World Journal of Pharmaceutical and Life Sciences
WJPLS
www.wjpls.org
SJIF Impact Factor: 6.129

USES OF BIOINFORMATICS IN THE DIFFERENT DISCIPLINES AND PLATFORMS USED

Julia San Miguel Rodríguez*[1], Julita Rodríguez Barbero, Angel San Miguel Hernández[2], Angel San Miguel Rodríguez[3] and María San Miguel Rodríguez[1]

[1]Research Service. Río Hortega University Hospital. Valladolid. Spain.
[2]Clinical Analysis Service. Río Hortega University Hospital. Valladolid. International University of La Rioja. Spain.

Corresponding Author: Julia San Miguel Rodríguez
Clinical Analysis Service. Río Hortega University Hospital. Valladolid. Spain.

Article Received on 30/07/2021 **Article Revised on 20/08/2021** **Article Accepted on 10/09/2021**

ABSTRACT

Bioinformatics is a scientific science that uses information technologies to organize, analyze and distribute biological information to solve complex questions in the area of biology. It is a multidisciplinary research area, as it can be conceived as the interface between computer science and biology. It allows to research, develop and apply computer and computational tools to allow and improve the handling of biological data. One of its applications is the management of the automation of diagnostic technologies. DNA repeats are located in both gene and intergenic regions and are classified into sparse and tandem repeats. Interleaved repeat DNA with broad repeats is made up of sequences that are repeated thousands of times in the genome but in a sparse manner. And tandem repeat DNA is about repeats of identical sequences that are arranged one after the other. These repetitions will be classified according to the length of the unit that is repeated and the number of repetitions that occur of that unit. Comparative Genomics is based on the study of the relationship between the structure and function of the genome through different species or strains. It is responsible for comparing gene and protein sequences from different genomes to explain functional and evolutionary relationships. DNA sequencing is the set of biochemical methods and techniques that have the purpose of determining the order of nucleotides (A, C, G and T) in a DNA oligonucleotide. The DNA sequence makes up the heritable genetic information that forms the basis for the development of living things. Signal analysis is based on the identification of characteristic sequence motifs of the elements that make up genes. One of the most widely used databases in Bioinformatics is the sequence database. It is a collection of DNA, protein and other sequences, which are stored in computers. These databases can include sequences from a single organism or can include sequences from all organisms. Cross-references are marks that reference or link to another place in the document at a certain point in the document. The benefits of cross-referencing long documents are significant and numerous.

KEYWORDS: Bioinformatic, Comparative genomics, signal analysis, databases, platforms.

INTRODUCTION

Bioinformatics is defined as the application of computational technologies and statistics to the management and analysis of biological data.[1]

The terms bioinformatics, computational biology, biological informatics and biocomputing are used in many situations as synonyms.[2,3] and refer to closely linked interdisciplinary fields of study that require the use or development of different techniques such as applied science of the discipline. Computing.[4] Among them can be highlighted, applied mathematics, statistics, computer science, artificial intelligence, chemistry and biochemistry.[5-10] with which problems are solved when analyzing data, or simulating systems or mechanisms, all of them of a biological nature, and usually, but not exclusively) at the molecular level.[11]

The main thing about these techniques is the use of computational resources to solve or investigate problems on scales of such magnitude that they exceed human judgment. Computational biology research often overlaps with systems biology.[12]

Major efforts in these fields include sequence alignment, gene prediction, genome assembly, protein structural alignment, protein structure prediction, gene expression prediction, protein-protein interactions, and evolution modeling.[13]

In bioinformatics and computational biology, the use of mathematical tools is important to extract useful

information from data produced by high-throughput biological techniques, such as genome sequencing. In particular, the assembly or assembly of high-quality genomic sequences from fragments obtained after large-scale DNA sequencing is an area of interest.[13,14]

Other goals include the study of gene regulation to interpret gene expression profiles using data from DNA chips or mass spectrometry.[15]

The NIH (National Institutes of Health. USA) recognized that no definition could completely eliminate the overlap between activities of the different techniques, it explicitly defines the terms bioinformatics and computational biology.[16]

Thus, bioinformatics is the research, development or application of computational tools and approaches for the expansion of the use of biological, medical, behavioral or health data, including those tools that serve to acquire, store, organize, analyze or visualize such data. And computational biology would be the development and application of theoretical and data analysis methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral and social systems. In this way, bioinformatics would have more to do with information, while computational biology would do with hypotheses.

The term biocomputing is usually framed in current research with biocomputers and is defined as the construction and use of computers that contain biological components or function like living organisms.[17]

Apart from the definitions of reference bodies or institutions, the manuals on this subject provide their own operational definitions, logically linked to a greater or lesser extent with those already seen. As an example, Mount, in his widespread text on bioinformatics,[18] points out that bioinformatics focuses more on the development of practical tools for data management and analysis, for example, the presentation of genomic information and sequential analysis, but with less emphasis on efficiency and precision.[19-21]

In addition, computational biology is related to the development of new and efficient algorithms, which can be shown to work on a difficult problem, such as the multiple alignment of sequences or the assembly of genome fragments.

Bioinformatics is the science of using information to understand biology. It is a subset of the larger field of computational biology, this being the application of quantitative analytical techniques in the modeling of biological systems. And it allows to investigate, develop and apply computer and computational tools to allow and improve the handling of biological data. One of its applications is the management of the automation of diagnostic technologies.

It is one of the disciplines that has had the most prominence in recent years, such as in the management and interpretation of data on SARS-CoV-2. Its work consists of researching, developing and applying computer and computational tools to allow and improve biological data, with the use of tools that gather, store, organize, analyze and interpret the data.

Bioinformatics was born in the 1960s, with the application of computational methods to the analysis of protein sequences. Its growth was linked to the development of Molecular Biology, the discovery of DNA, and advances in computing. The concept of bioinformatics today is somewhat different, since it is considered an emerging discipline that has become necessary for the management of the enormous volume of data generated by new omic technologies, such as genomics, proteomics, metabolomics..., making Big data a fundamental asset in current biomedicine.

Among these technologies, one of the most relevant is high-throughput sequencing or massive sequencing, which since 2004, with the sequencing of the human genome, has made it possible to obtain the genomic sequence of many organisms.

The use of computing, programming languages and large computational infrastructures are the pillars used by bioinformatics to collect, handle, store and analyze biological data, from those derived from genomic, proteomic, metabolomic sequencing, to image, clinical, epidemiological data ..., developing algorithms or mathematical models to extract the maximum knowledge from the data and apply it directly to solving biological or biomedical problems (22-25).

Among the most relevant problems that have benefited from the development of genomics and bioinformatics are, among many others, the study of rare diseases of genetic origin; identification of tumor-associated mutations; the identification of the pathogen causing an infectious outbreak or the discovery of new viruses, such as SARS-CoV-2.

