Chapter

Bioinformatics as a Tool for the Structural and Evolutionary Analysis of Proteins

Edna María Hernández-Domínguez, Laura Sofía Castillo-Ortega, Yarely García-Esquivel, Virginia Mandujano-González, Gerardo Díaz-Godínez and Jorge Álvarez-Cervantes

Abstract

This chapter deals with the topic of bioinformatics, computational, mathematics, and statistics tools applied to biology, essential for the analysis and characterization of biological molecules, in particular proteins, which play an important role in all cellular and evolutionary processes of the organisms. In recent decades, with the next generation sequencing technologies and bioinformatics, it has facilitated the collection and analysis of a large amount of genomic, transcriptomic, proteomic, and metabolomic data from different organisms that have allowed predictions on the regulation of expression, transcription, translation, structure, and mechanisms of action of proteins as well as homology, mutations, and evolutionary processes that generate structural and functional changes over time. Although the information in the databases is greater every day, all bioinformatics tools continue to be constantly modified to improve performance that leads to more accurate predictions regarding protein functionality, which is why bioinformatics research remains a great challenge.

Keywords: computational biology, databases, proteomics, transcriptomics, functional genomics, phylogeny

1. Introduction

The study to understand the functioning of the cell, as well as the molecules and processes that are carried out within it, originated the use of various disciplines and sciences to facilitate the progress in research for its characterization over time. In the 1950s, the sequencing of small biological molecules began, and in 1956, the sequencing of the first protein was achieved. Thus, Margaret O. Dyhoff determined that bovine insulin is a small peptide of 51 amino acids. With these advances and the constant production of biological information, there was a need to collect and organize all the information generated from these sequencing projects [1]. In 1965, the first biological sequence database was created, in which all the DNA and protein sequences described up to that time were stored and made available to the scientific community. Eight years later, the oldest known database was created, which is still in force today, *Protein Data Bank* (PDB) [2].

In the 80s, bioinformatics had already gained a new meaning in scientific research, so several research groups such as Theoretical Biology and Biophysics Group attached to the American Institute The Alamos National Laboratory, together with Stanford University, gave rise to the best-known database in the world called GenBank. Almost at the same time, in 1981, Temple Smith and Michael Waterman extensively reviewed the mathematical algorithms for comparing biological sequences. As a result of their analysis, they generated the well-known local alignment algorithm that allowed to optimize the comparison of biological sequences, being the most important contribution for the direct comparison of sequences and cornerstone of the alignment by sequence pair [3].

A few years after the creation of *GenBank*, its European and Asian versions were generated, known as the EMBL database (*European Molecular Biology Laboratory*) and DDBJ (*DNA Data Bank of Japan*) in 1981 and 1984, respectively. In 1985 the FASTA algorithm (*FAST-AII*) of sequence comparisons was reported, which operated as a search engine for similar sequences within the *GenBank* [4]. During the years from 1987 to 1990, databases for protein sequences were propelled which resulted in the creation of Swiss-Prot and PIR (Protein Information Resource). In 1990, another of the most important milestones in bioinformatics originated the BLAST algorithm (Basic Local Alignment Tool) that completely revolutionized the exploration and search of biological sequences in databases [5].

The National Center for Biotechnology Information (NCBI) makes the following definition:

Bioinformatics is a field of science in which various disciplines such as applied mathematics, statistics, artificial intelligence, chemistry, biochemistry, computing and information technology converge, whose objective is to facilitate the discovery of new biological ideas, as well as create global perspectives from which unifying principles in biology can be discerned [6].

It consists of two complementary subfields with each other:

1. The development of computer tools and databases.

2. The application of these in the generation of biological knowledge to better understand living systems [7].

According to the *National Institute of Health* of the United States, bioinformatics or also called computational biology, deals with the development and application of analytical data and theoretical methods, mathematical modeling and computer simulation techniques to study biological, behavioral and social systems [8]. The programs use public or private databases (with restricted access or with economic value) that have been created with information that is constantly growing and managed by institutions from various sectors. The main databases used in computational biology are described below:

1.1 Biological databases

- *Primary databases* contain original biological data. They are raw sequence files or structural data (for example*, GenBankm y Protein Data Bank)* [6].

- *Secondary databases* contain information processed computationally based on primary data. Translated protein sequence databases contain the functional annotation belonging to this category (for example, *Swiss-Prot and PIR)* [6].

- *Specialized databases* are those that serve a particular research interest (for example, *Flybase*). The HIV sequence database and Ribosomal Database Project are examples of databases that specialize in a particular organism or a certain type of data. Many of the problems detected in scientific research lie in the need to connect secondary and specialized databases to primary databases. It is desirable that entries in a database be cross-referenced or linked to related entries in other databases that contain additional information [6].

There are primary databases, which contain direct information on the sequence, structure or pattern of DNA or protein expression, and secondary, which contains data derived from primary databases, such as mutations, evolutionary relationships, grouping by families or by functions, involvement in diseases, etc.

1.2 Databases for protein analysis (amino acid sequence databases)

Swiss-Prot: It contains annotated or commented sequences, that is, each sequence has been reviewed, documented and linked to other databases. External link: Swiss-Prot in the EBI (*http://www.ebi.ac.uk/swissprot/access.html*), Swiss-Prot in ExPASy (*http://us.expasy.org/sprot/*) [9].

TrEMBL: *Translation of EMBL Nucleotide Sequence Database* includes the translation of all coding sequences derived from (EMBL-BANK) and which have not yet been annotated in Swiss-Prot. External link: TrEMBL (*http://www.ebi.ac.uk/trembl/*) [9].

PIR: *Protein Information Resource* is divided into four sub-bases that have a decreasing annotation level. External link: PIR (*http://pir.georgetown.edu/*) [9].

ENZYME: It links the complete enzyme activity classification to the Swiss-Prot sequences. External link: ENZYME (*http://us.expasy.org/enzyme/*) [9].

PROSITE: It contains information on the secondary structure of proteins, families, domains, etc. External link: PROSITE (*http://us.expasy.org/prosite/*) [9].

INTERPRO: It integrates information from various secondary structure databases such as PROSITE, providing links to other databases and more extensive information. External link: INTERPRO (*http://www.ebi.ac.uk/interpro/index.html*) [9].

PDB: *Protein Data* Bank is the 3-D tertiary structure database of proteins that have been crystallized. External link: PDB (*http://www.rcsb.org/pdb/*) [9].

1.3 Data warehouse

A *Data Warehouse* (*DW*) is a set of integrated data oriented to a subject, which vary over time and are not transitory, which support the decision-making process of the administration [10]. From the review of the bioinformatics projects it is found that the requirements of this field require the storage of large volumes of data, with multiple dimensions, of extended periods of time and with heterogeneous formats as well as their sources. For example, *Ligand Depot* is an integrated data source for finding information about small molecules, proteins and nucleic acids. It focuses on providing chemical and structural information for small molecules. Accepts keyword-based queries, also provides a graphical interface for conducting chemical substructure searches, and allows access to a wide variety of web resources [11].

1.4 Data mining in bioinformatics

Data mining is oriented towards the study of techniques to extract valuable information from a large amount of biological data. For this, efficient software tools

are necessary to recover data, compare biological sequences, discover patterns and visualize the discovery of knowledge [8].

Among the most common data mining techniques in bioinformatics can be highlighted [8]:

KDD is the complete process of extracting knowledge, not trivial, previously unknown and potentially useful from a data set.

KDT is oriented to the extraction of knowledge from data (unstructured in natural language) stored in textual databases, is identified with the discovery of knowledge in the texts.

1.5 Applications of bioinformatics

The areas in which bioinformatics is currently developed are many and varied, ranging from simple tasks such as direct acquisition of data from DNA or protein sequencing assays (when techniques such as mass spectrophotometry are used), until the development of software for the storage and analysis of the data, which implies in many cases, the generation of algorithms that require both mathematical and biological knowledge. Within the areas in which bioinformatics takes place are genomics, proteomics, pharmacogenetics and phylogeny. The plant genome databases and gene expression analysis of this profile have played an important role in the development of new crop varieties that have higher productivity and more disease resistance [7].

Specifically, bioinformatics encompasses the development of databases or knowledge to store and retrieve biological data, algorithms to analyze and determine their relationships with biological data, and the statistical tools to identify and interpret data sets. The following describes in detail what refers to metabolomics, transcriptomics, proteomics, comparative genomics, functional genomics, phylogeny and protein modeling.

2. Metabolomic data analysis

The metabolomics was originally proposed as a tool of functional genomics, but its use has been extended much more, as it has had great advances like other omics sciences, such as transcriptomics and proteomics; because the metabolomic work is determined by physical-chemical characteristics of organic molecules unlike the genes, mRNA and proteins that come from a specific sequence, so the success of the characterization of these biopolymers is thanks to bioinformatics technology and tools that help sequence characterization [12]. Its objective is to detect, quantify and interpret the overall analysis of all metabolites; these studies are used in various areas and, like proteomics, one of its main contributions is biomarkers, helping to identify metabolites that are correlated with diseases and environmental exposures [13]. Metabolites are chemical entities that do not come from a transfer of information within the cell, coupled with this, they are also characterized by being diverse as they are substrates and metabolism products that drive essential cellular functions, such as energy production and storage, signal transduction and cell apoptosis; in this great diversity of chemical structures we find endogenous and exogenous metabolites, the former are produced naturally by an organism and the latter come from interaction with the outside. The great diversity of molecules reflects in a wide range of polarities, molecular weights, functional groups, stability and chemical reactivity, etc. [12, 13].

