Applied Microbiology and Biotechnology
https://doi.org/10.1007/s00253-020-11056-2

MINI-REVIEW

Bioinformatics: new tools and applications in life science and personalized medicine

Iuliia Branco[1] · Altino Choupina[1]

Received: 29 November 2020 / Revised: 29 November 2020 / Accepted: 9 December 2020

Abstract

While we have a basic understanding of the functioning of the gene when coding sequences of specific proteins, we feel the lack of information on the role that DNA has on specific diseases or functions of thousands of proteins that are produced. Bioinformatics combines the methods used in the collection, storage, identification, analysis, and correlation of this huge and complex information. All this work produces an "ocean" of information that can only be "sailed" with the help of computerized methods. The goal is to provide scientists with the right means to explain normal biological processes, dysfunctions of these processes which give rise to disease and approaches that allow the discovery of new medical cures. Recently, sequencing platforms, a large scale of genomes and transcriptomes, have created new challenges not only to the genomics but especially for bioinformatics. The intent of this article is to compile a list of tools and information resources used by scientists to treat information from the massive sequencing of recent platforms to new generations and the applications of this information in different areas of life sciences including medicine.

Key points

- *Biological data mining*
- *Omic approaches*
- *From genotype to phenotype*

Keywords Sequencing · Bioinformatics · Tools · Applications · Life science · Personalized medicine

Introduction

The knowledge derived from genomic and computational technologies increases in geometric progression. The understanding of this avalanche of data is closely linked to the formidable development in the bioinformatics area. By enabling the overall assessment of this extraordinary amount of data, bioinformatics has considerably accelerated scientific discoveries. This growth has as a consequence a large supply of products, services, and information, so that keep up to date, locate, and use the latest innovations; it has become a full-time activity.

Although we initially tried to create a complete profile of all the available bioinformatics resources, quickly, it was evident dynamism and constant updating of this field surpassed this goal. We create a division into four categories: sequence analysis software, software prediction of protein structures, resource servers "online," and finally left a list of places of interest on the Internet that can shorten the search time. We opted for the selection of these categories because we believe that analyze in a comprehensive way the molecular biology central dogma.

Bioinformatics, as a scientific area, gathering techniques and tools from the subjects: molecular biology, source of information to be analyzed; informatics or computer science, provides the hardware for analysis and networks to share the results; mathematics, the origin of the algorithms used in the data analysis. The interrelationship of the three areas creates the basis for bioinformatics applications in molecular biology, as can be seen in the following diagram (Li et al. 2013) (Fig. 1).

✉ Altino Choupina
albracho@ipb.pt

[1] Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Published online: 06 January 2021

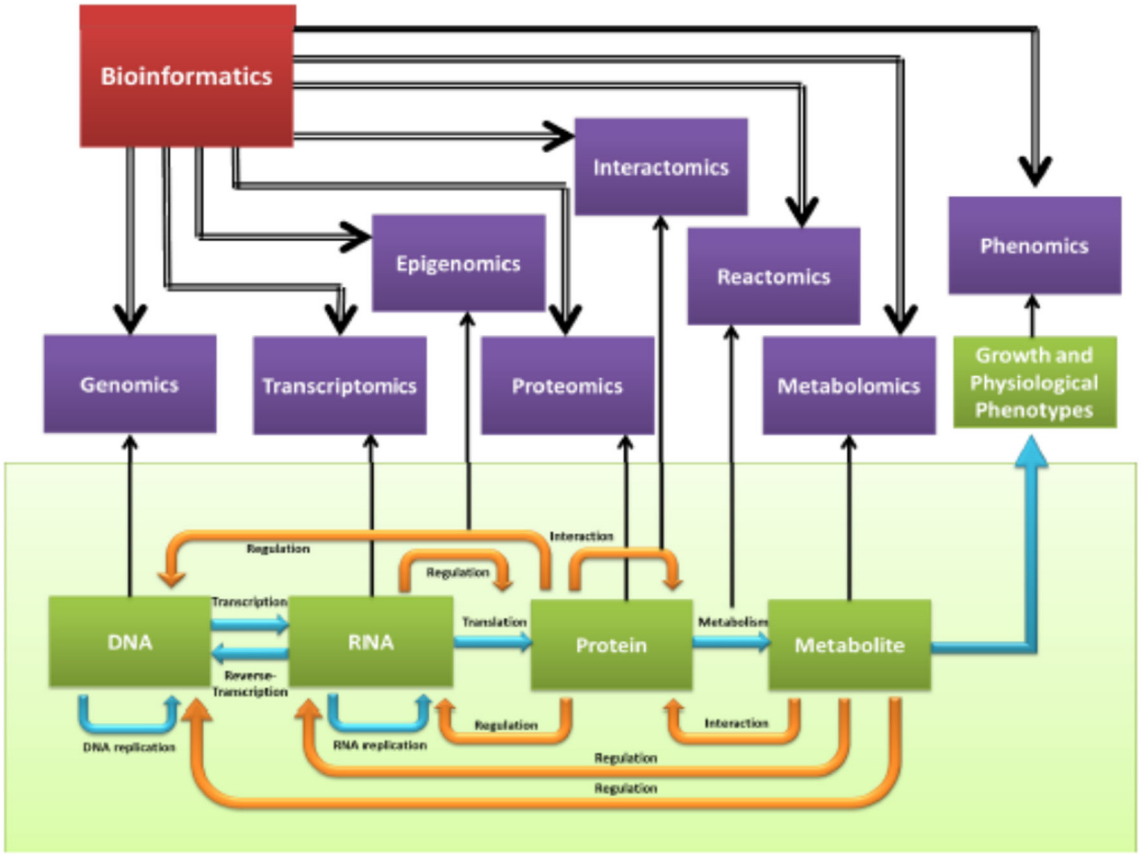

Fig. 1 Relationship of biological "-omics" with bioinformatics (Li et al. 2013)

We know that there have been many reviews of bioinformatics tools published, but in this review, we want to present in a simple way the fundamental and most useful tools for the life science researcher work, integrating several tools that can cover all omics. Tools were used since the design of the experiments, the obtaining of biological data, the deposit of these data, and their mineralization in order to deduce phenotypes and their interrelationships and applications both in molecular cloning in pharmacy and medicine. So this mini-review will be useful both for biotechnology researchers and for agronomists, zootechnics, ecologists, pharmacists, and medicine agents (Martins et al. 2014, 2019).

Deduce the order of DNA sequences is essential for basic biological research, with several important applications in biotechnology. The large capacity sequencing obtained with modern DNA sequencing technologies has been responsible for the immense, extraordinary, and complete sequencing of DNA sequences, or genomes, including the human genome. The first sequencing method "Sanger sequencing technique" based on the selective incorporation of chain-terminating dideoxynucleotides (ddNTP's) by DNA polymerase with capillary electrophoresis, automatic, was developed by Applied Biosystems (Namely AB370). These automated tools, with significant capacity sequencing, have been the main tool in the sequencing of various genomes and the human genome. These first genome projects were, in turn, a stimulus to the development of new and powerful platforms for sequencing called next-generation sequencing (NGS) (Heather and Chain 2016).