Furthermore, the involvement of bioinformatics in the resolution of human pathologies has led to the appearance of Clinical Bionformatics, which is a multidisciplinary specialty in which specialists in molecular biology, genetics, computing, and mathematics work side by side.

The great advance of Bioinformatics as an indispensable discipline in many fields such as Biomedicine, Agriculture, Food, among others, has led to a great increase in the demand for professionals and has led to integration in new environments. This has highlighted the need to generate pathways for the formation of bioinformatics.

Therefore, bioinformatics is a scientific science that uses

information technologies to organize, analyze and distribute biological information to solve complex questions in the area of biology. It is a multidisciplinary research area, since it can be conceived as the interface between computer science and biology. It is very promising and emerging, which is reaching other biological sciences and medicine, such as telemedicine, the discovery of oncogenes, nanotechnology and the Human Genome Project. In bioinformatics, it is important to know the procedures for the location and masking of DNA sequences, as well as the comparison methods and the analysis of both signals and nucleotide sequences. In addition to searching databases of expressed sequences. And classify the different types of biological databases.[26-29]

In figure 1, the relationship of bioinformatics with other sciences is shown.

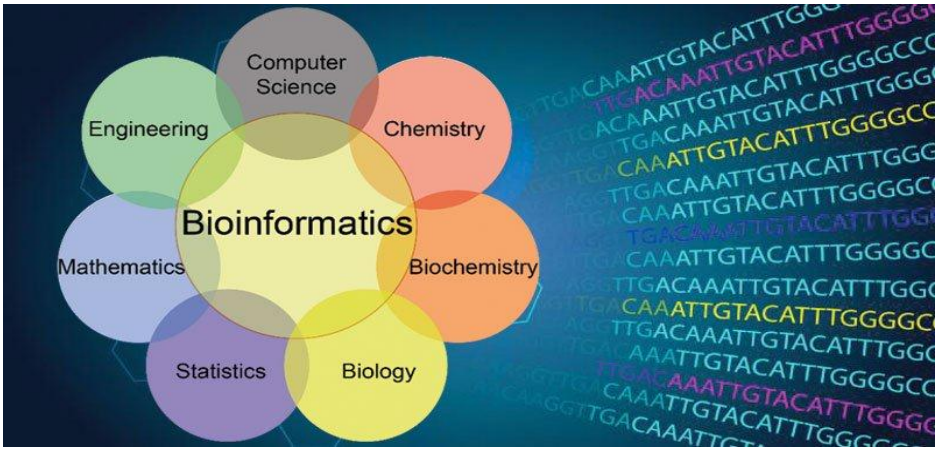

Figure 1: Relationship of bioinformatics with other sciences.

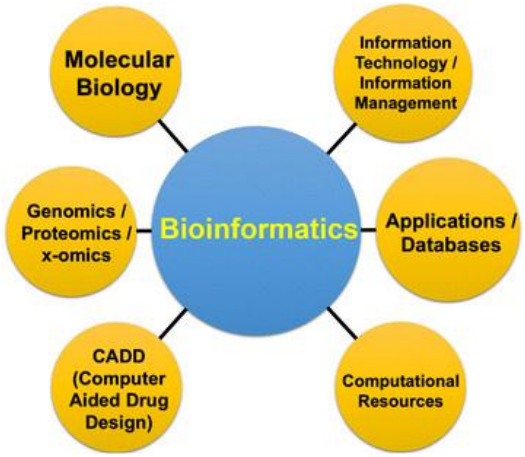

Figure 2: Shows the relationship of bioinformatics with the most related disciplines

Figure 2: Relationship of bioinformatics with the closest disciplines.

The table 1 lists the main fundamental objectives of bioinformatics.[6,11,15]

Table 1: Fundamental objectives of bioinformatics.[6,11,15]

•	**Information storage**: By creating databases, large volumes of biological data can be stored, organized and managed. These databases must be able to locate the required information quickly.
-	Likewise, these databases must be accessible from the Internet so that any researcher has access from any location and their use must be highly intuitive.
•	**Information analysis**: To analyze all the information stored in these databases, bioinformatics must be able to search for algorithms and other statistical and analysis tools that allow establishing different relationships between data to, for example, compare genetic sequences or other biological structures.
•	**Development of computer tools:** The objective is to develop and implement tools that allow analyzing and interpreting the large amount of data that is handled.
-	These tools are computer programs either in the cloud or installable under different platforms. The important thing is that they are easily accessible.

The table 2 shows the main applications of bioinformatics.[2,3,6,13]

Table 2: Main applications of bioinformatics.[2,3,6,13]

• **Sequence analysis**. The DNA sequences of different living organisms have been decoded. It is a large amount of information that must be stored in databases.
• Bioinformatics helps analyze this data to determine the genes that code for certain proteins. It is also used to compare DNA sequences between species in order to determine the similarities between these species.
• **Annotation of genomes.** Annotation is a process of marking the genes and other biological characteristics of the DNA sequence.
• This allows to perform the analysis of the genome and understand how they determine the characteristics of different living beings.
• **Drug design.** The process consists of the search for new drugs based on the knowledge of a biological target as an aberrant protein).
• We can use bioinformatics to design a drug (a small molecule) that activates or inhibits the function of the protein to which it binds. This translates into a therapeutic benefit for the patient.
• **Prediction of protein structure.** This is another of the applications of bioinformatics. Knowing the amino acid structure of proteins is feasible through the nucleotide sequence of the gene it encodes. This is very important to understand the function of the protein.
• **Search for counterparts.** The search for similarities between different molecules. This allows, for example, to find a function of a poorly characterized protein by looking for better-studied homologues. It is also used in genomics, to confirm coding regions in new sequenced genomic data.
• **Applications in medicine.** The application of information technology to medicine allows optimal administrative and hospital management with the use of electronic medical records.
• It can also allow the anticipated prediction of diseases by comparing evolutionary patterns of diseases, which will result in earlier diagnoses and therefore with a greater probability of cure.

Location and Masking Of Repeated Sequences

DNA repeats are located in both gene and intergenic regions and are classified into sparse repeats and tandem repeats.

Sparse repeat DNA is distributed throughout the genome, and there are different families of interleaved repeat elements.[1,13] (Tabla 3).

Repeats	Acronym	Characteristics
Repetitive palindromes	REP or PU	Between 21-65 bp Palindrome imperfect. Intragenic sequence
Bacterial dispersed mosaic elements	BIME	Between 40-500 bp. Tiled combination of REPs separated for other reasons
Short palindromic repeats arranged regularly forming a matrix	MITE	Between 100-400 bp. They are not autonomous. Flanked by inverted repeats
Intergenic repeating units	IRU or ERIC	Between 69-12 bp. Large palindromic sequences
Insert sequences	IS	Between 0.7-3.5 bp. Self-contained element flanked by inverted repeats
Bacteriophage elements		Phage Mu and transposable elements

Tabla Sparse repeat DNA families

Interleaved repeating DNA with broad repeats is made up of sequences that are repeated thousands of times in the genome but sparse, not in tandem.

