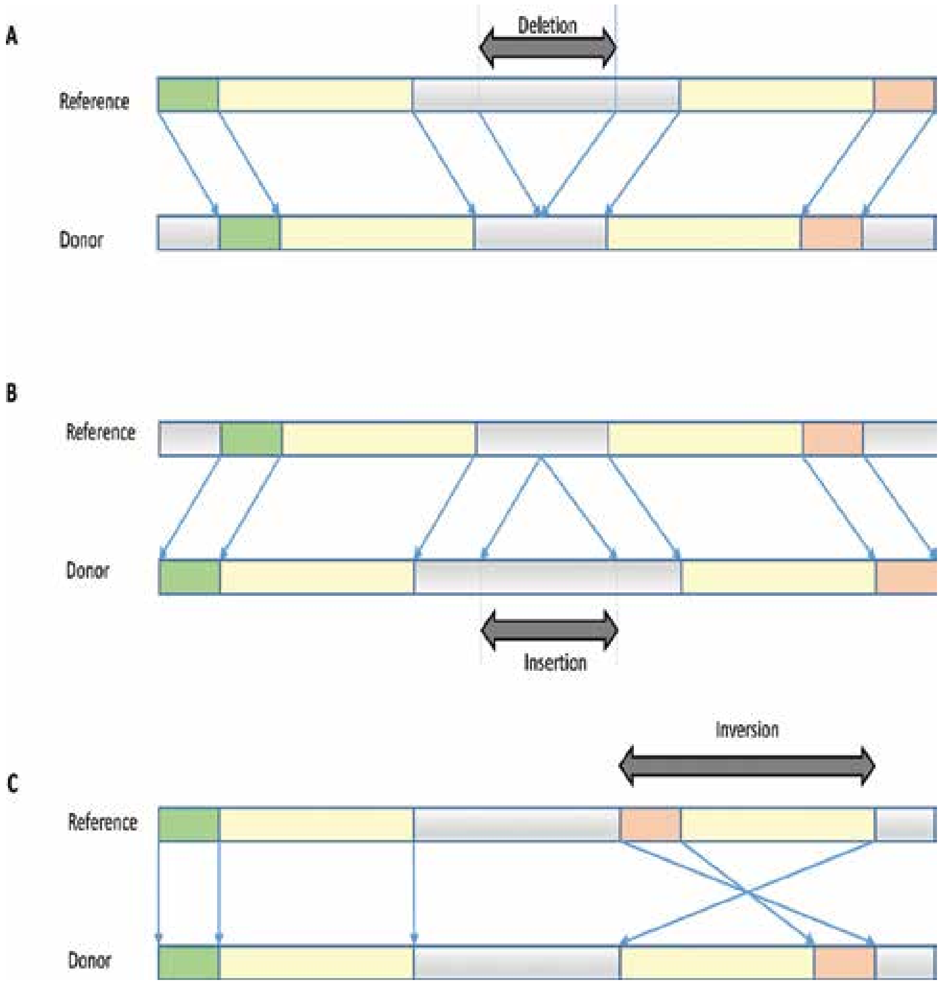


Figure 4. Examples of identification of deletion, insertion, and inversion using paired-end approach: (A) paired-reads are closer together than expected (deletion), (B) paired-reads are further apart than expected (insertion), (C) paired-reads are in inverted orientation (inversion).

into two parts (**Figure 5**). Parts of a read are independently aligned to the reference genome, so the reads should be long enough to be aligned uniquely. Therefore, algorithms based on split-reads are feasible only when the sequencing reads are sufficiently long.

4.4. Algorithms based on de novo assembly

Algorithms based on de novo assembly (AS) are able to detect all forms of structural variation. De novo assembly refers to reassembling the original sequence from which the fragments were sampled. When the sequenced genome is assembled, it is compared to the reference genome to identify SVs. The method enables discovery of novel sequence fragments (insertions). The approach is time-consuming, costly, and prone to assembly errors. In terms of computational efficiency and detection power, targeted SV assembly is more effective. They dissect a problem into a set of local assembly problems that can be more effectively solved.

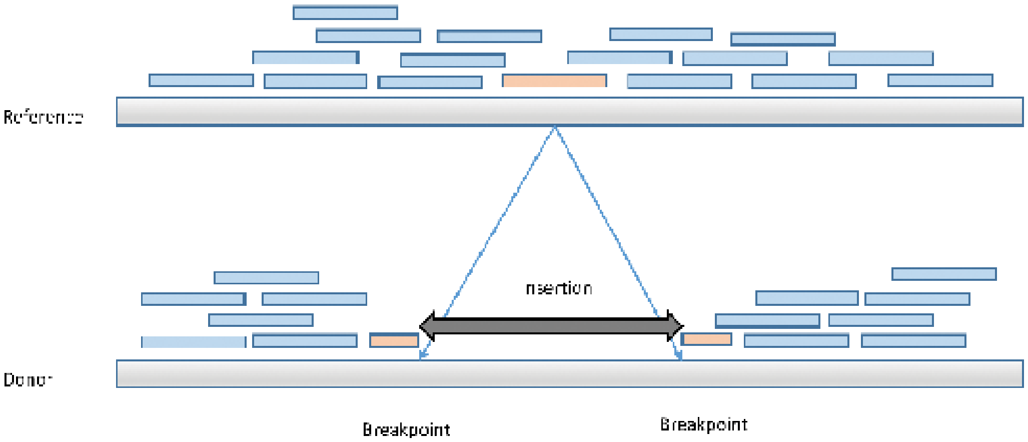


Figure 5. An example of deletion in an individual genome detected by split-read method.

Tool	SV type	Strategy	Released	Reference
PEMer	Indels, inversions	paired-reads	2009 Feb	[49]
VariationHunter	Transposon insertions	paired-reads	2010 Jun	[50]
SegSeq	CNVs	read-depth	2009 Jan	[51]
BreakDancer	Indels, inversions, and translocations	paired-reads	2009 Jul	[52]
Pindel	Breakpoints of large deletions and medium-sized insertions	split-read	2009 Nov	[53]
VariationHunter	Transposon insertions	paired-reads	2010 Jun	[50]
Cortex	simple and complex SVs	de novo assembly	2011 Apr	[54]
CNVnator	CNVs	read-depth	2011 Jun	[55]
GASVPro	Indels, inversions, interchromosomal translocations	read-depth, paired-end	2012 Mar	[56]
SVseq2	Indels with exact breakpoints	split-read, paired-end	2012 Apr	[57]
Breakpointer	Indels, mobile insertions and non-homologous recombinations	read-depth, split-read,	2012 Apr	[58]
DELLY	Copy-number variable deletions, tandem duplications, inversions, reciprocal translocations	split-read, paired-end	2012 Sep	[59]
SVM^2	Short insertions and deletions	paired-end, machine learning	2012 Oct	[60]
PeSV-Fisher	Deletions, gains, intra- and interchromosomal translocations, and inversions	paired-reads, read-depth	2013 May	[61]
LUMPY	Deletions, inversions, tandem duplications, and interchromosomal translocations	split-read, paired-end	2014 Jun	[62]

Tool	SV type	Strategy	Released	Reference
Gustaf	Deletions, inversions, dispersed duplications and translocations of ≥30 bp	split-read	2014 Dec	[63]
MetaSV	Indels, insertions, inversions, translocations, and CNVs	integration of SV callers (BreakSeq, Breakdancer, Pindel, CNVnator), local assembly	2015 Aug	[64]
Manta	Medium-sized indels, large insertions	split-read, paired-end	2016 Apr	[65]
SRBreak	CNV breakpoints	read-depth, split-read	2016 Sep	[66]
Seeksv	Deletion, insertion, inversion and interchromosomal transfer	split-read, paired-end, read-depth fragments with two ends unmapped	2017 Jan	[67]
SVachra	Large insertions-deletions, inversions, inter and intrachromosomal translocations	paired-end	2017 Oct	[68]

Table 1. The list of tools for different types of SV calling.

4.5. Hybrid-approaches for SV calling

SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. One single method cannot detect complete range of SVs, each is limited to specific type of SVs. Combined approaches can overcome limitations of a single method [44]. Two directions can be taken, combining strategies within one caller or combining SV callers [45]. A class of SV detection methods bases on machine learning. Variations are identified by various methods and are filtered against empirically derived training set data.

4.6. Bioinformatics tools for structural variation calling

A number of algorithms have been proposed to address the issue of structural variants calling from NGS data, but the structural variation calling remains challenging. The complete range of SVs cannot be discovered using one single method. The process of SV calling usually utilizes multiple variant callers to overcome limitations of individual approaches. Knowing advantages and drawbacks of various tools (**Table 1**) is important to make proper decisions when designing NGS data analysis pipelines. Different callers yield lists of identified SVs with limited overlap. Pipelines SVMerge [46], HugeSeq [47], iSVP [48], and IntanSV that integrate different SV callers, such as BreakDancer, CNVnator, SVseq2, Pindel, and DELLY and merge their results were published.

5. Conclusions

Using next-generation sequencing technologies, large volume of sequence data is produced with an unprecedented speed and constantly reducing cost. It allowed rapid progress in single-nucleotide polymorphisms detection. The awareness that structural variations represent a significant source of genotypic and phenotypic variation is permanently growing. However, the accurate detection of structural variants from NGS data is a daunting task. Relatively short reads, often repetitive character of SV, large amount of data, and large number of benign variants in complex genomes represent a major challenge for bioinformatics analysis of SVs. A number of algorithms have been proposed to address the issue of structural variants calling from NGS data. SV detection algorithms rely on different properties of the underlying data and vary in accuracy and sensitivity. SV detection process usually utilizes multiple variant callers. However, knowing advantages, drawbacks, and properties of different tools is inevitably required for proper decisions when designing NGS data analysis pipelines from publicly available tools. This chapter summarizes basic concepts of bioinformatics analysis of SV and introduces some rules for their assessment.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding P4-0220).

Conflict of interest

We have no conflict of interest to declare.

Author details

Minja Zorc*, Jernej Ogorevc and Peter Dovč

*Address all correspondence to: minja.zorc@bf.uni-lj.si

Biotechnical Faculty, Department of Animal Science, University of Ljubljana, Domzale,
Slovenia

References

[1] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;**409**(6822):928-933

[2] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;**305**(5683):525-528

[3] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nature Genetics. 2004;**36**(9):949-951

[4] Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, Szatmari P, et al. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. Journal of Medical Genetics. 2010;**47**(3):195-203

[5] Cho SC, Yim SH, Yoo HK, Kim MY, Jung GY, Shin GW, et al. Copy number variations associated with idiopathic autism identified by whole-genome microarray-based comparative genomic hybridization. Psychiatric Genetics. 2009;**19**(4):177-185

[6] Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, et al. Structural variation of chromosomes in autism spectrum disorder. American Journal of Human Genetics. 2008;**82**(2):477-488

[7] Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;**455**(7210):232-236

[8] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;**518**(7538):197-206

[9] Olsson LM, Nerstedt A, Lindqvist AK, Johansson SC, Medstrand P, Olofsson P, et al. Copy number variation of the gene NCF1 is associated with rheumatoid arthritis. Antioxidants & Redox Signaling. 2012;**16**(1):71-78

[10] Molokhia M, Fanciulli M, Petretto E, Patrick AL, McKeigue P, Roberts AL, et al. FCGR3B copy number variation is associated with systemic lupus erythematosus risk in Afro-Caribbeans. Rheumatology (Oxford, England). 2011;**50**(7):1206-1210

[11] de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nature Genetics. 2009;**41**(2):211-215

[12] Moon S, Hwang MY, Jang HB, Han S, Kim YJ, Hwang JY, et al. Whole-exome sequencing study reveals common copy number variants in protocadherin genes associated with childhood obesity in Koreans. International Journal of Obesity. 2017;**41**(4):660-663

[13] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Research. 2011;**39**(Database issue):D945-D950

[14] Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nature Genetics. 2008;**40**(6):722-729

[15] Weckselblatt B, Rudd MK. Human structural variation: Mechanisms of chromosome rearrangements. Trends in Genetics. 2015;**31**(10):587-599

[16] Weckselblatt B, Hermetz KE, Rudd MK. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. Genome Research. 2015;**25**(7):937-947

[17] Ravnan JB, Tepperberg JH, Papenhausen P, Lamb AN, Hedrick J, Eash D, et al. Subtelomere FISH analysis of 11 688 cases: An evaluation of the frequency and pattern of

subtelomere rearrangements in individuals with developmental disabilities. Journal of Medical Genetics. 2006;**43**(6):478-489

[18] Shao L, Shaw CA, Lu XY, Sahoo T, Bacino CA, Lalani SR, et al. Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: A study of 5,380 cases. American Journal of Medical Genetics. Part A. 2008;**146a**(17):2242-2251

[19] Ballif BC, Sulpizio SG, Lloyd RM, Minier SL, Theisen A, Bejjani BA, et al. The clinical utility of enhanced subtelomeric coverage in array CGH. American Journal of Medical Genetics. Part A. 2007;**143a**(16):1850-1857

[20] Zhang F, Carvalho CM, Lupski JR. Complex human chromosomal and genomic rearrangements. Trends in Genetics. 2009;**25**(7):298-307

[21] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in open microfabricated high density Picoliter reactors. Nature. 2005; **437**(7057):376-380

[22] de Ravel TJ, Devriendt K, Fryns JP, Vermeesch JR. What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). European Journal of Pediatrics. 2007;**166**(7):637-643

[23] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;**464**(7289):704-712

[24] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annual Review of Medicine. 2010;**61**:437-455

[25] Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics. 2010;**11**(Suppl 4):S7

[26] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;**27**(6):863-864

[27] Lassmann T, Hayashizaki Y, Daub CO. TagDust–A program to eliminate artifacts from next generation sequencing data. Bioinformatics. 2009;**25**(21):2839-2840

[28] Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS One. 2012;**7**(2):e30619

[29] Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nature Methods. 2012;**9**(4):357-359

[30] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;**25**(14):1754-1760

[31] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;**25**(16):2078-2079

[32] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012;**6**(2):80-92

[33] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;**20**(9):1297-1303

[34] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;**38**(16):e164

[35] Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genetic Epidemiology. 2013;**37**(6):622-634

[36] Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: Visualizing and analyzing exome-scale sequence variation data on a desktop computer. Bioinformatics. 2012;**28**(4):599-600

[37] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research. 2012;**40**(7):e53

[38] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;**129**(4):823-837

[39] Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011;**12**(5):363-376

[40] Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics. 2014;**15**(2):256-278

[41] Guan P, Sung WK. Structural variation detection using next-generation sequencing data: A comparative technical review. Methods. 2016;**102**:36-49

[42] Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in Bioengineering and Biotechnology. 2015;**3**:92

[43] Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;**330**(6004):641-646

[44] Gao J, Qi F, Guan R, editors. Structural variation discovery with next-generation sequencing. In: 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA); December 23-24, 2013. pp. 23-24

[45] Lin K, Smit S, Bonnema G, Sanchez-Perez G, de Ridder D. Making the difference: Integrating structural variation detection tools. Briefings in Bioinformatics. 2015;**16**(5):852-864

[46] Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biology. 2010;**11**(12):R128

[47] Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. Nature Biotechnology. 2012;**30**:226

[48] Mimori T, Nariai N, Kojima K, Takahashi M, Ono A, Sato Y, et al. iSVP: An integrated structural variant calling pipeline from high-throughput sequencing data. BMC Systems Biology. 2013;**7**(Suppl 6):S8

[49] Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biology. 2009;**10**(2):R23

[50] Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, et al. Next-generation VariationHunter: Combinatorial algorithms for transposon insertion discovery. Bioinformatics. 2010;**26**(12):i350-i3i7

[51] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature Methods. 2009;**6**(1):99-103

[52] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. Nature Methods. 2009;**6**:677

[53] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;**25**(21):2865-2871

[54] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics. 2012;**44**:226

[55] Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research. 2011;**21**(6):974-984

[56] Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biology. 2012;**13**(3):R22

[57] Zhang J, Wang J, Wu Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. BMC Bioinformatics. 2012;**13**(Suppl 6):S6

[58] Sun R, Love MI, Zemojtel T, Emde AK, Chung HR, Vingron M, et al. Breakpointer: Using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics. 2012;**28**(7):1024-1025

[59] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;**28**(18):i333-i3i9

[60] Chiara M, Pesole G, Horner DS. SVM(2): An improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data. Nucleic Acids Research. 2012;**40**(18):e145-e14e

[61] Escaramis G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martinez-Fundichely A, et al. PeSV-fisher: Identification of somatic and non-somatic structural variants using next generation sequencing data. PLoS One. 2013;**8**(5):e63377

[62] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biology. 2014;**15**(6):R84

[63] Trappe K, Emde AK, Ehrlich HC, Reinert K. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics. 2014;**30**(24):3484-3490

[64] Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, et al. MetaSV: An accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;**31**(16):2741-2744

[65] Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;**32**(8):1220-1222

[66] Nguyen HT, Boocock J, Merriman TR, Black MA. SRBreak: A read-depth and split-read framework to identify breakpoints of different events inside simple copy-number variable regions. Frontiers in Genetics. 2016;**7**:160

[67] Liang Y, Qiu K, Liao B, Zhu W, Huang X, Li L, et al. Seeksv: An accurate tool for somatic structural variation and virus integration detection. Bioinformatics. 2017;**33**(2):184-191

[68] Hampton OA, English AC, Wang M, Salerno WJ, Liu Y, Muzny DM, et al. SVachra: A tool to identify genomic structural variation in mate pair sequencing data containing inward and outward facing reads. BMC Genomics. 2017;**18**(Suppl 6):691

Chapter 7

Graphical Representation of Biological Sequences

Satoshi Mizuta

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74795

Abstract

Sequence comparison is one of the most fundamental tasks in bioinformatics. For biological sequence comparison, alignment is the most profitable method when the sequence lengths are not so large. However, as the time complexity of the alignment is the square order of the sequence length, the alignment requires a large amount of computational time for comparison of sequences of large size. Therefore, so-called alignment-free sequence comparison methods are needed for comparison between such as whole genome sequences in practical time. In this chapter, we reviewed the graphical representation of biological sequences, which is one of the major alignment-free sequence comparison methods. The notable effects of weighting during the course of the graphical representation introduced first by the author and co-workers were also mentioned.

Keywords: alignment-free, amino acid sequence, binary image, DNA sequence, mitochondria, phylogeny

1. Introduction

Comparison between biological sequences is one of the most fundamental tasks in the area of bioinformatics. For relatively short sequences, such as nucleotide sequences of genes or amino acid sequences of proteins, *alignment* is the most profitable method for the sequence comparison. However, as the dependency of the computational time of the alignment on the sequence length N is $O(N^2)$, the alignment is hard to be applied to comparison between sequences of large size, such as whole genome sequences. Therefore, the development of *alignment-free* methods is required to analyze the similarities between the sequences of large size in practical time. One of the most actively studied methods of the alignment-free sequence comparison is *graphical representation* [1, 2]. In addition to overcoming the time-consuming problem mentioned above, the graphical representation has the advantage that the similarities between sequences can be easily noticed visually.

Since the seminal paper by Hamori and Ruskin [3] was published, various kinds of sequence comparison methods based on the graphical representation have been proposed by many researchers. The basic procedure of the graphical representation is outlined as follows: first, each character in a biological sequence, which is expressed by the four-letter alphabet for nucleotide sequences and the 20-letter alphabet for amino acid sequences, is expressed by individual vectors in a certain dimensional space; next, the vectors are connected successively in a head-to-tail fashion, drawing a curve, or a *graph*, in the expression space; and last, if necessary, the distances between the graphs are calculated based on the predefined distance measures.

In this chapter, we briefly review the graphical representation methods for biological sequence comparison. In addition, we introduce our work recently published, in which weighting during the course of the graphical representation shows the notable effects.

2. Variations of graphical representations

The graphical representation methods are classified into some classes according to the target sequences and the dimension of the representation space. The target sequences of the graphical representation are amino acid sequences of proteins and nucleotide sequences of DNA (or RNA), including specific genes, mitochondrial genomes, and others. **Table 1** summarizes the classification of the graphical representation methods published so far.

2.1. Graphical representation of DNA sequences

Biological sequences stored in data archives are expressed by the four-letter alphabet for nucleotide sequences of DNA and the 20-letter alphabet for amino acid sequences of proteins.

Target sequence	Dimension of expression space	Work
DNA sequences		
Specific genes	2D	[4–22]
	3D ≤	[23–36]
Mitochondrial genomes	2D	[37–41]
	3D ≤	[42]
Others	3D ≤	[3, 43]
Proteins		
	2D	[44–49]
	3D ≤	[50–53]

Table 1. Classification of graphical representation methods.

To represent the biological sequences by graphs, it is necessary to express each character composing the sequences in numerical form.

The most popular strategy for the numerical expression is assigning vectors to respective characters in the alphabet. As for nucleotide sequences of DNA, the individual vectors of two, three, or higher dimension are assigned to four types of bases, A, T, G, and C.

2.1.1. Two-dimensional representation

Figure 1 is the two-dimensional vector assignment utilized by Gates [4]. Although many variations of the assignments are given according to the layout of the four bases, the number of the independent assignments is reduced to 3!/2 = 3, when the assignments that are transformed to each other by rotation on the *xy*-plane or inversion with respect to the *x*- or *y*-axis are assumed to be equivalent. The assignments of this type including the variations with some modifications are utilized in Refs. [5, 6, 10, 16, 20, 21, 40, 41].

By connecting the vectors successively in a head-to-tail fashion according to each base appearing in a nucleotide sequence, a graphical representation is generated. **Figure 2** shows, as an example, the graphical representation of sequence "TGAGTTC" generated by Gates' assignment.

The assignment of **Figure 1** may draw circuits in the graphical representation, leading to the loss of information that the original biological sequence has. To get rid of the degeneracy, Yau [9] introduced the assignment shown in **Figure 3**, which makes no circuit in the graphical representation; because the *x*-components of the vectors have all positive values, no backward motion along the *x*-axis exists in the graphical representation. For comparison, **Figure 4** illustrates the graphical representations of the first exon of the human β-globin gene represented by Gates' vector assignment (**Figure 1**) and Yau's vector assignment (**Figure 3**). There are many circuits in the graph by Gates' assignment; on the other hand, there is no circuit in the graph by

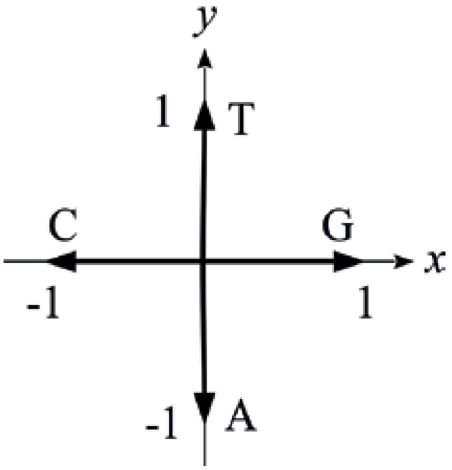


Figure 1. Two-dimensional vector assignment to bases utilized by Gates [4].

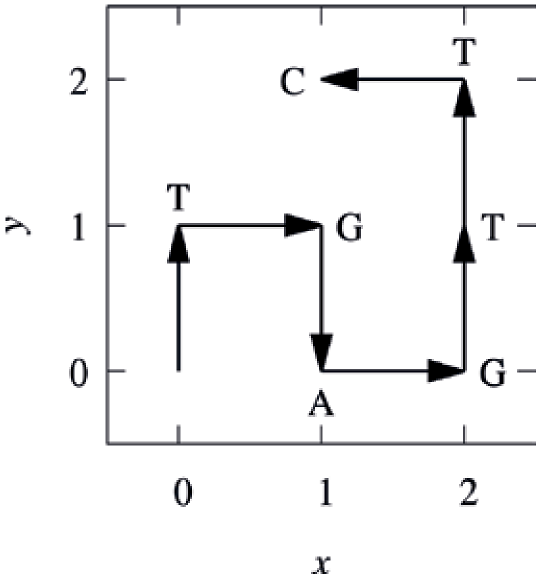


Figure 2. Graphical representation of sequence "TGAGTTC" generated by Gates' assignment (**Figure 1**).