Next-generation sequencing systems

Next-generation sequencing (NGS) is a high-throughput methodology that allows massive base-pair sequencing in DNA or RNA samples. Making a large number of applications possible, including full sequencing of numerous genomes, the study of gene expression profiles, the study of epigenetic changes, the study of mutations, and molecular analysis, to make the future of personalized medicine possible (Goldman and Domschke 2014).

NGS systems include multiple platforms, the so-called second generation of sequencers, with different approaches and sequencing capabilities such as Life Sciences' SOLiD/Ion Torrent PGM, Illome's Genome Analyzer/HiSeq 2000/MiSeq, and Roche GS FLX Titanium/GS Junior. In the third generation of sequencers, the most popular platform is the single-molecule real-time (SMRT) sequencing is a parallelized single-molecule DNA sequencing method. Each of the four DNA bases is attached to one of four different

fluorescent dyes. When a nucleotide is incorporated by the DNA polymerase, a detector detects the fluorescent signal of the nucleotide incorporation, and the base call is made according to the corresponding fluorescence of the dye. Other sequencing platforms already from the fourth generation of sequencers, based on nanoporous, are developing with more data generation capabilities in less time and lower costs. Whichever platform you use, millions of data points are generated in hours, so getting data is no longer a problem, leading to a paradigm shift, where data processing, storage, and analysis become the task most relevant task. It is at these points that bioinformatics, with its ability to analyze large amounts of data with diversified objectives, assumes its essential role, also considering that each of the mentioned platforms incorporates a series of bioinformatics tools for processing the output data (Goldman and Domschke 2014; Kulski 2016).

Primary analysis of DNA sequences

The primary analysis of DNA sequences is essential in the daily life of the biotechnology laboratory, in the detection of mutations and the establishment of phylogenies, elaboration of restriction maps to make cloning, and cassettes for silencing genes in order to see their role in the cell metabolism.

In the genome analysis software, several program packages can be found, which accompany the entire process from receiving the sequencer graphics to publishing the data in online databases. These features, along with free access to academics, file compatibility, and their date of conception are the main factors in the choices made.

We point out that many of the services provided by these programs are also provided by some programs available online, the disadvantage that each query requires a network connection, but with the advantage that these online resources are updated regularly.

Staden package (http://staden.sourceforge.net/) (Bonfield and Whitwham 2010; Rodger et al. 2003a, b)

The very complete program package for nucleotide sequence analysis, free for students and researchers, allows requests via mail or directly from the network. Staden is very powerful and lends itself to automated processing of data; it is not very intuitive as it requires some learning but it is certainly an excellent work tool.

The Staden package was developed at the Medical Research Council (MRC) Laboratory of Molecular Biology, Cambridge, England, by Rodger Staden's group. The package was converted to open source in 2004, and several new versions have been released since.

The authors describe the current version of the sequence analysis package developed at the MRC Laboratory of Molecular Biology, which has come to be known as the "Staden package": "the package covers most of the standard sequence analysis tasks such as restriction site searching, translation, pattern searching, comparison, gene finding, and secondary structure prediction, and provides powerful tools for DNA sequence determination."

This package contains the following programs:

- Gap4 and Gap5: This program is the main tool of this package; it performs compilation, sequence merging, compilation rectification, reads sequence pairs, and allows editing them (Fig. 2);
- Pregap4: Allows the reception and analysis of the information from the sequencers constituting the information input port for this program package;
- Trev: Fast and effective, allows the visualization of sequences in ABI, ALF, or SRF formats;
- Trace diff: Automatically localizes mutation points by comparing the sequence understudy with reference sequences. It supports any number of sequences and allows the visualization of results by gap4;
- Sip4: Compares sequence pairs in various ways, often displaying results graphically. It allows a comparison between nucleotides between proteins and between proteins and nucleotides.
- Nip4: Analyze nucleotide sequences to find genes, restriction sites; allows translation, etc.

pDRAW32 (https://www.acaclone.com/)

A program to be used on the Windows platform, with a nice and intuitive interface, available for free on the Internet at the website (https://www.acaclone.com).

With this program, it is passive to perform various operations, such as annotations for the DNA understudy, cloning DNA, editing sequences, selecting restriction enzymes, exporting graphics and text, calculating the optimal PCR temperature, calculating homologies between two DNA fragments, and containing scientific aid files. Possibly one of the best programs for cloning strategies, extremely intuitive, so easy-to-use, and produces very beautiful, simple, and complete images.

GenBeans (http://www.genbeans.org/)

GenBeans is an integrated stand-alone platform for bioinformatics based on NetBeans (developed by Apache Software Foundation) open-source software. It focuses on molecular biology and provides a fully integrated toolbox in a rich, easy-to-use graphical interface for analyzing and visualizing sequences. Another interesting program based on NetBeans is

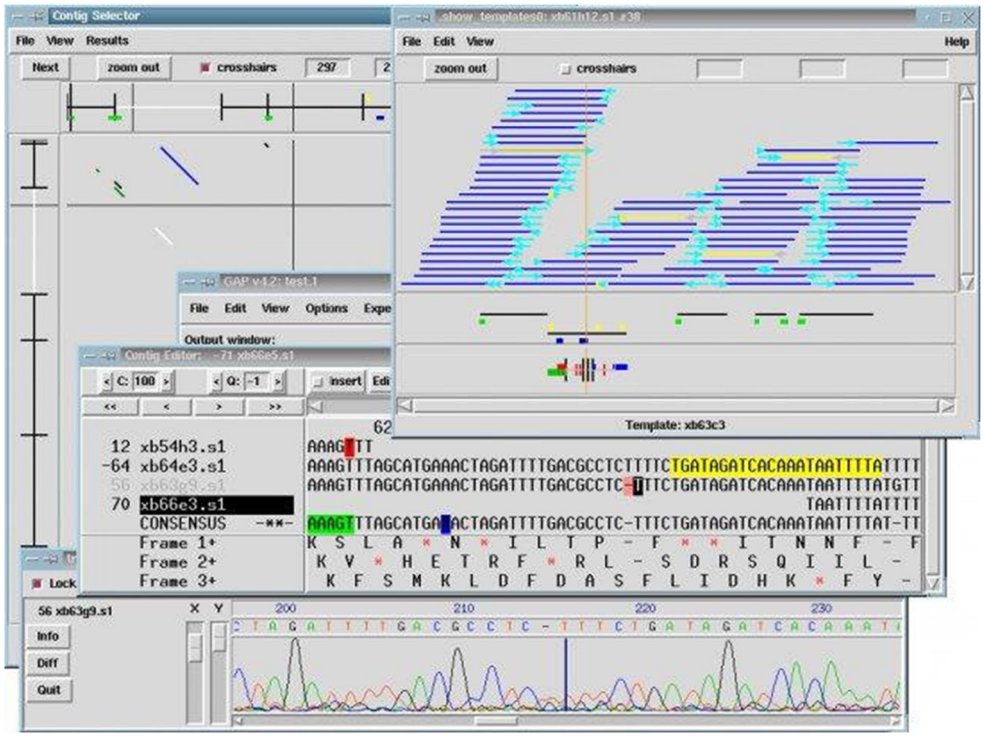

Fig. 2 Gap Interface (Staden Package Handbook) (Bonfield and Whitwham 2010; Rodger et al. 2003a, b)

geneinfinity (http://www.geneinfinity.org/) that we describe in Table 1.

