الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي

Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research

Ahmed Zabana Relizane University
Faculty of Natural and Life science
Department of Ecology and environment

جامعة أحمد زبانة-غليزان
Ahmed Zabana Relizane University

COURSE HANDOUT

Intended for 1 nd year Master all biology specialties (LMD).

Title:

Bioinformatics

Developed by:

Dr. Aouadj Sid Ahmed

Academic year : 2024/2025

The preface

Here is a new introductory on Bioinformatics. One more? No, because it is part of the evolution and continuous enrichment of this subject and it testifies to its vitality and its maturation. Far from being still a standardized field, it is no longer in its infancy and its exposition has been enriched as experience has been acquired by teaching it to: students and to varied audiences.

This book represents a pragmatic choice based on teaching carried out in interaction with students. It is an original choice of subjects and it reveals throughout the pages the extensions of the basic elements towards more specialized subjects. In this sense, it constitutes a stimulating introduction, which makes beginners want to go further and reserves surprises for specialists in the field.

It will especially appeal to students with a taste for discrete mathematics, but also to those who like algorithms. We will thus find there as well a little-known result.

The presentation is exceptionally clear and the proofs are given with great care. A sign of the times of quality research, the exercises are accompanied by solutions which are the only guarantee that the exercise is feasible. So, have a good trip to the reader who is tackling this subject and to whom I wish as much pleasure as I experienced reading Basics of bioinformatics which allowed me to discover it myself – I believe it was in 2021.

Dr Aouadj Sid Ahmed

Abstract

Databases dedicated to molecular biology are an essential complement to literature data. Today there is a very wide variety of heterogeneous databases. This diversity is, of course, explained by the variety of biological data, which are not limited to sequences, but also by the variety of objectives which governed their design. The major problem in the management of biological data therefore does not result so much from their volume as from this heterogeneity, both in terms of nature and format. The fundamental question is therefore how to integrate these biological data in order to make them accessible and usable as easily as if they appeared in a same database. Examination of the different technical solutions proposed highlights the need, in all cases, to explain and formally represent the entities concerned and their relationships. A simple but complete modeling example illustrates this approach.

Abbreviations

RNA: Ribonucleic acid
EBI: European Bioinformatics Institute
ENCODE: Encyclopedia of DNA Elements
NCBI: National Center of Biotechnology Informatics
UCSC: University of California, Santa Cruz

Glossary

Exons = coding part of DNA
in slilico = bioinformatics
Introns = non-coding part of DNA
Biological molecule = protein, sugar..
Nucleotide = nitric base+sugar+phosphorus

Table of contents

I. Introduction

The aim of bioinformatics is to produce new knowledge on the functioning of the cells of living organisms, their evolution, their healthy or pathological state, etc. To do this, it first limited itself to genomics, which studies the structure, functioning and evolution of genomes. But it appeared that the representation of the cell given by genomics is static, and does not allow us to account for its evolution over time. Thus, was born post-genomics, which seeks to know when and under what conditions genes will trigger the production of proteins, and how the proteins produced are involved in the functioning of the cell.

Bioinformatics will allow two types of comparatives analyzes based on sequencing data: between cancer cells and normal cells of an organism, and between cancer cells of one organism and cancer cells of other organisms.

How to do bioinformatics?

The bioinformatician has acquired knowledge in both fields (biology and computer science) thanks to dual training, he works alongside biologists or doctors, computer scientists and biostatisticians.

Creating applications in very evolving fields, he maintains an ongoing dialogue with members of the research team, as well as with public and private research partners.

Exercise: Pre-test

DNA sequencing: is the determination of the sequence of nucleotides composing it:

 Sequencing of all coding genes: exome sequencing and Whole genome sequencing: genome sequencing.

 Whole genome sequencing: genome sequencing.

 Sequencing of all coding genes: introme sequencing.

 all

II. Introduction to bioinformatics

II. Introduction to bioinformatics

1. Bioinformation

Bioinformation is information linked to biological molecules: their structures, their functions, their "kinship" links, their interactions and their integration into the cell.

There are two types of bioinformation: the nucleotide sequence and the amino acid sequence. Sequences are one of the main types of bioinformation analyzed in bioinformatics.

Various fields of study make it possible to obtain this bioinformation: structural genomics, functional genomics, proteomics, determination of the spatial structure of biological molecules, molecular modeling, etc.

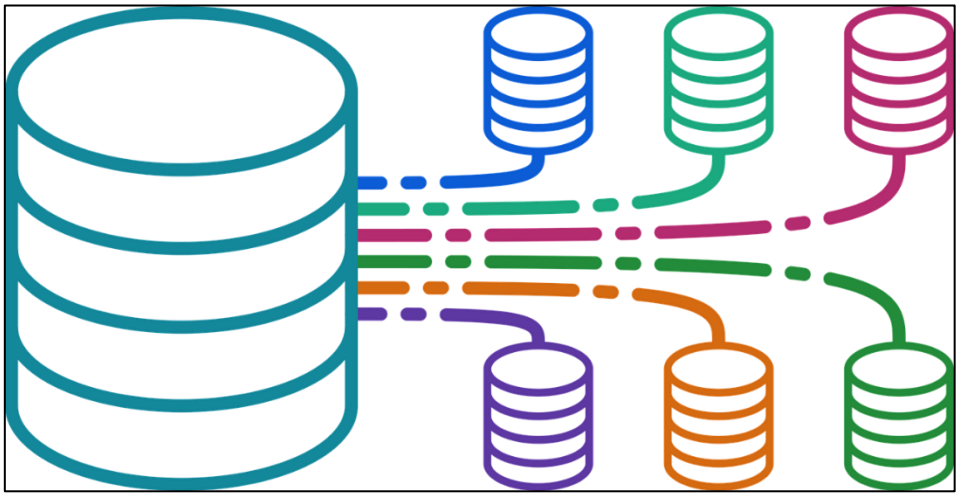


Figure 1. Bio-information.