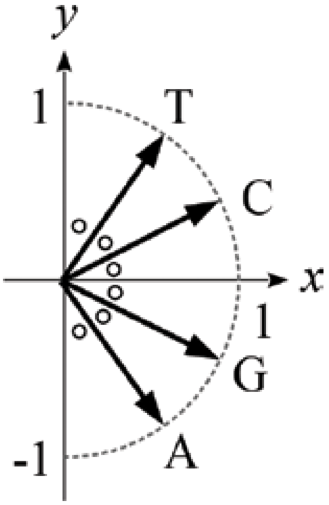


Figure 3. Two-dimensional vector assignment to bases utilized by Yau [9].

Yau's assignment. The assignments of Yau's type (including the variations with some modifications) are utilized in Refs. [9, 12, 15, 18, 19, 37–39].

Some researchers used another approach; they directly mapped bases on the *xy*-plane without vector assignment. Randić et al. plotted the *i*th base of a DNA sequence on the *xy*-plane at (*i*,0), (*i*,1), (*i*,2), and (*i*,3) for bases C, G, T, and A, respectively [7]. By connecting the plots, a *zigzag curve* is given. **Figure 5** demonstrates the zigzag curve for sequence "ATGGTGCACC" given

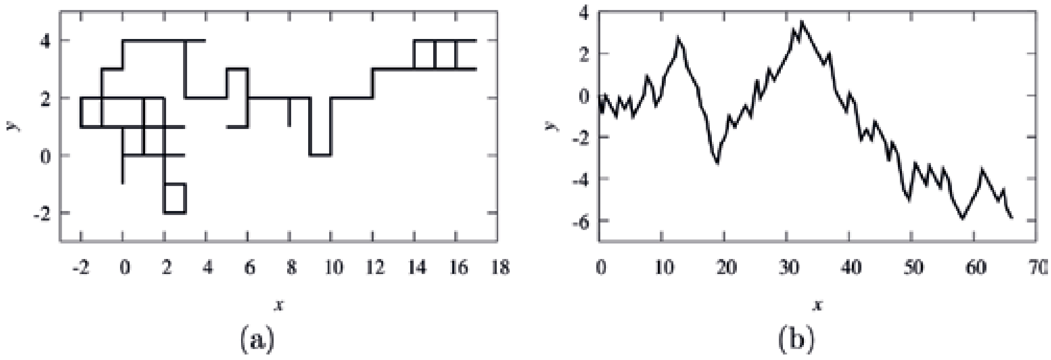


Figure 4. Graphical representations of the first exon of the human β-globin gene (GenBank: AF527577) represented by (a) Gates' vector assignment (**Figure 1**) and (b) Yau's vector assignment (**Figure 3**).

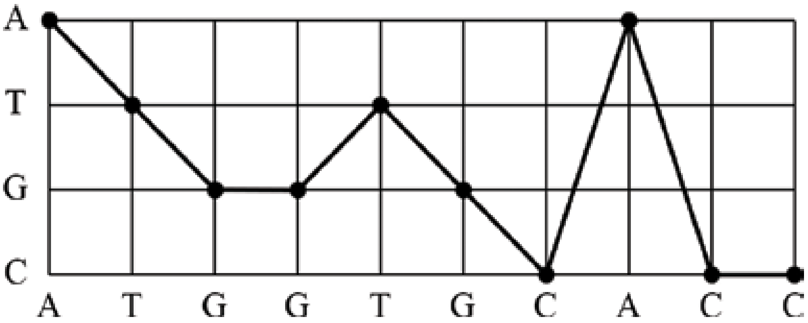


Figure 5. Zigzag curve for sequence "ATGGTGCACC" given by Randić's approach [7].

by Randić's approach. Similar to the graphical representation given by Yau's vector assignment (**Figure 4(b)**), the zigzag curve has no circuits. The approaches of this kind (including the variations with some modifications) are utilized in Refs. [8, 11, 13, 14, 17, 22].

2.1.2. Three-dimensional representation

Hamori and Ruskin [3] used a three-dimensional vector assignment to bases (**Figure 6**). Gates' approach (**Figure 1**) [4] is a simplified version of this assignment. However, unlike Gates' approach, Hamori's assignment does not make any circuit in the resultant curve (called *H-curve*), because the *z*-coordinate of the curve decreases monotonically with the positions of the bases in the original sequence. The assignments of this type (including the variations with some modifications) are utilized in Refs. [26, 27, 29, 31–36].

Zhang and Zhang [43] used another three-dimensional vector assignment shown in **Figure 7**. The resultant curve, called *Z-curve*, may have circuits therein like the curves generated by Gates' vector assignment (**Figure 1**). The assignments of this type (including the variations with some modifications) are utilized in Refs. [24, 42].

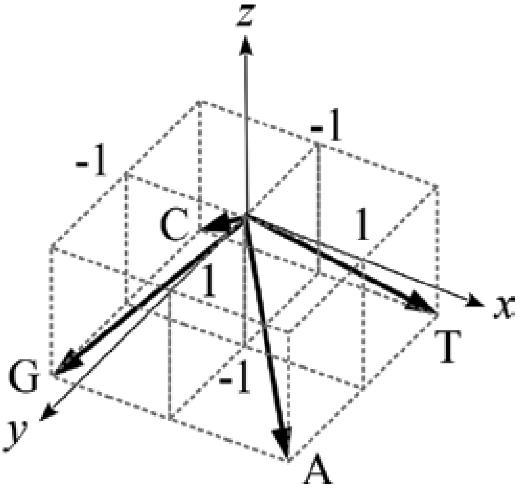


Figure 6. Three-dimensional vector assignment to four bases utilized by Hamori [3].

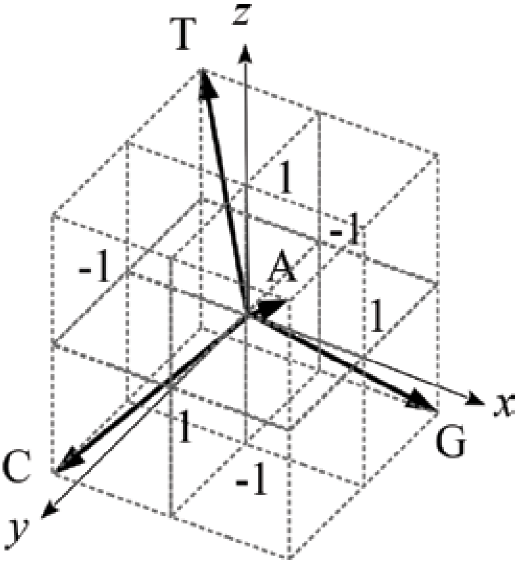


Figure 7. Three-dimensional vector assignment to four bases utilized by Zhang and Zhang [43].

2.1.3. Higher than three dimensions

The graphical representations in the space of higher than three dimensions cannot be visualized directly. Instead of direct visualization, they are expressed abstractly or projected on some spaces of lower dimensions. The approaches of this type (including the variations with some modifications) are utilized in Refs. [25, 30, 28].

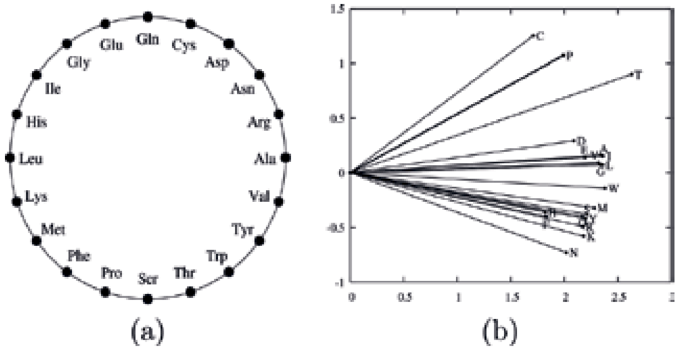


Figure 8. Two-dimensional vector assignments to 20 amino acids utilized by (a) Randić [44] and (b) Wen [45]. The 20 amino acids are indicated by three-letter codes and single-letter codes, respectively.

2.2. Graphical representation of proteins

A general strategy for graphical representation of protein sequences is common to that for DNA sequences, namely, numerical expression of characters followed by mapping on certain dimensional spaces, except for the fact that the number of character types is 20 instead of 4 for DNA sequences. A detailed review of graphical representation of protein sequences is given by Randić et al. [54]. Here, we briefly mention the variations of the graphical representation scheme of proteins.

Figure 8(a) and **(b)** presents two-dimensional vector assignments to 20 amino acids utilized by Randić et al. [44] and Wen and Zhang [45], respectively. In Randić's assignment, the 20 amino acids (indicated by three-letter codes) are arranged uniformly on a unit circle in alphabetical order. On the other hand, in Wen's assignment, the horizontal and the vertical coordinates of the vectors are given by pK_a values of COOH and NH_3^+ of the corresponding amino acid, respectively. The assignments of Randić's type and Wen's type (including the variations with some modifications) are utilized in Refs. [47] and [46, 48], respectively.

Yu and Huang [49] directly mapped 20 amino acids on a two-dimensional space and drew zigzag curves similar to the curve for the case of DNA sequences given by Randić's approach (**Figure 5**) [7].

He et al. [52] extended Randić's vector assignment (**Figure 8(a)**) to three dimensions by adding one extra coordinate corresponding to the position of the amino acid in the original sequence, with the modification of the arrangement of the 20 amino acids on the unit circle based on the 6-bit binary gray code assigned by the codon structure of the amino acids.

3. Numerical characterization of graphical representations

As well as the visual evaluation of the similarities between biological sequences through their graphical representations, the quantitative estimation of the similarities also can be done by the numerical characterization of the graphs. The general method of the quantitative estimation is to construct *feature vector*s based on the various kinds of characteristics of the graphs and, then, to calculate the distances between the feature vectors based on some sort of distance measures.

For the numerical characterization, there are two kinds of methods: geometrical methods and graph-theoretical ones [2].

3.1. Geometrical characterization

The most simple method of the geometrical characterization was proposed by Raychaudhury and Nandy [55], in which the graphs are numerically characterized by their geometrical centers. Let $\left(x_i, y_i\right)$ be the coordinate of the *i*th point of the graph, and then the geometrical center $\left(\overline{\mu}_x, \overline{\mu}_y\right)$ is computed by $\overline{\mu}_x = 1/N\sum_{i=1}^{N} x_i$ and $\overline{\mu}_y = 1/N\sum_{i=1}^{N} y_i$, where N is the total number of the points on the graph. The similarity/dissimilarity between the graphs of sequences, A and B, is measured by the Euclidean distance between their geometrical centers by

$$d_{AB} = \sqrt{\left(\overline{\mu}_x^A - \overline{\mu}_x^B\right)^2 + \left(\overline{\mu}_y^A - \overline{\mu}_y^B\right)^2}, \tag{1}$$

where A and B refer to the corresponding sequences.

A more accomplished geometrical characterization was proposed by Liao et al. [37], in which they constructed a two-component feature vector based on the 2×2 covariance matrix CM calculated from the two-dimensional graph by

$$CM = \begin{pmatrix} 1/N\sum_{i=1}^{N}\left(x_i - \overline{\mu}_x\right)^2 & 1/N\sum_{i=1}^{N}\left(x_i - \overline{\mu}_x\right)\left(y_i - \overline{\mu}_y\right) \\ 1/N\sum_{i=1}^{N}\left(y_i - \overline{\mu}_y\right)\left(x_i - \overline{\mu}_x\right) & 1/N\sum_{i=1}^{N}\left(y_i - \overline{\mu}_y\right)^2 \end{pmatrix}. \tag{2}$$

The two-component vector is given by the two eigenvalues of CM, λ_1, and λ_2, as (λ_1, λ_2). The similarity/dissimilarity between the graphs is measured by the Euclidean distance between the end points of their feature vectors.

The approach proposed by Qi et al. [18] is another example of the geometrical characterization. They constructed an eight-component feature vector from the averages of the y-coordinates of the eight different patterns of the two-dimensional graphical representations. The similarity/dissimilarity between the graphs is measured by the Euclidean distance between the end points of their feature vectors.

3.2. Graph-theoretical characterization

The graph-theoretical characterization that is most widely used is the method based on the D/D (distance/distance) matrix [56]. The off-diagonal (*i*, *j*) elements of the D/D matrix are defined as the quotient of the Euclidean distance between the *i*th and the *j*th vertices of the graph and the graph-theoretical distance between the two vertices. The D/D matrix is symmetric, and all the diagonal elements are zero by definition.

There are two variations of the D/D matrix. If the denominator (the graph-theoretical distance) is replaced by the sum of the geometrical lengths of the edges between the two vertices, the D/D matrix is denoted as the L/L matrix; if the denominator is replaced by the total number of the edges between the two vertices, the D/D matrix is denoted as the M/M matrix.

As an example, **Table 2** demonstrates the upper off-diagonal elements of the L/L matrix calculated for the graph of sequence "TGAGTTC" in **Figure 2**.

The feature vectors are constructed from the leading eigenvalues of the D/D matrix, which are the invariants of the matrix and can well describe the characteristics of the individual graphs. For example, Randić et al. [8] used 12-component vectors given by the first leading eigenvalues of the L/L matrices calculated from the 12 essentially different patterns of the graphical representations, and Liao and Wang [13] used three-component vectors constructed by the similar manner.

The similarity/dissimilarity between the sequences, A and B, is measured by the Euclidean distance between the end points of the corresponding feature vectors by

$$d_{AB} = \sqrt{\sum_{i=1}^{K} \left(\lambda_i^A - \lambda_i^B\right)^2}, \tag{3}$$

or the cosine of the angle between the feature vectors by

$$C_{AB} = \frac{\sum_{i=1}^{K} \lambda_i^A \cdot \lambda_i^B}{\sqrt{\sum_{i=1}^{K} \left(\lambda_i^A\right)^2 \cdot \sum_{i=1}^{K} \left(\lambda_i^B\right)^2}}, \tag{4}$$

where λ_i^A and λ_i^B are the ith components of the K-component feature vectors of the sequence A and B, respectively.

	G	A	G	T	T	C
T	1/1	$\sqrt{2}/2$	$\sqrt{5}/3$	2/4	$\sqrt{5}/5$	$\sqrt{2}/6$
G		1/1	$\sqrt{2}/2$	1/3	$\sqrt{2}/4$	1/5
A			1/1	$\sqrt{2}/2$	$\sqrt{5}/3$	2/4
G				1/1	2/2	$\sqrt{5}/3$
T					1/1	$\sqrt{2}/2$
T						1/1

Table 2. The upper off-diagonal elements of the L/L matrix for the graph in **Figure 2**.

4. Graphical representation based on binary images

The author and co-workers recently published the paper about a novel two-dimensional graphical representation of DNA sequences based on binary images [41]. In this section, we introduce our method and demonstrate the notable effects of *weighting* for the construction of the graphical representations introduced first by the author and co-workers [40].

4.1. Vector assignment

We used the two-dimensional vector assignment to four bases shown in **Figure 9**, which is a modified version of Gates' assignment (**Figure 1**). We located both G and C on the same side so that the GC-contents of the target sequences can be represented on the graphs; the graphs for the sequences with high GC-contents tend to grow in the downward direction, although the tendency is not rigid due to the weighting mentioned below.

4.2. Introducing weighting

In order to extract potential information conveyed by individual bases in DNA sequences, we introduced *weighting* into the process of generating graphical representations; we calculated the weighting factors based on a Markov chain model and multiplied them to the vectors assigned to the bases. As the weighting factors, we used self-information, which is the amount of information that we will receive when a certain event occurs [42]. The self-information is defined by

$$I(E) = -\log P(E), \tag{5}$$

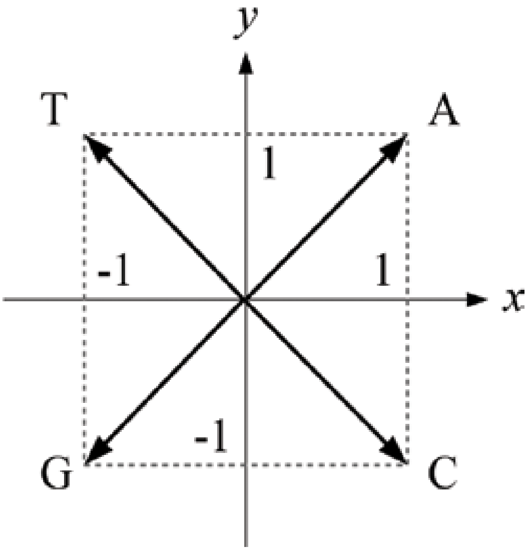


Figure 9. Two-dimensional vector assignment to four bases utilized by Kobori and Mizuta [41].

where $P(E)$ is the probability that event E occurs. We employed the conditional probability calculated based on the second-order Markov chain as $P(E)$ concerning about *codons*, which are triplets of bases in the coding regions of DNA sequences.

The conditional probability is calculated from the appearance frequencies of triplets of bases. For example, the probability that base A occurs after a pair of bases TC is given by

$$P(\mathrm{A}|\mathrm{TC}) = \frac{f(\mathrm{TCA})}{f(\mathrm{TCA}) + f(\mathrm{TCT}) + f(\mathrm{TCG}) + f(\mathrm{TCC})}, \tag{6}$$

where $f(S)$ is the number of appearances of triplet S. For the other combinations of bases, the conditional probabilities are calculated by a similar manner. The numbers of appearances of triplets were measured in all the DNA sequences analyzed.

Table 3 lists the weighting factors calculated with base 4 of the logarithm in Eq. (5) from 31 mammalian mitochondrial genomes. The weighting factor lower than 1.00 indicates that the pair of bases on the row tends to be followed by the base on the column, and on the other hand, the weighting factor higher than 1.00 indicates that, after the pair of bases on the row, the base on the column is hard to appear.

Let us illustrate the procedure of the graphical representation with weighting factors by a simple example. **Figure 10(a)** and **(b)** shows the graphical representations of sequence "ACATATG" by Kobori's vector assignment (**Figure 9**) without and with weighting, respectively. The weighting is not applied to the first two bases, because the weighting factors are not given for the first two bases by our weighting scheme. The weighting factors for the subsequent bases A, T, A, T, and G are 0.83, 0.90, 0.84, 0.92, and 1.42, respectively (see **Table 3**). The vectors for the bases are multiplied by the corresponding weighting factors. As a result, the graphical representation in **Figure 10(a)** is modified as shown in **Figure 10(b)**.

We demonstrate the notable effects of the weighting on the graphical representations by the real sequences. **Figure 11** depicts the graphical representations of three mammalian mitochondrial genomes without weighting and with weighting. Comparing the graphs with weighting (lower row) to the graphs without weighting (upper row), it can be recognized that the characteristics of the graphs are emphasized by the weighting and the individual species can be easily distinguished.

4.3. Generating binary images

A binary image is defined as a digitized image composed of the pixels with two possible values (e.g., 0 and 1), which are typically assigned by *white* and *black*, respectively, on the image. From the graphical representation, a binary image is generated in the following ways: if the pixels include at least a portion of a curve of the graphical representation, they are assigned 1; otherwise, they are assigned 0.

Preceding pair of bases	Third base			
	A	T	G	C
AA	0.82	0.92	1.47	0.95
AT	0.84	0.90	1.42	0.97
AG	1.04	1.11	1.11	0.79
AC	0.83	0.88	1.64	0.90
TA	0.86	0.93	1.28	1.01
TT	0.77	0.97	1.51	0.94
TG	0.73	1.14	1.16	1.06
TC	0.77	0.93	1.69	0.91
GA	0.79	1.08	1.15	1.03
GT	0.67	1.04	1.36	1.11
GG	0.79	1.14	1.22	0.93
GC	0.85	0.93	1.97	0.75
CA	0.84	0.90	1.44	0.96
CT	0.68	1.02	1.51	1.02
CG	0.91	1.00	1.22	0.91
CC	0.90	0.82	1.79	0.85

The 31 mammalian species are (with the accession numbers in the parentheses), human (V00662), pygmy chimpanzee (D38116), common chimpanzee (D38113), gorilla (D38114), gibbon (X99256), baboon (Y18001), Bornean orangutan (D38115), African green monkey (AY863426), cat (U20753), dog (U96639), wolf (EU442884), pig (AJ002189), sheep (AF010406), cow (V00654), buffalo (AY488491), tiger (EF551003), leopard (EF551002), Indian rhinoceros (X97336), white rhinoceros (Y07726), harbor seal (X63726), gray seal (X72004), African elephant (AJ224821), Asiatic elephant (DQ316068), black bear (DQ402478), brown bear (AF303110), polar bear (AF303111), rabbit (AJ001588), hedgehog (X88898), Norway rat (X14848), vole (AF348082), and squirrel (AJ238588).

Table 3. Weighting factors calculated from 31 mammalian mitochondrial genomes.

4.4. Numerical characterization by local pattern histograms

In this work, each graph is characterized by the frequency distributions of *local patterns* that appear on the graph. A local pattern is defined here as a small bitmap image of a certain size. Because each pixel of a binary image takes two variations (0 and 1), the number of the local patterns is 2^n, where n is the number of the pixels of the local pattern. Local patterns of large size are dominated by white pixels, while, on the other hand, those of small size do not have enough variations to express the characteristics of the local area of a graph. For this study, therefore, we chose 3×3 as the size of the local patterns (the number of the local patterns is 2^9=512). **Figure 12** shows the examples of the local patterns of window size 3×3. Excluding the pattern of which the pixels are all white, we construct a feature vector, or a *local pattern histogram*, of dimension 511 for each graph from the appearance frequencies of the local patterns on the graph.

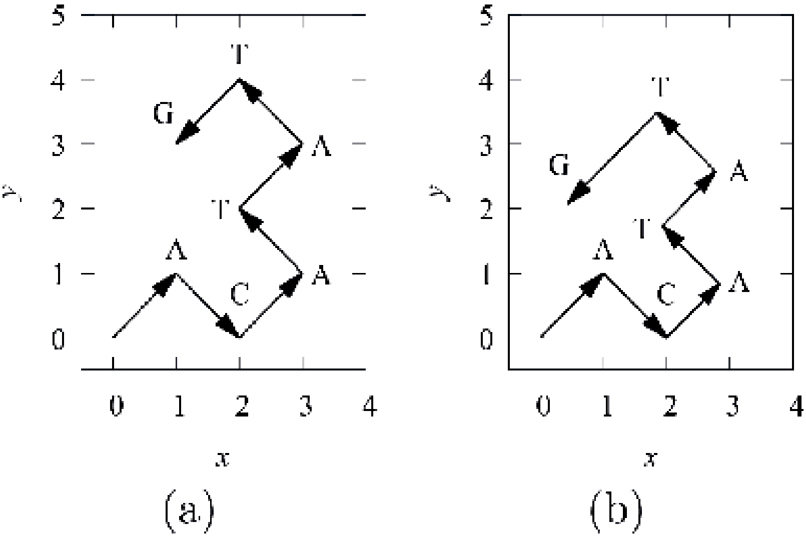


Figure 10. Graphical representation of sequence "ACATATG" by Kobori's vector assignment (Figure 9) without weighting (a) and with weighting (b).

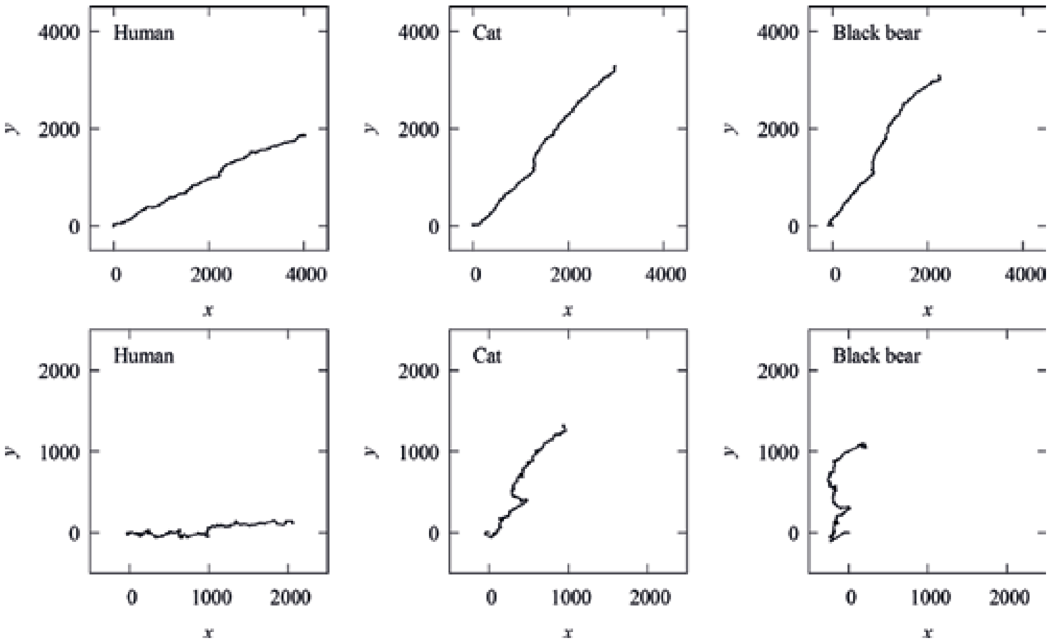


Figure 11. Graphical representations of mitochondrial genomes of three mammalian species without weighting (upper row) and with weighting (lower row). The arrow heads of the vectors are eliminated.