DNASTAR™ (https://www.dnastar.com/)

Another computer package whose utilization has been getting big is the DNASTAR™; this package has programs with which we can edit and compare sequences, deduce physico-chemical characteristics, and do genetic constructions, restriction maps, etc.

Serial Cloner (http://serialbasics.free.fr/Serial_Cloner.html)

This program was developed at Institut Curie by Franck Perez. Serial Cloner is designed to provide molecular biology software for Macintosh and Windows users. It reads and writes DNA Strider compatible files and imports and exports files in universal FASTA format. It consists of graphical display tools and simple interfaces that help you analyze and build in a very intuitive way.

"The user interface is relatively simple to operate, in that it is within the reach of any user with advanced biology knowledge, who should be especially impressed with the huge amount of options available. With Serial Cloner you can, among other things, join DNA fragments obtained through PCR, manipulate the shRNA, or simply assemble fragments of different chains."

Sequencher (https://www.genecodes.com/)

"Gene Codes Corporation is a privately-owned international firm, which specializes in bioinformatics software for genetic sequence analysis. Its flagship software product, Sequencher, is a sequencing software used throughout the world. Its targeted use is by researchers at academic and government labs as well as for biotechnology and pharmaceutical companies for DNA sequence assembly."

Sequencher is a simple but useful program that allows us to:

- Analyze nucleic acid sequences in editing modes;
- Alignment with possible visualization of the various chromatograms;
- Perform manipulations and restriction maps.

The latest release of Sequencher highlights Gene Codes' goal of providing researchers with powerful, easy-to-use DNA analysis software tools. Sequencher 5.3 adds RNA-Seq analysis to its long list of DNA sequence analysis features, as well as improvements to Sequencher Connections, its newest architecture for DNA sequence analysis.

FastPCR (https://primerdigital.com/fastpcr.html)

FastPCR is an integrated tool for PCR primers or probe design, in silico PCR, oligonucleotide assembly and analyzes alignment and repeat searching developed by PrimerDigital

Table 1 Tools for biological sequences characterization

Tool	Description	References/URL
BLAST: Basic Local Alignment Search Tool	Compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches	(Altschul et al. 1990; Lobo 2008)
FASTA	This tool provides sequence similarity searching against protein databases using the FASTA suite of programs	https://www.ebi.ac.uk/Tools/sss/fasta/ (Madeira et al. 2019)
HMMER	Sequence analysis identifies homologous protein or nucleotide sequences and performs sequence alignments	https://www.ebi.ac.uk/Tools/hmmer/search/phmmer (Potter et al. 2018; Finn et al. 2011)
Clustal Omega	A new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate multiple alignments	https://www.ebi.ac.uk/Tools/msa/clustalo/ (Madeira et al. 2019)
Sequerome	Integrating the results of a BLAST sequence-alignment report with external research tools and servers that perform sequence manipulations	https://www.bioinformatics.org/sequerome/wiki/Main/HomePage (Ganesan et al. 2005)
ProtParam	Prediction of multiple physicochemical properties of proteins	https://web.expasy.org/protparam/ (Gasteiger et al. 2005)
JIGSAW	To find genes and predict splicing locations in the DNA sequence selected	(Allen and Salzberg 2005)
novoSNP	Used to find the unique nucleotide variation in the DNA sequence	(Weckx et al. 2005)
ORF Finder	Searches for open-reading frames (ORFs) in the DNA sequence you enter and verify predicted protein using SMART BLAST or regular BLASTP	https://www.ncbi.nlm.nih.gov/orffinder/
GENEinfinity	Informational resources, tools, and calculators to facilitate work at the bench and analysis of biological data	http://www.geneinfinity.org/
PPP	Finds promoter regions and TFBS in prokaryotes and also finds sigma A binding sites in the upstream region	http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp_start.php
Virtual Foorprint	Designed to analyze transcription factor binding sites in whole bacterial genomes and their underlying regulatory networks	(Münch et al. 2005)
WebGeSTer	Database of intrinsic transcription terminators	http://pallab.serc.iisc.ernet.in/gester (Mitra et al. 2011)
Genscan	Used for predicting the locations and exon-intron structures of genes in genomic sequences	https://bio.tools/genscan (Burge and Karlin 1997)
Softberry Tools	Animal, plant, and bacterial genomes annotation and RNA and proteins structures prediction	http://www.softberry.com/
GeneID	Ab initio gene finding program used to predict genes along DNA sequences in a large set of organisms	(Parra et al. 2000)
SpliceView	Tools for prediction and analysis of protein-coding gene structure	http://bioinfo.itb.cnr.it/~webgene/wwwspliceview.html
GeneBuilder	Based on prediction of functional signals and coding regions by different approaches in combination with similarity searches in proteins and EST	http://bioinfo.itb.cnr.it/~webgene/genebuilder.html
GeneFinder	To predict putative internal protein-coding exons in genomic DNA sequences	http://rulai.cshl.org/tools/genefinder/
HCPolyA	Tools for prediction and analysis of protein-coding gene structure	http://bioinfo.itb.cnr.it/~webgene/wwwHC_polya.html

(Kalendar et al. 2017a, b, c). PrimerDigital is a biotechnology company specialized in high-quality primer, probe design service, and software development that delivers state-of-the-art PCR software. From the wide experience we have in using FastPCR, we agree with the description of this software made by the company that we summarize: "The FastPCR software is an integrated tools environment that provides comprehensive and professional facilities for designing any kind of PCR primers for standard, long-distance, inverse, real-time PCR (TaqMan, LUX-primer, Molecular Beacon, Scorpion), multiplex PCR, Xtreme Chain Reaction (XCR), group-specific (universal primers for genetically related DNA sequences) or unique (specific primers for each from genetically related DNA sequences), overlap extension PCR (OE-PCR) multi-fragments assembling cloning and Loop-mediated Isothermal Amplification (LAMP); single primer PCR (design of PCR primers from close located inverted repeat), automatically detecting SSR loci and direct PCR primer design, amino acid sequence degenerate PCR, Polymerase Chain Assembly (PCA), design multiplexed of overlapping and non-overlapping DNA amplicons that tile across a region(s) of interest for targeted next-generation sequencing (Molecular Tagging) and much more."