This DNA is classified according to the size of the repeating unit, and can be differentiated.[26,30-34]

The SINEs are short dispersed nuclear elements. They make up 13% of the human genome. They are short sequences repeated thousands of times in the human genome in a sparse manner. The main SINE is the family of ALU elements that constitutes 10% of our genome. An ALU element is composed of a 250-280 nucleotide sequence with about 1,500,000 copies per genome and one repeat every 4 kb on average. It is an element rich in guanines and cytosines. It is predominantly located in the R bands of human chromosomes. It is flanked by small direct repeats. It has the structure of a non-identical dimer. It contains poly-A tails at the end of each monomer, and is transcribed by RNA polymerase III from an internal Promoter, but does not encode any

protein. It acts as a retrotransposon, since it can be copied into other regions of the genome.

LINES, are long dispersed nuclear elements and constitute 20% of the human genome. They are sequences with a size of several Kb, grouped in different families. The main LINE is called 1 (L1), formed by a sequence of about 6 kb repeated about 800,000 times in the genome, constituting around 15% of the genome. These elements are located predominantly in the G bands of chromosomes. An L1 element encodes two proteins, an RNA-binding protein in the ORF1 reading frame and a protein with endonuclease and reverse transcriptase activity in the ORF2 reading frame.

It is flanked by a few small direct repeats and ends in a poly-A tail. LINE elements are retrotransposons, since they can copy themselves through an RNA intermediate and transpose to other genomic locations.

HERVs, are human endogenous retroviruses and represent copies of human retroviruses that have integrated into the human genome in the course of evolution and are usually the origin of cellular proto-oncogenes. They represent truncated copies of the genome of these viruses and constitute about 8% of the genome. They are usually called LTR-like repeats, since they usually retain some of the long terminal repeats of these genomes.

DNA transposons, are DNA sequences that can move self-sufficiently to different parts of the genome of a cell. They make up 3% of the total genome. These elements contain the transposase gene, flanked by inverted repeats. It is worth highlighting the MER1 or MER2 type and the mariner elements responsible for some important chromosomal rearrangements in human pathologies.

Tandem repeat DNA consists of repeats of identical sequences that are arranged one after the other. These repetitions will be classified according to the length of the unit that is repeated and the number of repetitions that occur of said unit.

Therefore, they can be classified into
Satellites. These are sequences of between 5 and hundreds of nucleotides that are repeated in tandem thousands of times, thus generating regions of between 100 kb at several megabases. The Human Genome has a total of 250 Mb of satellite DNA. This satellite DNA can be classified in turn into:

- Satellite DNA 1: 42 nucleotide sequence.
- Satellite DNA 2: the repeat sequence is (ATTCCATTCG).
- Satellite DNA 3: pentamer repeats (ATTCC).
- Alpha Satellite DNA: the repeating sequence is 171 nucleotides in size. It is part of the DNA of the centromeres of human chromosomes.
- Beta Satellite DNA: 68 nucleotide repeat.
- Gamma Satellite DNA: 220 nucleotide repeat. They are found in the centromeric chromatin of various chromosomes.

Minisatellites: they are composed of a basic unit of 6 to 25 nucleotides repeated in tandem that form regions of between 100 and 20,000 base pairs. Some repeats of this type are polymorphic and give rise to the VNTR-like markers.

Microsatellites: they are repeated sequences of 1 to 6 nucleotides that are repeated until creating blocks with a size not exceeding 150 nucleotides. There are repeats of this type throughout the human genome and many of them are very useful as genetic markers.

- To analyze repeated DNA sequences through computerized means, we can use public databases or application software aimed at locating and masking repeated sequences.
- The public databases that can be used are

Aclame: Is a database dedicated to the collection and classification of mobile genetic elements (MGE) from various sources, comprising all known genome phages, plasmids and transposons. In addition to providing information on complete genomes and genetic entities, it aims to construct a comprehensive classification of functional modules of MGE such as proteins and genes.

CBS Genome Atlas
CRISPRdb. It is a database in which the microbial genomes have been pre-processed in search of structures. It has a complementary software, FlankAlign that is designed to align the sequences that falter. It is useful for the identification and comparison of repeated sequences.

IS-Finder. This database provides a list of isolated insertion sequences of eubacteria. It is organized into individual files that contain its general characteristics, as well as its DNA and main protein sequences. Although most of the entries have been identified as individual elements, a growing number are included in the description of the sequence of bacterial genomes.

MICdb. MICAS is a user-friendly and interactive web-based analysis server dedicated to searching for non-yielding microsatellites from a selected bacterial genome sequence. MICdb, is the database on which this web server is based. It is a comprehensive relational database of perfect microsatellites extracted from sequenced genomes. The latest version is MICdb 3.0.

Prophage DB.
Tandem Repeats DB. This is a public repository of information on genomic DNA tandem repeats and contains a wide variety of tools for their analysis including query algorithms and filter capabilities.

Some of the applications used for the location and masking of repeated sequences are:

Censor (GIRI): this is a software tool that identifies sequences deposited in RepBase by similarities, and masks repeated sequences in human, rodent, plant and invertebrate sequences.

RepeatMasker (UWGC). Masks regions of low complexity and repeated sequences, due to similarities to the sequences found in RepBase, in primates, rodents, mammals, vertebrates, plants and Drosophila.

Apart from these softwares, we can find other applications such as: ADPLOT CRISPRFINDER, CRT, EULERALIGN, MREPATT, MREPS, PATTERN LOCATOR, PHOBOS, PILER, REAS RECON, REPEATFINDER REPEATOIRE EPEATSCOUT, REPET. REPSEEK, REPUTER, SPUTNIK, SSRIT, STAR, TRED AND TRF

Comparison Methods

Comparative Genomics is based on the study of the relationship between the structure and function of the genome through different species or strains. It is responsible for comparing gene and protein sequences from different genomes to explain functional and evolutionary relationships.

Comparison is a substantial part of science and are inherent in bioinformatics. Bioinformatics has its origins in the 1960s, with Dayhoff who, for the first time, collected all the amino acid sequences of proteins that were known until then and systematically compared them.

The close relationship that is created between sequence and function, and the ability to biologically quantify the resemblance between sequences has been one of the pillars of Modern Molecular Biology. That is why in the 90s the BLAST program stood out, by means of which a sequence, be it amino acids or nucleotides, can be compared with all the sequences in a database.