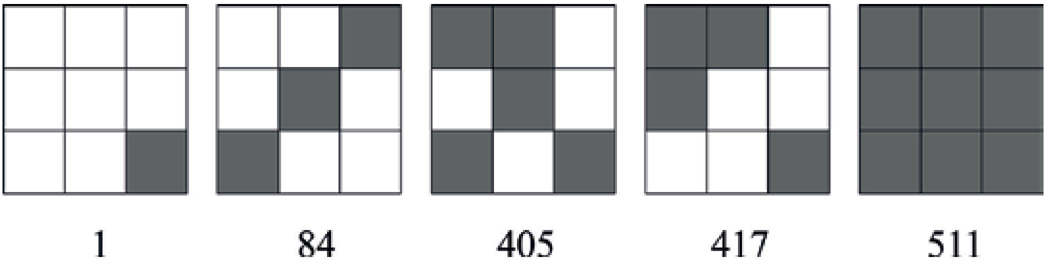


Figure 12. Examples of local patterns. The numbers below each local pattern are the serial numbers assigned to the local patterns.

4.5. Distance measures between local pattern histograms

There are several measures to estimate similarity/dissimilarity between two histograms. Here, we briefly mention five frequently used methods. In the following formulas, K is the number of the local patterns ($K = 511$), and p_i and q_i are the normalized appearance frequencies of the local pattern of serial number i in histograms P and Q, respectively $\left(\sum_{i=1}^{K} p_i = \sum_{i=1}^{K} q_i = 1\right)$.

4.5.1. Histogram intersection

Histogram intersection was proposed by Swain et al. [57] for color indexing of images, which is defined as

$$HI(P, Q) = \sum_{i=1}^{K} \min\ (p_i, q_i). \tag{7}$$

It ranges from 0 to 1, with 1 for P and Q being identical. It is converted to a distance by $D_{HI}(P, Q) = 1 - HI(P, Q)$.

4.5.2. Manhattan distance

Manhattan distance, also known as *city block distance* or L_1-norm, is defined as

$$D_{MD}(P, Q) = \sum_{i=1}^{K} |p_i - q_i|, \tag{8}$$

which ranges from 0 to 2, with 0 for P and Q being identical.

4.5.3. Bhattacharyya distance

Bhattacharyya distance [58] is defined between two probability distributions from a divergence

$$BD(P, Q) = \sum_{i=1}^{K} \sqrt{p_i q_i}, \tag{9}$$

which ranges from 0 to 1, with 1 for P and Q being identical. The Bhattacharyya distance is defined from the divergence by $D_{BD}(P, Q) = -\ln BD(P, Q)$.

4.5.4. Jensen-Shannon divergence

Jensen-Shannon divergence [59] is a symmetrized and smoothed version of Kullback–Leibler divergence [60], which is defined as

$$D_{JS}(P,Q) = \frac{1}{2}KL(P,M) + \frac{1}{2}KL(Q,M), \tag{10}$$

where $M = (P+Q)/2$ and $KL(\cdot,M)$ is the Kullback-Leibler divergence calculated by

$$KL(P,M) = \sum_{i=1}^{K} p_i \log_2 \frac{p_i}{m_i}, \tag{11}$$

$$KL(Q,M) = \sum_{i=1}^{K} q_i \log_2 \frac{q_i}{m_i}. \tag{12}$$

Here, $m_i = \left(p_i + q_i\right)/2$. Note that the local patterns having $p_i = p_i = 0$ are excluded from the calculation. The Jensen-Shannon divergence ranges from 0 to 1, with 0 for P and Q being identical.

4.5.5. Kendall's rank correlation coefficient

Kendall's rank correlation coefficient [61], also known as Kendall's τ, is defined as

$$\tau = \frac{X - Y}{\sqrt{X+Y+r}\sqrt{X+Y+s}}, \tag{13}$$

where X is the number of *concordant* $i, j\,(i > j)$ pairs, which are the i, j pairs that satisfy $\left(p_i - p_j\right)\left(q_i - q_j\right) > 0$; Y is the number of *discordant* pairs, which are the i,j pairs that satisfy $\left(p_i - p_j\right)\left(q_i - q_j\right) < 0$; r is the number of one kind of *tie* pairs, which are the i,j pairs that satisfy $p_i = p_j$ and $q_i \neq q_j$; and s is the number of the other kind of *tie* pairs, which are the i, j pairs that satisfy $p_i \neq p_j$ and $q_i = q_j$. The i, j pairs that satisfy both $p_i = p_j$ and $q_i = q_j$ are excluded from the calculation. Kendall's τ lies between −1 and 1, with 1 for the rank orders of p_is and q_is being completely in agreement with each other and with −1 for them being completely reversal with each other. The Kendall's τ is rescaled by

$$D_\tau(P,Q) = 1 - \frac{\tau + 1}{2}, \tag{14}$$

so that $D_\tau(P,Q)$ ranges from 0 to 1, with 0 for the rank orders of P and Q being identical.

4.6. Reconstruction of phylogenetic tree

Among the five distance measures mentioned above, histogram intersection and Manhattan distance showed the best performance. **Figure 13** shows the phylogenetic tree of 31 mammalian species reconstructed by our method using Unweighted Pair Group Method with

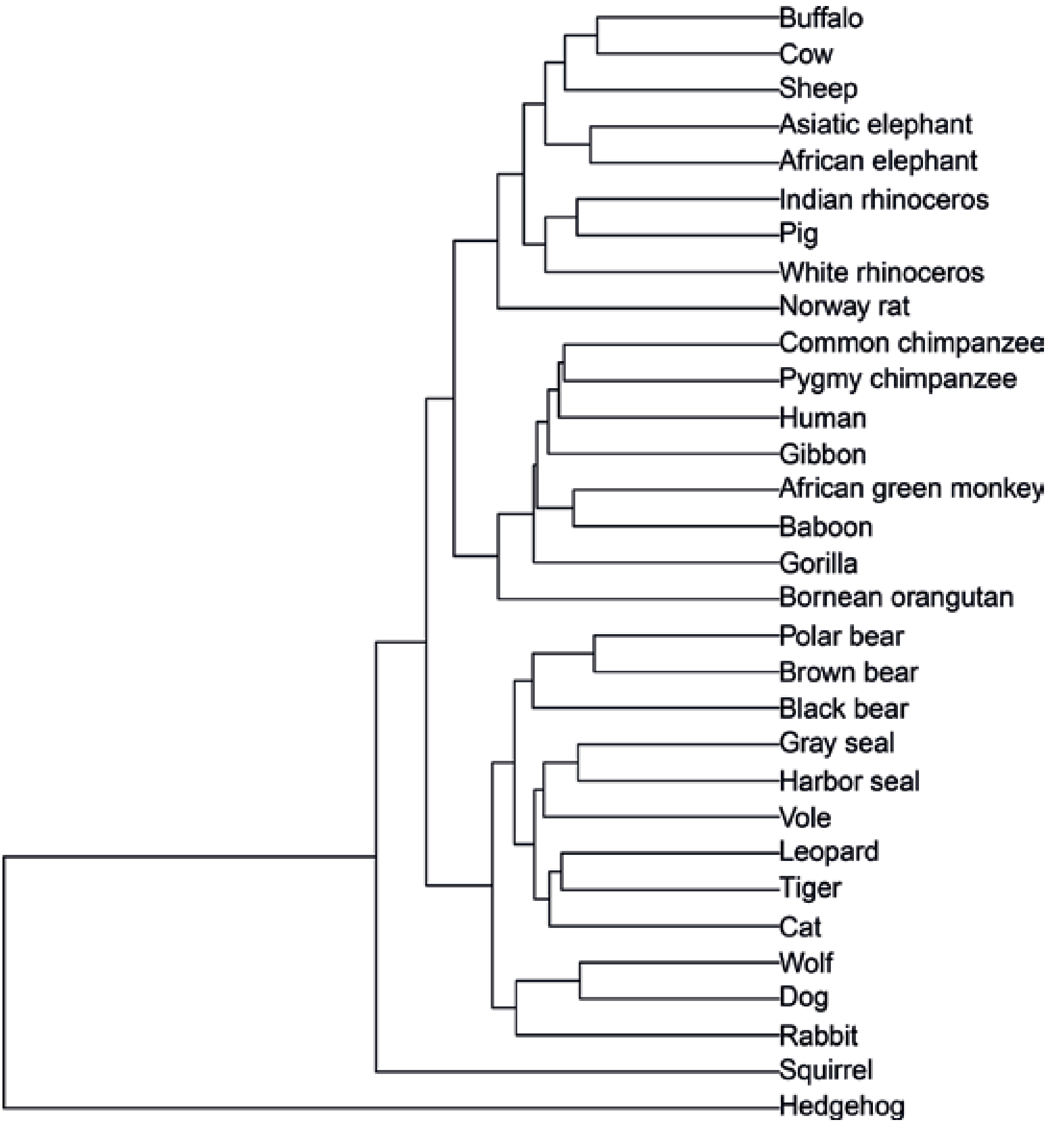


Figure 13. Phylogenetic tree of 31 mammalian species reconstructed by Kobori's method [41] using UPGMA based on the histogram intersection distance measure. The tree is generated by statistical analysis software R with package "ape".

Arithmetic mean (UPGMA) with the histogram intersection distance measure. The same tree is given by Manhattan distance.

5. Conclusion

With the rapid growth of the data size in the archives of biological sequences, the demand for the alignment-free sequence comparison methods is increasing. Graphical representation is one of the major alignment-free sequence comparison methods. In addition to the visual discrimination abilities of the sequences, the graphical representation has an advantage of requiring only small computational time. The similarity/dissimilarity between a pair of sequences is calculated from the feature vectors constructed based on the graphical representation. The time complexity of the calculation is estimated to be $O(K)$, where K is the dimension of the feature vector and K is usually independent of the sequence length (except for a few methods). Even though the computational time to make a graph, and to construct a feature vector from the graph, may depend on the sequence length N, typically $O(N)$, the construction of the graph and the feature vector is needed to be done for each sequence only once. Thus, the time complexity of the sequence comparison based on the graphical representation is regarded as $O(K)$, which is much less than that of the alignment, $O(N^2)$. From the above considerations, the graphical representation is expected to stay in the main stream of the alignment-free sequence comparison methods from now on, too.

Author details

Satoshi Mizuta

Address all correspondence to: slmizu@hirosaki-u.ac.jp

Graduate School of Science and Technology, Hirosaki University, Hirosaki, Aomori, Japan

References

[1] Roy A, Raychaudhury C, Nandy A. Novel techniques of graphical representation and analysis of DNA sequences—A review. Journal of Biosciences. 1998;**23**(1):55-71. DOI: 10.1007/BF02728525

[2] Nandy A, Harle M, Basak SC. Mathematical descriptors of DNA sequences: development and applications. ARKIVOC. 2006;**2006**(9):211-238. DOI: 10.3998/ark.5550190.0007.907

[3] Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry. 1983;**258**(2):1318-1327

[4] Gates MA. Simpler DNA sequence representations. Nature. 1985;**316**(6025):219. DOI: 10.1038/316219a0

[5] Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Current Science. 1994;**66**:309-314

[6] Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Bioinformatics. 1995;**11**(5):503-507. DOI: 10.1093/bioinformatics/11.5.503

[7] Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2003;**368**(1–2): 1-6. DOI: 10.1016/S0009-2614(02)01784-0

[8] Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters. 2003; **371**(1–2):202-207. DOI: 10.1016/S0009-2614(03)00244-6

[9] Yau SST, Wang J, Niknejad A, Lu C, Jin N, Ho YK. DNA sequence representation without degeneracy. Nucleic Acids Research. 2003;**31**(12):3078-3080. DOI: 10.1093/nar/gkg432

[10] Liu Y, Guo X, Xu J, Pan L, Wang S. Some notes on 2-D graphical representation of DNA sequence. Journal of Chemical Information and Modeling. 2002;**42**(3):529-533. DOI: 10.1021/ci010017g

[11] Liu XQ, Dai Q, Xiu Z, Wang T. PNN-curve: A new 2D graphical representation of DNA sequences and its application. Journal of Theoretical Biology. 2006;**243**(4):555-561. DOI: 10.1016/j.jtbi.2006.07.018

[12] Wu Y, Liew AWC, Yan H, Yang M. DB-curve: A novel 2D method of DNA sequence visualization and representation. Chemical Physics Letters. 2003;**367**(1–2):170-176. DOI: 10.1016/S0009-2614(02)01684-6

[13] Liao B, Wang TM. New 2D graphical representation of DNA sequences. Journal of Computational Chemistry. 2004;**25**(11):1364-1368. DOI: 10.1002/jcc.20060

[14] Song J, Tang H. A new 2-D graphical representation of DNA sequences and their numerical characterization. Journal of Biochemical and Biophysical Methods. 2005;**63**(3):228-239. DOI: 10.1016/j.jbbm.2005.04.004

[15] Zhang Y, Chen W. Invariants of DNA sequences based on 2DD-curves. Journal of Theoretical Biology. 2006;**242**(2):382-388. DOI: 10.1016/j.jtbi.2006.03.012

[16] Bielińska-Wąż D, Clark T, Wąż P, Nowak W, Nandy A. 2D-dynamic representation of DNA sequences. Chemical Physics Letters. 2007;**442**(1–3):140-144. DOI: 10.1016/j.cplett.2007.05.050

[17] Qi ZH, Qi XQ. Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chemical Physics Letters. 2007;**440**(1–3):139-144. DOI: 10.1016/j.cplett.2007.03.107

[18] Qi ZH, Li L, Qi XQ. Using Huffman coding method to visualize and analyze DNA sequences. Journal of Computational Chemistry. 2011;**32**(15):3233-3240. DOI: 10.1002/jcc.21906

[19] Zhang ZJ. DV-curve: A novel intuitive tool for visualizing and analyzing DNA sequences. Bioinformatics. 2009;**25**(9):1112-1117. DOI: 10.1093/bioinformatics/btp130

[20] Jafarzadeh N, Iranmanesh A. A novel graphical and numerical representation for analyzing DNA sequences based on codons. MATCH Communications in Mathematical and in Computer Chemistry. 2012;**68**:611-620

[21] Wąż P, Bielińska-Wąż D, Nandy A. Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences. Journal of Mathematical Chemistry. 2013;**52**(1):132-140. DOI: 10.1007/s10910-013-0249-1

[22] Zou S, Wang L, Wang J. A 2D graphical representation of the sequences of DNA based on triplets and its application. EURASIP Journal on Bioinformatics and Systems Biology. 2014;**2014**(1):1. DOI: 10.1186/1687-4153-2014-1

[23] Hamori E. Novel DNA sequence representations. Nature. 1985;**314**:585. DOI: 10.1038/314585a0

[24] Randić M, Vračko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. Journal of Chemical Information and Computer Sciences. 2000;**40**(5):1235-1244. DOI: 10.1021/ci000034q

[25] Randić M, Balaban AT. On a four-dimensional representation of DNA primary sequences. Journal of Chemical Information and Computer Sciences. 2003;**43**(2):532-539. DOI: 10.1021/ci020051a

[26] Liao B, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chemical Physics Letters. 2004;**388**(1–3):195-200. DOI: 10.1016/j.cplett.2004.02.089

[27] Liao B, Ding K. A 3D graphical representation of DNA sequences and its application. Theoretical Computer Science. 2006;**358**(1):56-64. DOI: 10.1016/j.tcs.2005.12.012

[28] Liao B, Li R, Zhu W, Xiang X. On the similarity of DNA primary sequences based on 5-D representation. Journal of Mathematical Chemistry. 2007;**42**(1):47-57. DOI: 10.1007/s10910-006-9091-z

[29] Yao YH, Nan XY, Wang TM. Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation. Chemical Physics Letters. 2005;**411**(1–3):248-255. DOI: 10.1016/j.cplett.2005.06.040

[30] Chi R, Ding K. Novel 4D numerical representation of DNA sequences. Chemical Physics Letters. 2005;**407**(1–3):63-67. DOI: 10.1016/j.cplett.2005.03.056

[31] Qi ZH, Fan TR. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2007;**442**(4–6):434-440. DOI: 10.1016/j.cplett.2007.06.029

[32] Yu JF, Sun X, Wang JH. TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. Journal of Theoretical Biology. 2009;**261**(3): 459-468. DOI: 10.1016/j.jtbi.2009.08.005

[33] Xie G, Mo Z. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. Journal of Theoretical Biology. 2011; **269**(1):123-130. DOI: 10.1016/j.jtbi.2010.10.018

[34] Jafarzadeh N, Iranmanesh A. C-curve: A novel 3D graphical representation of DNA sequence based on codons. Mathematical Biosciences. 2013;**241**(2):217-224. DOI: 10.1016/j.mbs.2012.11.009

[35] Wąż P, Bielińska-Wąż D. 3D–dynamic representation of DNA sequences. Journal of molecular modeling. 2014;**20**(3):2141. DOI: 10.1007/s00894-014-2141-8.

[36] Wąż P, Bielińska-Wąż D. Non-standard similarity/dissimilarity analysis of DNA sequences. Genomics. 2014;**104**:464-471. DOI: 10.1016/j.ygeno.2014.08.010

[37] Liao B, Tan M, Ding K. Application of 2-D graphical representation of DNA sequence. Chemical Physics Letters. 2005;**414**(4–6):296-300. DOI: 10.1016/j.cplett.2005.08.079

[38] Yu C, Liang Q, Yin C, He RL, Yau SST. A novel construction of genome space with biological geometry. DNA Research. 2010;**17**(3):155-168. DOI: 10.1093/dnares/dsq008

[39] Huang G, Zhou H, Li Y, Xu L. Alignment-free comparison of genome sequences by a new numerical characterization. Journal of Theoretical Biology. 2011;**281**(1):107-112. DOI: 10.1016/j.jtbi.2011.04.003

[40] Mizuta S, Yamaguchi K. A novel 2-dimensional graphical representation of DNA sequences using weighted vector assignments. In: The Proceedings of the 6th International Conference on Bioinformatics Computational Biology (BICoB2014); Las Vegas; 2014. pp. 33-38

[41] Kobori Y, Mizuta S. Similarity estimation between DNA sequences based on local pattern histograms of binary images. Genomics, Proteomics & Bioinformatics. 2016;**14**(2):103-112. DOI: 10.1016/j.gpb.2015.09.007

[42] Yamaguchi K, Mizuta S. A new graphical representation of DNA sequences using symmetrical vector assignment. Review of Bioinformatics and Biometrics. 2014;**3**:14-21

[43] Zhang R, Zhang CT. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. Journal of Biomolecular Structure & Dynamics. 1994;**11**(4):767-782. DOI: 10.1080/07391102.1994.10508031

[44] Randić M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. Chemical Physics Letters. 2006;**419**(4–6):528-532. DOI: 10.1016/j.cplett.2005.11.091

[45] Wen J, Zhang Y. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters. 2009;**476**(4–6):281-286. DOI: 10.1016/j.cplett.2009.06.017

[46] Wu ZC, Xiao X, Chou KC. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. Journal of Theoretical Biology. 2010;**267**(1):29-34. DOI: 10.1016/j.jtbi.2010.08.007

[47] He PA, Zhang YP, Yao YH, Tang YF, Nan XY. The graphical representation of protein sequences based on the physicochemical properties and its applications. Journal of Computational Chemistry. 2010;**31**(11):2136-2142. DOI: 10.1002/jcc.21501

[48] Yu C, Cheng SY, He RL, Yau SST. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011;**486**(1–2):110-118. DOI: 10.1016/j.gene.2011.07.002

[49] Yu HJ, Huang DS. Novel 20-D descriptors of protein sequences and it's applications in similarity analysis. Chemical Physics Letters. 2012;**531**:261-266. DOI: 10.1016/j.cplett.2012.02.030

[50] Abo el Maaty MI, Abo-Elkhier MM, Abd Elwahaab MA. 3D graphical representation of protein sequences and their statistical characterization. Physica A: Statistical Mechanics and its Applications. 2010;**389**(21):4668-4676. DOI: 10.1016/j.physa.2010.06.031

[51] He P, Wei J, Yao Y, Tie Z. A novel graphical representation of proteins and its application. Physica A: Statistical Mechanics and its Applications. 2012;**391**(1–2):93-99. DOI: 10.1016/j.physa.2011.08.015

[52] He P, Li D, Zhang Y, Wang X, Yao Y. A 3D graphical representation of protein sequences based on the gray code. Journal of Theoretical Biology. 2012;**304**(0):81-87. DOI: 10.1016/j.jtbi.2012.03.023

[53] Czerniecka A, Bielińska-Wąż D, Wąż P, Clark T. 20D-dynamic representation of protein sequences. Genomics. 2016;**107**(1):16-23. DOI: 10.1016/j.ygeno.2015.12.003

[54] Randić M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D. Graphical representation of proteins. Chemical Reviews. 2011;**111**(2):790-862. DOI: 10.1021/cr800198j

[55] Raychaudhury C, Nandy A. Indexing scheme and similarity measures for macromolecular sequences. Journal of Chemical Information and Computer Sciences. 1999;**39**(2):243-247

[56] Randić M, Kleiner AF, De Alba LM. Distance/distance matrixes. Journal of Chemical Information and Modeling. 1994;**34**(2):277-286. DOI: 10.1021/ci00018a008

[57] Swain MJ, Ballard DH. Color indexing. International Journal of Computer Vision. 1991;**7**(1):11-32. DOI: 10.1007/BF00130487

[58] Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of Calcutta Mathematical Society. 1943;**35**(1):99-109

[59] Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991;**37**(1):145-151. DOI: 10.1109/18.61115

[60] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics. 1951;**22**(1):79-86. DOI: 10.1214/aoms/1177729694

[61] Kendall MG. A new measure of rank correlation. Biometrika. 1938;**30**(1–2):81-93. DOI: 10.1093/biomet/30.1-2.81

Section 2

Data Models and Network-based Systems

Data Models in Neuroinformatics

Elishai Ezra Tsur

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.73516

Abstract

Advancements in integrated neuroscience are often characterized with data-driven approaches for discovery; these progressions are the result of continuous efforts aimed at developing integrated frameworks for the investigation of neuronal dynamics at increasing resolution and in varying scales. Since insights from integrated neuronal models frequently rely on both experimental and computational approaches, simulations and data modeling have inimitable roles. Moreover, data sharing across the neuroscientific community has become an essential component of data-driven approaches to neuroscience as is evident from the number and scale of ongoing national and multinational projects, engaging scientists from diverse branches of knowledge. In this heterogeneous environment, the need to share neuroscientific data as well as to utilize it across different simulation environments drove the momentum for standardizing data models for neuronal morphologies, biophysical properties, and connectivity schemes. Here, I review existing data models in neuroinformatics, ranging from flat to hybrid object-hierarchical approaches, and suggest a framework with which these models can be linked to experimental data, as well as to established records from existing databases. Linking neuronal models and experimental results with data on relevant articles, genes, proteins, disease, etc., might open a new dimension for data-driven neuroscience.

Keywords: databases, hierarchy-based data models, integrated neuroscience, LEMS, layer-oriented data models, NeuroML, object-based data models

1. Introduction

Integrated neuroscience (IN) is an emerging field of research with implications that range from the derivation of neural networks motifs [1] to approaching one of the most important questions ever tackled: the nature of consciousness [2]. IN has emerged from the aspiration for insights, which could only be inferred from data obtained across multiple spatial scales

(Ångströms to centimeters) and temporal scales (milliseconds to years). An integrated approach toward neuroscience requires multiscale neural data—from molecular regulations (S1) and the dynamics of individual synapses (S2) to information processing in neural networks (S3) and to the orchestrated function of brain maps (S4) and systems (S5) (**Figure 1**).