The design of the primer has to be done very rigorously to guarantee the future of the PCR project, errors in the design can be noticed only long after much effort, and money has already been spent on the project. That is why it is recommended to use a software that allows us in silico to previously establish the conditions of reaction and design of the primers; FastPCR is a free (free) software, friendly, and extremely versatile to avoid spending time and money.

Biological sequences characterization

Annotation is the process of characterizing genes and their biological products in a DNA sequence. This process had to be automated because the number of genes is too large to be written down by hand. The annotation was made possible by the fact that the genes have recognizable start and stop regions. Sequence analysis refers to the study of different characteristics of molecules such as nucleic acids or proteins, which guarantee their specific functions. In the first instance, the sequences of the molecules are deposited in public biological databases (Mehmood et al. 2014).

Then, several tools can be used to predict their characteristics related to their function, structure, evolutionary history, or identification of counterparts with high precision. These analyses are quite popular due to the many applications in science biological factors, simplicity, and quantity of information about the gene/protein under study. Table 1 presents a list of tools for the characterization of biological sequences (Mehmood et al. 2014).

An orthodox procedure to deduce genetic information consists of obtaining fragments of genomic DNA by mechanical fragmentation or with restriction endonucleases or of cDNA obtained with the enzyme reverse transcriptase from messenger RNA; these fragments can be cloned into cloning vectors for sequencing with the help of programs like pDRAW32 (https://www.acaclone.com/). The cloned sequences obtained are separated from the vector sequences (which served as a reference for the design of the sequencing primers) with a simple but useful VecScreen program (https://www.ncbi.nlm.nih.gov/tools/vecscreen/). The partial sequences obtained by sequencing multiple clones can be assembled in programs such as Sequencher (https://www.genecodes.com/) to obtain larger contigs. The information contained in these sequences can begin to be worked on in order to search for open-reading frames (ORFs) through programs such as the ORF finder described in Table 1 (https://www.ncbi.nlm.nih.gov/orffinder/). The homology of the proteins deduced from the ORFs, with protein sequences deposited in the databases, will then be searched using programs such as BLAST, FASTA, or CLUSTAL. Then, the physical-chemical characteristics, the 3D structure, and the fate of proteins in the cell, as well as their function, will be established with the programs and methodologies that are described below.

Prediction of subcellular protein location

The prediction of the subcellular location of proteins predicts the fate of a protein in the cell, using computational methods with the protein sequence.

There are several publicly available software, using different methods to predict the location of proteins (amino acid composition, signal peptide composition, physical-chemical composition, among others), which is a very important part of the bioinformatics prediction of protein function and genome annotation (Nielsen et al. 2019).

Software used for protein location predictions can be accessed via URL addresses as follows:

SignalP 3.0: http://www.cbs.dtu.dk/services/SignalP-3.0/

As written on the website "SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models" (Bendtsen et al. 2004a, b).

CELLO2GO: http://cello.life.nctu.edu.tw/cello2go/

"Cello2go is a publicly available, web based system for screening various properties of a targeted protein and its subcellular localization" (Yu et al. 2014).

LocTree2: https://rostlab.org/services/loctree2/

"The method, LocTree2, predicts the location of all proteins in all areas of life. Similar to the previous method, LocTree, incorporates a system of Support Vector Machines organized hierarchically to mimic the mechanism of protein trafficking in cells" (Goldberg et al. 2012).

EuK-mPLoc 2.0: http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/

Euk-mPLoc 2.0, Predicting subcellular localization of eukaryotic proteins including those with multiple sites (Chou and Shen 2010).

ESLpred: http://www.imtech.res.in/raghava/eslpred/index.html

ESLpred is a tool for predicting the subcellular localization of proteins using support vector machines. The predictions are based on dipeptide and amino acid composition and physicochemical properties (Horler et al. 2009).

SecretomeP: http://www.cbs.dtu.dk/services/SecretomeP/

The SecretomeP 2.0 server produces ab initio predictions of non-classical, i.e., not signal peptide-triggered protein secretion. The method queries a large number of other feature prediction servers to obtain information on various post-translational and localizational aspects of the protein, which are integrated into the final secretion prediction (Bendtsen et al. 2004a, b).

Proteins characterization

After decoding the open-reading frame of a gene, a series of bioinformatics tools can be used to characterize the deduced sequence of the protein. A search on the Expasy Proteomics Server website (http://expasy.org/tools) and a nucleotide sequence allows us to identify and characterize proteins; identify motifs, patterns, and profiles; infer their stability, cell location, or function; make predictions of secondary and tertiary structures; look for similar sequences deposited in databases and compare them; and establish phylogenetic relationships.

The detection of the physical-chemical characteristics of proteins can be carried out in PROSITE (http://prosite.expasy.org/scanprosite/), in the neural network system of the Pôle BioInformatique Lionnais/Network Protein Sequence Analysis or in the DiANNA 1.1 application (http://clavius.bc.edu/~clotelab/DiANNA/), for the prediction of post-translational modifications on the Center of Biological Sequence Analysis website (http://www.cbs.dtu.dk/services).

ProDom is a comprehensive database of protein domain families generated from the global comparison of all available protein sequences. Recent improvements include the use of three-dimensional (3D) information from the SCOP database, a completely redesigned web interface (http://www.toulouse.inra.fr/prodom.html).

Phylogenetic analysis

Phylogenetic analyses are procedures used to rebuild relations evolutionary between a group with molecules and related organisms for prediction of certain characteristics of a molecule in which its functions are not known (Mehmood et al. 2014). The underlying principle of phylogeny is to group living organisms according to the degree of similarity. Phylogenetic comparison analysis is used usually to control the lack of statistical independence between species. Phylogenetic tools are usually used to test various hypotheses evolutionary, and they are indispensable to functional genomics (Mehmood et al. 2014; Khan et al. 2014).

MEGA-Molecular Evolutionary Genetics Analysis (MEGA) is computer software for conducting statistical analysis of molecular evolution and for constructing phylogenetic trees, very recommended for sequence alignment and phylogeny inference (https://www.megasoftware.net/). We share the opinion of Kumar et al. in 2016 about the Molecular Evolutionary Genetics Analysis (MEGA) that MEGA includes a large repertoire of programs for assembling sequence alignments, inferring evolutionary trees, estimating genetic distances and diversities, inferring ancestral sequences, computing time trees, and testing selection. Over the last 25 years, MEGA's use in evolutionary analysis has been cited in over one hundred thousand studies in diverse biological fields (Kumar et al. 2018).

MOLPHY: Molecular phylogenetic analysis tool (https://sbgrid.org/software/titles/molphy).

PAML: Package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood (https://bio.tools/paml).

PHYLIP: PHYLogeny Inference Package (PHYLIP). One of the most useful and used free computational phylogenetic package of programs for inferring evolutionary trees (phylogenies). The author is Joseph Felsenstein, Professor at the University of Washington, Seattle. It consists of 35 programs that include methods for, distance matrix and maximum likelihood, including calculating statistical support for clades (bootstrapping) and consensus trees based on the following types of data: molecular sequences, gene frequencies, restriction sites and fragments, matrices from distance (http://evolution.genetics.washington.edu/phylip.html).