Two of the softwares used for the comparison are.[1-6,35-37]

- **BLAST (NCBI).** It uses various programs to search for related sequences in protein and DNA sequence databases, including ESTs.
- **Procrustes (USC).** It is based on the modeling of genes in eukaryotes through the spliced alignment algorithm that uses the amino acid sequence of proteins similar to those encoded by the analyzed genomic sequence, to reconstruct the structure of genes in terms of introns and exons.

Analysis of The Dna Sequence At The Nucleotide Level

DNA sequencing

It is the set of biochemical methods and techniques that have the purpose of determining the order of nucleotides (A, C, G and T) in a DNA oligonucleotide. The DNA sequence makes up the heritable genetic information that forms the basis for the development of living things. The DNA sequence can be used to determine somatic mutations generated between organisms. Current techniques have enabled faster sequencing, which has influenced the Human Genome Project.

The classical chain termination method requires a single stranded DNA template strand, a DNA primer, a DNA polymerase with radioactively labeled nucleotides, and modified nucleotides terminating the DNA strand.

The DNA sample is divided into four separate sequencing reactions containing the four standard deoxynucleotides (dATP, dGTP, dCTP, and DTTP) and a DNA polymerase.

The incorporation of a dideoxynucleotide into the nascent DNA strand terminates its extension, producing several DNA fragments of varying length. The dideoxynucleotides are added at concentrations low enough to produce all possible fragments and at the same time to be sufficient for sequencing. The synthesized and re-labeled DNA fragments are heat denatured and size separated by polyacrylamide-urea gel electrophoresis. Each of the four synthesis reactions is run in individual lanes and the DNA bands are visualized by autoradiography or ultraviolet light, and the DNA sequence can be read directly from the X-ray plate or the gel image.

Technical variations of the chain termination sequencing method still exist. DNA fragments are marked with radioactive phosphorus-labeled nucleotides. The DNA fragments labeled with dyes are easy to read through an optical system, this allows a faster and cheaper analysis, sequencing by means of dyes coupled to the primer.

The various methods of DNA strand termination have simplified the amount of work and planning required for DNA sequencing.

The Sequenase kit is based on the chain termination method and contains most of the reagents necessary for sequencing pre-aliquoted and ready to use.

Some sequencing problems in the Sanger method such as non-specific junctions of the DNA primer that affect the interpretation of the sequence. It can also affect the fidelity of the sequence.

The three main nucleotide databases are

EMBL. The European nucleotide file provides a comprehensive record of information on nucleotide sequencing. Captures and presents information regarding experimentation workflows that are based around nucleotide sequencing. Data comes from a wide variety of sources: raw data presentations, assembled sequences, and annotations from small-scale sequencing efforts.

GenBank. It is the NIH genetic sequence database. It is

a collection where all publicly available DNA sequences are annotated. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory, and GenBank. These three organizations exchange data in a database. It is designed to provide and promote access by the scientific community to the most up-to-date and complete information on DNA sequences. It does not impose restrictions on the use or distribution of the data, although some may apply for patents, copyrights or other intellectual property rights on all or part of the data presented.

DDBJ. It is responsible for collecting nucleotide sequence data as a member of INSDC, which is an International collaboration on nucleotide sequence databases and provides nucleotide sequence data to support research activities.
Some of the software used are:

CodonW (Pasteur). This is a program used to calculate various statistics related to the use of codons in a sequence. It is designed to simplify multivariate analysis of codon and amino acid usage. Calculate standard indices of codon usage and with menus and line interfaces.

Codon Usage Database (KDRI). It is a database made up of codon usage tables.

Sequence Manipulation Suite (Bioinformatics.org). Collection of JavaScript programs for the generation, formatting and analysis of DNA and short protein sequences. It is commonly used by molecular biologists for teaching, program, and algorithm testing. Use other platforms, including Codon Plot and Codon Usage.

- **Codon Plot.** Which accepts a DNA sequence and generates a graphical representation consisting of a horizontal bar for each codon. The length of the bar is proportional to the frequency of the codon in the frequency table. It is used to find portions of DNA sequence that may be mis-expressed, or to view a graphical representation of a codon usage table.
- **Codon Usage.** Which accepts one or more DNA sequences and returns the number and frequency of each type of codon. From this program the frequencies of codons that code for the same amino acid are compared. It can be used to assess whether a sequence shows a preference for certain codons.

Signal Analysis
It is based on the identification of characteristic sequence motifs of the elements that make up genes such as promoters, start and stop codons, etc.

The degree of preservation of these motifs varies considerably. Those that have been conserved can be identified by searching with consensus sequences that represent the motif sequence for a certain majority of examples.

Consensus sequences do not contain information on the frequency with which each nucleotide occurs at each position. Said information can be expressed in the form of profiles, that is, in tables where the frequency with which each nucleotide appears throughout the sequence is recorded.

There are numerous applications capable of analyzing the signals from DNA sequences. Among the most prominent we can talk about:

- **WWW Promoter Scan (BIMAS),** which is based on the prediction of promoter regions that are identified by the existence of a series of possible binding sites for transcription factors, normally associated with promoters recognized by RNA polymerase.
- **Promoter 2.0 Prediction Server (CBS**), is based on the prediction of promoters recognized by RNA polymerase II in vertebrates.
- **Promoter prediction (BDGP),** works on the prediction of promoters through neural networks, in prokaryotes and eukaryotes.
- **NetStart 1.0 (CBS),** based on the prediction of start codons in vertebrates and Arabidopsis, by means of a neural network.
- **AUG (ITBA),** estimates the prediction of start codons.
- **MatInspector (GBF),** based on the prediction of transcription factor binding sites, based on the Transfac database.
- **CpG Plot (EBI),** works on the identification of CpG islands, which are unmethylated regions of the genome, associated with the 5 'end of genes in vertebrates. They normally overlap with the promoter and approximately the initial 1000 bp of the transcription unit.
- **Fuzznuc (Pasteur),** is based on the identification of motifs in sequences, specified as a consensus sequence with possible ambiguities and failures.
- **DPInteract (Harvard U),** it is a database of binding sites for transcriptional regulatory proteins in *Escherichia coli.*
- **Baylor College of Medicine's BCM Search Launcher,** is a web resource that compiles multiple applications for DNA and protein sequence analysis. It is a starting point to carry out a specific type of analysis such as sequence or protein multi-alignment, or the prediction of the structure of a protein. Sequence retrieval systems or search engines are:
- **Sequence Retrieval System,** system used by the European Bioinformatics Institute to access those databases found on its servers. They are publicly accessible.
- **Entrez,** which is the system used by the NCBI to access the GenBank, OMIM, snpBD, PubMed or Gene databases.
- **Google** that, in the case of having the access code to

the sequence of a gene, it is possible to use it to locate the record that contains that sequence.