Among the first reports of metabolite detection are those where mass spectrometry (MS) was used to separate a wide range of metabolites present in urine and

tissue extracts [14]. In addition, multicomponent analyzes were described to obtain the metabolic profile for three types of urinary constituents: steroids, acids, drugs and drug metabolism [15]. On the other hand, there are reports where physical, chemical or psychological changes can cause biological responses such as oxidative stress and inflammation; among the biomarkers that are the result of a chemical reaction are lipoperoxides or oxidized proteins that are the result of the reaction of molecules with reactive oxygen species (ROS) and those that represent the biological response to stress, such as the transcription factor NRF2 or inflammation and inflammatory cytokines [16]. Among the best known and clinically used examples we find glucose as a marker of diabetes [17] and phenylalanine as a marker of congenital metabolic disorder [18].

Because metabolites play important roles in the biological pathways; its differential flow or regulation can reveal new knowledge about diseases and environmental influences, so one of the most important objectives of the metabolic analysis has been to assign the identity of the metabolite within a metabolic pathway [19, 20]; generating a large amount of data; requiring for its processing an arduous mathematical, statistical and bioinformatic work [12, 21, 22], this last area is crucial for the development of metabolomics as it helps in the handling of data and information, analytical data processing, metabolomic standards, ontology, statistical analysis, mining and data integration, and mathematical modeling of metabolomic networks with antecedents of biological systems [12], it is also necessary to decide which metabolites are biologically more significant. This can be achieved by helping the identification process, reducing the redundancy of characteristics, presenting better candidates for the MS, accelerating or automating the workflow, recovering data through characteristics through meta-analysis or multigroup analysis, or using stable isotopes and mapping of pathways. For all the above, in recent years, the technologies for analyzing metabolites have undergone improvements, establishing more efficient protocols for experimental design, as well as better sample extraction techniques and data acquisition that have been worthwhile in providing sets of complex and solid data [20].

The database management system for metabolomics requires the collection of raw and processed metadata, some important aspects for comparing data and obtaining results in different laboratories and reproducing experimental conditions are: The nature and treatment of samples prior to study. Among the bases and tools for the analysis and visualization of available data are: Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/) [23] and Metabolic Pathways From all Domains of Life (MetaCyc; http://metacyc.org/) [24].

3. Transcriptome data analysis

The genes response to intracellular or extracellular stimuli includes a hierarchy of signals that allows genes encoded in the DNA to be expressed or repressed by the transcription process. The total set of transcripts (RNA molecules) produced by a cell under a given condition and time, is defined as a *transcriptome* [25]. Unlike the genome, the transcriptome is highly dynamic and actively changes as a consequence of factors that influence the stage of development of organisms, as well as the surrounding environmental conditions. In this sense, transcriptomics is an essential tool to interpret the functional elements of the genome, having as object of study, all species of transcripts, messenger RNA, non-coding RNA and small RNAs [26]. Its main purpose being to determine transcriptional structure of genes, that is, where a gene begins and ends (start sites 5′ and 3′ end), posttranscriptional modifications, splicing patterns and differential expression analysis [27].

The RNA molecules synthesized by a cell have a specific function in a given cellular process, the transcripts include: (a) messenger RNA (mRNA) that is the intermediary between the gene information and the proteome. In this way, the amount of mRNA molecules makes it possible to elucidate expression patterns and in turn correlate the abundance of mRNA molecules with changes in protein abundance [28]; (b) non-coding RNA (cRNA) that is responsible for the regulation of gene expression [29]. Determining where, how and when a transcript is generated is essential to know the biological activity of a gene [28]. Analyzing the transcripts that coexist at any given time gives us global information on the cellular state under a certain condition, which has allowed us to establish patterns of gene regulation coordinated with the consequent identification of promoter elements common to several genes [30].

3.1 RNA study technologies and tools in bioinformatic analysis

The RNA study approach has changed from the sequencing of the first determined RNA molecule, to the sequencing of the transcriptome using new generation technologies [25]. *Northern blot* is a technique based on hybridization and radioactive labeling, cDNA microarrays (complementary DNA obtained from mRNA) and cDNA-AFLP tools widely used in studies of expression levels and serial analysis of gene expression (SAGE), at the time they provided relevant information, being Microarrays widely used today [31–35]. However, these techniques require prior knowledge of the genome, have low coverage and are based on hybridization, in this sense the abundance of transcripts is inferred by the intensity of hybridization and the results obtained are noisy, which directly interferes with the reproducibility of the results, besides being insufficient techniques to detect new transcripts [25].

The growing importance of DNA sequencing in model organisms, as well as in the quest to understand the dogma of biology, the NGS technologies (Next Generation Sequencing) arise, which have high yields in the treatment of the sample, are reproducible and highly reliable, as well as accessible and economical, to the point of being more profitable than sequencing by SANGER. These next-generation technologies are based on sequencing by synthesis (SBS) known as pyrosequencing, the transcriptomic variant of pyrosequencing technology is known as short-reading massive parallel sequencing (RNA-seq). The availability of this technology has revolutionized the approach of transcriptome study, having commercially available Roche/454; Applied Biosystems SOLID; HeliScope e Illumina [36].

From the first RNA studies based on sequencing by SANGER to NGS technologies, bioinformatics has been a key tool in the analysis process. Initially the differential expression based on the analysis by Microarrays presented its own computational challenges [36], currently while the reads are shorter than those created by sequencing by SANGER, NGS has a higher performance and generates data set of up to 50 gigabases per run [37], this requires algorithms capable of processing this amount of data in the shortest time possible and with a high degree of reliability.

The study of the transcriptome by RNAseq involves different stages ranging from RNA extraction, library construction, sequencing and data analysis. In this last step four main stages are distinguished (a) *Quality analysis of the reads,* this allows to determine possible problems in the reads. FastQC is a next-generation data quality control tool, which reports graphs and tables providing quality information based on the reads (per base sequence quality); check the quality of subsets of reads (per sequence quality scores); it also shows the proportion of each nucleotide base of the DNA in each base of the reads (per base sequence content); presents the average GC content in the reads and compares that content with the normal distribution (per sequence GC content); shows the proportion of N, that is, unknown

nucleotide observed in each reading position (per base N content); shows the size distribution of reads (sequence length distribution); detects adapters in the reads (adapter content); detects possible sequencing problems introduced in the reads after the adapter (k-mer content) https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/ [38]. Is advisable that the length of the reads to be analyzed is the same, also if there is a poor quality in the reads, the procedure to follow is to cut those bases where there is poor quality. Tools such as Fastx-toolkit (https://bio.tools/fastx-toolkit) [39], Trimmomatic [40], PRINSEq [41], Flexbar [42] and others can be used to cut or filter reads, ensuring reliable data for alignment. (b) *Mapping and identification of transcripts*: at this stage the location of the reads with respect to a reference genome is known or a Novo assembly is made. There are three study strategies: (1) the reads are aligned with allocator with gaps to a reference genome (example TopHat, STAR) which allows the identification of new transcripts [43, 44]; (2) If the discovery of new transcripts is not sought, the reads can be aligned to the reference genome using an aligner without gaps for example RSEM [45]; (3) When the genome is not available, the reads are mounted on transcripts what is known as *Novo* assembly (example TRINITY) [46]. In the transcription level analyzes, the isoforms that a gene presents are considered separately. On the contrary, in the level analyzes of gene, the isoforms that it presents form a unit [47]. (c) *Quantification of reads*: Sample reads are quantified in relation to the transcripts that appear in the reference genome or by *Novo* assembly. The tools used in quantification can be based on alignment or without alignment. Alignment-based tools map all reads of a sample, to a genome or to transcriptome. Subsequently, quantify the reads that are assigned to a transcript, in the case of TopHat and RSEM [43, 45]. Tools that skip sequence alignment like HTSEq and featureCounts [48, 49], use the k-mer count, that is, they count all the k-mer in a sequencing library without aligning them to any reference, in this way the k-mer are counted and the unique k-mer are selected to quantify the expression and finally, these unique k-mer are assigned to the transcriptome to identify the transcription. (d) *Differential Expression Analysis*: At this stage, it is analyzed if the expression of a gene is different between different conditions. To determine if in a specific gene there are significant differences in the number of mapped reads corresponding to that gene, there are a large number of tools that are based on the comparison of the reading count for each transcript/gene under different biological conditions, by statistical analysis, which implies normalization methods since transcripts are synthesized at different levels (genes or transcripts with low or high level of expression), probabilistic models, modeling of reading counts at given distribution etc. In the differential expression analysis by RNA-seq, should be considered that the longer transcripts generate more reads compared to shorter transcripts. In addition, the technical noise introduced into the data during the sequencing process, as part of the variability in the number of reads produced by execution causes fluctuations in the number of mapped elements in the sample. To reduce the technical noise introduced into the data during the sequencing process, the number of reads must be normalized in order to obtain significant estimates of the expression. Among the statistical parameters used for this process are the metric of reads per kilobase per million mapped reads (RPKM), fragments per kilobase per million mapped reads (FPKM) [50, 51]. With these parameters it is possible to quantify transcription levels and make the comparison between samples. On the other hand, fold change allows us to evaluate the rate of change of a transcript in both conditions [52]. Within the challenges of transcriptome analysis, it is important to understand how the levels of expression differ in each situation studied, to achieve this objective, different methods try to model the biological variability such as EdgeR, DESeq, Cuffdiff [48, 53, 54]. In this way, there are currently different computational tools suitable for the overall study of

the transcriptome suitable for each stage of analysis and specialized for each type of transcript under study (**Table 1**).