In their seminal paper "Neuroscience on the NET" [3], Peter Fox and Jack Lancaster draw parallels between neuroinformatics and the "genome informatics community" that have gained remarkable insights leveraging the Web to generate federated frameworks for "collective wisdom." Fox and Lancaster called the "prospective developers of neuroscience databases" to "absorb the collective wisdom of these network pioneers," handle the challenge of "sematic compatibility," and develop a neuroscientific database federation to realize the field's potential of "scientific exploration." The increased attention over the past decade to data-driven neuroscience is attested by the number of published papers having these terms as keywords. Tracking the number of published papers on the subject (retrieved from PubMed) follows an exponential curve, where the "knee" of the curve is in 2010 (**Figure 2**, left). A combination of an integrated approach to neuroscience with the establishment of a federated framework for "collective wisdom" of neuroscientists and engineers can fuel the celebration of the "era of the brain."

1.1. The data tail

Neuroscientific data flow from various resources, ranging from government funded consortiums of laboratories, to individual laboratories spread worldwide.

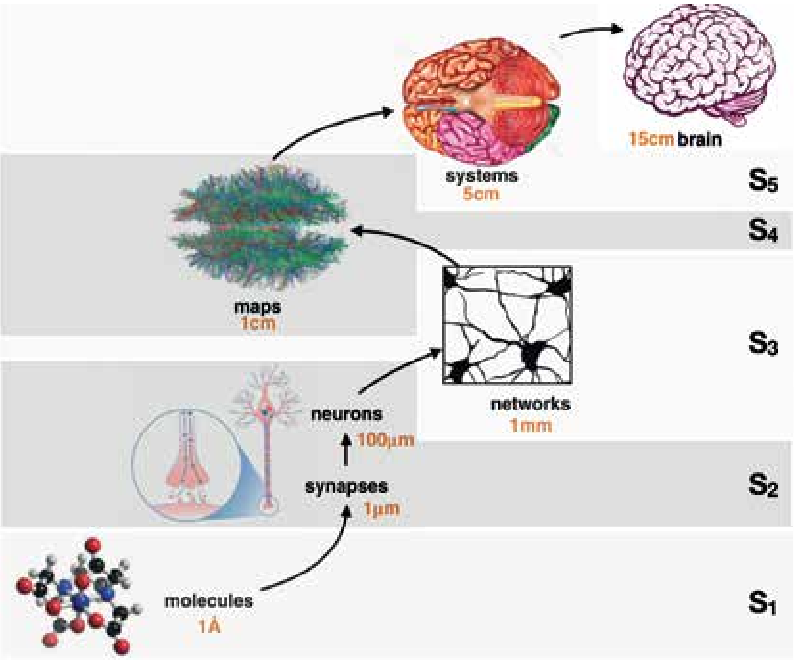


Figure 1. Schematics of the spatial scales (molecules to complete nervous systems) of integrated neuroscience.

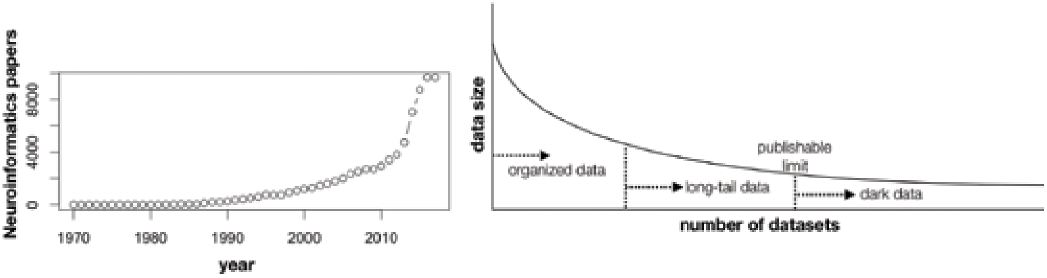


Figure 2. Number of papers linking neuroscience and data across the last 4 decades (right), the emergence of the long-tail and dark data volumes from exploring the size and number of neuroscientific data sets.

1.1.1. "Big science" initiatives

Today, one of the most ambitious endeavors aiming at integrated neuroscience is the human brain project (HBP) [4]. HBP is a multinational EU-funded research initiative, aimed at advancing multiscale brain-inspired information technology. Neuroinformatics lies at the core of HBP and orchestrated by COLLAB, a Web-based collaborative cloud-based system, developed within HBP's neuro informatics platform (NIP). COLLAB is fueling the project's other platforms (brain simulation, neurorobotics, medical informatics, and neuromorphic computing) with immense upstream and downstream data flows. It is distributed as a software as a service (SaaS) by HBP's high-performance analytics and computing platform (HPAC), enabling massive data archiving and distribution of virtual machines (VM) to collaborators, empowering them with high-end supercomputing capabilities for simulation and data analytics. COLLAB's mission is not a trivial one: it must be interfaced with heterogeneous data types and ontologies to manage metadata storage and provide a query system with which rodent and human brain atlases can be constructed and populated using different data modalities (anatomy, physiology). Moreover, COLLAB should link its data with foreign maps, databases, and atlases. HBP precedent is the human genome project (HGP) [5], a project that radically changed the ways research in molecular biology is carried out and how we perceive it. HGP has new disciplines as heirs, ranging from personalized genomic-based medicine to comparative genomics. It has established innovative approaches to biological database creation and maintenance, such as the construction of public small-molecule libraries with which biological pathways can be standardized. HBP approach aims to do the same for neuroscience.

Inspired by HGP and HBP, a new scientific endeavor termed "BRAIN" was initiated in the US by the White House, "aimed at revolutionizing our understanding of the human brain" [6] and like the other initiatives to "empower individual labs by providing…open-access databases." Another ambitious project is the NIH-funded human connectome project (HCP), which aims to characterize the human brain connectivity and functions. In this project, colossal amount of data will be gathered from many hundreds of patients with state-of-the-art 3D fMRI machines, EEG and MEG. Full-genome sequencing from all subjects will be performed as well. Behavioral measures in different domains (cognition, emotion, perception, and motor function) will also be recorded [7]. Other governmentally funded integrated neuroscience programs include the "Brain Canada" [8] and the "China Brain Project" [9]. All aforementioned acknowledge the fact that establishing standardized data collection and processing, as well as mechanisms for data sharing and credit allocation, are fundamental to their project's success.

1.1.2. The long tail data

Enormous "big-science" initiatives such as the HBP, HGP, and the BRAIN have large coordination teams, and as mentioned above, great emphasis is given within their scope to data and copyrights. Moreover, they are usually required (by the funding agency) to share their results with the community. However, routine scientific work in individual labs or small consortiums generates the majority of scientific data. Although each lab produces relatively limited amount of data, together they constitute the bulk of neuroscientific information. These granular, individually assembled data sets (usually given as publishable units) are referred to in the literature as "long-tail data." The tail of data also includes "dark-data," which is comprised of unpublished information, sitting aimlessly in personal hard drives or in restricted shared folders (**Figure 2**, right). Within this tail of neuroscientific data lies a unique opportunity—the possibility of assembling these scattered pieces of knowledge into "deep" data collections [10]. Ferguson and colleagues reviewed "data sharing" in the long tail of neuroscience [11]. While describing the limitations of data sharing among individual labs, they demonstrated the impact such an attempt would make through the success of the IMPACT consortium [12]. IMPACT collected tailed clinical data from over 43,243 patients who have suffered from traumatic brain injury (TBI) over the span of 20 years into a "deep" database. Their data were mined to derive a prognostic model with unprecedented precision for predicting recovery, ushering a new era for TBI precision medicine [13]. IMPACT demonstrated the way "deepening" long-tail data can provide incredible insights and even revolutionize treatment. Another example is the recently established data sharing community for spinal cord injury (SCI) research [14].

1.1.3. Deepening the long tail data

The main challenges of deepening tailed neuroscientific data encompass all levels of data handling and association including acquisition, quality control, representation, system implementation, user interface and documentation, data analysis, budget and maintenance, and federation [15]. Among all these dimensions, data representation is the most extensively discussed, as it stands as a prominent bottle neck in the definition of data sharing standards. Recently, a group of thought leaders, comprised of scholars, librarians, archivists, publishers and research funders, came together to provide the research community with guidelines toward the creation of standards for data sharing, which they termed the "FAIR Data Principles" [16]. The FAIR guidelines dictate that data should be (1) findable, with a rich assigned standardized metadata and persistent identifier; (2) accessible, via an identifier and an open, free, and universally implementable communications protocol; (3) interoperable, via broadly applicable language for knowledge representation; and (4) reusable, via domain-relevant community standards. A great emphasis is therefore given in the FAIR guidelines to carefully constructed metadata.

Following the importance of data standardization in computational modeling in biology, and particularly in neuroscience, the COMBINE consortium has been initiated in 2009 [17]. COMBINE aims to coordinate and facilitate different community-based standardization efforts in the field of computational biology. COMBINE's neuro-related standardization efforts include computational neuroscience ontology (CNO) [18], NeuroML [19], and spiking neural markup language (SpineML) [20].

One of the most prominent database federations for the neuroscientific community is the neuroscience information framework (NIF) [21], which has been cataloging and surveying the neuroscience resource landscape since 2006. NIF currently gives access to over 250 data sources categorized to different subjects ranging from software tools to funding resources. NIF provides a distributed query engine to tailed data, which is independently created and curated. This type of distributed search among independent databases is enabled through NIF's DISCO registry tool with which a Web resource can send automatic, or manual, data updates to the NIF system [22].

1.2. Models for computational neuroscience

Linking neuroscientific data with simulation environments has deep roots in the origins of neuronal modeling and databases. Starting with the seminal works of Alan Hodgkin, Andrew Huxely, and Wilifrid Rall during the 1970s, which established today's most utilized models for neuronal dynamics, the scale of simulating neural networks has picked up. As computing resources became abundant, neuronal simulations began to be carried out by an increasing number of labs, creating the need for a database in which already established models could be realized and build upon.

Models of neuronal dynamics span over all scales abstraction, where each abstraction level encapsulates an increasing amount of details (**Figure 3**) [23].

Increasing level of complexity entails increasing amount of required data. Databases for computational models are therefore well integrated with simulation platforms such as NEURON [24] and GENESIS [25].

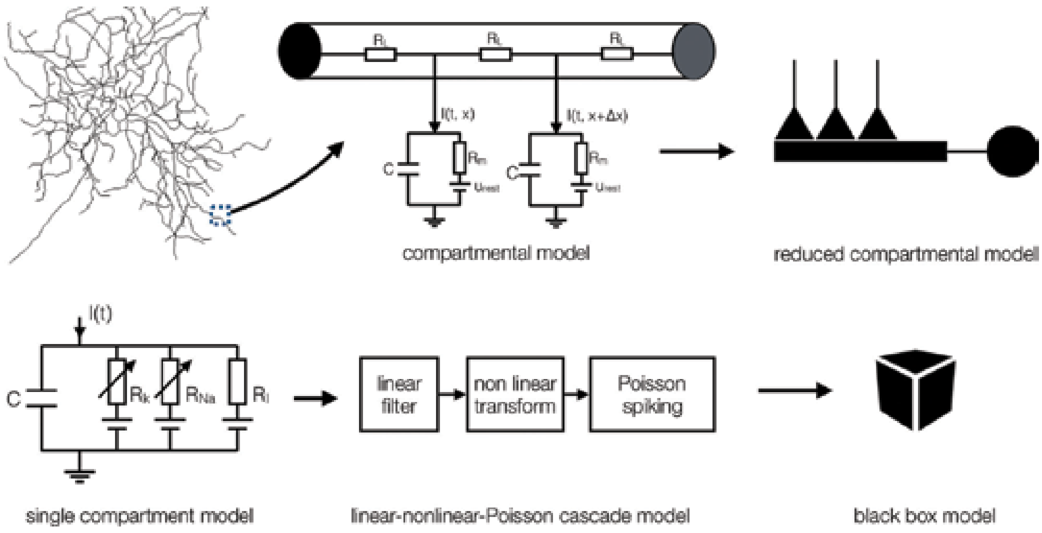


Figure 3. Models' schematics in computational neuroscience.

PyNN [26] and NeuroML are independently developed approaches to allow standardization of neuronal modeling, enabling models' utilization across simulators. While NeuroML took the declarative approach for modeling, explicitly specifying the model using in a structured format (with XML), PyNN took the procedural approach, specifying the models using functions and procedures, in this case, executing python scripts on different simulators.

Neuronal modeling usually requires morphological, connectivity, and physiological data. Neuromorph.org is the largest federated collection of 3D neuronal reconstructions and associated metadata [27]. For each neuron, a rich metadata is gathered, including miscellaneous information ranging from file format to the source specie, sex, age, weight, etc. Forward automatic analysis, ranging from size to topology, is also made for each morphology, leading to a range of morphological insights [28]. NeuroMorpho.Org is carefully curated and administrated, with a team responsible for file transfer, conversion, annotation and curation, minimizing the burden on the data submitter. A model can be submitted to NeuroMorpho.Org only when it is associated with published results—a curation decision that on one hand ensures data quality but on the other hand rejects the wealth of information residing on the dark side of the data tail. Since neuronal modeling incorporates morphological data, as well as physiological data, interoperability between the two is essential. Indeed, the NeuroMorpho.Org database can be utilized with other complementary resources such as the CellPropDB, NeuronDB [29], ModelDB [30], and MicrocircuitDB (all four are curated by SenseLab at Yale university). While CellPropDB is comprised of data regarding receptors, channels, and transmitters, NeuronDB distributes these elements across a specific neuron. ModelDB comprised of computational models of neurons derived from NeuronDB. MicrocircuitDB contains circuit modeling, which was built upon data from ModelDB. Today, all SenseLab databases are tightly coupled with Neuron [31].

2. Data models

Neuroscientific data models must encompass the different levels of neuronal scales: starting at the molecular regime, going up to the membrane and synapse levels, moving through the dendritic tree and axonal branches, and finishing at the circuit and system levels. Each level encapsulate further details. For example, at the circuit level, data on proteins and ions is 'hidden' at the encapsulated lower levels of representation. Various data models exist for each scale—here I chose a representative for each model, which in my opinion reflects its main properties. Please note that the schematics shown below for each data model, particularly for Neuron's object-based representation schemes, do not aim to accurately specify the objects hierarchy scheme in terms of inheritance or composition. They are given here to purely illustrate the general approach for modeling.

2.1. Structuring data

Following samples acquisition, data must be structurally organized. It can be structured in either a "flat file," a tabular formation, a structured file (such as XML), an object based, or a layer-oriented scheme (**Figure 4**). Data in a flat file are stored in an unstructured manner and therefore manipulating it would require reading it entirety into memory. Data can be

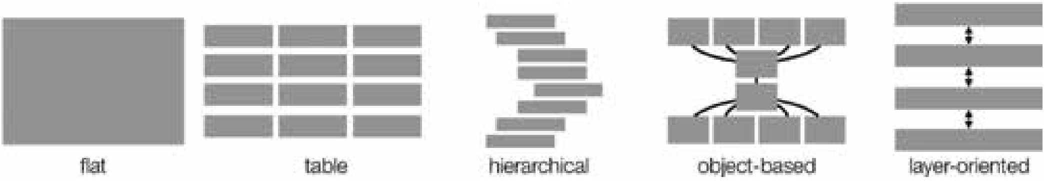


Figure 4. Schematics of data structuring paradigms.

structured as a table, where each value is headed with a type and usually also with a size identifier. eXtensible markup language (XML) is a different approach for data structuring, in which data is arranged in schemes, where each subsequent level increases the scope of the previous one. XML gained industry momentum due to its simplicity and flexibility, enabling declarative specifications rather than coding. This facilitates automated transformation of model specifications into multiple other formats. One of the main alternatives to data modeling is object-based representation of information, in which entities are defined with a set of properties and connected as attributes. Object-based representation allows the encapsulation of internal details of the data associated with the heterogeneity of the underlying data sources. Another approach is the layer-oriented approach (LOA), in which interlinked declarative languages (or layers) specify the model. The rationale behind the LOA is the premise that computational models are not a "flat collection of equations" but rather a hierarchical structure from which the underlying biological concept is reflected.

2.2. Models of morphological data

Before data can be modeled, it needs to be abstracted. The level of neuromorphological details with today's advanced imaging techniques, such as the two photons microscopy, is staggering. Moreover, since image stacks cannot be directly used for computational modeling due to their nontrivial interpretability and size, morphology must be reconstructed from them. Encapsulation of the details of neuromorphological data needs to consider its application, which in our case is computational modeling. Since different environments such as Neurolucida, NEURON, and GENESIS use a different representation of morphological data (**Figure 5**), a generalized representation, such as the MorphoML, is required to enable easy conversion to each format.

2.2.1. Flat structuring of morphological data

Neuromantic [32] is a semiautomatic stand-alone freeware reconstruction application, in which serial image stacks (JPEG to TIFF) are used to reconstruct dendritic trees. Reconstructions are stored in the SWC data format. SWC is one of the most widely used data models for neuromorphological data, for which a standardized version is used by Neuromorpho.org (not to be confused with Adobe file format). It is ASCII encoded text, where each line represents a single morphological sample point, which is represented by seven data items: id, structure identifier, 3D location, radius, and parent id. For example, the data entry:

```
2 1 -2 -3.33 0 7.894 1
```

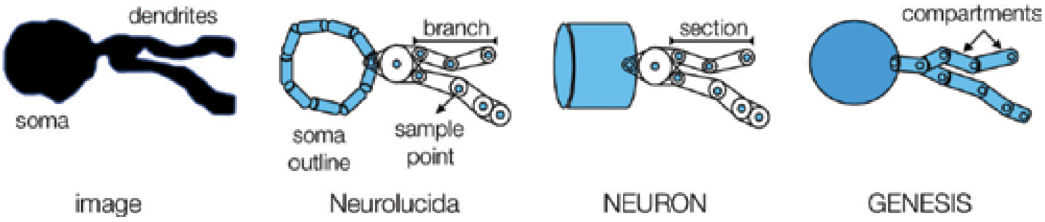


Figure 5. Representation of morphological data across different environments.

signifies a sample point with id number 2, connected to sample point number 1, identified as being located at the soma (structure identifier 1), located at (x = −2, y = −3.33, z = 0), in a compartment with a 7.894 radius. SWC files are generally small in size, trivial to read, and widely adopted across applications.

2.2.2. Hierarchical structuring of morphological data

Another approach for neuromorphological data modeling is using XML. One example is the MorphoML [33], which is a part of NeuroML. For example, defining soma and a dendrite can be written as:

```
<cells>
   <cell name = "Example">
     <meta:notes>A Simple cell</meta:notes>
     <segments>
       <segment id = "0" name = "Soma" cable = "0">
         <proximal x = "0.0" y = "0.0" z = "0.0" diameter = "16.0"/>
         <distal x = "0.0" y = "0.0" z = "0.0" diameter = "16.0"/>
      </segment>
      <segment id = "1" name = "Dend" parent = "0" cable = "1">
         <proximal x = "8.0" y = "0.0" z = "0.0" diameter = "5.0"/>
         <distal x = "28.0" y = "2.0" z = "0.0" diameter = "6.0"/>
     </segment>
     </segments>
     <cables>
       <cable id = "0" name = "SomaCable" />
       <cable id = "1" name = "DendriteCable" />
     </cables>
   </cell>
</cells>
```

This XML-based neuromorphological specification can be verified using a dedicated software, as well as be converted to GENESIS or NEURON readable formats. Schematics of hierarchy-based representations of neuromorphological data are illustrated in **Figure 6** (left).

2.2.3. Object-based structuring of morphological data

NEURON, one of the dominant players in computational neuroscience, has a dedicated file type termed "HOC." It has C-like syntax with an additional object-oriented expressability. One of the uses for "HOC" is defining a neuronal morphology by constructing an array of "section" objects, each defined by a series of four points (using neuron's "pt3dadd" function): three coordinates and a radius. Sections can be connected to one another (using neuron's "connect" function). For example, two connected sections can be characterized by sample points: (109.72, 125.39, 19.28) and (109.93, 125.85, 19.01) with radiuses 3.96136 and 3.88, respectively, for the first section and (115.42, 125.23, 15.19) and (115.69, 125.16, 15.05) with radiuses 0.752 and 0.64, respectively, for the second section:

```
create section[703]

section[0] {
  pt3dclear()
  pt3dadd(109.721,125.39,19.2812,3.96136,0)
  pt3dadd(109.93,125.285,19.0172,3.88406,0)
}

section[1] {
  pt3dclear()
  pt3dadd(115.427,125.239,15.19,0.752,0)
  pt3dadd(115.695,125.161,15.0518,0.649936,0)
}
connect section[1](0.0), section[0](1.0)
```

A list of sections can be linked as attributes in a "cell" class, enabling treating them in a unified (abstracted) manner. Schematics of object-based representation of neuromorphological data are illustrated in **Figure 6** (right).

2.2.4. Tabular structuring of morphological data

One of the prevalent platforms for morphological reconstruction is Neurolucida (http://www.mbfbioscience.com/neurolucida), which provides different data models for representation, including ASC, DAT, and XML, for which format specification is not publicly available. However, reversed engineered specification for Neurolucida's DAT data format (available

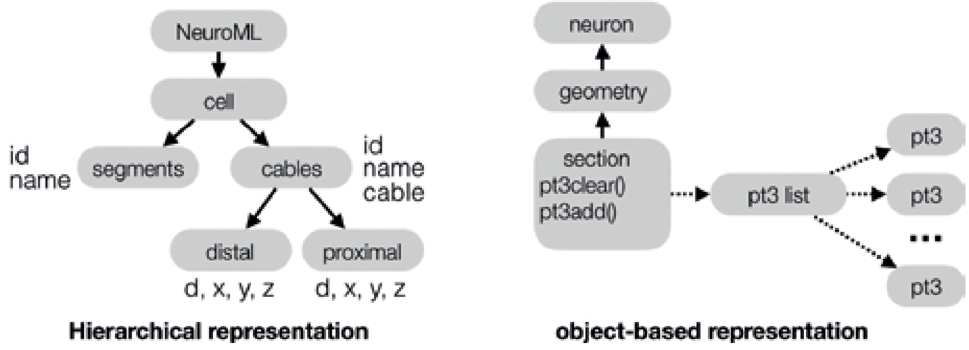


Figure 6. Hierarchical and object-based representation of neuromorphological data.

through: neuronland.org) reveals a hierarchy of data blocks, each identified by a Hexadecimal-encoded header (specifying the block type and size), followed by ASCII encoded data. For example, name and sample data are encoded using:

```
% header : size       : String
  0x0001 : 0x0000000A : 'Name'
% header : size       : x    : y     : z     : d    : section id
0x0101   : 0x00000018 : 2.15 : -3.25 : 18.55 : 0.54 : 0x0000
```

The type of block determines the data which follow the header including the Tree and Sub-Tree types to define the topology and connections of the samples. Data is therefore organized as a table.

Frameworks such as neuroconstruct [34] can import morphology files in all of the above formats and use them in conjunction with network specification and cellular mechanisms to generate script files for various simulation platforms, such as NEURON, GENESIS, and PyNN. While Neuromorpho.org adopted SWC and NEURON's data model as their data-sharing standard, the Human Brain Project adopted the Neurolucida data model as the format of choice.

2.3. Models of biophysical data

The establishment of the Hodgkin–Huxley-type compartments modeling and the development of experimental methods such as patch-clamp recording and imaging techniques are two complementary advancements which have transformed the field of neuroscience. Molecular aspects of neuroscience could be precisely measured and then used for computational modeling. Modeling neuronal behavior at the molecular level is a crucial aspect of modern neuroscience. Standardizing and modeling neurophysiological data, which often include mechanisms as a set of nonlinear equations, differential equations, or kinetic reaction schemes, are critical for utilization of computational models across simulators.