For sequence alignment, and to determine whether a DNA sequence is identical or not at all like any other sequence in the database, various programs can be used that allow sequence alignment. These types of programs use algorithms that they use to compare nucleotide sequences that are treated as if they were texts or words. Some compare large pieces of text and others use other terms that allow ambiguity; these are known as FASTA or BLAST.

Other useful tools for DNA and protein sequence analysis include:

- **Translate. I**s an application that allows you to search for open reading frames in the DNA sequences that you provide. It allows the visualization of the ORFs in the 3 reading frames of the codons in a sequence.
- **ProtParam.** Application that predicts the theoretical physicochemical parameters of a protein from a sequence.
- **MotifScan.** Which allows searching for protein domains in a protein sequence.
- **InterPro Scan,** uses 12 different applications to find the supplied sequence, protein domains, protein fingerprints and signals, and protein structure profiles in various databases.
- **Prediction of the secondary structure by the HNN method.** It is one of the programs that predicts the secondary structure of proteins based on their sequence.

Search in Sequence Databases

One of the most widely used databases in Bioinformatics is the sequence database. It is a collection of DNA, protein and other sequences, which are stored in computers. These databases can include sequences from a single organism or can include sequences from all organisms.

An expressed sequence marker is a small subsequence of a transcribed nucleotide sequence. These markers are used for the identification of transcribed genes and for the determination of sequences.

These markers are produced through the execution of sequences on a cloned mRNA. The resulting sequence is a poor quality fragment, between 500 and 800 nucleotides in length.

The source of the data for the expressed sequence markers is dbEST. Is a division of GenBank that contains sequence data and other information on expressed DNA sequences or sequence markers that come from a number of different organisms.

There are primary databases, which contain direct information on the sequence, structure or expression pattern of DNA or protein; and the secondary databases that contain data and hypotheses derived from the analysis of the primary databases, such as mutations, evolutionary relationships, groupings (by families or functions), involvement in diseases, etc (2,3,5,6).

The tools most commonly used are the following:

- **BLAST (NCBI).** It is a tool whose function is to search for related sequences in databases of ESTs and cDNAs. It is a local sequence alignment computer program (DNA, RNA or proteins). This program is capable of comparing a problem sequence with a large number of sequences that are found in a database. It uses a heuristic algorithm so it cannot guarantee that the correct answer has been found. Used to locate homologous genes
- **The mamalian Gene Collection.** It is a database of human and mouse cDNAs. Its goal is to provide researchers with the encoding of valid protein sequences from human mice and rats.
- **HUNT (human novel transcripts).** This database aims to generate publicly available full-length clones. Contains more than 4,300 long-lived DNA clones.

The German Human cDNA Project. It is a database of tissue or chromosome specific human cDNAs

Figure 3 shows schematically how bioinformatics works.

Figure 3: Schematic representation of how bioinformatics works.

Types of Biological Databases

Depending on the type of information and the elements, different types of databases can be distinguished, being able to distinguish the following:

- **Primary databases (databases**), are those that save and store the data as it was entered by those who generated it.
- **Secondary databases (derived**), are those that start from the analysis of the information stored in the primary databases to establish new relationships between the data or even discover new properties of these.
- **Composite databases,** these are databases that have been merged. This type of database avoids conducting multiple searches.
- **Specific databases** are those databases that only contain information related to a specific organism, or are dedicated to the study of a specific type of molecule.

1. Cross references with other databases

These refer to an item that appears elsewhere in the document. Cross references are marks that at a certain point in the document refer to or link to another place in the document. The advantages of cross-referencing long documents are numerous and important.

A cross reference helps us to help the reader in his understanding of the document that we present to him, pointing out at a certain moment parts of the document, its physical location, relevant to the one in question.

To understand the operation of databases it is necessary to know the concepts of relational databases. A relational database is one that complies with the relational model, which is the most used today to implement databases that have already been planned. They allow to establish relationships between the data, and through connections to relate the data of both tables.

In biological databases, the content of the data that is included in it, have gene sequences, textual descriptions, attributes, ontological classifications, annotations and data in tabular form.

This data is often described as semi-structured data and can be represented by tables, key-delimited records, and XML structures. In this type of tables it is very common to use cross references between databases using access numbers.

SWISS-PROT is a database rich in cross references to other databases and has a format similar to the EMBL Nucleotide Sequence Database.

2. Sequence databases

These databases contain the sequences that characterize genes from humans and other species. Apart from the data pair sequences, they contain information related to the gene name, references to articles that have studied those gene sequences and the characteristics of the gene structure.

The main databases of DNA sequences are: GenBank. EMBL and DDBJ.

These databases share and complement their information in such a way that in practice, any search carried out on any of them includes the search on the others.

Protein databases contain information on their protein or enzyme function and classification, the size of the molecules, their structure, their component regions, and the variants in the population. The main databases in this case are: SWISSPROT, PIR, MAPview and Genome Browser.

3. Main databases

Sen groups in databases of nucleotides, proteins and genomes.

*** Nucleotides**

The collaboration of the three most important databases makes it possible to access almost all DNA sequence information from any of its three locations:

- **EMBL-BANK** at the European Bioinformatics Institute (EBI).
- **DNA Data Bank of Japan (DDBJ)** at the Center for Biological Information (CIB).
- **GenBank** at the National Center for Biotechnology Information (NCBI).

Although they are maintained by different organizations in different countries, there is coordination between the different bases. A sequence sent to any of the bases will be reflected in the other two in about a week, since that is the update frequency between the different genetic bases. For this reason it is indistinct which base is used to send new sequences, although normally Europeans use EMBL-BANK and Americans GenBank.

EMBL-BANK captures and presents information regarding experimentation workflows that are based on nucleotide sequencing. The typical workflow involves isolation and preparation of material for sequencing. ENA (European Nucleotide Archive) records the information obtained in a data model that collects input information such as sample, experimental setup and machine setup, output machine data such as sequence traces, quality scores and interprets the information as montage, cartography and functional annotation.

The data arriving at ENA comes from various sources and includes raw data presentations, assembled sequences, small-scale sequences, provision of main data from European sequencing centers, and exchange in the nucleotide sequence database. This provision of data to ENA is a mandatory step for the dissemination of the results originated by the research of the scientific community. ENA has a group of editors and scientists who seek the guarantee that leads to providing optimal presentation systems and tools for accessing databases that work perfectly.

DNA Data Bank of Japan (DDBJ), plays an important role in bioinformatics research. The purpose of DDBJ is to improve the quality of INSD as public domains. When researchers make data available to the public through INSD and it is shared around the world, DDBK focuses its efforts on richly describing the data.