3.2 Bioinformatics tools in the study of coding RNA, non-coding RNA and microRNAs

The identification of non-coding RNAs and small RNAs is a vital issue in genetic analysis [29], in this sense algorithms have been developed for the analysis of this type of RNAs in particular (**Table 1**). Currently, the tools used to classify

Process	Tools	Objective	References
Quality analysis of reads	FastQC Fastx-toolkit Trimmomatic, PRINSEq, Flexbar	It analyzes the quality of the reads It debugs poor quality reads	[38–42]
Assembly	Trinity, Trans-ABySS, Oases, IDBA-Tran TOPHAT, STAR, IDBA-Tran, HISAT	Assembly of reads without genome or reference transcriptome Assembly of reads with genome or reference transcriptome	[46, 55, 56] [43, 44, 57, 58]
Classification of transcripts	BLAST, BLAT, GMAT, AUGUSTUS CPAT, FEELnc, NRC, lncRScan-SVM	It identifies coding transcripts by homology or by known transcript characteristics	[5, 59–61] [62–65]
Mapping	TOPHAT, STAR, HISAT, HISAT2, Bowtie	It aligns reads with a reference genome or transcriptome	[43, 44, 58, 66]
Quantification	RSEM, Feature Count StringTie, Salmon, Kallisto	It estimates the number of transcripts with or without their alignment	[45, 49, 67–69]
Classification of coding and non-coding transcripts	BEDTools, glbase	It determines the coordinates of the reference genome	[70]
	BLAST, BLAT, GMAP, AUGUSTUS	Through homology it manages to determine known sequences of transcripts found in databases	[5, 59–61]
	CPAT, FELLnc, lncRScan-SVM, NRC	It evaluates characteristics of coding and non-coding transcripts	[62–65]
Small RNA analysis	miRDeep Pic Tar	It quantifies known micro RNAs and identify new RNAs	[71–73]
	PiPMir	It identifies new micros RNAs in plants	[74]
	DARIO	It allows the recognition of micro RNAs, snoRNA and tRNA	[73]
	IntaRNA	It analyzes micro RNAs in eukaryotes and small bacterial RNAs	[75, 76]
	CopraRNA	It makes comparative predictions that include functional enrichment analysis	[76, 77]

Table 1.
Computational tools in the study of the transcriptome.

coding and non-coding sequences have two aspects, those that classify transcripts according to similarity and those that use known coding and non-coding properties [47]. Similarity-based tools classify transcripts, taking as reference the amino acid sequences of their transcripts translated with known protein coding genes, for example BLAST [5], BLATS [59], GMAP [60]. On the other hand, tools focused on coding and non-coding characteristics are based on the properties of known transcripts to predict whether a transcript encodes or not for a protein. The coding potential can be estimated using automatic learning approaches such as CPAT [62], FEELnc [63], lncRScan-SVM [64] and NRC [65]. These exclude transcripts based on properties such as transcription length, length of open reading frame (ORF), ORF coverage, k-mer frequency, codon usage bias, in addition to being optimized for different techniques [47]. In the choice of the tool to be used to evaluate the coding potential of a transcript, it will depend on what is sought in the study, if there is a good annotation and reference genome the tools based on similarity are practical and feasible in the analysis. However, in organisms that lack good gene annotations it is advisable to use tools based on coding and non-coding characteristics, which also allow to identify new genes. On the other hand, the availability of small readings opened a new field of study for small RNAs such as microRNAs (miRNAs), small RNAs of interference (siRNA) and piwiRNAs (piRNAs); Currently there are specialized tools for this type of RNA that provide additional biological knowledge. In this case miRDeep and its varieties are widely used to quantify known and novel RNA (miRNA), from the sequencing of small RNA by RNAseq [71, 72]; PiPMir [74] has been used for the detection of miRNA in plants. DARIO (http://dario.bioinf.uni-leipzig.de/index.py) is a web service that allows not only the recognition of new microRNAs but also small RNAs derived from other types of parental RNAs, such as snoRNA and tRNA [73]. Pic Tar is an algorithm for the identification of micro RNAs, which is based on functional interactions of micro RNA [78, 79]. IntaRNA has been designed for the study of micro RNAs in eukaryotes and small bacterial RNAs (RNAs) [75, 76]. CopraRNA is a comparative prediction algorithm that is complemented by post-processing methods that includes functional enrichment analysis [76, 77]. Finally, after analyzing the data, the biological conclusions must be carefully interpreted.

4. Proteomics data analysis

Transcriptome sequences provide resources for gene expression profile studies, as well as for the identification of mutations, sequence aberrations and RNA editing events [25], the above is possible to the existence of the open reading frame (ORF), however, in genomic data this does not imply the existence of a functional gene; despite the great advances in bioinformatics that facilitate the analysis and prediction of genes with the help of comparative genomics, and although they are years of development of molecular simulation methods, attempts to improve models that are already relatively close to the structure native, they have had little success, which may be due to inaccuracies in the potential functions used in simulations, such as the treatment of electrostatic and solvation effects or it may be necessary to improve sampling strategies due to the relatively long folding time scale of proteins; the combination of chemistry and physics with the large amount of information in known protein structures could provide a better route for the development of enhanced potential functions. Currently, it is difficult to accurately predict protein structures from genes, the success rate for the correct prediction of structures remains low [25, 80, 81]. Proteomics involves various technologies for deep proteome analysis, thus achieving quantification and identification of these proteins;

covering the part of functional analysis of genetic products, interaction studies, and protein localization, which helps explain the identity of an organism's proteins to know the structure and function. However, considering that the proteome is highly dynamic due to the complex regulatory systems that control the levels of protein expression, its use is limited, since in addition to the use of specialized personnel, facilities and equipment, software is also included for equipment, and databases, which increases costs [80, 82, 83]. Proteomics is constantly updated, generating challenges ranging from sample preparation to data collection. A large amount of information is generated from protein folding models, three-dimensional structures, prediction of unknown protein structures and functions, data obtained from the separation of proteins by electrophoresis in two-dimensional gels, isoelectric focusing, 2D protein visualization, peptide mass fingerprinting (PMF), MS, MS in tandem, etc., the above generates high performance proteomes with the help of bioinformatics, which introduces new algorithms to handle a large amount of heterogeneous data [84–86].

Some of the most used platforms in proteomics are: The Basic Local Alignment Search Tool (BLAST), Expert Protein Analysis System (ExPASy) and Protein Data Bank (PDB); BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). It is one of the most used and updated platforms, which uses simple but powerful methods for protein analysis comparing amino acid sequences, which makes it possible to determine homology between proteins, where the algorithms used to perform this procedure guarantee the best possible alignment, however, it does not guarantee the best structure [5, 86–90]. ExPASy gives access to a wide variety of databases and analytical tools dedicated to proteins and proteomics. On the other hand, PDB (https://www.wwpdb.org/) is the global repository of three-dimensional structures of macromolecules that is updated weekly and contains more than 153,000 protein structures, resulting from crystallographic studies, X-rays or nuclear magnetic resonance (NMR) created by modeling software, all these platforms contain various servers that help classify proteins according to their sequence, structure and function [86, 91, 92].

All this information is of great help, since it is used in different research areas, such as detection of diagnostic markers, candidates for vaccine production, understanding the mechanisms of pathogenicity, alteration of expression patterns in response to different signals and interpretation of functional protein pathways in different diseases [93–98].

5. Comparative genomics

Comparative genomics is a broad field of study that identifies differences between genomes and elucidates which of them are responsible for phenotypic changes in organisms [99]. In contrast to 'traditional' genomic studies that focus on a single genome per study [100], comparative genomics provides additional detailed information to that obtained from the analysis of a single genome, which can reveal the encoded functional potential of an organism compared to another [101–103]. Comparisons between different genomes of organisms lead to more rapid identification of different underlying mechanisms are shared between organisms and others that are different among them [104–106]. Likewise, comparative genomics allows a better understanding of how species have evolved [107]. In this sense, the concept of pangenome (**Figure 1**) refers to the set of genes in a particular species [106]. The commonly used partition of a pangenome considers three main parts: the central genome, the expendable or accessory genome and the singleton genome [108]. The central genes are responsible for the basic aspects

of the biology of the species and its main phenotypic features; while accessory genes and singletons generally belong to supplementary biochemical pathways and functions that can confer selective advantages such as ecological adaptation [108]. While the global analysis of gene content (as in pangenome studies) provides information on differences in functional potential and possible phenotypic differences between organisms, specific central gene analyzes have also been used for studies of phylogenetic diversity [99, 108].

Initially, the concept of pangenome was used to refer to bacterial genomes, however, over time it has been used to refer to genomes of eukaryotic organisms such as yeasts [106, 109], plants [108, 110, 111], and viruses [108, 112]. Different organisms can be compared despite their phenotypic differences and with respect to their relationship of kinship (phylogenetic distances) [105, 113]. The assembly of genomes from sequencing data by Illumina or PacBio methods [114] involves five important stages, these steps are described in **Figure 2**, as well as some of the tools used [106].