2.3.1. Object-based structuring of biophysical data

Over the years, NEURON has been extended to include a library of biophysical mechanisms, which were developed using its dedicated high-level programming language: NMODL (which was also adopted later by GENESIS). For example, a model for a leak current using the canonical electrical model of a current channel, with i (leak current), e (equilibrium potential), and g (conductance) can be defined using NMDOL with [35]:

```
NEURON {                    % interface
  SUFFIX leak               % density mechanism
  NONSPECIFIC_CURRENT I     % i in charge of the balance equations
  RANGE i, e, g             % are functions of position
}
PARAMETER {
  g = 0.001 (siemens/cm2) < 0, 1e9 >
  e = -65 (millivolt)
}
ASSIGNED {
  i (milliamp/cm2)
  v (millivolt)
}
BREAKPOINT {        % to be incrementally executed by the simulator
  i = g * (v - e)
}
```

In this modeling paradigm for physiological data, its type is encapsulated with a "template" class (following the object-based data structuring) and instantiate as objects where appropriate. For example, to instantiate a leakage current (with specific values for i and g) and attribute it to a NEURON's cable segment, one can write:

```
cable {
  nseg = 5
  insert leak
  g_leak = 0.002 % S/cm2
  e_leak = -70 % mV
}
print cable.i_leak(0.1) % show leak current density near 0 end of cable
```

Schematics of object-based representation of biophysical data are illustrated in **Figure 7** (right).

2.3.2. Hierarchical structuring of biophysical data

ChannelML is the second layer of NeuronML, enabling specifying biophysical data with XML. For example, specifying a Na + channel in ChannelML can be written as:

```
<channelml>
    <channel_type name="NaChannel" density="yes">
        <current_voltage_relation
          cond_law="ohmic" ion="na" default_erev="50" default_gmax="120">
            <gate name="m" instances="3">
                <closed_state id="m0"/>
                <open_state id="m"/>
                <transition name="h" from="m0" to="m" expr_form="exp_linear"
                           rate="1" scale="10" midpoint="-40"/>
                <transition name="beta" from="m" to="m0" expr_form="exponential"
                           rate="4" scale="-18" midpoint="-65"/>
            </gate>
            <gate name="h" instances="1">
                <closed_state id="h0"/>
                <open_state id="h"/>
                <transition name="alpha" from="h0" to="h" expr_form="exponential"
                           rate="0.07" scale="-20" midpoint="-65"/>
                <transition name="beta" from="h" to="h0" expr_form="sigmoid"
                           rate="1" scale="-10" midpoint="-35"/>
            </gate>
        </current_voltage_relation>
    </channel_type>
</channelml>
```

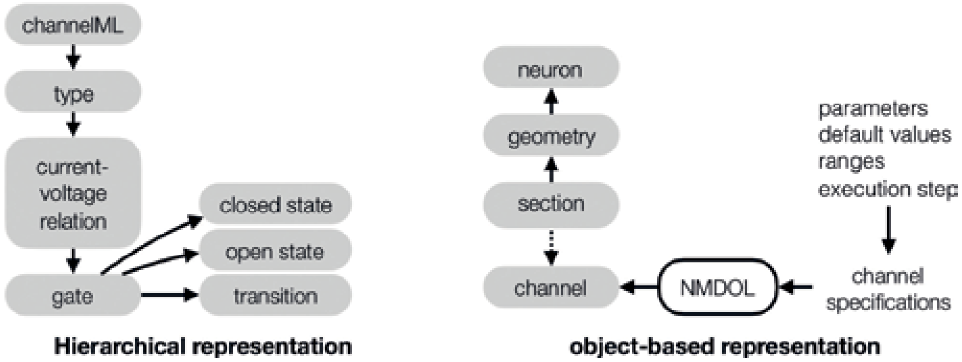


Figure 7. Hierarchical and object-based representation of biophysical data.

Neuroconstruct support both data models. Moreover, scripts for converting ChannelML specification to NEURON are also available. Schematics of hierarchy-based representation of biophysical data are illustrated in **Figure 7** (right).

2.3.3. Layer-oriented structuring of biophysical data

Another approach for physiological modeling is the layer-oriented approach (LOA) [36], in which the mathematical model (usually a set of differential equations) is governed by interlinked aspects of its structure. The LOA rationale is that biophysiological models such as the Hodgkin–Huxley model for ion channels have a hierarchical structure from which the underlying biological concept is reflected. Layer structure and relations are described in **Figure 8**.

By structuring mathematical behavior in a layered-structure manner, modules can be reused where different parameters are incorporated. One can utilize for example the same computational mechanism for membrane potential with either Hodgkin-Huxley model or GHK model or utilize the same gating dynamics for different dynamic models. Here, each layer is defined using a XML-like definition language (similarly to what was shown above), where connections between layers are defined separately in a meta-data file.

2.4. Models of network data

A model of a neural network must indicate at the very least the following specifications: connectivity scheme, as well as neuron and synapse models (typically by a set of differential equations, spike generation criteria, and refractory periods) [37].

2.4.1. Hierarchical structuring of network data

NetworkML is NeuroML's third specification level, which allows positioning neurons in 3D, as well as defining their connectivity pattern, and synaptic specifications to other neurons. It uses three core elements for network description: population (cells of a specific type),

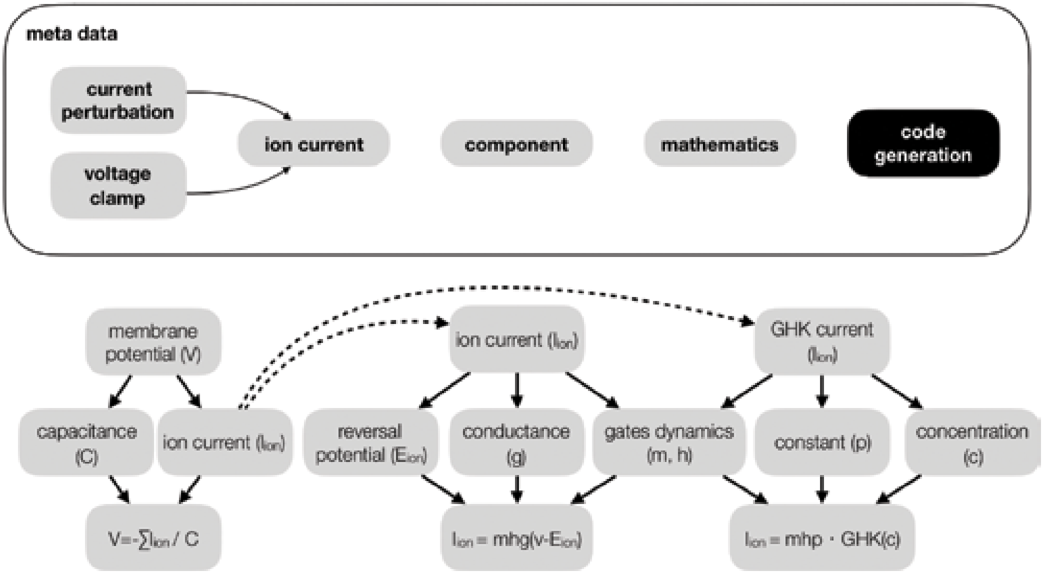


Figure 8. Layer-oriented representation of biophysical data.

projection (set of synaptic connections between populations), and input (describes an external electrical input into the network). Networks can be described with either instance-based (explicit list of positions and synaptic connections) or template-based (e.g., placing and connecting N cells randomly in a particular rectangular region) representation. For example, placing two populations of neuron PopA and PopB in 3D can be specified in NetworkML with [19]:

```
<populations>
  <population name="PopA" cell_type="CellA">
    <instances size="2">
       <instance id="0"> <location x="0" y="0" z="0"/> </instance>
       <instance id="1"> <location x="10" y="0" z="0"/> </instance>
    </instances>
  </population>
  <population name="PopB" cell_type="CellB">
    <instances size="3">
      <instance id="0"> <location x="0" y="100" z="0"/> </instance>
      <instance id="1"> <location x="10" y="100" z="0"/> </instance>
      <instance id="2"> <location x="20" y="100" z="0"/> </instance>
    </instances>
  </population>
</populations>
```

PopA and PopB can be connected with “projection”:

```
<projections units="Physiological Units">
  <projection name="NetworkConnection" source="PopA" target="PopB">
    <synapse_props synapse_type="DoubExpSynA" internal_delay="5" weight="1" threshold="-20"/>
      <connections>
        <connection id="0" pre_cell_id="0" pre_segment_id = "1"
        post_cell_id="1" post_segment_id = "0" post_fraction_along = "0.25"/>
        <connection id="1" pre_cell_id="1" pre_segment_id = "1"
        post_cell_id="0 post_segment_id = "0" post_fraction_along = "0.25"/>
        </connection>
      </connections>
  </projection>
</projections>
```

Schematics of hierarchy-based representation of network data are illustrated in **Figure 9** (left).

```
for i in range(N):                     % N cells
    src = cells[i]                     % select source cell
    tgt = cells[(i + 1) % N]           % select target cell
    syn = h.ExpSyn(tgt.dend(0.5))      % place a synapse in the middle of the target
    nc = h.NetCon(src.soma(0.5)._ref_v, syn, sec=src.soma)
% Connect source soma to target synapse
    nc.weight[0] = .05
    nc.delay = 5
```

2.4.2. Object-based structuring of network data

In NEURON, neurons can be interconnected to form networks using the object-based approach. For example, giving an array of "cell" objects (each encapsulates its defining sections, such as a soma and dendrites), they can be connected (e.g., circle topology) using Neuron's ExpSyn and NetCon object using (written in NEURON-Python):

Schematics of object-based representation of network data are illustrated in **Figure 9** (right).

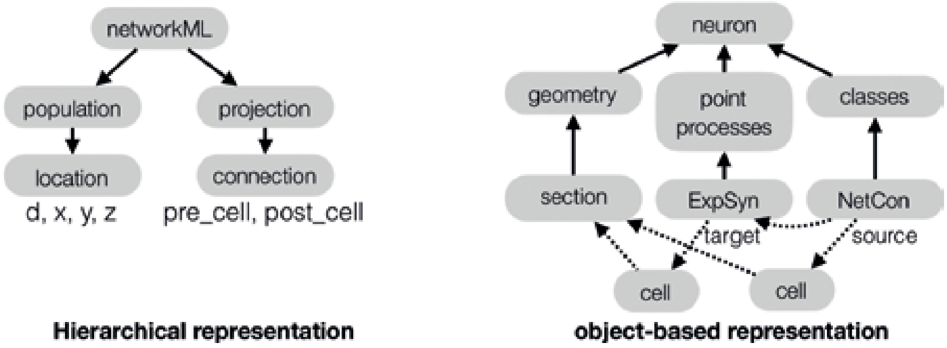


Figure 9. Hierarchical and object-based representation of network data.

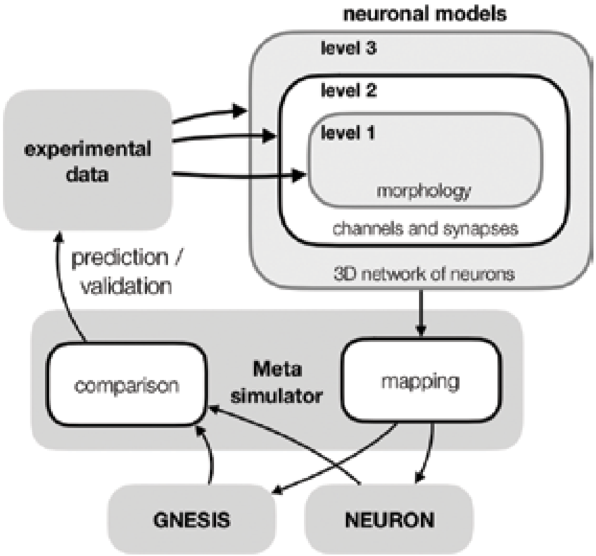


Figure 10. NeuroML 1 integrated approach to morphological, biophysical and network modeling.

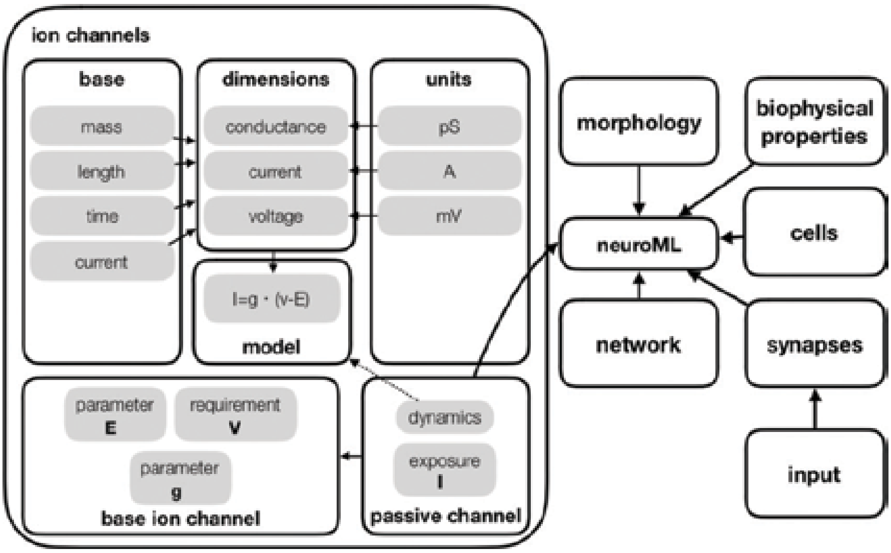


Figure 11. NeuroML 2 hybrid approach to morphological, biophysical and network modeling.

2.5. Integrated models

When it comes to integrated structuring of neuromorphic data, NeuroML is a prominent standard. It is defined using MorphML, ChannelML, and NetworkML, as they were described above. This integrated approach for neuroinformation standardization enables such models to be directly converted and mapped into different simulation frameworks. When integrating standard representation models with a "Meta Simulator" such as the NeuroConstruct or PyNN, a powerful framework is established. With such an approach, data can be distributed across multiple simulators, compared, and then validated with experimental data (**Figure 10**) [19].

In the second version of NeuroML, a new holistic approach is being developed for modeling, termed Low Entropy Model Specification (LEMS). LEMS is a hierarchical, XML-based language in which ion channels, synapses, neurons, and networks can be specified together. It combines a hybrid hierarchical object-based approach to modeling. An illustration is given in **Figure 11**. Detailed example is given in [38].

3. Rapid development of specialized neurocentric databases

In contrary to primary and secondary databases, specialized databases are mostly curated by individual laboratories or consortiums. They are characterized with a research-specific relational schema and specialized data types. Specialized databases are under constant development, aiming at supporting the rapid advancements in experimental techniques, which often produce vast amount of heterogeneous data. Most specialized databases are comprised of both new results and datasets–derived entries, constituting a hybrid approach of the new and the established. This stands as a major challenge to specialized data base designer, which have to support data querying, acquiring, and parsing from established data sources, as well as to integrate (or link) the results, with their own data model.

Specifically, the curation of specialized databases for neuroinformatics is an ever-growing challenge due to the need for organizing, structuring, and interconnecting vast amount of data, with standardized data structures. Here, an open-source framework for the curation of specialized databases is proposed. Our framework has the potential of realizing two complementary needs in the context of neuroinformatics: (1) structuring experimental data with standardized models which can be used for cross-simulations and (2) incorporating the experimental data and models with other data such as relevant diseases, articles, and biological models.

3.1. Framework

Databases often use a stable URL syntax, which renders a standard set of input parameters into the information needed to search and fetch the requested data. The proposed framework supports the generation of URL structured interface to local and remote data sets, including NCBI's databases, Malacards, and Biomodels. It was implemented with Java, extended to support objects' persistency with EclipseLink. I chose Apache Derby (part of the Apache DB Project) for data management. Derby is written in Java, and it is suitable for code embedding due to its small footprint and ease of use. Syntactic analysis was based on the w3c.dom open libraries, Apache Commons, J3D, and jsoup. The framework is described in length and exemplified for the curation of a database dedicated for aneurysms in [39].

In the context of neuroinformatics, the user can therefore take her morphology, biophysical, and connectivity experimental data, encapsulate them into interconnected classes (thus, creating a schema), and then link each of them to a structured data model (such as the ones described above). Each data model can be connected to articles, biological models, and diseases, which can be derived from existing databases and deposited in a specialized local database. Data can be retrieved later for further analysis. See schematics in **Figure 12**.

The proposed framework can be implemented with different packages and programming environments. For example, Java was utilized to map data entities to NCBI's PubChem schema and to provide functions to invoke NCBI eUtilities and PubChem web services [40]. Similarly, objects persistency can be attained with either Python, Java, or C++. Python's standard library for example supports a family of hash-based file formats and objects serialization. The Java Persistence API (JPA) was also implemented by various development groups, including Apache OpenJPA, Hibernate, and EclipseLink, offering metadata-based automatic creation of data models. Providers of database management frameworks are likewise varied and include Apache Derby and the cloud-based MongoDB.

3.2. Implementation

I have recently proposed a framework for the development of specialized databases [39]. In this framework, Java was chosen as the development environment, with which interfaces to online databases such as MalaCards (to retrieve disease information), Biomodels (to retrieve biological models), and NCBI's databases (to retrieve gene, taxonomy, protein, and articles data) were designed. By integrating these interfaces with EclipseLink (JPA provider), Apache Derby (database

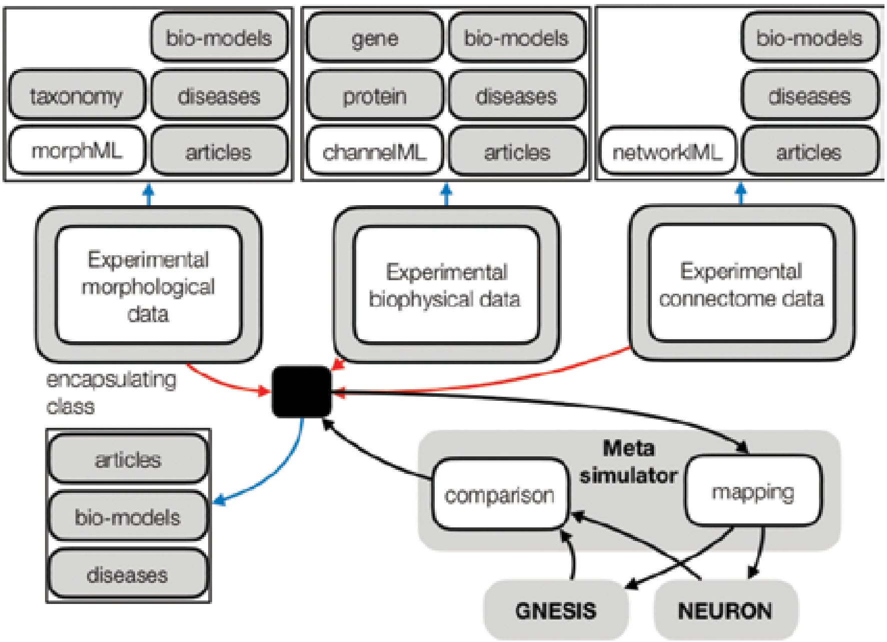


Figure 12. Database integrated approach to morphological, biophysical and network modeling.

manager), and a range of data parsers, a versatile framework for the curation of specialized databases is provided. This framework can be used to integrate new data and database-derived information into a user-defined data model. A schematic of the implementation is presented in **Figure 13**.

In the framework's main data flow, structured URL interfaces are used to establish connections between the user-defined data model to online data sets. Here, I used Entrez to interface with NCBI's data sets. NCBI's Entrez Programming Utilities provide a structured URL interface to their dozens of databases covering a variety of biomedical data, including gene and protein sequences, gene records, three-dimensional molecular structures, and biomedical literature [41].

Efforts to provide a similar utility for the neuroscientific community were also made. For example, Samwald and colleagues developed the "Neuron Entrez" [42], which integrates several neuroscientific ontologies: NeuronDB and ModelDB, subcellular anatomy ontology (SAO), and an OWL conversion of the cell centered database (CCDB). Once matured, this type of integrated neurocentric retrieval of data can greatly enhance frameworks, such as the one being proposed here.

A series of data processing tools were utilized to implement parsers for syntactic analysis of the retrieved data. The w3c.dom package provides the document object model (DOM) interfaces, which were used as the API for XML processing. This is essential for handling NeuroML structured data. The Apache Commons' libraries, the jsoup library, and the org.j3d library of the Java 3D Community were utilized for CSV, html, and STL parsing, respectively.

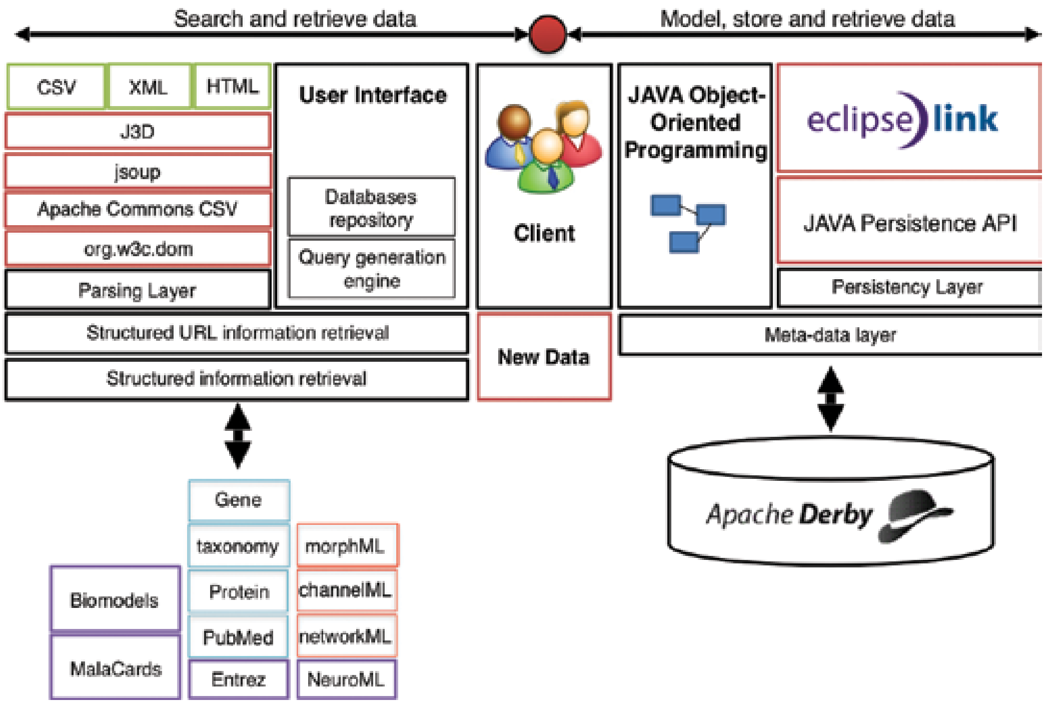


Figure 13. Realization of the database integrated approach to morphological, biophysical and network modeling.

The user utilizes Java object-oriented approach to encapsulate the retrieved data and to integrate it with her own data model. Object-relational mapping (converting Java objects to relational tables) is defined via persistence metadata. Metadata is defined via annotations embodied in the Java class and with an accompanying XML file. This allows EclipseLink to statically and dynamically query the database with SQL-like syntax. Apache Derby supports SQL data storing and querying in a client/server operation mode (commonly used database architecture). Suggested implementation for the above is provided via NBEL-lab.com and distributed under the creative common agreement.

4. Conclusions

Recent developments in Integrated Neuroscience (IN) are often characterized with efforts to up-scale data production and to provide frameworks from which new insights can emerge [43]. Since insights from integrated neuronal models often rely on the combination of experimental and computational approaches [44], simulations and modeling have a key role. Moreover, sharing neuroscientific data in the heterogeneous environment of IN drove the momentum for standardizing data models for neuronal morphologies, biophysical properties, and connectivity. Here, I propose a framework with which standardized models can be structured with experimental data, as well as with established data from existing databases. A combination of an integrated approach to neuroscience with the establishment of a federated framework for "collective wisdom" of neuroscientists and engineers might open a new dimension for data-driven neuroscience and fuel the celebration of the "era of the brain."