The nucleotide sequence records the evolution of organs more directly than other biological materials, making it of great value to the life sciences and to human well-being in general. This database could be described as a common treasure of human beings. Related to this precept, DDBJ can be considered as an online database accessible to anyone.

DDBJ operates within the National Institute of Genetics of Japan and is endorsed by the Japanese Ministry of

Education, Culture, Science and Technology.

DDBJ contributes internationally as a member of INSDC in the collection and proportion of nucleotide sequence data with ENA and NCBI. It is certified to collect nucleotide sequences from researchers and issue the internationally recognized accession number to the senders of such data. It primarily collects sequence data from Japanese researchers, but accepts data and issuance of accession numbers from other countries.

GenBank (NCBINational Center for Biotenbology Information) is a public collection of annotated nucleotide sequences. It includes sequences of mRNA with coding regions, genomic DNA, corresponding to one or more genes and ribosomal RNA. It is a store of sequences that cannot be modified without first obtaining the consent of the authors.

To facilitate the search, they have segmented the sequences into different divisions. There are taxonomic divisions for high quality sequences, such as primate, rodent, other mamalian, invertebrate, etc; and other divisions that are reserved for the lower quality sequences (EST, GSS, HTGS).

The sequences that come from new sequencing systems are not deposited individually in the GenBank Database, but there is a special database where the complete files (SRA) are stored.

Refseq (Reference Sequence) is another nucleotide database maintained by the NCBI. It is a secondary database that has a collection of DNA, RNA and protein sequences that is maintained and reviewed manually. It has independent entries for genomic, transcript, and protein DNA. It only has one sequence per gene and organism. At the entrance of transcripts and proteins there will be as many sequences as transcripts and proteins are capable of coding for the gene. In the 2014 version it had 36,335 of the main agencies, compared to more than 160,000 of GenBank.

*** Protein.**

The amino acid sequence databases

- **Swiss-Prot**, contains annotated or annotated sequences, that is, each sequence has been reviewed, documented and linked to other databases.
- **TrEMBL by Translation of EMBL Nucleotide Sequence Database** includes the translation of all coding sequences derived from (EMBL-BANK) and that have not yet been annotated in Swiss-Prot.
- **PIR for Protein Information Resource,** is divided into four sub-bases that have decreasing annotation level. -ENZYME links the complete enzyme activity classification to the Swiss-Prot sequences.
- **PROSITE,** contains information on the secondary structure of proteins, families, domains, etc.
- InterPro integrates the information from various secondary structure databases such as PROSITE, providing links to other databases and more extensive information.

Protein Data Bank (PDB), is the 3-D tertiary structure database of proteins that have been crystallized.

Uniprot. It is a protein database that was created by the union of the Swiss-Prot, TrEMBl and PIR-PSD databases.

It was born from the union of three previous databases: Swiss-prot, TrEMBL and PIR-PSD. The Swiss-Prot and TrEML bases continue to be two sections at Uniprot. Swiss-Prot is made up of manually annotated records and the proteins contain high-quality information. On the other hand, TrEMBL is a machine translation of the EMBL sequences. Uniprot has a selection of reference proteins called UniRef100, 90, 50.

PIR (Protein Information Resource)

It is a public bioinformatics resource that supports genomics, proteomics research, and systems biology.

PIR was created in 1984 by the National Biomedical Research Foundation to serve as a resource to assist researchers in identifying and interpreting protein sequence information. Before PIR, the Foundation compiled the first complete collection of macromolecular sequences in the Atlas of Protein Sequence and Structure, which was made public in 1965. This group was the first to develop computer methods for sequence comparison. of proteins and duplications within the sequences. Barker and Ledley assumed subsequent leadership of the project. In 1999 Wu joined the Foundation, and later Georgetown University Medical Center (GUMC).

For many years, PIR has provided, from the Atlas of Protein Sequence and Structure, databases and protein analysis tools freely accessible to the scientific community, including the Protein Sequence Database (PSD).

In 2002, PIR, together with EBI and SIB (Swiss Bioinformatics Institute), were awarded a grant from the NIH for the creation of Uniprot, which provided a single database of protein sequence worldwide, through the Unification of the PIR-PSD, Swiss-Prot, TrEMBL Databases.

Enzyme

It is a repository of information related to the nomenclature of enzymes. It is based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology and describes each type of enzyme characterized by which, the EC number (Commission on Enzymes) has been provided. This database contains the following data for each type of enzyme characterized: EC number, recommended name, alternative names, catalytic activity,

cofactors, Swiss-Prot pointers, and associated human disease pointers. Some related tools and databases are: BRENDA, IUBMB ExplorEnz Enzyme Database, KEGG, MetaCyc, IUBMB Enzime Nomenclature, and BioCarta.

Prosite
It consists of documentation entries that describe protein domains, families, and functional sites, as well as associated patterns and identification profiles. It is based on the observation that, while there are a large number of different proteins, most of them can be grouped into a limited number of families. Proteins or protein domains that belong to a particular family generally share functional attributes. Currently PROSITE contains specific patterns and profiles for more than a thousand protein families. Each of these signatures comes with documentation that provides background information on the structure and function of these proteins.

InterPro (Protein sequence analysis and classification)
It provides a functional analysis of proteins by classifying them into families, and the prediction of important domains and locations. It combines protein signatures from a number of member databases from a single search resource - capitalizing on their individual strengths to produce a powerful integrated database and diagnostic tool.

Is a resource that provides a functional analysis of protein sequences. To classify proteins into families, it uses predictive models known as signatures provided by various databases that make up the InterPro consortium.

Combine signatures from multiple databases in a single search, reducing redundancy and helping users efficiently interpret the results of sequence analysis.

PDB (RCSB, Protein Data Bank)
Includes information on protein and nucleic acid structure. In addition to the structure, it includes information on the sequence, crystallization conditions, other proteins of similar structure and 3D images.

***Of genomes**

- **Ensemb.** Integrating large eukaryotic genomes, it currently contains human, mouse, rat, fugu, zebrafish, mosquito, *Drosophila, C. elegans,* and *C. briggsae* genomes.
- **Genomes server and TIGR.** They are portals with information or links to all genomes sequenced at the moment, from viruses to humans.
- **Wormbase.** It is the portal of the C. elegans worm genome.
- **Flybase**. It is the portal of the vinegar fly Drosophila melanogaster.

Genomes server and TIGR
The European Nucleotide Archive has a server called Genomes server from which it is possible to access the complete genomes of viruses, phages and organelles, which were deposited in the EMBL database during the 1980s.

Since then, molecular biology has undergone a change trying to obtain the complete sequences of as many genomes as possible, combining them with important sequencing technology developments, which have resulted in hundreds of complete genome sequences being added to the database. These web pages facilitate access to a large number of complete genomes.