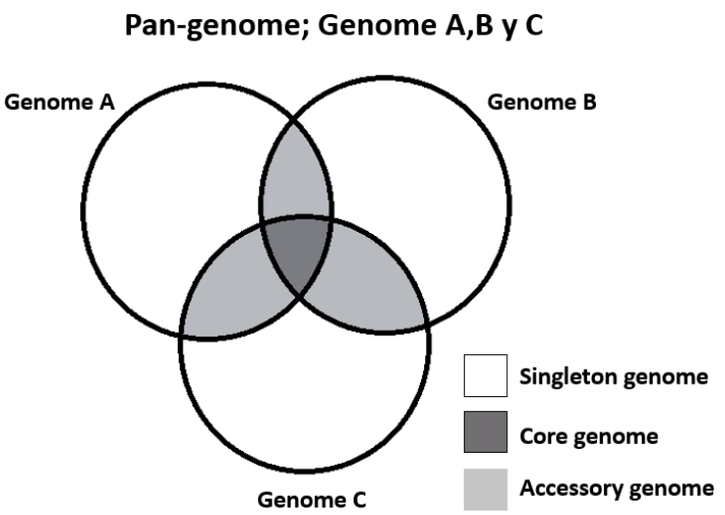

Figure 1.
Pangenome diagram of three different genomes.

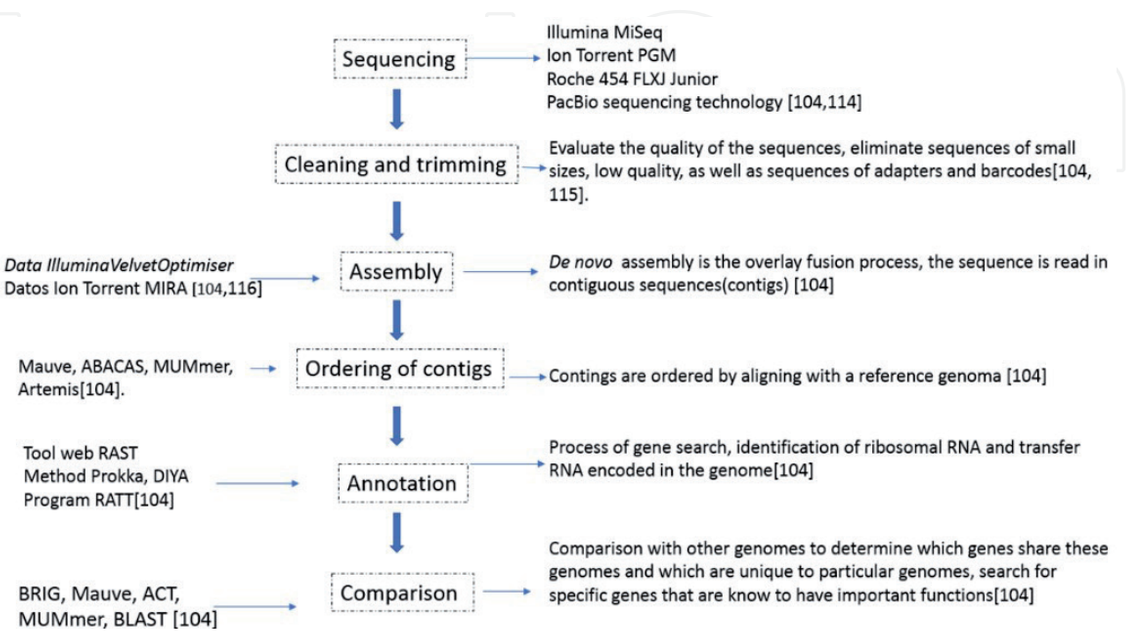

Figure 2.
Workflow for the de novo genome comparative analysis.

For gene comparisons databases with different characteristics are used, for example, to obtain gene families and identify their orthology the EDGAR database [108, 115] is used, as well as, the prokaryotic-genome analysis tool (PGAT) for the analysis of bacterial genomes [108, 116]. There are independent applications such as the Pan-genome analysis pipeline (PGAP) that have specific modules to perform the functional analysis of genes, the analysis and determination of each of the components of the pangenome, the detection of genetic variation as well as the analysis of Species evolution [108, 117], PanFunPro is a tool that allows pangenome analysis in protein prediction from genetic information [96]. There are tools that allow you to work with large amounts of data such as PanGP [118] and the large acale BSR [119].

The bacterial pan genome analysis tool (BPGA) [120] is a recently published package for pangenome analysis with seven functional modules; In addition to routine analysis, it presents a series of novel features for subsequent analyzes such as phylogeny, as well as tools that allow determining the presence and absence of certain genes in specific strains, another module to perform subset analysis, content analysis atypical G+C and KEGG & COG mapping of central, accessory and unique genes [108, 121–124].

6. Functional genomics

Functional genomics studies and assigns functions to the genome of an organism, including genes and non-genetic elements [125, 126], with the support of molecular and cellular biology studies, focused on the dynamic aspects of transcriptomics, proteomics and metabolomics [127], that allow to know the relationship of genes, their transcription, translation and protein-protein interactions [128, 129], that promote the phenotypic characteristics of each organism [125, 126]. A functional genomic approach can use multiple techniques for data analysis in a single study [129]. Apart from the tools of transcriptomics and proteomics, functional genomics needs of studies that allow us to know gene interactions [130, 131], genetic variations (polymorphisms) in different individuals through the study of SNPs [126, 132]. Likewise, it is important to know the regulation of genes in the expression of proteins that first carries out the analysis of promoter sequences, followed by the expression of the promoters and subsequently the expression of proteins [126, 133, 134]. Another study used for a rapid and systematic analysis of the expression of a large number of genes is the microarrays, which make it easier to observe the differential expression of genes from DNA or cDNA, as well as, allowing the finding gene functions novel and unexpected [135]. In addition, compare the pattern of gene expression under different conditions [136]. SAGE serial analysis of gene expression based on the study of cDNA allows to examine gene expression in a cell [126]. To perform a functional genomic observation, an assembled and identified genome must be had, which does not contain gaps, to avoid erroneous annotations. Subsequently, the assembled genome is compared with a reference genome, which together allows to predict genes. Next, the mapped elements are combined, and the biological information that allows to define an optimal set of annotations or functions is assigned. At the end, the data will have to be validated, this is achieved through manual inspections, experimental checks and quality measures [137]. To perform the genome annotation there are computational tools, one of the most used and friendly is Blast2GO which is a bioinformatics platform for high quality functional annotations and analysis of genomic data sets [138]. The data obtained can be shared with the public through databases so that other researchers can access them. Currently, GEO of NCBI is the public functional genomics database

Reference organisms	Databases	References
Escherichia coli	https://www.genome.jp/kegg-bin/show_organism?org=eco	[141]
Saccharomyces cerevisiae	https://www.yeastgenome.org/	[142]
Arabidopsis thaliana	https://www.arabidopsis.org/	[143]
Caenorhabditis elegans	https://wormbase.org/#012-34-5	[144]
Drosophila melanogaster	http://www.flybase.org/	[145]
Danio rerio	http://zfin.org/	[146]
Mus musculus	http://www.informatics.jax.org/	[147]
Homo sapiens: variation in humans	https://www.genome.jp/kegg-bin/show_organism?org=hsa	[148]

Table 2.
Databases of reference organisms used for genomic analysis.

that provides tools that help users in the consultation and download of data [139]. Likewise, KEGG is a database that is used as a tool to understand the high-level functions and utilities of the biological system, such as the cell, the organism or the ecosystem, based on molecular level information, generated by sequencing of the genome and other high performance [140]. There are also databases that store specific information on each of the most important model organisms (**Table 2**).

7. Phylogeny in the protein evolutionary process

The sequencing of the genome of an organism, has allowed to know the set of all its genes, elucidating the functions and products that they express, as well as the mechanisms of regulation in different metabolic processes, where endless proteins participate. To determine their possible functions, biochemical and genetic analyzes are used in a classical way, however, sequencing has contributed to the knowledge about the type of amino acids that make it up, and through the use of software multiple sequences have been aligned, where they have those that have been fully characterized as well as proteins where their biochemical characteristics are unknown and by homology between amino acids can be inferred in the functions that these proteins can present [149]. The use of bioinformatics, in protein analysis is a challenge, in recent years, phylogenetic profiles have been fundamental to relate homologous proteins by aligning their sequences, where it has been revealed that many share highly conserved regions and similar structures [150]. Phylogeny analyzes the changes that occur within the sequences and groups them in a diagram with ramifications, called a phylogenetic tree, all those sequences that belong to the same family can be grouped into a clade and in turn into subfamilies, providing data on their evolution and functional diversity [151].

Eukaryotic cells during their evolution have captured microorganisms that originated mitochondria, chloroplasts and other organelles, where their genes have been transferred to the nuclear genome, allowing the transport of encoded proteins in the nucleus. The different locations of proteins in the cell, and the different proteins that participate in cellular processes, have originated phylogenetic analyzes on the location of proteins in the cell, finding that they are closely related to prokaryotic proteins that have eukaryotes. The proteins of chloroplasts and mitochondria have a composition of amino acids, length, sequences and conserved regions very similar to those of prokaryotes [152, 153]. One of the limitations to analyze proteins among

related organisms is that genomes must be complete, in order to determine the presence or absence of genes in these species [154].