Author details

Elishai Ezra Tsur

Address all correspondence to: elishai85@gmail.com

Neuro-Biomorphic Engineering Lab, Jerusalem College of Technology, Israel

References

[1] Kashtan N, Alon U. Spontaneous evolution of modularity and network motifs. Proceedings of the National Academy of Sciences of the United States of America. 2005;**102**(39): 13773-13778

[2] Crick F, Koch C. A framework for consciousness. Nature Neuroscience. 2003;**6**(2):119-126

[3] Fox PT, Lancaster JL. Neuroscience on the net. Science. 1994;**266**(5187):994-997

[4] Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T. The human brain project: Creating a European research infrastructure to decode the human brain. Neuron. 2016; **92**(3):574-581

[5] Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from large-scale biology. Science. 2003;**300**(5617):286-290

[6] Insel TR, Landis SC, Collins FS. The NIH brain initiative. Science. 2013;**340**(6133):687-688

[7] Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, et al. The human connectome project: A data acquisition perspective. NeuroImage. 2012;**62**(4): 2222-2231

[8] Jabalpurwala I. Brain Canada: One brain one community. Neuron. 2016;**92**(3):601-606

[9] Poo MM, Du JL, Ip NY, Xiong ZQ, Xu B, Tan T. China brain project: Basic neuroscience, brain diseases, and brain-inspired computing. Neuron. 2016;**92**(3):591-596

[10] Heidorn PB. Shedding light on the dark data in the long tail of science. Library Trends. 2008;**57**(2):280-299

[11] Ferguson AR, Nielson JL, Cragin MH, Bandrowski AE, Martone ME. Big data from small data: Data-sharing in the 'long tail' of neuroscience. Nature Neuroscience. 2014;**17**(11): 1442-1447

[12] Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, Steyerberg EW, Maas AI. IMPACT database of traumatic brain injury: Design and description. Journal of Neurotrauma. 2007;**24**(2):239-250

[13] Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JDF, Maas AIR. Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics. PLoS Medicine. 2008;**5**(9):e165

[14] Callahan A, Anderson KD, Beattie MS, Bixby JL, Ferguson AR, Fouad K, Jakeman LB, Nielson JL, Popovich PG, Schwab JM, Lemmon VP. Developing a data sharing community for spinal cord injury research. Experimental Neurology. 2017;**295**:135-143

[15] Kötter R. Neuroscience databases: Tools for exploring brain structure–function relationships. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2001;**356**(1412):1111-1120

[16] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2016;**3**:160018

[17] Hucka M, Nickerson DP, Bader GD, Bergmann FT, Cooper J, Demir E, Garny A, et al. Promoting coordinated development of community-based information standards for modeling in biology: The COMBINE initiative. Frontiers in Bioengineering and Biotechnology. 2015;**3**

[18] Yann LF, Davison AP, Gleeson P, Imam FT, Kriener B, Larson SD, Ray S, Schwabe L, Hill S, Schutter ED. Computational neuroscience ontology: A new tool to provide semantic meaning to your models. BMC Neuroscience. 2012;**13**(1):P149

[19] Gleeson P, Crook S, Cannon RC, Hines ML, Billings GO, Farinella M, Morse TM, et al. NeuroML: A language for describing data driven models of neurons and networks with a high degree of biological detail. PLoS Computational Biology. 2010;**6**(6):e1000815

[20] Richmond P, Cope A, Gurney K, Allerton DJ. From model specification to simulation of biologically constrained networks of spiking neurons. Neuroinformatics. 2014;**12**(2):307-323

[21] Gardner D, Akil H, Ascoli GA, Bowden DM, Bug W, Donohue DE, Goldberg DH, et al. The neuroscience information framework: A data and knowledge environment for neuroscience. Neuroinformatics. 2008;**6**(3):149-160

[22] Marenco LN, Wang R, Bandrowski AE, Grethe JS, Shepherd GM, Miller PL. Extending the NIF DISCO framework to automate complex workflow: Coordinating the harvest and integration of data from diverse neuroscience information resources. Frontiers in Neuroinformatics. 2014;**8**

[23] Herz AV, Meier R, Nawrot MP, Schiegel W, Zito T. G-node: An integrated tool-sharing platform to support cellular and systems neurophysiology in the age of global neuroinformatics. Neural Networks. 2008;**21**(8):1070-1075

[24] Hines M. NEURON – A program for simulation of nerve equations. Neural Systems: Analysis and Modeling. 1993;**127**:136

[25] Wilson MA, Bhalla US, Uhley JD, Bower JM. GENESIS: A system for simulating neural networks. Advances in Neural Information Processing Systems. 1989:485-492

[26] Davison AP, Brüderle D, Eppler JM, Kremkow J, Muller E, Pecevski D, Perrinet L, Yger P. PyNN: A common interface for neuronal network simulators. Frontiers in Neuroinformatics. 2009;**2**:11

[27] Ascoli GA, Donohue DE, Halavi M. NeuroMorpho.Org: A central resource for neuronal morphologies. The Journal of Neuroscience. 2007;**27**(35):9247-9251

[28] Shepherd GM, Stepanyants A, Bureau I, Chklovskii D, Svoboda K. Geometric and functional organization of cortical circuits. Nature Neuroscience. 2005;**8**(6):782-790

[29] Marenco L, Nadkarni P, Skoufos E, Shepherd G, Miller P. Neuronal database integration: The Senselab EAV data model. Proceedings of the AMIA Symposium. 1999:102

[30] Hines ML, Morse T, Migliore M, Carnevale NT, Shepherd GM. ModelDB: A database to support computational neuroscience. Journal of Computational Neuroscience. 2004;**17**(1): 7-11

[31] McDougal RA, Morse TM, Carnevale T, Marenco L, Wang R, Migliore M, Miller PL, Shepherd G, Hines ML. Twenty years of ModelDB and beyond: Building essential modeling tools for the future of neuroscience. Journal of Computational Neuroscience. 2017;**42**(1):1-10

[32] Myatt DR, Hadlington T, Ascoli GA, Nasuto SJ. Neuromantic–from semi-manual to semi-automatic reconstruction of neuron morphology. Frontiers in Neuroinformatics. 2012;**6**

[33] Crook S, Gleeson P, Howell F, Svitak J, Silver RA. MorphML: Level 1 of the NeuroML standards for neuronal morphology data and model specification. Neuroinformatics. 2007;**5**(2):96-104

[34] Gleeson P, Steuber V, Silver RA. neuroConstruct: A tool for modeling networks of neurons in 3D space. Neuron. 2007;**54**(2):219-235

[35] Hines ML, Carnevale NT. Expanding NEURON's repertoire of mechanisms with NMODL. Neural Computation. 2000;**12**(5):995-1007

[36] Raikov I, Schutter ED. The layer-oriented approach to declarative languages for biological modeling. PLoS Computational Biology. 2012;**8**(5):e1002521

[37] Nordlie E, Gewaltig M-O, Plesser HE. Towards reproducible descriptions of neuronal network models. PLoS Computational Biology. 2009;**5**(8):e1000456

[38] Cannon RC, Gleeson P, Crook S, Ganapathy G, Marin B, Piasini E, Silver RA. LEMS: A language for expressing complex biological models in concise and hierarchical form and its use in underpinning NeuroML 2. Frontiers in Neuroinformatics. 2014;**8**

[39] Tsur EE. Rapid development of entity-based data models for bioinformatics with persistence object-oriented design and structured interfaces. BioData Mining. 2017;**10**(1):11

[40] Southern MR, Griffin PR. A Java API for working with PubChem datasets. Bioinformatics. 2011;**27**(5):741-742

[41] NCBI. "Entrez programming utilities help," 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK25501/

[42] Samwald M, Lim E, Masiar P, Marenco L, Chen H, Morse T, Mutalik P, Shepherd G, Miller P, Cheung K-H. Entrez neuron RDFa: A pragmatic semantic Web application for data integration in neuroscience research. Studies in Health Technology and Informatics. 2009;**150**:317-321

[43] Narasimhan K. Scaling up neuroscience. Nature Neuroscience. 2004;**7**:425

[44] Markram H. The blue brain project. Nature Reviews Neuroscience. 2006;**7**(2):153-160

Chapter 9

Developing Network-Based Systems Toxicology by Combining Transcriptomics Data with Literature Mining and Multiscale Quantitative Modeling

Alain Sewer, Marja Talikka, Florian Martin,
Julia Hoeng and Manuel C Peitsch

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75970

Abstract

We describe how the genome-wide transcriptional profiling can be used in network-based systems toxicology, an approach leveraging biological networks for assessing the health risks of exposure to chemical compounds. Driven by the technological advances changing the ways in which data are generated, systems toxicology has allowed traditional toxicity endpoints to be enhanced with far deeper levels of analysis. In combination, new experimental and computational methods have offered the potential for more effective, efficient, and reliable toxicological testing strategies. We illustrate these advances by the "network perturbation amplitude" methodology that quantifies the effects of exposure treatments on biological mechanisms represented by causal networks. We also describe recent developments in the assembly of high-quality causal biological networks using crowdsourcing and text-mining approaches. We further show how network-based approaches can be integrated into the multi-scale modeling framework of response to toxicological exposure. Finally, we combine biological knowledge assembly and multiscale modeling to report on the promising developments of the "quantitative adverse outcome pathway" concept, which spans multiple levels of biological organization, from molecules to population, and has direct relevance in the context of the "Toxicity Testing in the 21st century" vision of the US National Research Council.

Keywords: omics data, systems toxicology, biological networks, backward reasoning, literature mining, multiscale modeling, adverse outcome pathways

1. Introduction to network-based systems toxicology

The ongoing public debates on the impact on human health of glyphosate, bisphenol A, or electronic cigarettes have underlined the importance of performing reliable toxicological assessments [1–3]. In this context, regulatory authorities need to require evidence packages to assess the health risks associated with chemical compounds of uncertain safety risk contained in consumer products or present in the environment. In order to make the authorities' decisions persuasive to the public, it is critical to support them with objective evidence obtained using the latest scientific and technological advances. The US National Research Council's (NSR) "Toxicity Testing in the 21st Century: A Vision and a Strategy" manifesto, issued in 2007, was a noteworthy response to this critical need [4]. It fostered innovative, interdisciplinary approaches *(i)* to scale up the experimenting capacities by favoring in vitro screening to whole-animal testing, *(ii)* to deepen the interpretation of the experiments in terms of biological mechanisms by integrating the pathway-based approaches used in biomedical research, and *(iii)* to process the extensive data generated using adequate statistical and modeling tools to provide quantitative answers and informative predictions.

Developments in systems toxicology during the last 10 years have been driven largely by the goal of concretizing the NSR's vision. Simply stated, systems toxicology can be seen as the application of the systems biology mindset and approaches to toxicity testing. Thus, an essential feature of systems toxicology is the holistic perspective used in systems biology, in which a biological system is viewed as a complex assembly of interacting, often numerous parts rather than the simple union of individual elements, which corresponds to the reductionist standpoint [5, 6]. The first consequence of the holistic perspective is the fundamental role played by molecular omics profiling technologies, as they enable the simultaneous quantification of the abundances of all the (detectable) elements of a given class of biomolecules. The technology used most frequently is transcriptome profiling, which has become an almost routine operation thanks to its numerous advantages (technical, practical, and economical). In the current post-genomic era, its coverage exceeds 20,000 genes, and the resulting large data volume requires proper Bioinformatics processing to be exploited adequately. The second consequence of the holistic perspective is the introduction of a modeling approach for the interactions between the system parts to produce the system-level properties. In cases where transcriptome profiles are available, the modeling approach builds upon the relationships between genes to achieve a bottom-up description of the biological mechanisms taking place in cells, tissues, or organs. This inherent modeling aspect implies that systems toxicology positions itself at the final end of the "gene sets < pathways < networks" sequence, which results from the ordering of the transcriptomics interpretation approaches according to increasing structural complexity and informational richness [7]. In that sense, systems toxicology can be distinguished from toxicogenomics, for which the gene interaction modeling aspect is not an essential component. It is important to stress, however, that complete descriptions in terms of interacting genes are not (yet) available for all the system-level biological mechanisms. Inversely, not all genes measured by transcriptomics have been shown to be involved in system-level biological mechanisms.

In this chapter, we will focus on the developments of network-based systems toxicology [8], as networks have turned out to be the most suitable description framework for systems biology [5, 6]. In this case, the complex interaction map between the system parts accompanying the holistic view reduces to a (large) series of pairwise relationships encoded by edges of the networks, which connect two nodes representing the interacting system parts. Importantly, networks have been shown to constitute a suitable framework for not only representing but also understanding systems-level biological mechanisms [9]. Network-based approaches have been subsequently extended to achieve a novel understanding of disease effects in healthy systems [10–12] as well as to integratively examine the main and secondary effects of drugs (**Figure 1a**) [13, 14]. From the point of view of toxicological assessments, it is very reasonable to expect that network-based approaches would provide an appropriate framework for examining the system responses to test exposure in terms of perturbed biological mechanisms, in perfect alignment with the NSR's vision. As system-level biological mechanisms result from the interactions of multiple nodes, the network-based modeling framework enables elegant collection of the distributed effects of a test exposure on individual nodes into the perturbation of a single entity (**Figure 1a**) [8, 15].

In the remaining part of this section, several essential features of applying network-based systems toxicology are briefly explained. First, it is important to note the fundamental difference between systems biology and systems toxicology or, more broadly, between the inves-

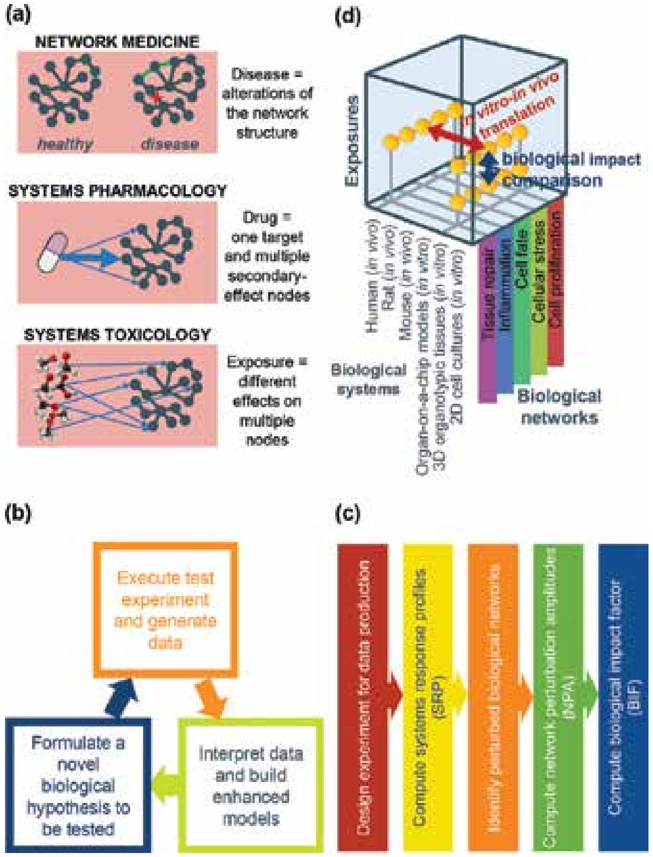

(a)
NETWORK MEDICINE
healthy
disease
Disease = alterations of the network structure
SYSTEMS PHARMACOLOGY
Drug = one target and multiple secondary-effect nodes
SYSTEMS TOXICOLOGY
Exposure = different effects on multiple nodes
(d)
Exposures
in vitro-in vivo translation
biological impact comparison
Human (in vivo)
Rat (in vivo)
Mouse (in vivo)
Organ-on-a-chip models (in vitro)
3D organotypic tissues (in vitro)
2D cell cultures (in vitro)
Biological systems
Tissue repair
Inflammation
Cell fate
Cellular stress
Cell proliferation
Biological networks
(b)
Execute test experiment and generate data
Interpret data and build enhanced models
Formulate a novel biological hypothesis to be tested
(c)
Design experiment for data production
Compute systems response profiles (SRP)
Identify perturbed biological networks
Compute network perturbation amplitudes (NPA)
Compute biological impact factor (BIF)

Figure 1. Features of network-based systems toxicology. **(a)** Schematic representation of network-based view of disease, drug, and exposure effects. **(b)** the iterative discovery cycle in systems biology [5]. **(c)** the linear five-step assessment workflow underlying network-based systems toxicology [8]. **(d)** the tridimensional representation "biological systems-exposure treatments-biological networks" illustrating the mechanism-based comparative assessment of exposure effects (blue arrow) and in vitro-in vivo or interspecies translatability (red arrow).

tigation of novel biological mechanisms and the biological mechanism-based assessment of exposure effects in a toxicity testing context. When investigating a biological system to discover novel mechanisms, the goal of the experimental data analysis and interpretation is to identify the most promising candidate mechanisms compatible with the observations, which would eventually lead to a novel, refined hypothesis to be tested. The implementation of this iterative process has been facilitated by systems biology, as the rich system-wide omics data allow for both confirmatory and exploratory investigations (**Figure 1b**) [5]. On the other hand, in the toxicity testing context, the biological mechanisms and their models must be determined a priori and remain "locked" when evaluating the test exposure experimental data in a systematic and least subjective manner. The outcome of the experiment is therefore the comparative assessment of the effects of the test exposure with varied parameters, such as the tested compounds, their doses, and the exposure durations. This can be represented by a linear assessment workflow (**Figure 1c**) [8], in contrast to the circular systems biology discovery cycle mentioned before. Interestingly, this difference between discovery and assessment approaches possesses an analogy in the context of transcriptomics gene set analysis: the competitive "Q1" statistic enabling the identification of the best associated gene sets corresponds to the discovery mode, whereas the self-contained "Q2" statistic quantifying the relevance of a given gene set corresponds to the assessment mode [16, 17].

Another advantageous aspect of applying network-based systems toxicology is the fact that it offers an explicit framework for "mechanistic translatability" between test systems. As the resemblances between exposure responses in human subjects and test systems (in vivo animal or, more recently, in vitro human) are fundamental in toxicity testing, the network-based approach enables establishing the validity of intersystem associations using the mappings of the biological mechanism-specific networks (red arrow on **Figure 1d**). This intersystem mechanistic translatability supports the use of in vitro test systems, such as cellular cultures, organotypic tissues, or organ-on-a-chip models, to reduce animal testing (typically rodents), in agreement with the NSR vision and the "3Rs" principles (i.e., "reduce the number of animals," "refine the experiments," and "replace the animals with nonanimal systems") [18–21]. The tridimensional representation "system-exposure-network" also contains the setup for performing a comparative, mechanism-based assessment of exposure responses, which obviously remains the primary goal of network-based systems toxicology (blue arrow on **Figure 1d**, which results from the completion of the workflow on **Figure 1c**) [8, 15, 22]. A biologically sound impact assessment between two considered exposures therefore consists of multicriteria comparisons between the mechanism-specific responses or "network perturbations", based on an appropriate selection of biological mechanisms.

The two concepts of network perturbation quantification and biological network selection are central to network-based systems toxicology and will be deepened further in this chapter. The methodology for calculating network perturbation amplitudes (NPAs) will be presented as a biologically driven complexity reduction scheme delivering valuable, structured information about the impact of toxicological exposure (Section 2). The related endeavor to ensure the quality of the biological networks will discussed afterward and illustrated by two innovative approaches based on crowdsourcing and literature mining (Section 3). The modeling perspective will be broadened beyond networks of interacting molecules to present other components of the multiscale modeling framework of an organism response to exposure (Section 4). Finally, emerging concepts from the quantification of adverse outcome pathways (qAOPs) will illustrate how extended multiscale modeling and biological knowledge assembly can combine to develop the predictive aspect of network-based systems toxicology. Throughout this chapter, our intention will not be to present a comprehensive review nor an abstract synthesis; rather we will coherently pick out concepts that are relevant for the past, current, and future developments of network-based systems toxicology as well as appealing in the context of "Bioinformatics in the Era of Post Genomics and Big Data."

2. Quantification of network perturbation amplitudes

In this section, we describe in more detail a core element of network-based systems toxicology: the quantification of NPAs, which amounts to calculating the exposure-induced response of biological mechanisms modeled by a network using transcriptomics data. As shown in **Figure 1c**, it represents a key ingredient of the five-step workflow for toxicity assessment and constitutes a concrete application of network-based systems toxicology [8, 15, 22]. Here we focus on the particular type of "causal networks" for which a mathematically and statistically sound methodology has been recently developed [23, 24]. Given a suitably organized collection of causal networks selected for a priori relevant biological mechanisms, the structure of the associated NPA results can be seen as a complexity reduction scheme starting from large experimental transcriptomics data. It provides a quantification of the exposure-induced impact on the considered biological mechanisms, which is used to comparatively assess toxicity in concrete applications. Additionally, it constitutes the starting point for the network-based systems toxicology developments that will be discussed later in this chapter.

Concretely, the implementation of the NPA methodology applicable to causal networks requires three distinct inputs in terms of experimental data and biological knowledge:

1. The differential gene expression values obtained from the transcriptomics data. Although we will consider them as resulting from "treatment versus control "pairwise comparisons, other types of contrasts can be used in the case of less trivial designs. These data are obtained by applying the suitable statistical models at the individual gene level and extend over the full transcriptome, in line with the first aspect of the holistic perspective of systems biology discussed above. We used to call them "systems response profiles" [8, 22, 25].
2. A suitably organized collection of causal networks covering the essential biological mechanisms of the test system response to the applied exposure treatment. Unlike other types of networks, causal biological networks contain nodes that not only describe molecular concentrations but also represent functions such transcriptional, enzymatic, or kinase activities. The network edges encode causal (i.e., directed) relationships between their nodes, which is attributed a positive sign when the activities of the connected nodes are changing similarly (e.g., increase in start node causes increase in end node) or a negative one when they change oppositely (e.g., increase in start node causes decrease in end node). The underlying biological knowledge in these networks has been manually extracted from the scientific literature and encoded in the biological expression language (BEL), an ontology developed specifically

for causal biological networks. The current version of the causal biological network collection is publicly available on the causal biological network (CBN) database website [26, 27]. The recent developments around the causal networks are discussed in Section 3.

3. "Transcriptional footprints" for a large fraction of the nodes contained in the causal networks. Transcriptional footprints are transcript abundance nodes that are connected to the causal network nodes via signed directed edges, similar to the ones in the causal networks. They follow the "backward reasoning" approach, in which changes in molecular mechanisms encoded by causal network nodes (e.g., the activity of a transcription factor) can be deduced from the expression changes of their downstream-regulated genes. Clearly, these edges allow the transcriptomics data to connect to the mechanistic networks, and the NPA calculations will consist of the experimental differential gene expressions "propagating through the networks" to obtain the corresponding node- and network-level perturbations. In our assessment applications, we licensed the Selventa Knowledgebase to get a good coverage of the nodes of the causal network collection in terms of transcriptional footprints [28]. Other options are possible: the small "BEL corpus" derived from the Selventa Knowledgebase [29], the networks contained in our publications [23, 24, 30], or the commercial IPA® "causal analysis" knowledgebase [31].

Given these three inputs, the NPA methodology performs the following computational steps to quantify the treatment-induced perturbations across a network (**Figure 2**):

1. Calculation of the "raw" perturbations for the nodes connected to the transcriptional footprints. Essentially, this consists of performing an edge-based, weighted average of the differential gene expressions attached to the transcriptional footprint nodes [23]. Optionally, this calculation can be applied to a complete "aggregated" network if it is "causally consistent" (or "balanced" in the graph-theoretic language). This property means that the edge-based relative sign between any two nodes must be unambiguous (i.e., must not depend of the specific path relating the two nodes). As most networks do not satisfy this condition (e.g., negative feedback loops are not causally consistent), the aggregation option would require additional processing to be operative [32].

2. Calculation of the perturbations for all network nodes based on a constraint optimization problem. This is obtained by searching for node values that are "smooth" over the network and the transcriptional footprint edges (i.e., that have the smallest edge sign-corrected differences between connected nodes) while matching the differential gene expression values for the transcriptional footprint nodes. This problem has an exact solution that can be expressed in terms of the inverse of the adapted, signed Laplacian matrix of the network graph and of the "raw" node perturbations obtained previously.

3. Calculation of the NPAs using an edge-based summation. The summed values are the squared edge sign-corrected mean of the corresponding node (smoothed) perturbation values. As this value is always positive, it is important to examine the node-level perturbation values to determine whether the underlying biological mechanism is activated or inhibited as a consequence of the exposure treatment.