Wormbase
It is an international Consortium of biologists and computer scientists dedicated to providing accurate, up-to-date and accessible information regarding the genetics, genomics and biology of C. elegans and related nematodes.

Flybase
It consists of a consortium of *Drosophila* researchers and computer scientists at: Harvard University, Cambridge University (England), Indiana University, and the University of New Mexico.

Software Tools
In bioinformatics they range from simple command line tools to much more complex graphical programs and autonomous web services located in bioinformatics companies or public institutions. The best known computational biology tool among biologists is probably BLAST, an algorithm to determine the similarity of arbitrary sequences with other sequences probably resident in protein or DNA sequence databases. The NCBI (National Center for Biotechnology Information, USA), provides a widely used, web-based implementation that works on its databases.

For multiple sequence alignments, the classic ClustalW, 70 currently in version 2, is the reference software. You can work with an implementation of the same in the EBI (European Institute of Bioinformatics).

BLAST and ClustalW, are just two examples of the many sequence alignment programs available. On the other hand, there is a multitude of bioinformatic software with other objectives: structural alignment of proteins, prediction of genes and other motifs, prediction of protein structure, prediction of protein-protein coupling, or modeling of biological systems, among others. In Annex: Software for sequence alignment and Annex: Software for structural alignment can be found respective relationships of programs or web services suitable for each of these two objectives in particular.

Free software in bioinformatics
The need for new algorithms for the analysis of new data of biological origin, in combination with the potential for innovative experiments in silico and the availability of free repositories for free software have helped create

opportunities for research groups to contribute to bioinformatics and to the free code available, regardless of your funding sources. Open source tools often act as incubators for ideas, or as a complement to commercial applications. They can also provide standards and models or structures that contribute to the challenge of integration in bioinformatics.

Free software includes Bioconductor, BioPerl, Biopython, BioJava, BioJS, BioRuby, Bioclipse, EMBOSS, NET Bio, Orange with its bioinformatics add-ons, Apache Taverna, UGENE and GenoCAD. To maintain this tradition and create new opportunities, the nonprofit Open Bioinformatics Foundation has sponsored the Bioinformatics Open Source Conference (BOSC) annually since 2000.

An alternative method of building public databases is to use MediaWiki wiki software with the WikiOpener extension. This system allows access and updating of the database to all experts in the field.

Web services in bioinformatics

1. Interfaces based on SOAP and REST (Representational State Transfer) have been developed for a wide variety of bioinformatics applications, allowing an application, running on a computer anywhere in the world, to use algorithms, data and computing resources hosted on servers. in any other part of the planet. The main advantages are that the end user does not worry about updates and modifications in the software or in the databases.
2. The basic bioinformatics services, according to the EBI classification, can be classified as: Online information gathering services, with database queries. Analysis tools as services that give access to EMBOSS. Searches for similarities between sequences, FASTA or BLAST access services. Multiple sequence alignments, with access to ClustalW or T-Coffee. Structural analysis with access to protein structural alignment services. And access services to specialized literature and ontologies.
3. Since 2009, basic bioinformatics services are classified by the EBI into three categories, such as sequence similarities (SSS), multiple sequence alignments (MSA) and bioinformatics sequence analysis (BSA).
4. The availability of these SOAP-based web services through systems such as registry services demonstrates the applicability of web-based bioinformatics solutions. These tools range from a collection of stand-alone tools with a common data format, and under a single stand-alone or web-based interface, to integrative and extensible systems for managing bioinformatics workflow.

Bioinformatics workflow management systems

1. It is a specialized form of workflow management specifically designed to compose and execute a series of computational or data manipulation steps, or a workflow, in a Bioinformatics application. Such systems are designed to:
2. Where an easy-to-use environment is provided for individual application scientists themselves to create their workflows. They provide interactive tools for scientists that allow them to run their workflows and view their results in real time. They simplify the process of sharing and reusing workflows among scientists. And it enables scientists to trace the origin of the workflow execution results and the steps of creating the workflow.
3. Some of the platforms that offer this service: Galaxy, Kepler, Taverna, UGENE, Anduril, HIVE. BioCompute and BioCompute Objects (BCO)
4. In 2014, the US Food and Drug Administration sponsored a conference held at the National Institutes of Health in Bethesda to discuss reproducibility in bioinformatics. Over the next several years, a consortium of stakeholders met regularly to discuss what would become the BioCompute paradigm. These stakeholders included representatives from government, industry, and academia. Session leaders represented numerous chapters of the FDA and NIH Institutes and Centers, non-profit entities such as the Human Varioma Project and the European Federation for Medical Informatics, and research institutions such as Stanford, the New York Genome Center, and George Washington University.
5. It was decided that the BioCompute paradigm would be in the form of digital laboratory notebooks that allow the reproducibility, replication, review and reuse of bioinformatics protocols. This was proposed to allow for greater continuity within a research group in the course of the normal staff flow while encouraging the exchange of ideas between groups. The FDA funded this work to make the information more transparent and accessible to its staff.
6. In 2016, the group met again at the NIH in Bethesda and discussed the potential of a BioCompute Object, an instance of the BioCompute paradigm. Its objects allow records to be shared between regulatory partners and employees.
7. In the future, the different biological research projects will require knowledge of the domain, the elasticity of cloud computing, and efficient access to data that can create scalable applications (Figure 4).

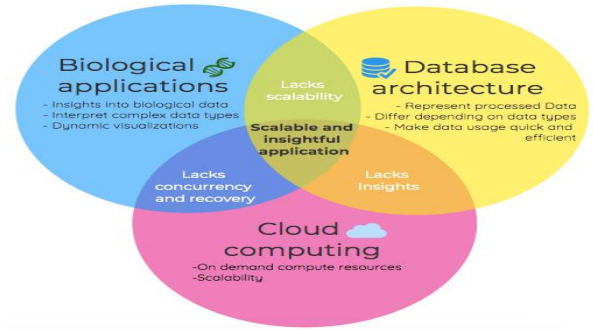

Figure 4: Future applications of the different research projects.

CONCLUSIONS

Data warehousing appears in bioinformatics to support the discovery of biological knowledge and to facilitate research and information exchange. So biological web applications have revolutionized research.

Bioinformatics makes a more correct use and takes advantage of all the technical characteristics and it is necessary to define the platforms and the implementation of the processes. There is no single approach to all the problems that we encounter in molecular biology and it is necessary to apply and develop existing methodologies to specific problems and generate new methodologies.