The high number of sequences that are stored in the different databases, have allowed to infer in the evolutionary relationships of different proteins, which when presenting homology retain their function during long evolutionary times, however, homologous proteins can perform the same activity, but the substrates they use can come from different routes [155]. When organisms adapt to different environmental conditions they cause mutational changes in genome sequences, causing amino acid substitutions in enzymes, making them improve their efficiency and specificity, to maintain their catalytic function. Not all genes that code for proteins are susceptible to mutation, due to the presence of essential amino acids in function, stability and folding, and therefore a restriction is generated. Many of the mutations are usually random and, in those proteins, where these changes have been observed, it is due to an evolutionary pressure. If the protein plays an important role in the functions of the organism and the mutation brings improvements in activity, the change in the genome is maintained and optimized, favored by selective pressure, otherwise, when the function of the protein is not relevant. In the cell, the mutant gene is removed from the genome by random deletions. Evolutionary mechanisms have given rise to homologous protein families, which share a common ancestor [155]. The study of ancestral enzymes has suggested that these presented a high thermostability, due to the Precambrian era that was thermophilic, in addition to the fact that most microorganisms and other organisms adapted to these environments with high temperatures. The ancestral protein alignments with the current ones show evidence of a slow evolution in structure, but not in amino acids [156]. Therefore, enzymes are the product of years of evolution, where they have undergone changes to obtain a specific function, as well as greater affinity with the substrate and/or act on multi-substrates. Therefore, the genetic variability has generated homologous genes (they descend from a common ancestor and are called orthologs) that encode adapted proteins to perform their catalysis in extreme conditions. However, there are also paralogous genes, which have diverged, to encode proteins with different activities [157], many times a particular characteristic is preserved, such as the binding of a molecule or reaction mechanism, but they specialize in carrying out the same reaction but on different substrates, different regulation mechanisms, as well as cell localization. On the other hand, orthologous proteins tend to have the same function and their sequences have a high conservation [155].

To analyze these changes in the sequences, bioinformatics programs use algorithms and mathematical models, based on empirical matrices of amino acid substitution, as well as those that incorporate structural properties of the native state, such as secondary structure and accessibility [158]. Protein phylogeny studies are currently necessary to know protein-protein interactions in biological systems. Molecular or structural analyzes on proteins will require more information to respond if a protein is present in one or several species, as well as to predict the common ancestor and evolution times [159]. There are different methods to estimate the genetic distance of proteins, among the most used are the minimum distance, which predicts the phylogenetic relationship minimizing the total distance of the pairs of sequences adjacent nodes tree. While those of maximum parsimony and maximum likelihood, use the multiple sequence alignment, however, the maximum parsimony maximum builds a tree minimizing the total evolutionary changes between adjacent proteins and the maximum likelihood tries to minimize the probability of making such changes. The bioinformatics tools that use these algorithms are: TOPAL, Hennig86 and PAML, the computational packages that allow to occupy any of these are PHYLIP and PAUP, as well as MOLPHY, PASSML, PUZZLE, TAAR [160].

8. Protein modeling

One of the challenges of protein engineering and biology is to improve industrial processes, to achieve this it is necessary to determine the tertiary structure of proteins from the amino acid sequence, in order to design new proteins and even new medicines. Many of the protein structures that we know today have been obtained through experimentation by X-ray crystallography, NMR spectroscopy or cryo-EM, however, the large amount of proteins, makes these processes require more time and increase costs [161]. Modeling through bioinformatics programs has managed to predict the atomic structure of several proteins from their amino acid sequence, by comparison with known protein structures, commonly called templates, although these do not present an accuracy with traditional techniques, the processes are faster and more economical in addition to providing low resolution data during sequence comparison [162, 163]. If the protein studied presents a homolog of known structure, the analysis is easier and the generated model is of higher resolution, but if the homologs do not exist or are not identified, the modeling is constructed from scratch [164]. De novo modeling is based on the assembly of proteins using short peptide fragments, originating from known proteins based on similarity, although advances have been made using this process, it has only worked on proteins that contain less than 100 amino acids, on large proteins size is difficult to analyze due to lack of information, as well as the type of software used [161, 165].

The 3D protein structures provide data at the molecular level, functions and properties, among which are the study of the catalytic mechanism, design and improvement of ligands, union of macromolecules with proteins, functional relationships through structural similarity and identification of conserved residues [55]. The interest in finding new protein models is generating a large amount of data, which is being stored in different databases, including Protein Data Bank, where the coordinates of the experimentally obtained atoms are stored; until 2014 this base contained more than 80 million sequences and more than 100,000 experimentally obtained 3D structures [166, 167]. These data have allowed the classification of proteins in different hierarchical levels as family, superfamily and fold in relation to their structure and evolution. All those that are grouped into a family are evolutionarily related to high sequence similarity. It is suggested that the different families that maintain a structure and function, present a common ancestor and are grouped into superfamilies and the difference between these is due to the folds or secondary structure that they possess [160]. In the last decade, the predictions by computational models have revealed the structure and function of many proteins, but the advances have been in some cases slow and expensive, due to the programming methods used and the precision of these during modeling. Currently working on automated bioinformatics servers that will generate models with a high percentage of accuracy [168, 169]. One of the most used servers worldwide is SWIIS-MODEL, which was the first to model proteins through homology, and in recent years has been automated allowing complex modeling, as well as the introduction of the modeling engines ProMod3 and QMEAN [167, 170, 171].
Most modeling algorithms use the following steps: (1) Identification of related structures, (2) template choice, (3) target sequence alignment with templates, (4) molding construction, (5) model evaluation. However, one of the limitations during homology protein modeling is the choice of model proteins or templates as well as alignments against the problem sequences [172, 173]. When the similarity of the sequences between the problem protein and that of the databases is low, the relationship and alignment can be improved if structural information is included during the analysis [166]. Advances in biocomputing have allowed the generation of tools for modeling proteins that are more reliable and easier to use, reducing

time and cost in the analysis. However, it is necessary to carry out experimentation to confirm that the prediction is correct, in addition to improving the efficiency of the techniques and with more known protein sequences and stored in the databases, therefore the different bioinformatics tools will play an important role in the postgenomic era [160].

9. Conclusions

Bioinformatics has evolved with daily work, which has allowed us to know how the biological molecules of a cell interact for their proper functioning, in addition to predicting various biological phenomena. In the last decade, the omic sciences have generated a great amount of data increasing the knowledge of the biological functions so that in the future they are able to predict diseases or formulate drugs with greater efficiency, however it is still necessary, to have a higher percentage of sequenced genes of the different organisms, as well as protein sequences, that allow enriching the databases, and with this more precise mathematical models are generated, which will benefit the computer programs so that they are more efficient, reliable, easy to use, reducing time and cost in the analyzes. This discipline becomes an essential part of biological studies every day, so its expansion and growth will be infinite, due to the evolutionary changes that are taking place in the cells caused by the different environmental phenomena.

Conflict of interest

The authors declare no conflict of interest.

Author details

Edna María Hernández-Domínguez[1], Laura Sofía Castillo-Ortega[1], Yarely García-Esquivel[1], Virginia Mandujano-González[2], Gerardo Díaz-Godínez[3] and Jorge Álvarez-Cervantes[1*]

1 Universidad Politécnica de Pachuca, Mexico

2 Universidad Tecnológica de la Corregidora, QRO, Mexico

3 Centro de Investigación en Ciencias Biológicas, Universidad Autónoma de Tlaxcala, Ixtacuixtla, Tlaxcala, Mexico

*Address all correspondence to: jorge_ac85@upp.edu.mx

References

[1] Benitez A, Cárdenas S. Bioinfomática en Colombia: Presente y futuro de la investigación biocomputacional. Biomédica. 2010;**3**:170-177. DOI: 10.7705/biomedica.v30i2.180

[2] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR. The protein data bank: A computer-based archival file for macromolecular structures. Journal of Molecular Biology. 1977;**112**:535-542. DOI: 10.1111/j.1432-1033.1977.tb11885.x

[3] Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981;**147**:195-197. DOI: 10.1016/0022-2836(81)90087-5

[4] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;**227**:1435-1441. DOI: 10.1126/science.2983426

[5] 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

[6] Meneses-Escobar CA, Rozo Murillo LV, Franco SJ. Tecnologías bioinformáticas para el análisis de secuencias de ADN. Scienctia et Technica. 2011;**16**:116-121

[7] Bustos RLS, Moreno LRD, Néstor D. Modelo de una bodega de datos para el soporte a la investigación bioinformática. Scientia et Technica. 2011;**16**:145-152

[8] Quíceno AHV. Bioinformática un Campo por conocer. Revista Electrónica de Veterinaria. 2006;**7**:1-9

[9] Harjinder SG, Prakash CR. Data Warehousing. La Integración de Información para la Mejor Toma de Decisiones. México: Prentice Hall; 1996. 382p

[10] Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, et al. Ligand depot: A data warehouse for ligands bound to macrolecules. Bioinformatics. 2004;**20**:2153-2155. DOI: 10.1093/bioinformatics/bth214

[11] Judice LYK, Vladimir B. Database warehousing in bioinformatics. In: Bioinformatics Technologies. Berlin Heidelberg: Springer-Verlag; 2005. pp. 45-62. DOI: 10.1007/b138246

[12] Shualev V. Metabolomics technology and bioinformatics. Briefings in Bioinformatics. 2006;**7**:128-139. DOI: 10.1093/bib/bbl012

[13] Patti G, Yanes O, Siuzdak G. Metabolomics: The apogee of the omic triology. NIH Public Access. 2013;**13**:263-269. DOI: 10.1038/nrm3314

[14] Dalgliesh C, Horning E, Horning M, Knox K, Yarger K. A gas-liquid-chromatographic procedure for separating a wide range of metabolites occurring in urine or tissue extracts. The Biochemical Journal. 1966;**101**:792-810. DOI: 10.1038/nrm3314

[15] Horning E, Horning M. Metabolic profiles: Gas-phase methods for analysis of metabolites. Clinical Chemistry. 1971;**17**:802-809