4. Calculation of three accompanying statistics to decide whether the obtained NPA value represents a true or a false positive. The first statistic is based on the biological variability propagated from the uncertainties of the differential gene expression values: the 95% confidence

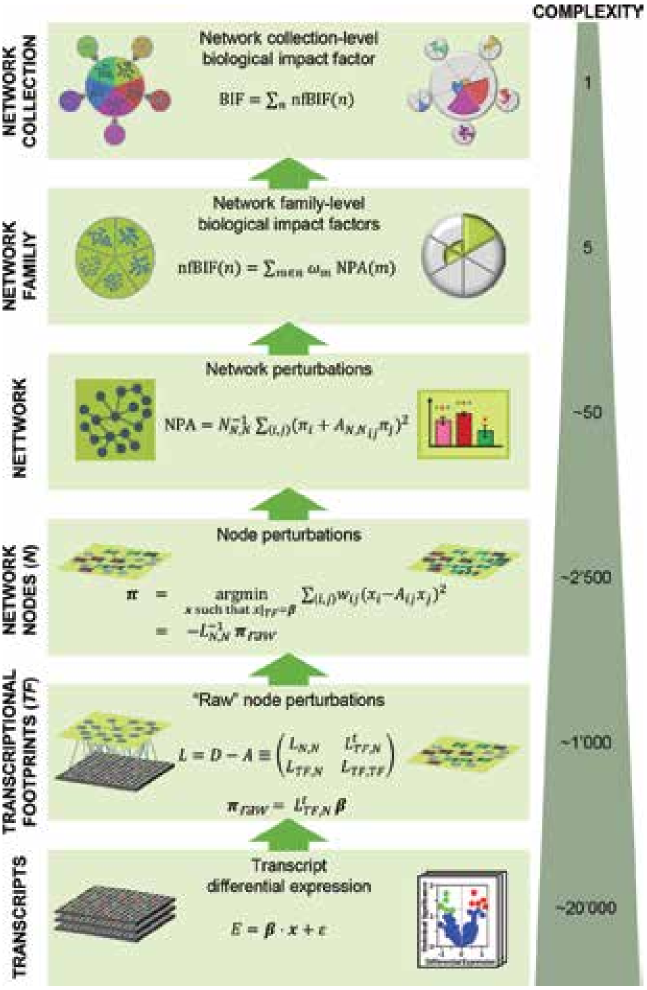

COMPLEXITY
NETWORK COLLECTION
Network collection-level biological impact factor
$\mathrm{BIF} = \sum_n \mathrm{nfBIF}(n)$
1
NETWORK FAMILIY
Network family-level biological impact factors
$\mathrm{nfBIF}(n) = \sum_{m \in n} \omega_m \, \mathrm{NPA}(m)$
5
NETTWORK
Network perturbations
$\mathrm{NPA} = N_{N,N}^{-1} \sum_{(i,j)} (\pi_i + A_{N,N_{ij}} \pi_j)^2$
~50
NETWORK NODES (N)
Node perturbations
$\pi = \underset{x \text{ such that } x|_{TF}=\beta}{\operatorname{argmin}} \sum_{(i,j)} w_{ij} (x_i - A_{ij} x_j)^2 = -L_{N,N}^{-1} \pi_{raw}$
~2'500
TRANSCRIPTIONAL FOOTPRINTS (TF)
"Raw" node perturbations
$L = D - A \equiv \begin{pmatrix} L_{N,N} & L_{TF,N}^{t} \\ L_{TF,N} & L_{TF,TF} \end{pmatrix}$
$\pi_{raw} = L_{TF,N}^{t} \beta$
~1'000
TRANSCRIPTS
Transcript differential expression
$E = \beta \cdot x + \varepsilon$
~20'000

Figure 2. The calculations of the network perturbation amplitudes (NPAs) and biological impact factor (BIF) in a bottom-up representation. The six layers correspond to the six steps **(1–6)** explained in the main text. Their respective inputs, mathematical processing, and results are schematically displayed from left to right. The "complexity" column gives an order of magnitude of the corresponding data size and illustrates the associated complexity reduction scheme.

interval around the NPA value should not contain zero. The other two statistics test the relevance of the biological mechanism(s) encoded in the network by randomly reshuffling the network edges or the transcriptional footprints. This yields two null distributions for the network-level perturbation values. If the actual NPA value lies above the 95% quantile of a null distribution, it is considered to be statistically significant and labeled as "K-specific" or "O-specific," respectively. Significant network perturbations correspond to the cases where all three statistical tests are successful.

By extending the calculation of NPAs to the full network collection contained in the CBN database, we take advantage of its hierarchical structure to complete a useful, pyramidal, bottom-up complexity reduction scheme (**Figure 2**). The grouping of networks into network families, themselves constituting the overarching collection, allows quantification and displays the exposure-induced biological impact in a more concise way, which is particularly useful in a comparative approach to toxicity assessment. These final two steps are the following:

5. Calculation of network family-level biological impact factors (BIF). The network families distribute the ~50 networks into five families based on their biological similarities: cell proliferation, cellular stress, cell fate, pulmonary inflammation, and tissue repair/angiogenesis. The evaluation of their BIF consists first of filtering out the networks that are not significantly perturbed and then summing the remaining NPA values with weights that take into account the number of network in each family and the nodes overlapping between networks.
6. Calculation of network collection-level BIF. This aims at providing balanced relative weights between the five network families so the main features of the biological systems response to the exposure treatment can be perceived easily. In that sense, the BIF represents a pan-mechanistic, quantitative metric for the exposure-induced effects measured at the molecular transcript level and "shaped" by the *a priori* chosen network collection from the CBN database [15, 30]. It represents the starting point for investigating the impacted biological mechanisms in a top-down approach.

Having presenting the NPA methodology, it is instructive to see how it compares to existing approaches providing network-level quantification. The causal biological networks used in the NPA calculations are usually composed of several molecular signaling pathways assembled around common nodes. Generally, pathways have a simpler and somewhat more linear structure than networks, so their structure is not as important. As a consequence, it has been often disregarded in the published methodologies, which were primarily aimed at dealing with pathways. In a recent review recapitulating the network- and pathway-based methodologies developed over the last decade, only one category (out of three) takes into account the structure: the so-called pathway topology (PT) group [7]. We further observe a recurrent difference between the NPA and most PT methodologies: the goal of the quantification is the determination of the most relevant pathways or networks (compared to the other ones in the collection) to support the biological interpretation. This is achieved by sorting either abstract scores or enrichment p-values [33]. This approach corresponds to the abovementioned competitive Q1 statistic, which suits the discovery rather than the assessment perspective, corresponding to systems toxicology [16, 17]. This also indicates that the NPA approach is closer to the self-contained Q2 statistic in the sense that it allows meaningful comparison of several treatments. In short, the NPA methodology provides an explicitly network-based quantification scheme that inherently incorporates the self-contained Q2 statistic, allowing meaningful comparisons between the exposure effects on the same biological mechanism.

The NPA approach has been successfully employed across a range of toxicological questions of concern:

- Comparative assessment of biologically active substances to complement the standard toxicological endpoints. This was used for the preclinical assessment of a candidate modified-risk tobacco product in comparison with conventional cigarettes [34–36].
- In vitro screening of multiple compounds in combination with the capacity of the high-content screening technologies. This was applied to selections of environmental toxicants and nutraceuticals [37, 38].

- Investigation of in vivo-in vitro translatability (red arrows in **Figure 1c**). The case of the xenobiotic metabolism response to cigarette smoke exposure was investigated and supported the validity of in vitro testing [39].
- Classification of individual human subjects. A proof-of-principle application of the NPA methodology to individual subjects has been published [24], and the approach was benchmarked during the sbvIMPROVER diagnostic signature challenge [40].
- Exploratory investigations of transcriptomics data. Examining the biological process activities contained in the collection of causal networks provides an additional point of view already used several times [41–43].

In this section, we explained the NPA methodology as a core element of network-based systems toxicology. However, its validity also depends on the quality of input from causal network collection available in the CBN database. In the following section, we discuss several innovative ways to ensure constant quality in order to consolidate the acceptance of the network-based systems toxicology.

3. Enhancements of the causal network collection

The application of network-based systems toxicology requires the *a priori* identification of the biological mechanisms involved in the test systems response to the applied exposure (**Figure 1c** and **d**). This led to gradually assemble a structured collection of causal networks of high-quality standards, which has been deposited in the CBN database to be accessible to run the NPA calculations in concrete situations. The validity of the whole network-based systems toxicology approach depends heavily on the biological pertinence of the retained mechanisms and of the networks encoding them. In this section, we examine these validity conditions more closely and describe two recent efforts aimed at augmenting the biological pertinence and extending the biological contexts of the causal networks: a crowdsourced review of their content and the use of semi-automated text-mining tools.

Over last two decades, the ever-increasing use of transcriptomics technologies has resulted in compilations of a number of pathway resources aimed at associating biological insight to sets of differentially expressed genes: KEGG [44], Reactome [45], BioCarta [46], Wiki-pathways [47], SPIKE [48], UCSD signaling gateway [49], NCI pathway interaction database [50], or NetPath [51]. The parallel assembly of the CBN database was decided and justified by the requirement to satisfy higher-quality standards, which were not always met by the available pathway resources (Table 1 in [27]):

1. Explicitly accounting for the biological context by setting mechanistic boundaries in terms of species, tissue or cell type, and disease state
2. Supporting all the causal relationships encoded in the network edges by (at least) one explicit, literature-based statement
3. The use of BEL to encode the manually curated literature statements into a format that is both human-readable and computable and that stores the rich mechanistic and contextual information accurately

4. Application of data-driven enhancement by analyzing suitable public or dedicated datasets using a complementing source of prior biological knowledge, such as the Selventa Knowledgebase, which contains more than two million curated relationships [28]

Note that the last feature is also relevant for augmenting the ensemble of transcriptional footprint edges, which were also extracted from the Selventa Knowledgebase in our assessment applications (see Section 2). Typically, the public dataset GSE44747 investigates the gene expression regulation by the activation of protein kinase C ("PKC"), and the molecular changes in this datasets can be causally related to the node *act(p(SFAM:"PRKC Family"))* [52]. Whenever a sizable fraction of the genes regulated in this dataset are changed in response to an exposure treatment, the activation or inhibition of PKC can be inferred [28]. This example illustrates the transcriptional footprint-based "backward reasoning" necessary to connect the causal biological networks and the transcriptomics data in order to apply the NPA methodology.

In 2011, we published our first biological networks "Cell Proliferation" that are still part of the collection that serves as the input for NPA and BIF calculations [53]. The initial mechanistic interest focused on the lung biology, and version 1.0 of the collection consisted of 108 assembled causal networks regrouped into five high-level functional families (cell proliferation 15 [53], cellular stress 7 [54], cell fate 34 [55], pulmonary inflammation 24 [56], and tissue repair/angiogenesis 9 [57]). The design and assembly processes were the same for all the networks, each of them having been defined by biological boundaries chosen to globally cover all of the essential biological processes and responses of healthy lung tissues (**Figure 3**). Since 2015, the CBN database website has provided free access to the full collection [27]. In addition to the original focus of inhalation toxicology covering the non-diseased respiratory tract tissues, causal networks for non-diseased vascular tissues, chronic obstructive pulmonary disease, and atherosclerosis plaque destabilization have been assembled and published to enrich the covered biological contexts [58–60].

As mentioned above, the scientific acceptability was the main requirement during the assembly of the causal networks collection, which is freely available to the scientific community in the CBN database. This motivated additional and innovative crowdsourced verification initiatives to consolidate the accuracy of the biological mechanisms encoded in the networks. They took place in the framework of the network verification challenges of the systems biology verification initiative (sbvIMPROVER NVC) [59, 61–63]. These challenges were based on a novel crowdsourcing approach by a large community of more than 50 contributors who were given tools to vote on various edges and nodes of the causal networks via a dedicated web interface [64]. A moderator supervised the votes for each network and made decisions to include or exclude nodes and edges based on community choices. The resulting 46 causal networks were made publicly available through the CBN website and constituted version 2.0 of the causal network collection organized along the same five high-level functional families as version 1.0 (cell proliferation 15, cellular stress 7, cell fate 34, pulmonary inflammation 26, and tissue repair/angiogenesis 11). Currently, the NVC platform supports a third crowdsourced network verification challenge for the liver xenobiotic metabolism [64]. Eventually, the new models will be shared via the CBN website [26].

As the original network assembly process involved significant efforts in manual literature curation (**Figure 3**), the development of text-mining-based capabilities appeared as an appropriate solution to increase the quantity of assembled causal networks while preserving their quality. A novel, semi-automated biological knowledge extraction workflow called the "BEL information extraction workflow" (BELIEF) was developed, which incorporates

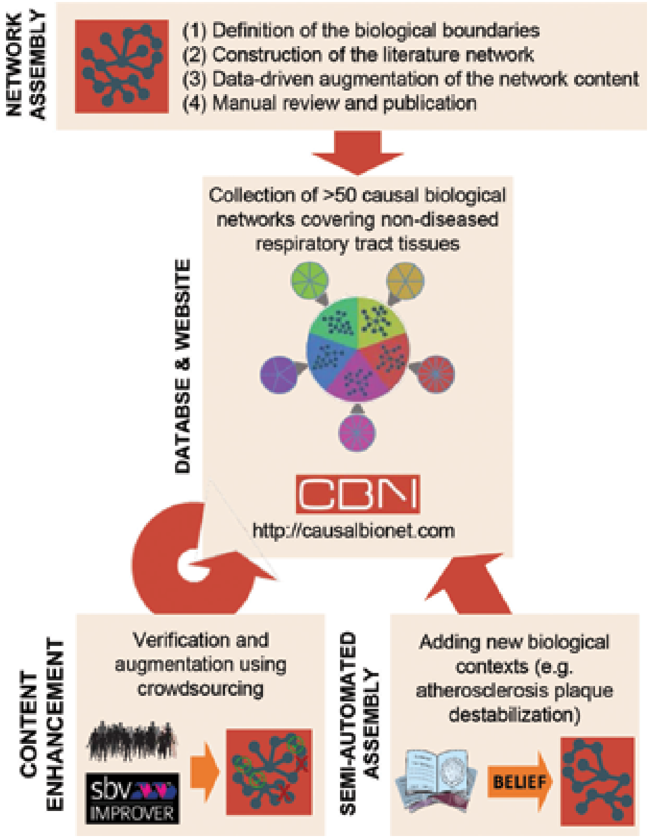

NETWORK ASSEMBLY
(1) Definition of the biological boundaries
(2) Construction of the literature network
(3) Data-driven augmentation of the network content
(4) Manual review and publication
DATABSE & WEBSITE
Collection of >50 causal biological networks covering non-diseased respiratory tract tissues
CBN
http://causalbionet.com
CONTENT ENHANCEMENT
Verification and augmentation using crowdsourcing
sbv IMPROVER
SEMI-AUTOMATED ASSEMBLY
Adding new biological contexts (e.g. atherosclerosis plaque destabilization)
BELIEF

Figure 3. Overview of the causal biological network assembly and enhancements. The CBN database website contains the initial hierarchically structured collection of biological networks describing the essential biological processes and responses of healthy lung tissues. The enhanced network versions resulting from the sbvIMPROVER network verification challenges are integrated in CBN, as well as the networks describing relevant response mechanisms in other biological contexts, which were obtained by the BELIEF semi-automated literature mining workflow.

state-of-the-art linguistic tools for recognition of specific entities [65, 66]. It mines preselected, unstructured scientific literature and enables its users to extract causal and correlative relationships that are subsequently transcribed into the computable and human-readable BEL format used in the CBN network collection. A web interface has been developed, as well, to facilitate its practical application [67]. The usefulness of the BELIEF workflow was assessed during the assembly of a network describing atherosclerotic plaque destabilization and containing 304 nodes and 743 edges supported by 33 PubMed literature references [65]. The comparison between the semi-automated and conventional curation processes showed similar results but with significantly reduced curation effort for the semi-automated process. It is currently applied to a variety of biological mechanisms extending beyond the initial focus of pulmonary biology (e.g., vascular tissues).

The high quality of CBN causal network collection provides a solid foundation for the network-based systems toxicology approach. Supplementing its essentially manual assembly process, innovative crowdsourced verification initiatives have consolidated and updated the biological content of the networks. The development of the semi-automated BELIEF workflow has been beneficial not only directly, by speeding up the maintenance of the CBN collection, but also indirectly, by popularizing the use of causal networks in biomedical contexts beyond toxicological assessment [27].

4. Integration into the multiscale modeling of exposure responses

In the previous section, we saw that the enhancements of the causal network collection were opening new development opportunities for the approaches used in network-based systems toxicology. This leads to similar reconsideration of the molecular holistic approach underlying the systems biology approach from a broader perspective—that of modeling an organism response to exposure in the toxicological context. Indeed, the organism response to exposure is a complex process, covering multiple space and time scales, for which modeling approaches of diverse complexities have been used. In this section, we discuss how the causal networks used in our holistic systems biology approach can be integrated into the quantitative toxicology/pharmacology frameworks of absorption, distribution, metabolism, and excretion toxicity (ADMET) and physiologically based toxicokinetics (PBTK)/physiologically based pharmacokinetics (PBPK) modeling. This will not only reveal the approximations and limitations of the respective approaches but also eventually indicate where bridges between causal molecular networks and other modeling approaches can be built and which efforts would be required to achieve them. Paving the road for multiscale approaches constitutes a promising development perspective for improving the understanding of how potentially toxic substances interact with the human body.

ADMET belongs to the basic principles of pharmacology and toxicology and describes the kinetics, dynamics, and toxicity of compounds within the human body following an exposure. The objective of such an approach is to estimate the toxicokinetic and metabolic profiles (**Figure 4a**). Obviously, a molecular dynamics approach resolving the trajectories of individual molecules from absorption to excretion is not achievable because of our insufficient understanding of the interplay between the numerous molecular mechanisms involved and, from a practical perspective, limited computational power. As a consequence, the replacement of the individual molecular trajectories by the corresponding mean density distributions and velocity fields—the so-called continuum approximation—appeared to be the most suitable approach to perform quantitative modeling in the toxicokinetic context. In the specific case of inhalation toxicology, the inclusion of additional assumptions about the interplay between liquid, vapor, and aerosol phases lead to a well-defined computational fluid dynamics (CFD) scheme, which quantitatively describes the deposition of aerosol particles in the nasal and other respiratory cavities by calculating the airflows and velocities [68]. Therefore, a fine description of the dose reaching respiratory tissues (**Figure 4b**) can be achieved through CFD partial differential equation systems in space and time variables.

The description of the dynamics of each molecule when it reaches a cell can be done, for example, using a stochastic description of enzymatic activities through the chemical master equation (**Figure 4b**). While appealing on a local level, those approaches are not straightforward in global application to a whole-body model. To that end, simplifying the complex human body into a limited number of connected compartments underlying PBTK/PBPK modeling is usually explored for evaluating levels of a given substance in various tissues or organs (**Figure 4c**). PBTK/PBPK can also be linked to deposition CFD models, as discussed by some authors [69–71]. Metabolism is

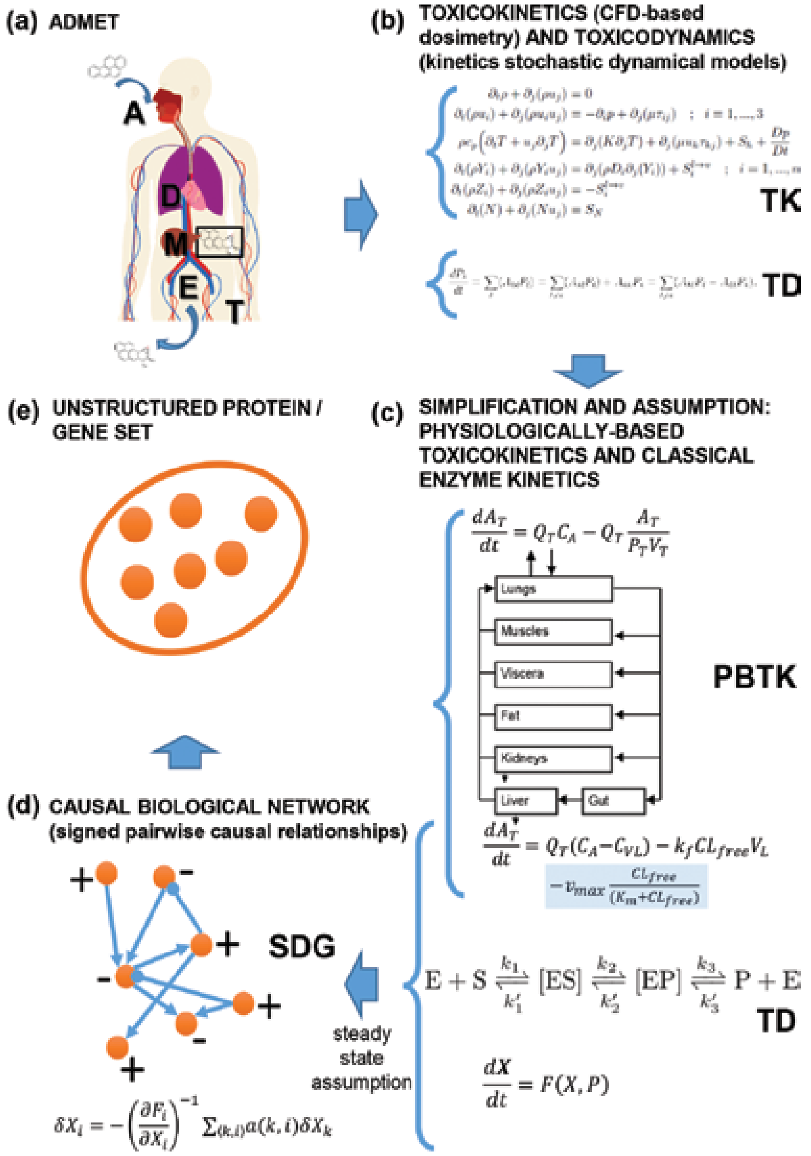

(a) ADMET
A
D
M
E
T
(b) TOXICOKINETICS (CFD-based dosimetry) AND TOXICODYNAMICS (kinetics stochastic dynamical models)
TK
TD
(c) SIMPLIFICATION AND ASSUMPTION: PHYSIOLOGICALLY-BASED TOXICOKINETICS AND CLASSICAL ENZYME KINETICS
$\frac{dA_T}{dt} = Q_T C_A - Q_T \frac{A_T}{P_T V_T}$
Lungs
Muscles
Viscera
Fat
Kidneys
Liver
Gut
$\frac{dA_T}{dt} = Q_T(C_A - C_{VL}) - k_f CL_{free} V_L - v_{max} \frac{CL_{free}}{(K_m + CL_{free})}$
PBTK
$E + S \underset{k'_1}{\overset{k_1}{\rightleftharpoons}} [ES] \underset{k'_2}{\overset{k_2}{\rightleftharpoons}} [EP] \underset{k'_3}{\overset{k_3}{\rightleftharpoons}} P + E$
TD
$\frac{dX}{dt} = F(X, P)$
steady state assumption
(d) CAUSAL BIOLOGICAL NETWORK (signed pairwise causal relationships)
SDG
+
-
+
-
+
-
+
$\delta X_i = -\left(\frac{\partial F_i}{\partial X_i}\right)^{-1} \Sigma_{(k,i)} a(k,i) \delta X_k$
(e) UNSTRUCTURED PROTEIN / GENE SET

Figure 4. The multiscale modeling framework of the human body response to toxicological exposure. The sequence from panels (a) to (e) spans several space and time scales, for which multiple modeling approaches are used. In order to make them applicable, simplifying assumptions are necessary at each step (blue arrows), and the resulting model parameters must be determined experimentally. From this perspective, the signed directed graphs (SDGs) underlying the network-based system toxicology approach can be integrated into a broader multiscale response modeling framework.

further simplified by assuming a well-stirred volume or that conversion of an enzyme-substrate complex to an enzyme-product complex is instantaneous. Such a description of the enzymatic and metabolic dynamics reduces them to a set of ordinary differential equations (ODE) in time.

In general, PBTK/PBPK-derived ODE systems involve many parameters that are not necessarily accessible to researchers, and an analytical study of the system may be required to estimate them. For that purpose, assuming steady state, ODEs can be represented semiquantitatively by a signed directed graph (SDG) derived from the Jacobian matrix of the ODE system evaluated at its steady-state solution (**Figure 4d**) [72]. In this context, the organismal response to an exposure is viewed as a perturbation of its steady state, which is characterized by the associated SDG encoding the time directionality and relative signs of the perturbations between all pairs of connected nodes. Although such an SDG is derived in the PBTK/PBPK context, in principle, other SDGs can be obtained similarly [73]. This is accomplished when a higher resolution of the description of molecular mechanisms involved in the response can be obtained from the scientific literature. This is exactly the case for the biological processes contained in the causal networks presented in the previous section, as we know how they integrate into the broad quantitative toxicology/pharmacology modeling frameworks built around ADMET and aimed at describing the organismal response to exposure in its full complexity.