Jalview: Program for multiple sequence alignment editing (https://www.jalview.org/).

Biological sequence databases

Biological sequence databases are a vast collection of biological information data, such as sequences of nucleotides, proteins, and macromolecular structures. The information stored in these databases has no only important for future applications but also serves as a tool for primary sequence analysis. The submission and storage of this information to be freely available to the scientific community led to the development of several bases worldwide. The bases of data contain varied information; therefore, they are classified as primary and

secondary, through information stored. Primary databases are composed of derived information directly from basic scientific research on sequencing. SWISSPROT, UniProt, GenBank, and PDB are examples of primary databases. Secondary databases contain information derived from the interpretation of information stored in the database's primary. SCOP, CATH, PROSITE, and eMOTIF are examples of secondary databases (Koonin and Galperin 2003) (Table 2).

Proteins structure prediction tools

Proteins are composed of polypeptides, which in turn are polymers composed of amino acids that fold together creating a three-dimensional structure (3D). Protein folding in its form correctly is a prerequisite for any protein that can perform its biological function; therefore, in order to understand the functions of a specific protein, information is needed about their three-dimensional structures, see Table 3.

Molecular interactions

Proteins rarely perform their functions in isolation and therefore interact with other molecules to run a particular process. Understand how biomolecules interact with other molecules could be used in purification techniques as well as drug development. It is also essential to understand the interactions between molecules in order to elucidate the biological functions of a molecule. For example, interactions between proteins have a key role in cellular activities such as signaling, transport, metabolism, and various biochemical processes (Table 4).

Molecular dynamics simulations

Biological activity is the result of molecular interactions. This behavior of molecules can be studied with the use of bioinformatics tools, usually referred to as simulation tools for molecular dynamics. They aim to provide detailed information on the dynamics of processes that occur in biological systems (refer to Table 5).

Medicines concession

Before bioinformatics tools, scientists resorted to chemistry, pharmacology, and clinical sciences to discover new compounds. Traditional processes are time-consuming and costly. Bioinformatics came to facilitate this complex process and has a vital role in the discovery of new drugs and its design due to the quick analysis of molecules in a computer when compared with experimentation (Lekamwasam and Liyanage 2013; Chordia and Kumar 2018) (refer to Table 6).

Integrative bioinformatics modules

As already mentioned, the amount of biological data grows exponentially and these data are spread over infinity of public and private repositories and are stored in different formats. This makes it difficult to search for these data and carry out the analyses necessary to deduce new knowledge from the set of deposited data. Integrative bioinformatics attempts to solve this problem by providing unified access to life science data.

The several directions which may lead to breaking the bottleneck of Integrative Bioinformatics are described by Chen et al. (2019) and include:

- "Integration of multiple biological data towards systems biology. Different omics data is reflecting different aspects of the biological problem. Often, to solve a problem, there are many different methods developed by many groups. These methods may perform differently, some good, some bad. Combing with big data, and other approaches, artificial intelligence (AI) has been successfully applied in bioinformatics, especially in the field of biomedical image analysis;
- Computing infrastructure development. Integrative Bioinformatics in the big data era requires a more advanced IT environment. To facility the related computing and visualization demands, both hardware (e.g. GPU) and software (e.g. Tensor flow) are developing. Supercomputers are used. Cloud services are provided by more and more institutes and big companies."

Many ready-made professional commercial bioinformational programs are presented by development companies using modern sequencing technologies. Bioinformation groups usually prefer to use ready-made modules and write scripts to bind data between them.

Therefore, in parallel, two ways for data analysis are ready-made commercial products and scripts for linking different ready-made mini-programs. Both ways are necessary. Therefore, the most important thing is the support and updating of ready-made programs and modules. In the following list, we present the most important modules for integrated bioinformatics:

- Uniprot UGene: Ugene is free bioinformatics software for multiple sequence alignment, genome sequencing data analysis, and amino acid sequence visualization. Unipro UGENE is a multiplatform open-source software with the main goal of assisting molecular biologists without much expertise in bioinformatics to manage, analyze, and

Table 2 Biological databases

Database	Description	References/URL
Nucleotide database		
DNA Data Bank of Japan (DDBJ)	Member of the International Nucleotide Sequence Databases (INSD) is one of the largest sources of information regarding nucleotide sequences.	https://www.ddbj.nig.ac.jp/index-e.html (Miyazaki et al. 2003)
European Nucleotide Archive (ENA)	European primary repository for nucleotide sequences	https://www.ebi.ac.uk/ena (Leinonen et al. 2011)
GenBank	Member of the International Nucleotide Sequence Databases (INSD) is one of the largest sources of information regarding nucleotide sequences.	http://www.ncbi.nlm.nih.gov/genbank/ (Benson et al. 2010; NCBI Resource Coordinators 2018)
Rfam	A collection of RNA families, each represented by multiple alignment sequences	https://rfam.xfam.org/ (Kalvari et al. 2018)
Protein databases		
Uniprot	The UniProt Consortium is a collaboration between the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). Containing the manually annotated protein sequences section SWISS PROT.	https://www.uniprot.org/ (UniProt Consortium 2008)
Protein Data Bank	Database for biomolecule structures like proteins. Structures determined by experimentation.	https://www.rcsb.org/pdb/home/sitemap.do (Kinjo et al. 2012)
Prosite	A large collection of biologically meaningful signatures that are described as patterns or profiles.	https://prosite.expasy.org/ (Sigrist et al. 2012)
Pfam	Large collection of protein families, represented by multiple sequence alignments and hidden Markov models (HMMs).	https://pfam.xfam.org/ (El-Gebali et al. 2019)
InterPro	Provides functional analysis of proteins by classifying them into families and predicting domains and important sites.	https://www.ebi.ac.uk/interpro
Proteomics Identification Database (PRID)	Contains information on functional characterization and post-translation modifications.	https://www.ebi.ac.uk/pride/ (Perez-Riverol et al. 2019)
Genomic databases		
Ensembl	Contains eukaryotic genomes, including human, rat and other vertebrates, and many tools to work and compare genomes.	https://www.ensembl.org/index.html (Hunt et al. 2018; Spooner et al. 2018)
PIR	Integrated tool that supports genomic and proteomic research.	https://proteininformationresource.org/ (Chen et al. 2011)

visualize their data. It provides visualization modules for biological objects such as annotated genome sequences, next-generation sequencing (NGS) assembly data, multiple sequence alignments, phylogenetic trees, and 3D structures. Availability and implementation: UGENE binaries are freely available for MS Windows, Linux, and Mac OS X. (Okonechnikov et al. 2012);

- Vista: "Vista is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species" (Frazer et al. 2002). http://genome.lbl.gov/vista/index.shtml;
- Qlucore: Qlucore Omics Explorer (QOE) is next-generation bioinformatics software for research in the life sciences. Qlucore Omics Explorer is built for fast and easy analysis of many different types of data and a wide range of application areas are supported:

With Qlucore Omics Explorer, you can examine and analyze data from gene expression experiments, DNA methylation data, proteomics data, and next-generation sequencing (NGS) data (https://www.qlucore.com/);

Table 3 Tools for analyzing protein structures and functionality

Tool	Description	References/URL
SWISS-MODEL	A fully automated protein structure homology-modeling server. ... the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer)	https://swissmodel.expasy.org/ (Waterhouse et al. 2018)
GOpenMol	Tool for the visualization and analysis of molecular structures and their chemical properties	https://www.softpedia.com/get/Science-CAD/gOpenMol.shtml
Cn3d	Application for a web browser that allows you to view 3-dimensional structures from GenBank Entrez Structure database	https://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml (Wang et al. 2000)
CATH	Publicly available online resource that provides information on the evolutionary relationships of protein domains	https://www.cathdb.info/ (Dawson et al. 2017)
RaptorX	Prediction of protein structures	http://raptorx.uchicago.edu/ (Källberg et al. 2012)
JPRED	Protein secondary structure prediction server	http://www.compbio.dundee.ac.uk/jpred/ (Drozdetskiy et al. 2015)
PHD	Prediction of protein structures	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html (Rost et al. 1994)
HMMSTR	Prediction of structural correlations in proteins	http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php (Bystroff and Shao 2002)
APSSP2	Predict the secondary structure of protein's from their amino acid sequence	http://crdd.osdd.net/raghava/apssp2/ (Raghava 2002)
MODELLER	Prediction of 3D protein structures based on comparative models	https://salilab.org/modeller/ (Webb and Sali 2016)
Phyre and Phyre2	Prediction of protein structures on the web	http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index (Kelley et al. 2015)
Prot pi	Web application for calculating physicochemical parameters of proteins and peptides	https://www.protpi.ch/

- CIBI: The CRCM's Integrative BioInformatics (Cibi) is a technological platform of the Centre de Recherche en Cancérologie de Marseille. The Cibi platform offers a wide range of expertise in bioinformatics (large-scale data integration, development of specific analysis) and develops state-of-the-art bioinformatics pipelines as NGS (next-generation sequencing) data analysis and integration (Chip-Seq, RNA-Seq, SC-RNA-Seq, variant analysis for research and cancer diagnostics) (https://cibi.marseille.inserm.fr/);
- iMAP: iMAP is an integrated bioinformatics and visualization pipeline for microbiome data analysis. According to Buza et al., the iMAP tool wraps functionalities for metadata profiling, quality control of reads, sequence processing and classification, and diversity analysis of operational taxonomic units. This pipeline is also capable of generating web-based progress reports for enhancing an approach referred to as review-as-you-go (RAYG). The iMAP pipeline integrates several functionalities for better identification of microbial communities present in a given sample. The pipeline performs in-depth quality control that guarantees high-quality results and accurate conclusions. The vibrant visuals produced by the pipeline facilitate a better understanding of the complex and multidimensional microbiome data (Buza et al. 2019;
- MIGenAS: Migenas is a versatile and extensible integrated bioinformatics toolkit for the analysis of biological sequences over the Internet. The toolkit is part of the Max-Planck Integrated Gene Analysis System (MIGenAS) of the Max-Planck Society available at www.migenas.org (Rampp et al. 2006);
- Methy-Pipe: Methy-Pipe is an integrated bioinformatics pipeline for whole-genome bisulfite sequencing data analysis. According to Jiang et al., Methy-Pipe uses Burrow-Wheeler transform (BWT) algorithm to directly align bisulfite sequencing reads to a reference genome and implements a novel sliding window-based approach with statistical methods for the identification of differentially methylated regions (DMRs). Methy-Pipe is a useful pipeline that can process whole-genome bisulfite sequencing data in an efficient, accurate, and user-friendly

Table 4 Tools for studying molecular interactions

Tool	Description	References/URL
SMART	Provides varied information about the protein in question	http://smart.embl-heidelberg.de/help/latest.shtml
AutoDock	Predicts protein-ligand interactions	http://autodock.scripps.edu/downloads/autodock-registration/autodock-4-2-download-page/
HADDOCK	Describes the interaction between protein-protein, protein-DNA	https://haddock.science.uu.nl/
BIND	Biomolecular Interaction Network Database	(Bader et al. 2003)
STRING	Database of known protein interactions and prediction	https://string-db.org/ (Szklarczyk et al. 2019)
MIMO	Tool for molecular interactions	http://www.mybiosoftware.com/mimo-1-0-molecular-interaction-maps-overlap.html (Di-Lena et al. 2013)
IntAct	Provides a freely available, open-source database system and analysis tools for molecular interaction data	https://www.ebi.ac.uk/intact/ (Orchard et al. 2014)
PathBLAST	Research of interactions between proteins	http://www.pathblast.org/ (Ideker et al. 2004)

manner. Software and test dataset are available at http://sunlab.lihs.cuhk.edu.hk/methy-pipe/ (Jiang et al. 2014);

- IGV: The Integrative Genomics Viewer (IGV) is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources (https://igv.org/);
- Bioconductor: Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and Docker images (https://www.bioconductor.org/);
- Geneious: Geneious is a very useful and popular DNA, RNA and protein sequence alignment, assembly and analysis software platform, integrating bioinformatic and molecular biology tools into a simple interface. These tools are created by the company Biomatters, headquartered in New Zealand with offices in the USA, and users in 125 countries worldwide, yours solutions enhance productivity in more than 4000 universities, research institutes, and businesses. Biomatters create powerful, integrated, and visually appealing bioinformatics solutions, with a strong emphasis on ease of use and overall user experience (https://www.geneious.com/).

Trends and future of bioinformatics

Bioinformatics has developed significantly from the development and establishment of molecular cloning methodologies and the automation of DNA sequencing methods. With the development and application of the new generation sequencing platforms, large-scale sequencing of genomes and transcriptomes began, which contributes to the development of bioinformatics methodologies and tools at a level that went beyond academic centers and which includes medical biotechnology, gene therapy, agriculture biotechnology, animal biotechnology, environmental biotechnology, and forensic biotechnology. Currently, bioinformatics has a great application in genomics, proteomics, metabolomics, transcriptomics, and molecular phylogenomics. The development of biomarkers for the creation of safer and more personalized drugs is leading to more to greater development and use of bioinformatics. The sequencing of personal genomes and metagenomics projects

Table 5 Molecular dynamics simulation tools

Tool	Description	References/URL
Abalone	Program focused on molecular dynamics of biopolymers.	http://www.biomolecular-modeling.com/Abalone/index.html
Ascalaph	Program for modeling for molecular design and simulations	http://www.biomolecular-modeling.com/Ascalaph/index.html
Amber	Package of programs for molecular dynamics simulations of proteins and nucleic acids	https://ambermd.org/ (Salomon-Ferrer et al. 2013)
FoldX	Provides quantitative estimates of molecular interactions	http://foldxsuite.crg.eu/