[16] Ghezzi P, Floridi L, Boraschi D, Cuadrado A, Manda G, Levic S, et al. Oxidative stress and inflammation induced by environmental and psychological stressors: A biomarker perspective. Antioxidants & Redox Signaling. 2018;**20**:852-872. DOI: 10.1089/ars.2017.7147

[17] Kovatchev B. Diabetes technology: Markers, monitoring, assessment, and control of blood glucose fluctuations in diabetes. Scientifica (Cairo). 2012;**2012**:1-14. DOI: 10.6064/2012/283821

[18] Pourfarzam M, Zadhoush F. Newborn screening for inherited metabolic disorders; news and views. Journal of Research in Medical Sciences. 2013;**18**:801-808

[19] Jan S, Ahmad P. Ecometabolomics. Metabolic Fluxes versus Environmental Stoichiometry. Introducing Metabolomics. 1st ed. Cambridge: Academic Press; 2019. pp. 1-56

[20] Johnson C, Ivanisevic J, Benton H, Siuzdak G. Bioinformatics: The next frontier of metabolomics. Analytican Chemistry. 2015;**18**:801-808. DOI: 10.1021/ac5040693

[21] Johnson C, Patterson A, Idle J, González F. Xenobiotic metabolomics: Major impact on the metabolome. HHS Public Access. 2012;**52**:37-56. DOI: 10.1146/annurev-pharmtox-010611-134748

[22] Oliver S, Winson M, Kell D, Baganz F. Systematic functional analysis of the yeast genome. Trends in Biotechnology. 1998;**16**:373-378. DOI: 10.1016/S0167-7799(98)01214-1

[23] Kanehisa M, Goto S. KEGG: Kyptp encyclopedia of genes and genomes. Nucleic Acids Research. 2000;**28**:27-30. DOI: 10.1093/nar/28.1.27

[24] Caspi R, Billington R, Fulcher C, Keseler I, Kothari A, Krummenacker M, et al. The MateCyc database of metabolic pathways and enzymes. Nucleic Acids Research. 2018;**46**:D633-D339. DOI: 10.1093/nar/gkx935

[25] Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. Annual Review of Genomics and Human Genetics. 2009;**10**:135-151. DOI: 10.1146/annurev-genom-082908-145957

[26] de Carvalho LM, Borelli G, Camargo AP, de Assis MA, Ferraz SMF, Fiamenghi MB, et al. Bioinformatics applied to biotechnology: A review towards bioenergy research. Biomass and Bioenergy. 2019;**123**:195-224. DOI: 10.1016/j.biombioe.2019.02.016

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

[28] Sedano JCS, Carrascal CEL. RNA-seq: herramienta transcriptómica útil para el estudio de interacciones planta-patógeno. Fitosanidad. 2012;**16**(2):101-113. DOI: 10.1093/bioinformatics/btr026

[29] Santana CIB. Buscando agujas en un pajar: viajes de RNAs pequenos in silico e in vitro. Acta Biológica Colombiana. 2011;**16**(3):103-113

[30] Peng M, Aguilar-Pontes MV, Hainaut M, Henrissat B, Hildén K, Mäkelä MR, et al. Comparative analysis of basidiomycete transcriptomes reveals a core set of expressed genes encoding plant biomass degrading enzymes. Fungal Genetics and Biology. 2018;**112**:40-46. DOI: 10.1016/j.fgb.2017.08.001

[31] Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proceedings of the National Academy of Sciences. 1977;**74**:5350-5354. DOI: 10.1073/pnas.74.12.5350

[32] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;**270**(5235):467-470. DOI: 10.1126/science.270.5235.467

[33] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science.

1995;**270**:484-487. DOI: 10.1126/science.270.5235.484

[34] Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, et al. Functional annotation of a full-length Arabidopsis cDNA collection. Science. 2002;**296**:141-145. DOI: 10.1126/science.1071006

[35] Marguerat S, Bähler J. RNA-seq: From technology to biology. Cellular and molecular life sciences. Reino Unido. 2010;**67**:569-579. DOI: 10.1007/s00018-009-0180-6

[36] Parkinson J, Blaxter M. Expressed sequence tags. In: Parasite Genomics Protocols. Totowa: Humana Press; 2004. pp. 93-126. DOI: 10.1385/1-59259-793-9:075

[37] Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. Eukaryotic Cell. 2010;**9**:1300-1310. DOI: 10.1128/EC.00123-10

[38] Notes T, FAQ F. FastQC Tutorial & FAQ [Internet]. Rtsf.natsci.msu.edu. 2019. Available from: https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/ [cited 30 August 2019]

[39] FASTX-Toolkit [Internet]. Bio.tools. 2019. Available from: https://bio.tools/fastx-toolkit [cited 30 August 2019]

[40] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics. 2014;**30**:2114-2120. DOI: 10.1093/bioinformatics/btu170

[41] Schmieder R, Edwards R. Qality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;**27**(6):863-864. DOI: 10.1093/bioinformatics/btr026

[42] Dodt M, Roehr J, Ahmed R, Dieterich C. FLEXBAR—Flexible barcode and adapter processing for next-generation sequencing platforms. Biology. 2012;**1**:895-905. DOI: 10.3390/biology1030895

[43] Strickler SR, Bombarely A, Mueller LA. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. American Journal of Botany. 2012;**99**:257-266. DOI: 10.3732/ajb.1100292

[44] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bíoínformatícs (Oxford, England). 2013;**29**:15-21. DOI: 10.1093/bioinformatics/bts635

[45] 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

[46] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;**8**:1494. DOI: 10.1038/nprot.2013.084

[47] Babarinde IA, Li Y, Hutchins AP. Computational methods for mapping, assembly and quantification for coding and non-coding transcripts. Computational and Structural Biotechnology Journal. 2019;**17**:628-637. DOI: 10.1016/j.csbj.2019.04.012

[48] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;**11**:R106. DOI: 10.1186/gb-2010-11-10-r106

[49] Liao Y, Smyth GK, Shi W. Feature counts: An efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;**30**:923-930. DOI: 10.1093/bioinformatics/btt656

[50] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods. 2008;**5**:621-628. DOI: 10.1038/nmeth

[51] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2009;**26**:493-500. DOI: 10.1093/bioinformatics/btp692

[52] Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. Genetics. 2010;**185**:405-416. DOI: 10.1534/genetics.110.114983

[53] edgeR: Differential expression analysis of digital gene expression data [Internet]. 1st ed. 2008. Available from: http://chagall.med.cornell.edu/RNASEQcourse/edgeRUsersGuide-2018.pdf [cited 30 August 2019]

[54] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013;**31**:46. DOI: 10.1038/nbt.2450

[55] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nature Methods. 2010;7:909. DOI: 10.1093/gigascience/giz039

[56] Marcel H, Schulz Daniel R, Zerbino MV, Ewan B. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;**28**:1086-1092. DOI: 10.1093/bioinformatics/bts094

[57] 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**:26-334. DOI: 10.1093/bioinformatics/btt219

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

[59] Kent WJ. BLAT—The BLAST-like alignment tool. Genome Research. 2002;**12**:656-664. DOI: 10.1101/gr.229202

[60] Wu TD, Watanabe CK. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;**21**:1859-1875. DOI: 10.1093/bioinformatics/bti310

[61] Hoff KJ, Stanke M. Predicting genes in single genomes with augustus. Current Protocols in Bioinformatics. 2019;**65**:57. DOI: 10.1002/cpbi.57

[62] Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Research. 2013;**41**:e74-e74. DOI: 10.1093/nar/gkt006

[63] Wucher V, Legeai F, Hedan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research. 2017;**45**:e57-e57. DOI: 10.1093/nar/gkw1306

[64] Sun L, Liu H, Zhang L, Meng J. lncRScan-SVM: A tool for predicting long non-coding RNAs using support vector machine. PLoS One. 2015;**10**(10):e0139654. DOI: 10.1371/journal.pone.0139654

[65] Fiannaca A, La Rosa M, La Paglia L, Rizzo R, Urso A. nRC: Non-coding RNA classifier based on structural features. BioData Mining. 2017;**10**:1-27. DOI: 10.1186/s13040-017-0148-2

[66] Langmead B. Aligning short sequencing reads with bowtie. Current

Protocols in Bioinformatics. 2010;**32**:11-17. DOI: 10.1002/0471250953.bi1107s32

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

[68] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017;**14**:417. DOI: 10.1038/nmeth.4197

[69] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;**34**:525. DOI: 10.1038/nbt.3519

[70] Hutchins AP, Jauch R, Dyla M, Miranda-Saavedra D. A framework for combining, analyzing and displaying heterogeneous genomic and high-throughput sequencing data. Cell Regeneration. 2014;**3**:1-15. DOI: 10.1186/2045-9769-3-1

[71] Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnology. 2008;**26**:407. DOI: 10.1038/nbt1394

[72] An J, Lai J, Lehman ML, Nelson CC. miRDeep*: An integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Research. 2012;**41**:727-737. DOI: 10.1093/nar/gks1187

[73] Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: A ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Research. 2011;**39**:112-117. DOI: 10.1093/nar/gkr357

[74] Breakfield NW, Corcoran DL, Petricka JJ, Shen J, Sae-Seaw J, Rubio-Somoza I, et al. High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. Genome Research. 2012;**22**:163-176. DOI: 10.1101/gr.123547.111

[75] Busch A, Richter AS, Backofen R. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics. 2008;**24**:2849-2856. DOI: 10.1093/bioinformatics/btn544

[76] Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, et al. CopraRNA and IntaRNA: Predicting small RNA targets, networks and interaction domains. Nucleic Acids Research. 2014;**42**:119-123. DOI: 10.1093/nar/gku359