REFERENCES

1. Allen GK. Bioinformatics: new technology models for research, education, and services. Educause Centre for Applied Research Bulletin, 2005; 8: 1-9.
2. Xion J. Essential Bioinformatics. EEUU: Cambridge University Press, 2006. ISBN 978-0-511-16815-4.
3. Phoebe Chen YP. Bioinformatics Technologies. Alemania: Springer-Verlag Berlin Heidelberg, 2005; 396. ISBN 3-540-20873-9.
4. Harjinder S G, Prakash CR. Data Warehousing. La Integracion de Informacion para la mejor toma de decisiones. México: Prentice Hall, 1996. ISBN 968-880-792-3.
5. Attwood T.K., Gisel A., Eriksson N-E. and Bongcam-Rudloff E. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. *Bioinformatics - Trends and Methodologies*. InTech., 2011.
6. Attwood TK., Parry-Smith, DJ. Introducción a la Bioinformática. Prentice Hall, 2002. ISBN 84-205-3551-6.
7. Aluru, Srinivas, ed. Handbook of Computational Molecular Biology. Computer and Information Science Series. Chapman & Hall/Crc, 2006. ISBN 1-58488-406-1.
8. Baldi P, Brunak, S). Bioinformatics: The Machine Learning Approach, 2nd edition edición. MIT Press, 2001. ISBN 0-262-02506-X.
9. Barnes, M.R. and Gray, I.C., eds. Bioinformatics for Geneticists. First edition edición. Wiley, 2003. ISBN 0-470-84394-2.
10. Baxevanis D, Ouellette BFF, eds. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. En: Tthird edition edición. Wiley, 2005. ISBN 0-471-47878-4.
11. Cristianini, N. and Hahn, M. Introduction to Computational Genomics. Cambridge University Press, 2006. ISBN 978-0-521-67191-0 y 0-521-67191-4.
12. Baxevanis AD, Petsko GA, Stein LD, Stormo, GD, eds. Current Protocols in Bioinformatics. Wiley, 2007. ISBN 0-471-25093-7.
13. Durbin, R., S. Eddy, A. Krogh and G. Mitchison. Biological sequence analysis. Cambridge University Press. ISBN 0-521-62971-3, 1998.
14. Michael S W. Introduction to Computational Biology: Sequences, Maps and Genomes. CRC Press, 1995. ISBN 0-412-99391-0.
15. Mount DW. Bioinformatics: Sequence and Genome Analysis. 2ª ed. edición. Spring Harbor Press, 2004. ISBN 0-87969-712-1.
16. Pevzner, Pavel A. Computational Molecular Biology: An Algorithmic Approach. The MIT Press, 2000. ISBN 0-262-16197-4.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol., 1990; 215(3): 403-10.
18. Degrave,W, Leite L, Huynh CH. Fiocruz distance-learning website. Oswaldo Cruz Foundation, Oswaldo Cruz Institute, 2001. Dept. of Biochemistry and Molecular Biology, Río de Janeiro, Brazil. Available in: http://www.dbbm.fiocruz.br/helpdesk
19. Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3 - new capabilities and interfaces. Nucleic Acids Research, 2012; 40(15): 115.
20. National Institute of Health, NIH Working Definition of Bioinformatics and Computational Biology. Available in: http://www.bisti.nih.gov /CompuBioDef.pdf, 2000.
21. Hall, T.A. BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 1999; 41: 95-98.
22. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 2004; 32(5): 1792-1797.
23. Geer RC, Sayers ErW. Entrez: Making use of its power. En: Briefings in Bioinformatics, 2003; 4(2): 179.
24. Jena RK, et al. Soft computing Methodologies in Bioinformatics. En: European Journal of Scientific Research, 2009; 26(2): 193.
25. Bandyopadhyay S. An efficient technique for super family classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection. En: Journal Fuzzy Sets and Systems, 2005; 152(1): 5-16.

26. Cordón O, et al. Ten years of genetic fuzzy systems. En: Fuzzy Sets and Systems, 2004; 141(1): 5-31.
27. Tomida S, et al. Analysis of expression profile using fuzzy adaptive resonance theory. En: Bioinformatics, 2002; 18(8): 1073-1083.
28. Fu L. Knowledge Discovery Based on Neural Networks. En: Communications of the ACM (CACM), 1999; 42(11): 47-50.
29. Schlosshauer M Ohlsson M. A novel approach to local reliability of sequence alignments. En: Bioinformatics, 2002; 18(6): 847-854.
30. Uberbacher E, Mural R. Locating Protein Coding Regions in Human DNA Sequences Using a Multiple Sensor-Neural Network Approach. En: Proceedings of the National Academy of Sciences of United States of America, 1991; 88: 11261-11265.
31. Feng, Z, et al. Ligand Depot: a data warehouse for ligands bound to macromolecules. En: Bioinformatics Applications Note [on line], 2004; 20(13). Available in: http://bioinformatics. oxfordjournals.org/content/20/13/2153.full.pdf+ html?sid=5fbc13fd-7bee-4364-829bef27e2d53032.
32. Ressom H, Reynolds R, Varghese R. Increasing the efficiency of fuzzy logic based gene expression data analysis. En: Physiological Genomics, 2003; 13(2): 107-117.
33. Woolf, Peter y Wang, Yixing. A fuzzy logic approach to analyzing gene expression data. En: Physiological Genomics, 2000; 3(1): 9-15.
34. BioStar models of clinical and genomic data for biomedical data warehouse design [on line]. Wang L; Ramanathan M, Zhang A. State University of New York at Buffalo: New York, EEUU, 2005. Available in: http://www.cse.buffalo.edu /DBGROUP/bioinformatics/papers/ijbra05.pdf.
35. Torres A, Nieto J. The Fuzzy polynucleotide space: basic properties. En: Bioinformatics, 2003; 19(5): 92.
36. Nisbet, R. Bioinformatics. Handbook of Statistical Analysis and Data Mining Applications. John Elder IV, Gary Miner. Academic Press, 2009. ISBN 9780080912035.
37. Simonyan, Vahan; Goecks, Jeremy; Mazumder, Raja. Biocompute Objects A Step towards Evaluation and Validation of Biomedical Scientific Computations». PDA Journal of Pharmaceutical Science and Technology, 2017; 71(2): 136-146.

Other web pages

- National Center for Biotechnology Information. NCBI. NCBI/BLAST Home.
- Instituto Europeo de Bioinformática - EBI EMBL-EBI: ClustalW2.
- Open Bioinformatics Foundation: BOSC. Official website. Open Bioinformatics Foundation.
- Brohée S, Barriot R, Moreau Y. Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases. Bioinformatics, 2010; 26(17): 2210-2211. doi:10.1093/bioinformatics/btq348.
- European Bioinformatics Institute. 2006. EBI Web Services.
- https://www.biorxiv.org/
- BioCompute Object (BCO) project is a collaborative and community-driven framework to standardize HTS computational data. 1. BCO Specification Document: user manual for understanding and creating B., biocompute-objects, 2017.