Table 6 Databases for target drugs

Database	Description	References/URL
Potential Drug Target Database (PDTD)	Target drug database	http://www.dddc.ac.cn/pdtd/ (Zhang et al. 2018)
Drug Bank	Online database containing information on drugs and drug targets	https://www.drugbank.ca/ (Wishart et al. 2018)
Therapeutic Target Database (TTD)	Collection of therapeutic proteins	http://db.idrblab.net/ttd/ (Yunxia et al. 2019)
TDR Target Database	Identification and prioritization of molecular targets for drug development, focusing on pathogens responsible for neglected human diseases	https://tdrtargets.org/ (Magariños et al. 2012)
MATADOR: Manually Annotated Targets and Drugs Online Resource	Resources for exploring drug-target relationships	http://matador.embl.de/ (Günther et al. 2008)
TB Drug Target Database	Database specialized in drugs and target proteins for tuberculosis (TB) treatment	https://www.bioinformatics.org/tbdtdb/
DrugPort	Structure information available in the PDB for molecules of drugs	http://www.ebi.ac.uk/thornton-srv/databases/drugport/
ChEMBL	Database of bioactive molecules with drug-like properties	https://www.ebi.ac.uk/chembl/ (Mendez et al. 2019)

will increase significantly in the coming years with the consequent intervention of bioinformatics. We think that the future of bioinformatics will involve specialization in different areas that go down more in scientific depth at the level of nanopores and even the atom itself.

Conclusion

Bioinformatics is a discipline relatively new that in recent years progressed very quickly. It is discipline that makes it possible to test hypotheses virtually what allows you to have better knowledge before proceeding with expensive studies. Despite the development of the most many tools for analysis genomics, proteomics, inference of structures, drug design, and simulations of molecular dynamics, none can be considered the "perfect" tool. Bioinformatics tools provide results that are more accurate what allows reliable interpretations. Perspectives in the field of bioinformatics include contributions to understanding the human genome, leading to the discovery of new drugs and specific therapies. It is essential that bioinformatics and other disciplines move side by side to understand biological systems and the consequent development of human well-being. During the first years of biotechnology, the most important was to obtain biological data. With the development of the methods and techniques of the new generation of sequencing (NGS), the paradigm shifted to the ability to analyze such a large amount of data resulting from the sequencing of genes and genomes. However, with the development of bioinformatics tools in recent years, the most important is to know what we want in research, choose the appropriate tool, work hard with it, and know-how to correctly interpret the results provided.

Authors' contributions I.B. and A.C. designed prepared the manuscript; A.C. wrote the manuscript; and I.B. made the corrections including the English text and confirmation of the bibliography and URLs.

Funding The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) and FEDER under Programme PT2020 for financial support to CIMO (UID/AGR/00690/2019).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights No human participants or animals were involved in this research.

Informed consent This manuscript is original and submitted with the consent of all authors.

References

Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21:3596–3603

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bader GD, Betel D, Hogue CW (2003) BIND: the biomolecular interaction network database. Nucleic Acids Res 31(1):248–250. https://doi.org/10.1093/nar/gkg056

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004a) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795

Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004b) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17(4):349–356

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. Nucleic Acids Res 38(Database issue):D46–D51. https://doi.org/10.1093/nar/gkp1024

Bonfield JK, Whitwham A (2010) Gap5–editing the billion fragment sequence assembly. Bioinformatics 26(14):1699–1703

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Buza T, Tonui T, Stomeo F, Tiambo C, Katani R, Schilling M, Lyimo B, Gwakisa P, Cattadori IM, Buza J, Kapur V (2019) iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. BMC Bioinformatics 20:374

Bystroff C, Shao Y (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics 18(Suppl 1):S54–S61

Chen C, Natale DA, Finn RD, Huang H, Zhang J, Wu CH, Mazumder R (2011) Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. PLoS One 6(4):e18910

Chen M, Hofestädt R, Taubert J (2019) Integrative bioinformatics: history and future. Journal of Integrative Bioinformatics 16. https://doi.org/10.1515/jib-2019-2001

Chordia N, Kumar A (2018) Bioinformatics in drug discovery. SF Protein Sci J 1:1

Chou K, Shen H (2010) A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I (2017) CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45: D289–D295. https://doi.org/10.1093/nar/gkw1098

Di-Lena P, Wu G, Martelli PL, Casadio R, Nardini C (2013) MIMO: an efficient tool for molecular interaction maps overlap. BMC Bioinformatics 14:159 10.1186/1471-2105-14-159. 10.1093/bioinformatics/btn596

Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43(Web Server issue):W389–W394. https://doi.org/10.1093/nar/gkv332

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD (2019) The Pfam protein families database in 2019: nucleic acids res. https://doi.org/10.1093/nar/gky995

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2002) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32(Web Server issue):W273–W279

Ganesan N, Bennett NF, Velauthapillai M, Pattabiraman N, Squier R, Kalyanasundaram B (2005) Web-based interface facilitating sequence-to-structure analysis of BLAST alignment reports. Biotechniques 39(186):188

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) In: The proteomics protocols handbook. Protein identification and analysis tools on the ExPASy server. Springer, pp 571-607

Goldberg T, Hamp T, Rost B (2012) LocTree2 predicts localization for all domains of life. Bioinformatics 28(18):i458–i465. https://doi.org/10.1093/bioinformatics/bts390

Goldman D, Domschke K (2014) Making sense of deep sequencing. Int J Neuropsychopharmacol 17(10):1717–1725. https://doi.org/10.1017/S1461145714000789

Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ, Schneider R, Skoblo R, Russell RB, Bourne PE, Bork P, Preissner R (2008) SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res 36(Database issue):D919–D922. https://doi.org/10.1093/nar/gkm862

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107(1):1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Horler RS, Butcher A, Papangelopoulos N, Ashton PD, Thomas GH (2009) EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations. Bioinformatics 25(2):163–166

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, Cunningham F (2018) Ensembl variation resources. Database Volume 2018 https://doi.org/10.1093/database/bay119

Ideker T, Kelley, Shamir R, Karp R (2004) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data Proceedings: RECOMB 2004, pp. 282-289; J Comput Biol 12: 835–846, 2005

Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA, Rossa WK, ChiuY M, Lo D, Sun H (2014) Methy-Pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. PLoS One 9(6):e100360. https://doi.org/10.1371/journal.pone.0100360

Kalendar R, Khassenov B, Ramankulov Y, Samuilova O, Ivanov KI (2017a) FastPCR: an *in silico* tool for fast primer and probe design and advanced sequence analysis. Genomics 109:312–319

Kalendar R, Muterko A, Shamekova M, Zhambakin K (2017b) *In silico* PCR tools a fast primer, probe and advanced searching. Methods Mol Biol 1620:1–31. https://doi.org/10.1007/978-1-4939-7060-5_1