In this short excursion aimed at broadening the modeling scope beyond molecular systems biology, we saw several approaches to quantify the response to exposure. Their validity ranges covered specific space and time scales, while a higher complexity often demanded more and more parameters to be experimentally determined to make the model applicable. As a consequence, building bridges between modeling scales represents appealing development directions to achieve a more integrated understanding of an organism response to exposure. However, the effort required to preserve the applicability of the resulting models are substantial, and in the last section, we will examine a tentative, multiscale approach that is acquiring an increasing interest the context of modern (twenty-first century) toxicology.

5. Development of quantitative adverse outcome pathways

In the previous sections, we have described NPA as a core element of network-based systems toxicology. We then saw two extensions: new networks contexts and extended modeling framework. In this final section, we propose a combination of these elements in terms of a network-based approach to qAOPs. This direction offers a novel development opportunity that needs to incorporate the predictive aspect at population level, which is to be contrasted to the a posteriori approach of test system data-driven assessment discussed up to now (**Figure 1c**). This feature requires an adapted approach to select the relevant biological mechanisms as well as the development of quantitative, multiscale modeling approaches of the suitable complexity.

Starting from ecotoxicology and quickly gaining popularity in human toxicology, adverse outcome pathways (AOP) have become a valuable means to model exposure effects. Similar to the network models, AOPs organize existing, scattered literature knowledge into a structured representation with the aim to construct a linear sequence of "key events" (KE) from the initial interaction between a chemical and the biological system—the molecular initiating event (MIE)—to the individual and population-level adverse outcome [74] (**Figure 5**). We have contributed to the development of two AOPs for the common disorders resulting from long-term smoking and published them in the AOP wiki [75]. The first AOP maps the events from epidermal growth factor receptor activation by oxidative stress to decreased lung function [76], and the second AOP illustrates the different steps that are required for oxidative stress to lead to disruption in endothelial nitric oxide bioavailability and, finally, to hypertension [77]. These AOPs were built following the requirements by the Organization for Economic Co-operation and Development (OECD) [78].

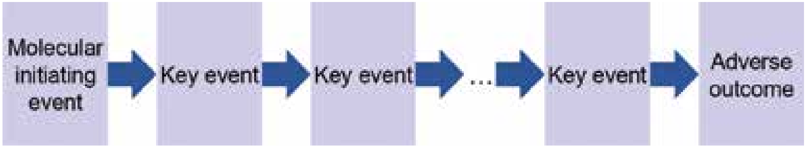


Figure 5. The structure of an adverse outcome pathway (AOP).

One avenue to network-based systems toxicology is to build BEL models that represent these events. The first BEL model suite is underway and describes the biological processes involved in impaired mucociliary clearance. It is foreseen to be published under an SBVimprover NVC in 2018 [64].

While the above effort aims at identifying the mechanistic biological knowledge underlying the chosen AOPs, the parallel development of the associated quantitative modeling approaches needs to be moved forward. It was anticipated to follow three steps to yield a "dynamic adverse outcome pathway" [79]:

1. Assembly and quantification of causal mechanistic networks
2. Development of dynamic models linking exposure to the organ-level responses
3. Simulation of the population-level effects of an exposure

The importance of this endeavor was underlined by the fact that the achievement of Step [3] was explicitly promoted as "the ultimate goal of systems toxicology." As Step [1] has been completed with the CBN database and the development of the NPA methodology, the attention now focuses on Steps [2] and [3], which have to incorporate the predictive capacity of the future qAOP. Typical useful resources in this context are the BioModels database containing hundreds of computational models of biological processes (Step [2], [80]) as well as the "mechanistic axes population ensemble linkage" algorithm, which enables the creation of large sets of mechanistically distinct virtual humans that, upon simulated exposure, statistically match the prevalence of phenotypic variability reported in human population sample studies (Step [3], [81]).

Given the network-based system toxicology components presented in this chapter, several directions could be considered to support the qAOP development. Typically, appropriate transcriptomics datasets could be identified and used for applying NPA quantification to causal networks representing the biological mechanisms underlying one or more KE and their relationships. Although the time dependence is not explicit in the SDGs associated to the networks, their causal characteristic can provide information about the time direction based on the sequence of causally related perturbations. As during the assembly of the CBN network collection, the use of transcriptomics data is expected to improve the accuracy of the networks. In the qAOP context; the usual "treatment vs. control" experimental design might be advantageously replaced by a time course design, which can reveal (part of) the time evolution of the relevant perturbed mechanisms. We may also consider the possibility of calculating NPA at individual level, which, as a consequence of its complexity reduction property, yields better between-class separations in classification contexts [24]. This could be used to more accurately model the population-level distributions of the exposure-induced perturbations.

To conclude on a more concrete note, we show the "real-life" example of a simple qAOP developed for risk assessment in ecotoxicology: the connection between the inhibition of cytochrome P450 19A aromatase (the MIE) and the population level decreases of the fish fathead minnow (the adverse outcome) [82]. Concretely, the easily collected measures of chemical inhibition of the rate-limiting steroidogenic aromatase enzyme are used to predict reductions in egg production and, subsequently, population size of the fish. The quantitative modeling of the associated sequence of events was achieved by linking three discrete models describing different components of the AOP, from the MIE (aromatase inhibition) through five intermediate KEs, to impacts of regulatory interest (fecundity, population size). While the qAOP was developed based on experiments with fish exposed to the aromatase inhibitor fadrozole, a toxic equivalence calculation allowed to predict the effects of another untested aromatase inhibitor, iprodione.

This example showed that as long as their main elements are well chosen, qAOPs do not need to be "complicated," as it would have been expected from a pathway covering multiple levels of biological organization (i.e., from molecules to population levels). This observation effectively illustrates the trade-off that needs to be found during qAOP development between biological accuracy, modeling complexity, and practical value in terms of predictive capacity. All three aspects are equally important for the validity of the outcome, as qAOPs are meant to play a central role in regulatory decision-making based on twenty-first-century toxicology approaches to risk assessment.

6. Conclusions

In this incursion into the field of network-based systems toxicology, we have seen how original approaches were used and developed to provide innovative tools for assessing the health risks associated with the exposure to chemical compounds of uncertain safety. The application of systems biology principles to the assessment of exposure-induced responses involved the generation of genome-wide transcription profiles. These large datasets were processed using a combination of standard bioinformatics tools and ad hoc methodologies following a network-based framework reflecting the holistic perspective of systems biology. This approach provided an implementation of the NSR principles and, in particular, supported the 3Rs initiative aimed at reducing animal use in research. We described in more detail the NPA methodology suitable for the particular type of causal networks using the "backward reasoning" approach. Combined with the collection of causal networks available on the CBN website, NPA enables the quantification of exposure-induced perturbations of the mostly molecular biological mechanisms described by the networks. This provided a quantitative assessment of the biological impact resulting from toxicological exposure treatments and offered multiple application possibilities. Turning to the current developments of network-based systems toxicology, we first mentioned the quality improvement of the CBN causal network collection using crowdsourcing initiatives (SBVimprover) and the extension to new biological contexts enabled by the application of literature mining tools that partially replace the manual curation process needed to assemble high-quality causal networks. After integrating the network-based systems biology approach into the multiscale modeling of exposure responses, we discussed the qAOP as a promising development avenue for network-based systems toxicology. Its expected advantageous use in the regulatory decision-making context represents an appealing perspective that justifies the past, current, and certainly future efforts deployed in the development and applications of systems toxicology.

Acknowledgements

We thank our colleagues from the Systems Toxicology Department and from the Research Technologies group of PMI R&D Biomedical, our long-standing collaborators from the Fraunhofer Institute for Algorithms and Scientific Computing, and all the members of the SBV-improver community. We also acknowledge Nicholas Karoglou for the editing of the manuscript.

Conflict of interest

All authors are employees of Philip Morris International.

Author details

Alain Sewer*, Marja Talikka, Florian Martin, Julia Hoeng and Manuel C Peitsch

*Address all correspondence to: Alain.Sewer@pmi.com

PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud, Neuchâtel, Switzerland

References

[1] Cressey D. Debate rages over herbicide's cancer risk. Nature News. 2015

[2] Ledford H. Bisphenol a Linked to Disease in Humans. Nature Publishing Group; 2008

[3] Iskandar AR, Gonzalez-Suarez I, Majeed S, Marescotti D, Sewer A, Xiang Y, et al. A framework for in vitro systems toxicology assessment of e-liquids. Toxicology Mechanisms and Methods. 2016;**26**(6):392-416

[4] Council NR. Toxicity Testing in the 21st Century: A Vision and a Strategy: National Academies Press; 2007

[5] Institute for Systems Biology 2017. Available from: https://www.systemsbiology.org/about/what-is-systems-biology

[6] Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems biology. Annual Review of Genomics and Human Genetics. 2001;**2**(1):343-372

[7] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Computational Biology. 2012;**8**(2):e1002375

[8] Hoeng J, Deehan R, Pratt D, Martin F, Sewer A, Thomson TM, et al. A network-based approach to quantifying the impact of biologically active substances. Drug Discovery Today. 2012;**17**(9):413-418

[9] Barabasi A-L, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nature Reviews. Genetics. 2004;**5**(2):101-113

[10] Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. Nature Reviews. Genetics. 2011;**12**(1):56-68

[11] Del Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. Current Opinion in Biotechnology. 2010;**21**(4):566-571

[12] Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009;**461**(7261):218-223

[13] Schadt EE, Friend SH, Shaywitz DA. A network view of disease and compound screening. Nature Reviews. Drug Discovery. 2009;**8**(4):286-295

[14] Csermely P, Korcsmáros T, Kiss HJ, London G, Nussinov R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacology & Therapeutics. 2013;**138**(3):333-408

[15] Hoeng J, Talikka M, Martin F, Sewer A, Yang X, Iskandar A, et al. Case study: The role of mechanistic network models in systems toxicology. Drug Discovery Today. 2014;**19**(2):183-192

[16] Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. Bioinformatics. 2007;**23**(8):980-987

[17] Nam D, Kim S-Y. Gene-set approach for expression pattern analysis. Briefings in Bioinformatics. 2008;**9**(3):189-197

[18] Wikipedia. The Three Rs (animals). Available from: http://en.wikipedia.org/wiki/The_Three_Rs_(animals)

[19] The European Commission. Alternative testing strategies - progress report 2009 - replacing, reducing and refining use of animals in research. In: Office for Official Publications of the European Communities. 2009

[20] Edwards SW, Preston RJ. Systems biology and mode of action based risk assessment. Toxicological Sciences. 2008;**106**(2):312-318

[21] Russell WM, Burch RL, Hume CW. The principles of humane experimental technique. London, UK: Methuen; 1959

[22] Sewer A, Hoeng J, Deehan R, Westra JW, Martin F, Thomson TM, et al. Systems Biology Approaches for Compound Testing. Data Mining in Drug Discovery: Wiley-VCH Verlag GmbH & Co. KGaA; 2013. pp. 291-316

[23] Martin F, Thomson TM, Sewer A, Drubin DA, Mathis C, Weisensee D, et al. Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks. BMC Systems Biology. 2012;**6**(1):1

[24] Martin F, Sewer A, Talikka M, Xiang Y, Hoeng J, Peitsch MC. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. BMC Bioinformatics. 2014;**15**(1):238

[25] Sewer A, Martin F, Schlage WK, Hoeng J, Peitsch MC. Quantifying the biological impact of active substances using causal network models. Computational Systems Toxicology. 2015:223-256

[26] Boué S, Westra JW, Hayes W, Di Fabio A, Park J, Schlage WK, Sewer A, Fields B, Ansari S, Martin F, Veljkovic E, Kenney R, Peitsch MC, Hoeng J. The Causal Biological Networks (CBN) database 2015. Available from: http://causalbionet.com.

[27] Boué S, Talikka M, Westra JW, Hayes W, Di Fabio A, Park J, et al. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. Database. 2015;**2015**:bav030.

[28] Catlett NL, Bargnesi AJ, Ungerer S, Seagaran T, Ladd W, Elliston KO, et al. Reverse causal reasoning: Applying qualitative causal knowledge to the interpretation of high-throughput data. BMC Bioinformatics. 2013;**14**(1):1

[29] Consortium To. Summary of Large and Small BEL Corpuses 2014. Available from: https://wiki.openbel.org/display/home/Summary+of+Large+and+Small+BEL+Corpuses.

[30] Thomson TM, Sewer A, Martin F, Belcastro V, Frushour BP, Gebel S, et al. Quantitative assessment of biological impact using transcriptomic data and mechanistic network models. Toxicology and Applied Pharmacology. 2013;**272**(3):863-878

[31] Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis (ipa). Bioinformatics. 2013:btt703

[32] Vasilyev DM, Thomson TM, Frushour BP, Martin F, Sewer A. An algorithm for score aggregation over causal biological networks based on random walk sampling. BMC Research Notes. 2014;**7**(1):1

[33] Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in Physiology. 2013;**4**:278

[34] Gonzalez-Suarez I, Martin F, Marescotti D, Guedj E, Acali S, Johne S, et al. In vitro systems toxicology assessment of a candidate modified risk tobacco product shows reduced toxicity compared to that of a conventional cigarette. Chemical Research in Toxicology. 2015;**29**(1):3-18

[35] Kogel U, Suarez IG, Xiang Y, Dossin E, Guy P, Mathis C, et al. Biological impact of cigarette smoke compared to an aerosol produced from a prototypic modified risk tobacco product on normal human bronchial epithelial cells. Toxicology in Vitro. 2015;**29**(8):2102-2115

[36] Phillips B, Veljkovic E, Boué S, Schlage WK, Vuillaume G, Martin F, et al. An 8-month systems toxicology inhalation/cessation study in Apoe−/− mice to investigate cardiovascular and respiratory exposure effects of a candidate modified risk tobacco product, THS 2.2, compared with conventional cigarettes. Toxicological Sciences. 2015:kfv243

[37] Gonzalez-Suarez I, Sewer A, Walker P, Mathis C, Ellis S, Woodhouse H, et al. Systems biology approach for evaluating the biological impact of environmental toxicants in vitro. Chemical Research in Toxicology. 2014;**27**(3):367-376

[38] Gonzalez-Suarez I, Martin F, Hoeng J, Peitsch MC. Mechanistic network models in safety and toxicity evaluation of Nutraceuticals. Nutraceuticals: Efficacy, Safety and Toxicity. 2016:287

[39] Iskandar AR, Martin F, Talikka M, Schlage WK, Kostadinova R, Mathis C, et al. Systems approaches evaluating the perturbation of xenobiotic metabolism in response to cigarette smoke exposure in nasal and bronchial tissues. BioMed Research International. 2013;**2013**

[40] Tarca AL, Lauria M, Unger M, Bilal E, Boue S, Dey KK, et al. Strengths and limitations of microarray-based phenotype prediction: Lessons learned from the IMPROVER diagnostic signature challenge. Bioinformatics. 2013;**29**(22):2892-2899

[41] Mathis C, Gebel S, Poussin C, Belcastro V, Sewer A, Weisensee D, et al. A systems biology approach reveals the dose-and time-dependent effect of primary human airway epithelium tissue culture after exposure to cigarette smoke in vitro. Bioinformatics and Biology Insights. 2015;**9**:19

[42] Poussin C, Laurent A, Peitsch MC, Hoeng J, De Leon H. Systems biology reveals cigarette smoke-induced concentration-dependent direct and indirect mechanisms that promote monocyte–endothelial cell adhesion. Toxicological Sciences. 2015:kfv137

[43] Titz B, Sewer A, Schneider T, Elamin A, Martin F, Dijon S, et al. Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects. Journal of Proteomics. 2015;**128**:306-320

[44] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000;**28**(1):27-30

[45] Haw R, Stein L. Using the reactome database. Current Protocols in Bioinformatics. 2012:8.7. 1-8.7. 23

[46] Nishimura D. BioCarta. Biotech Software & Internet Report: The Computer Software Journal for Science. 2001;**2**(3):117-120

[47] Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: Building research communities on biological pathways. Nucleic Acids Research 2011;40(D1):D1301-D13D7

[48] Paz A, Brownstein Z, Ber Y, Bialik S, David E, Sagir D, et al. SPIKE: A database of highly curated human signaling pathways. Nucleic Acids Research. 2010;**39**(suppl_1):D793-D7D9

[49] The UCSD Signaling Gateway Molecule Pages. Available from: http://www.signaling-gateway.org/molecule/

[50] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: The pathway interaction database. Nucleic Acids Research. 2008;**37**(suppl_1):D674-D6D9

[51] Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, Venugopal AK, et al. NetPath: A public resource of curated signal transduction pathways. Genome Biology. 2010;**11**(1):R3

[52] Kim H ZR, Bai X, Liu M. PKC Activation Induces Inflammatory Response and Cell Death in Human Bronchial Epithelial Cells; 2013. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44747

[53] Westra JW, Schlage WK, Frushour BP, Gebel S, Catlett NL, Han W, et al. Construction of a computable cell proliferation network focused on non-diseased lung cells. BMC Systems Biology. 2011;**5**(1):105

[54] Schlage WK, Westra JW, Gebel S, Catlett NL, Mathis C, Frushour BP, et al. A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. BMC Systems Biology. 2011;**5**(1):168

[55] Gebel S, Lichtner RB, Frushour B, Schlage WK, Hoang V, Talikka M, et al. Construction of a computable network model for DNA damage, autophagy, cell death, and senescence. Bioinformatics and Biology Insights. 2013;**7**:BBI. S11154

[56] Westra JW, Schlage WK, Hengstermann A, Gebel S, Mathis C, Thomson T, et al. A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue. Bioinformatics and Biology Insights. 2013;**7**:BBI. S11509

[57] Jennifer S, Park WKS, Frushour BP, Talikka M, Toedter G, Gebel S, Deehan R, Veljkovic E, Westra JW, Peck MJ, Boue S, Kogel U, Gonzalez-Suarez I, Hengstermann A, Peitsch MC, Hoeng J. Construction of a computable network model of tissue repair and angiogenesis in the lung. Journal of Clinical Toxicology; 2013

[58] De León H, Boué S, Schlage WK, Boukharov N, Westra JW, Gebel S, et al. A vascular biology network model focused on inflammatory processes to investigate atherogenesis and plaque instability. Journal of Translational Medicine. 2014;**12**(1):185

[59] Boue S, Fields B, Hoeng J, Park J, Peitsch MC, Schlage WK, et al. Enhancement of COPD biological networks using a web-based collaboration interface. F1000Research. 2015;**4**

[60] Szostak J, Martin F, Talikka M, Peitsch MC, Hoeng J. Semi-automated Curation allows causal network model building for the quantification of age-dependent plaque progression in ApoE−/− mouse. Gene regulation and systems biology. 2016;**10**:95

[61] Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, De La Fuente A, et al. Verification of systems biology research in the age of collaborative competition. Nature Biotechnology. 2011;**29**(9):811-815

[62] Meyer P, Hoeng J, Rice JJ, Norel R, Sprengel J, Stolle K, et al. Industrial methodology for process verification in research (IMPROVER): Toward systems biology verification. Bioinformatics. 2012;**28**(9):1193-1201

[63] Ansari S, Binder J, Boue S, Di Fabio A, Hayes W, Hoeng J, et al. On crowd-verification of biological networks. Bioinformatics and Biology Insights. 2013;**7**:307

[64] The sbvIMPROVER Network Verification Challenge 2015. Available from: https://bionet.sbvimprover.com/

[65] Szostak J, Ansari S, Madan S, Fluck J, Talikka M, Iskandar A, et al. Construction of biological networks from unstructured information based on a semi-automated curation workflow. Database. 2015;**2015**

[66] Madan S, Hodapp S, Senger P, Ansari S, Szostak J, Hoeng J, et al. The BEL information extraction workflow (BELIEF): Evaluation in the BioCreative V BEL and IAT track. Database. 2016:**2016**

[67] Madan S, Hodapp S, Senger P, Ansari S, Szostak J, Hoeng J, Peitsch M, Fluck J. BELIEF - a Semi-Automated Workflow for BEL Network Creation 2016. Available from: http://belief.scai.fraunhofer.de/BeliefDashboard/

[68] Nordlund M, Belka M, Kuczaj AK, Lizal F, Jedelsky J, Elcner J, et al. Multicomponent aerosol particle deposition in a realistic cast of the human upper respiratory tract. Inhalation Toxicology. 2017;**29**(3):113-125

[69] Bush ML, Frederick CB, Kimbell JS, Ultman JS. A CFD–PBPK hybrid model for simulating gas and vapor uptake in the rat nose. Toxicology and Applied Pharmacology. 1998;**150**(1):133-145

[70] Frederick CB, Gentry PR, Bush ML, Lomax LG, Black KA, Finch L, et al. A hybrid computational fluid dynamics and physiologically based pharmacokinetic model for comparison of predicted tissue concentrations of acrylic acid and other vapors in the rat and human nasal cavities following inhalation exposure. Inhalation Toxicology. 2001;**13**(5):359-376

[71] Campbell JL, Andersen ME, Clewell HJ. A hybrid CFD-PBPK model for naphthalene in rat and human with IVIVE for nasal tissue metabolism and cross-species dosimetry. Inhalation Toxicology. 2014;**26**(6):333-344

[72] Jamshidi N, Palsson BØ. Formulating genome-scale kinetic models in the post-genome era. Molecular Systems Biology. 2008;**4**(1):171

[73] DasGupta B, Enciso GA, Sontag E, Zhang Y. Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. Bio Systems. 2007;**90**(1):161-178

[74] Worth A, Munn S, Whelan M. Wittwehr C. 2.1 the adverse outcome pathway (AOP) concept. Alternative methods for regulatory toxicology–a state-of-the-art review. 2014:3

[75] The Collaborative Adverse Outcome Pathway Wiki (AOP-Wiki); 2013. Available from: https://aopwiki.org/

[76] Luettich K, Talikka M, Lowe FJ, Haswell LE, Park J, Gaca MD, et al. The adverse outcome pathway for oxidative stress-mediated EGFR activation leading to decreased lung function. Applied In Vitro Toxicology. 2017;**3**(1):99-109

[77] Lowe FJ, Luettich K, Talikka M, Hoang V, Haswell LE, Hoeng J, et al. Development of an adverse outcome pathway for the onset of hypertension by oxidative stress-mediated perturbation of endothelial nitric oxide bioavailability. Applied In Vitro Toxicology. 2017;**3**(1):131-148

[78] OECD. Users' handbook supplement to the Guidance document for developing and assessing adverse outcome pathways. https://aopkborg/common/AOP_Handbookpdf. 2014

[79] Sturla SJ, Boobis AR, FitzGerald RE, Hoeng J, Kavlock RJ, Schirmer K, et al. Systems toxicology: From basic research to risk assessment. Chemical Research in Toxicology. 2014;**27**(3):314-329

[80] Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, et al. BioModels database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research. 2006;**34**(suppl_1): D689-DD91

[81] Schmidt BJ, Casey FP, Paterson T, Chan JR. Alternate virtual populations elucidate the type I interferon signature predictive of the response to rituximab in rheumatoid arthritis. BMC Bioinformatics. 2013;**14**(1):221

[82] Conolly RB, Ankley GT, Cheng W, Mayo ML, Miller DH, Perkins EJ, et al. Quantitative adverse outcome pathways and their application to predictive toxicology. Environmental Science & Technology. 2017;**51**(8):4661-4672

Edited by Ibrokhim Y. Abdurakhmonov

Bioinformatics has evolved significantly in the era of post genomics and big data. Huge advancements were made toward storing, handling, mining, comparing, extracting, clustering and analysis as well as visualization of big macromolecular data using novel computational approaches, machine and deep learning methods, and web-based server tools. There are extensively ongoing world-wide efforts to build the resources for regional hosting, organized and structured access and improving the pre-existing bioinformatics tools to efficiently and meaningfully analyze day-to-day increasing big data. This book intends to provide the reader with updates and progress on genomic data analysis, data modeling and network-based system tools.

Published in London, UK
© 2018 IntechOpen
© 123dartist / iStock

IntechOpen

ISBN 978-1-83881-563-9
9 781838 815639