[77] Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, et al. Comparative genomics boosts target prediction for bacterial small RNAs. Proceedings of the National Academy of Sciences. 2013;**110**:487-496. DOI: 10.1073/pnas.1303248110

[78] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nature Genetics. 2005;**37**:495. DOI: 10.1038/ng1536

[79] Lall S, Grün D, Krek A, Chen K, Wang YL, Dewey CN, et al. A genome-wide map of conserved microRNA targets in *C. elegans*. Current Biology. 2006;**16**:460-471. DOI: 10.1016/j.cub.2006.01.050

[80] Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000;**405**:837-846. DOI: 10.1038/35015709

[81] Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001;**294**:93-96. DOI: 10.1126/science.1065659

[82] Seaton D, Graf K, Baerenfaller M, Stitt A, Millar A, Gruissem W. Photoperioric control of the Arabidopsis proteome reveals a translational coincidence mechanism. Molecular Systems Biology. 2018;**14**:e7962. DOI: 10.15252/msb.20177962

[83] Yanovsky M, Kay S. Molecular basis of seasonal time measurement in Arabidopsis. Nature. 2002;**419**:308-312. DOI: 10.1038/nature00996

[84] Blueggel M, Chamrad D, Meyr H. Bioinformatics in proteomics. Current Pharmaceutical Biotechnology. 2004;**5**:79-88. DOI: 10.1201/9781420027524

[85] Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. BMC Systems Biology. 2014;**8**:1-7. DOI: 10.1186/1752-0509-8-S2-S3

[86] Popov I, Nenov A, Petrov P, Vassilev D. Bioinformatics in proteomics: A review on methods and algorithms. Biotechnology and Biotechnological Equipment. 2009;**23**:1115-1120. DOI: 10.1080/13102818.2009.10817624

[87] Smoot M, Guerlain S, Pearson W. Visualization of near-optimal sequence alignments. Bioinformatics. 2004;**20**:953-958. DOI: 10.1371/journal.pone.0178059

[88] Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;**48**:443-453. DOI: 10.1016/0022-2836(70)90057

[89] Barton G. Sequence alignment for molecular replacement. Acta Crystallographica. 2007;**64**:25-32. DOI: 10.1107/S0907444907046343

[90] Johnson M, Zaretskaya I, Raytselis Y, Merezhuj Y, McGinnis S, Madden T. NCBI BLAST: A better web interface. Nucleic Acids Research. 2008;**36**:5-9. DOI: 10.1093/nar/gkn201

[91] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel R, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. 2003;**31**:3784-3788. DOI: 10.1093/nar/gkg563

[92] Rose P, Bojan B, Chunxiao B, Wolfgang B, Dimitris D, David G, et al. The RCSB protein data bank: Redesigned web site and web services. Nucleic Acids Research. 2011;**39**:392-401. DOI: 10.1093/nar/gkg1021

[93] Aslam B, Basit M, Nisar M, Khurshid M. Proteomics: Technologies and their applications. Journal of Chromatographic Science. 2017;**55**:182-196. DOI: 10.1093/chromsci/bmw167

[94] Stroggilos R, Mokou M, Latosinska A, Makridakis M, Lygirou V, Mavrogeorgis E, et al. Proteome-based classification of non-muscle invasive bladder cancer. International Journal of Cancer. 2019. DOI: 10.1002/ijc.32556

[95] Chaudhary H, Nameirakpam J, Kumrah R, Pandiarajan V, Suri D, Rawat A, et al. Biomarkers for Kawasaki disease: Clinical utility and the challenges ahead. Frontiers in Pediatrics. 2019;**7**:1-10. DOI: 10.3389/fped.2019.00242

[96] Yatoo M, Parray R, Bhat R, Nazir Q, Haq A, Malik U, et al. Novel candidates for vaccine development against *Mycoplasma capricolum* subspecies Capripneumoniae (Mccp)—Current knowledge and future prospects. Vaccine. 2019;**7**:2-21. DOI: 10.3390/vaccines703007

[97] Burgos-Canul Y, Canto-Canché B, Berezovski M, Mironov G, Loyola-Vargas V, Barba de Rosa A, et al. The cell wall proteome from two

strains of *Pseudocercospora fijiensis* with differences in virulence. World Journal of Microbiology and Biotechnology. 2019;**35**:105. DOI: 10.1007/s11274-019-2681-2

[98] Parolo S, Marchetti L, Lauria M, Misselbeck K, Scott-Boyer M, Caberlotto L, et al. Combined use of protein biomarkers and network analysis unveils deregulated regulatory circuits in Duchenne muscular dystrophy. PLoS One. 2018;**13**:e0194225. DOI: 10.1371/journal.pone.0194225

[99] Hu B, Xie G, Lo C, Starkenburg SR, Chain PSG. Pathogen comparative genomics in the next-generation sequencing era: Genome alignments, pangenomics and metagenomics. Briefings in Functional Genomics. 2011;**10**:322-333. DOI: 10.1093/bfgp/elr042

[100] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Nature Genetics. 2000;**25**:25-29. DOI: 10.1038/75556

[101] Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. The bacterial pan-genome: A new paradigm in microbiology. International Microbiology. 2010;**13**:45-57. DOI: 10.2436/20.1501.01.110

[102] Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. Nature Reviews. Microbiology. 2012;**10**:599-606. DOI: 10.1038/nrmicro2850

[103] Stahl PL, Lundeberg J. Toward the single-hour high-quality genome. Annual Review of Biochemistry. 2012;**81**:359-378. DOI: 10.1146/annurev-biochem-060410-094158

[104] Edwards DJ, Holt KE. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microbial Informatics and Experimentation. 2013;**3**:2. DOI: 10.1186/2042-5783-3-2

[105] Hardison RC. Comparative genomics. PLoS Biology. 2003;**1**:156-160. DOI: 10.1371/journal.pbio.0000058

[106] Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: The bacterial pan-genome. Current Opinion in Microbiology. 2008;**11**:472-477. DOI: 10.1016/j.mib.2008.09.006

[107] Mosquera-Rendón J, Rada-Bravo AM, Cárdenas-Brito S, Corredor M, Restrepo-Pineda E, Benítez-Páez A. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. BMC Genomics. 2016;**17**(45):1-14. DOI: 10.1186/s12864-016-2364-4

[108] Zekic T, Holley G, Stoye J. Pan-genome storage and analysis techniques. In: Setubal JC, Peter JS, Stadler F, editors. Comparative Genomics Methods and Protocols. Totowa: Humana Press; 2018. pp. 29-54. DOI: 10.1007/978-1-4939-7463-4. ch2

[109] Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, et al. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Research. 2001;**11**:1175-1186. DOI: 10.1101/gr.182901

[110] Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. The Plant Cell. 2014;**26**:121-135. DOI: 10.1105/tpc.113.119982

[111] Weigel D, Mott R. The 1001 genomes project for *Arabidopsis*

thaliana. Genome Biology. 2009;**10**:107. DOI: 10.1186/gb-2009-10-5-107

[112] Huang S, Zhang S, Jiao N, Chen F. Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic cyanopodoviruses. PLoS One. 2015;**10**: 1-17. DOI: 10.1371/journal.pone.0142962

[113] Rubin GM, Yandell MD, Wortman JR, Miklos GLG, Nelson CR, Hariharan IK, et al. Comparative genomics of the eukaryotes. Science. 2000;**287**:2204-2215. DOI: 10.1007/978-1-4939-7463-4_3

[114] Hassan YI, Lepp D, Zhou T. Next-generation whole-genome sequencing platforms and factors to consider for bacterial applications. Journal of Microbiology, Biotechnology and Food Sciences. 2015;**5**:29-33. DOI: 10.15414/jmbfs.2015.5.1.29-33

[115] Blom J, Kreis J, Sp€anig S, Juhre T, Bertelli C, Ernst C, et al. EDGAR 2.0: An enhanced software platform for comparative gene content analyses. Nucleic Acids Research. 2016;**44**:22-28. DOI: 10.1093/nar/gkw255

[116] Brittnacher MJ, Fong C, Hayden HS, Jacobs MA, Radey M, Rohmer L. PGAT: A multistrain analysis resource for microbial genomes. Bioinformatics. 2011;**27**:2429-2430. DOI: 10.1093/bioinformatics/btr418

[117] Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: Pan-genomes analysis pipeline. Bioinformatics. 2012;**28**:416-418. DOI: 10.1093/bioinformatics/btr655

[118] Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, et al. PanGP: A tool for quickly analyzing bacterial pan-genome profile. Bioinformatics. 2014;**30**:1297-1299. DOI: 10.1093/bioinformatics/btu017

[119] Sahl JW, Gregory Caporaso J, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR) pipeline: A method to rapidly compare genetic content between bacterial genomes. PeerJ. 2014;**2**:e332. DOI: 10.7717/peerj.332

[120] Chaudhari NM, Gupta VK, Dutta C. BPGA-an ultra-fast pan-genome analysis pipeline. Scientific Reports. 2016;**6**:1-10. DOI: 10.1038/srep24373

[121] Galperin MY, Koonin EV. Comparative genome analysis. In: Baxevanis AD, Francis Ouellette BF, editors. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd ed. Hoboken: John Wiley & Sons, Inc.; 2001. pp. 359-392. DOI: 10.1093/bib/bbk012

[122] Wattam AR, Thomas Brettin T, James J, Davis JJ, Svetlana Gerdes S, Kenyon R, et al. Assembly, annotation, and comparative genomics in PATRIC, the all bacterial bioinformatics resource center. In: Setubal JC, Peter JS, Stadler F, editors. Comparative Genomics Methods and Protocols. 1st ed. Totowa: Humana Press; 2018. pp. 79-102. DOI: 10.1007/978-1-4939-7463-4