Kalendar R, Tselykh T, Khassenov B, Ramanculov EM (2017c) Introduction on using the FastPCR software and the related Java web tools for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. Methods Mol Biol 1620:33–64. https://doi.org/10.1007/978-1-4939-7060-5_2

Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7:1511–152253

Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI (2018) Non-coding RNA analysis using the Rfam database. Curr Protoc Bioinformatics 62(1):e51. https://doi.org/10.1002/cpbi.51

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858

Khan FA, Phillips CD, Baker RJ (2014) Timeframes of speciation, reticulation, and hybridization in the bulldog bat explained through phylogenetic analyses of all genetic transmission elements. Syst Biol 63: 96–110

Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley MD, Nakagawa A, Nakamura H (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic Acids Res 40:D453–D460

Koonin EV, Galperin MY (2003) Sequence - evolution - function: computational approaches in comparative genomics. Chapter 3, Information Sources for Genomics. Kluwer Academic, Boston. https://www.ncbi.nlm.nih.gov/books/NBK20256/

Kulski JK (2016) Next-generation sequencing – an overview of the history, tools, and "Omic" applications, next generation sequencing-advances, applications and challenges. InTech

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549. https://doi.org/10.1093/molbev/msy096

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G (2011) The European nucleotide archive. Nucleic Acids Res 39:D28–D31

Lekamwasam S, Liyanage C (2013) Editorial. Galle Medical Journal 18(1). https://doi.org/10.4038/gmj.v18i1.5520

Li MW, Qi X, Ni M, Lam HM (2013) Silicon era of carbon-based life: application of genomics and bioinformatics in crop stress research. Int J Mol Sci 14(6):11444–11483

Lobo I (2008) Basic Local Alignment Search Tool (BLAST). Nature Education 1(1):215

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R (2019) The EMBL-EBI search and sequence analysis tools APIs. Nucleic Acids Res 47(W1):W636–W641. https://doi.org/10.1093/nar/gkz268

Magariños MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, Van Voorhis WC, Agüero F (2012) TDR Targets: a chemogenomics resource for neglected diseases. Nucleic Acids Res (Database issue):D1118–D1127. https://doi.org/10.1093/nar/gkr1053

Martins IM, Matos M, Costa R, Silva F, Pascoal A, Estevinho LM, Choupina AB (2014) Transglutaminases: recent achievements and new sources. Appl Microbiol Biotechnol 98:6957–6964

Martins IM, Meirinho S, Costa R, Cravador A, Choupina A (2019) Cloning, characterization, in vitro and *in planta* expression of a necrosis-inducing Phytophthora protein 1 gene *npp1* from *Phytophthora cinnamomi*. Mol Biol Rep 46:6453–6462

Mehmood MA, Sehar U, Ahmad N (2014) Use of bioinformatics tools in different spheres of life sciences. J Data Mining Genomics Proteomics 5:158. https://doi.org/10.4172/2153-0602.1000158

Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940. https://doi.org/10.1093/nar/gky1075

Mitra A, Kesarwani AK, Pal D, Nagaraja V (2011) WebGeSTer DB—a transcription terminator database. Nucleic Acids Res 39:129–135

Miyazaki S, Sugawara H, Gojobori T, Tateno Y (2003) DNA Data Bank of Japan (DDBJ) in XML. Nucleic Acids Res 31:13–16

Münch R, Hiller K, Grote A, Scheer M, Klein J, Schobert M, Jahn D (2005) Virtual footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics 21:4187–4189

Nielsen H, Tsirigos KD, Brunak S, von Heijne G (2019) A brief history of protein sorting prediction. Protein J 38:200–216. https://doi.org/10.1007/s10930-019-09838-3

Okonechnikov K, Golosova O, Fursov M, the UGENE team (2012) Unipro UGENE: a unified bioinformaticstoolkit. Bioinformatics. 28(8):11667. https://doi.org/10.1093/bioinformatics/bts091

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(D1):D358–D363. https://doi.org/10.1093/nar/gkt1115

Parra G, Blanco E, Guigó R (2000) GeneID in Drosophila. Genome Res 10(4):511–515. https://doi.org/10.1101/gr.10.4.511

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaíno JA (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47(D1):D442–D450. https://doi.org/10.1093/nar/gky1106

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204

Raghava GPS (2002) APSSP2 : a combination method for protein secondary structure prediction based on neural network and example based learning. CASP5. A-132

Rampp M, Soddemann T, Lederer H (2006) The MIGenAS integrated bioinformatics toolkit for web-based sequence analysis. Nucleic Acids Res 34:W15–W19. https://doi.org/10.1093/nar/gkl254

Resource Coordinators NCBI (2018) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46(D1):D8–D13. https://doi.org/10.1093/nar/gkx1095

Rodger S, David PJ, James KB (2003a) Analysing sequences using the Staden package and EMBOSS. In: Krawetz SA, Womble DD (eds) Introduction to Bioinformatics. A Theoretical and Practical Approach. Human Press Inc., Totawa, p 07512

Rodger S, David PJ, James KB (2003b) Managing sequencing projects in the GAP4 environment. In: Krawetz SA, Womble DD (eds) Introduction to Bioinformatics. A Theoretical and Practical Approach. Human Press Inc., Totawa, p 07512

Rost B, Sander C, Schneider R (1994) PHD–an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10:53–60

Salomon-Ferrer R, Case DA, Walker RC (2013) An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci 3:198–210

Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347. https://doi.org/10.1093/nar/gks1067

Spooner W, McLaren W, Slidel T, Finch DK, Butler R, Campbell J, Eghobamien L, Rider D, Kiefer CM, Robinson MJ, Hardman C, Cunningham F, Vaughan T, Flicek P, Huntington CC (2018) Haplosaurus computes protein haplotypes for use in precision drug design. Nat Commun 9:4128. https://doi.org/10.1038/s41467-018-06542-1

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1):D607–D613. https://doi.org/10.1093/nar/gky1131

UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195

Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH (2000) Cn3D: sequence and structure views for Entrez. Trends Biochem Sci 25(6):300–302

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46(W1):W296–W303

Webb B, Sali A (2016) Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics 54:5.6.1–5.6.37 John Wiley, Sons, Inc.

Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Jonghe P, Van Broeckhoven C, De Rijk P (2005) novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15:436–442

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx1037

Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK, Lu CH (2014) CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation.

PLoS One 9(6):e99368. https://doi.org/10.1371/journal.pone.0099368

Yunxia W, Song Z, Fengcheng L, Ying Z, Ying Z, Zhengwen W, Runyuan Z, Jiang Z, Yuxiang R, Ying T, Chu Q (2019) Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res 48(D1):D1031–D1041. https://doi.org/10.1093/nar/gkz981 ISSN 1362-4962

Zhang S, Zhang L, Wang Y, Liao M, Bi S, Xie Z, Ho C, Wan X (2018) TBC2target: a resource of predicted target genes of tea bioactive compounds. Front Plant Sci 9:211. Published 2018 Feb 22. https://doi.org/10.3389/fpls.2018.00211