[123] Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, et al. PANNOTATOR: An automated tool for annotation of pan-genomes. Genetics and Molecular Research. 2013;**12**:2982-2989. DOI: 10.4238/2013

[124] Angiuoli SV, Hotopp JCD, Salzberg SL, Tettelin H. Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinformatics. 2011;**12**:272-283. DOI: 10.1186/1471-2105-12-272

[125] Pevsner J. Bioinformatics and Functional Genomics. 3rd ed. Hoboken:

Wiley Blackwell; 2015. pp. 635-695. DOI: 10.1002/9780470451496

[126] Kaushik S, Sharma D. Functional genomics. Reference module in life sciences. Encyclopedia of Bioinformatics and Computational Biology. 2018. DOI: 10.1016/b978-0-12-809633-8.20222-7

[127] Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, et al. Potential of metabolomics as a functional genomics tool. Trends in Plant Science. 2004;**9**:418-425. DOI: 10.1016/j.tplants.2004.07.004

[128] Miller W, Makova KD, Nekrutenko A, Hardison RC. Comparative genomics. Annual Review of Genomics and Human Genetics. 2004;**5**:15-56. DOI: 10.1146/annurev.genom.5.061903.180057

[129] Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, et al. The functional genomics experiment model (FuGE): An extensible framework for standards in functional genomics. Nature Biotechnology. 2007;**25**:1127-1133. DOI: 10.1038/nbt1347

[130] Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, et al. From gene networks to gene function. Genome Research. 2003;**13**:2568-2576. DOI: 10.1101/gr.1111403

[131] Boucher B, Jenna S. Genetic interaction networks: Better understand to better predict. Frontiers in Genetics. 2013;**4**:1-16. DOI: 10.3389/fgene.2013.00290

[132] Karchin R. Next generation tools for the annotation of human SNPs. Briefings in Bioinformatics. 2009;**10**: 35-52. DOI: 10.1093/bib/bbn047

[133] Zhu J, Zhang MQ. SCPD: A promoter database of the yeast *Saccharomyces cerevisiae*. Bioinformatics. 1999;**15**:607-611. DOI: 10.1093/bioinformatics/15.7.607

[134] Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jiménez-Jacinto V, Martínez-Flores I, et al. Bioinformatics resources for the study of gene regulation in bacteria. Journal of Bacteriology. 2009;**191**:23-31. DOI: 10.1128/JB.01017-08

[135] Slonim K, Yanai I. Getting started in gene expresión microarray analysis. PLoS Computational Biology. 2009;**5**:e1000543. DOI: 10.1371/journal.pcbi.1000543

[136] Miller MB, Tang YW. Basic concepts of microarrays and potential applications in clinical microbiology. Clinical Microbiology Reviews. 2009;**22**:611-633. DOI: 10.1128/CMR.00019-09

[137] Alvarado VJ. Anotación de genoma. Conogasi.org 2019. Sitio web: http://conogasi.org/articulos/anotacion-de-genoma/ [cited 18 August 2019]

[138] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;**21**:3674-3676. DOI: 10.1093/bioinformatics/bti610

[139] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: Archive for high-throughput functional genomic data. Nucleic Acids Research. 2009;**37**:885-890. DOI: 10.1093/nar/gkn764

[140] KEGG: Kyoto Encyclopedia of Genes and Genomes. Available from: https://www.genome.jp/kegg/ [cited 17 August 2019]

[141] Brown SD, Jun S. Complete genome sequence of *Escherichia coli* NCM3722. Genome Announcements.

2015;**3**(4):00879-15. DOI: 10.1128/genomea.00879-15

[142] Saccharomyces genoma database. 2019. Available from: https://www.yeastgenome.org/ [17 August 2019]

[143] Tair Phoenix bioinformatics. 2019. Available from: https://www.arabidopsis.org [17 August 2019]

[144] WormBase versión: WS271. 2019. Available from: https://wormbase.org/#012-34-5 [17 August 2019]

[145] A Database of Drosophila Genes & Genomes. 2019. Available from: http://www.flybase.org [17 August 2019]

[146] The Zebrafish Information Network, University of Oregon. 2019. Available from: http://zfin.org/ [17 August 2019]

[147] Mouse Genome Informatics. The Jackson Laboratory. 2019. Available from: http://www.informatics.jax.org/. [17 August 2019]

[148] *Homo sapiens* (Human). 2019. Available from: https://www.genome.jp/kegg-bin/show_organism?org=hsa [17 August 2019]

[149] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates T, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. 1999;**285**:751-753. DOI: 10.1126/science.285.5428.751

[150] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates T. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of the National Academy of Sciences. 1999;**96**:4285-4288. DOI: 10.1073/pnas.96.8.4285

[151] Song L, Wu S, Tsang A. Phylogenetic analysis of protein family. In: de Vries R, Tsang A, Grigoriev I, editors. Fungal Genomics. Methods in Molecular Biology. New York, NY: Humana Press; 2018. pp. 267-275. DOI: 10.1007/978-1-4939-7804-5

[152] Margulis L. Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth. New Haven: Yale University Press; 1970. p. 349

[153] Marcotte EM, Xenarios I, Van der Bliek AM, Eisenberg D. Localizing proteins in the cell from their phylogenetic profiles. Proceedings of the National Academy of Sciences. 2000;**97**:12115-12120. DOI: 10.1073/pnas.220399497

[154] Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology. 2002;**12**:368-373. DOI: 10.1016/S0959-440X(02)00333-0

[155] Kaminska KH, Milanowska K, Bujnicki JM. The basics of protein sequence analysis. In: Bujnicki JM, editor. Prediction of Protein Structures, Functions, and Interactions. Hoboken: John Wiley & Sons, Ltd.; 2009. pp. 1-38. DOI: 10.1002/9780470741894

[156] Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. Biological Chemistry. 2016;**397**:1-21. DOI: 10.1515/hsz-2015-0158

[157] Tyzack JD, Furnham N, Sillitoe I, Orengo CM, Thornton JM. Understanding enzyme function evolution from a computational perspective. Current Opinion in Structural Biology. 2017;**47**:131-139. DOI: 10.1016/j.sbi.2017.08.003

[158] Bastolla U, Arenas M. The influence of protein stability on sequence evolution: Applications to phylogenetic inference. In: Sikosek T, editor. Computational Methods in Protein Evolution. New York, NY: Humana Press; 2019. pp. 215-231. DOI: 10.1007/978-1-4939-8736-8_11

[159] Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: Bridging the scales. Current Opinion in Structural Biology. 2018;**50**:26-32. DOI: 10.1016/j.sbi.2017.10.014

[160] Xu D, Xu Y, Uberbacher CE. Computational tools for protein modeling. Current Protein & Peptide Science. 2000;**1**:1-21. DOI: 10.2174/1389203003381469

[161] Cheung NJ, Yu W. De novo protein structure prediction using ultra-fast molecular dynamics simulation. PLoS One. 2018;**13**:e0205819. DOI: 10.1371/journal.pone.0205819.

[162] Bonneau R, Baker D. Ab initio protein structure prediction: Progress and prospects. Annual Review of Biophysics and Biomolecular Structure. 2001;**30**:173-189. DOI: 10.1146/annurev.biophys.30.1.173

[163] Hung L, Ngan S, Samudrala R. De novo protein structure prediction. In: Xu Y, Xu D, Liang J, editors. Computational Methods for Protein Structure Prediction and Modeling. New York: Springer; 2007. pp. 43-64. DOI: 10.1007/978-0-387-68825-1_2

[164] Lee J, Freddolino PL, Zhang Y. Ab initio protein structure prediction. In: Rigden DJ, editor. From Protein Structure to Function with Bioinformatics. Dordrecht: Springer; 2017. pp. 3-35. DOI: 10.1007/978-94-024-1069-3_1

[165] Shen Y, Bax A. Homology modeling of larger proteins guided by chemical shifts. Nature Methods. 2015;**12**:747. DOI: 10.1038/nmeth.3437

[166] Kelleym LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 2015;**10**:845. DOI: 10.1038/nprot.2015.053

[167] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Research. 2018;**46**:W296-W303. DOI: 10.1093/nar/gky427

[168] Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: Evaluation of the model quality estimates in CASP10. Proteins. 2014;**82**:112-126. DOI: 10.1002/prot.24347

[169] Yang J, Zhang Y. Protein structure and function prediction using I-TASSER. Current Protocols in Bioinformatics. 2015;**52**:5-8. DOI: 10.1002/0471250953.bi0508s52

[170] Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;**27**:343-350. DOI: 10.1093/bioinformatics/btq662

[171] Biasini M, Schmidt T, Bienert S, Mariani V, Studer G, Haas J, et al. OpenStructure: An integrated software framework for computational structural biology. Acta Crystallographica, Section D: Biological Crystallography. 2013;**69**:701-709. DOI: 10.1107/S0907444913007051

[172] Fiser A, Šali A. Modeller: Generation and refinement of homology-based protein structure models. In: Methods in Enzymology. Cambridge: Academic Press; 2003. pp. 461-491. DOI: 10.1016/S0076-6879(03)74020-8

[173] Song Y, DiMaio F, Wang RYR, Kim D, Miles C, Brunette TJ, et al. High-resolution comparative modeling with RosettaCM. Structure. 2013;**21**:1735-1742. DOI: 10.1016/j.str.2013.08.005