2. Bioinformatics

Bioinformatics is a recent discipline and a multidisciplinary field of research where biologists, computer scientists, mathematicians and physicists work together with the aim of solving a scientific problem posed by biology. It is a discipline which allows the analysis and interpretation of biological information contained either in the genome (DNA, RNA sequences) or in the proteome (all bio-synthesized proteins), or in the transcriptome (transcribed mRNAs). It can also be defined as the discipline of "*in silico*" analysis of biological information contained in nucleic and protein sequences.

According to the National Center of Biotechnology Informatics NCBI: science in wich biology, computer sciences and information technology merge into a single discipline. "bioinformatics is the science in which biology, computer science and technology merge in a single science" :

• Multi-disciplinary fields involving biology, computer science, mathematics, statistics whose objective is to analyze biological sequences and predict the structure and function of macromolecules.

• Increasingly, bioinformatics is being developed for application to agriculture, pharmacology and medicine.

• Discipline which evolves according to new problems posed by biology.

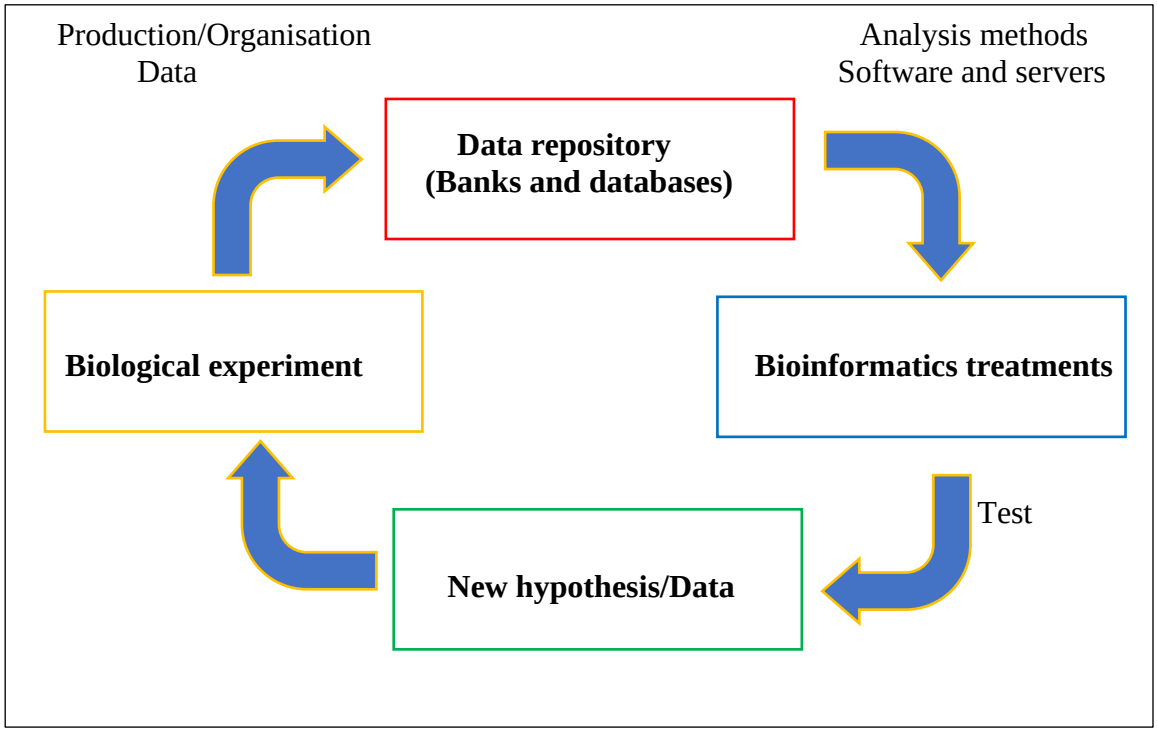


Figure 2. Bioinformatics.

3. Applications of bioinformatics

Bioinformatics has different objectives and different applications [1,amended] :

- Collect and store information in databases, accessible online.

- Provide sequence comparison tools (protein or nucleotide) in order to:

1. Identify a sequence in relation to a database

2. Determine the degree of similarities between two sequences (interest in taxonomy)

3. Identify structural patterns:

4. Genes, promoters, etc. for a nucleotide.

5. Fallback zone, active site, etc. for a polypeptide.

Figure 3. Sequencing (https://www.ddw-online.com/estonia-national-biobank-to-sequence-10000-whole-genomes-29158-202404/).

- Provide sequence translation tools in order to :

1. Simplify translation tasks

2. Propose several protein possibilities for the same sequence

3. Spot exons/introns

So Bioinformation is information linked to biological molecules (DNA / RNA / Proteins).

- Provide prediction tools (physiological and functional prediction) in order to :

1. Locate an operon

2. Spot an abnormal gene or protein

3. Predict the 3D structure of a protein

4. Spot mutations

5. Predict a pathology...

- Experimental prediction in order to:

1. Identify restriction sites

2. Predict the digestion of a nucleotide

3. Predict / simulate the migration of nucleotide or protein fragments during electrophoresis...

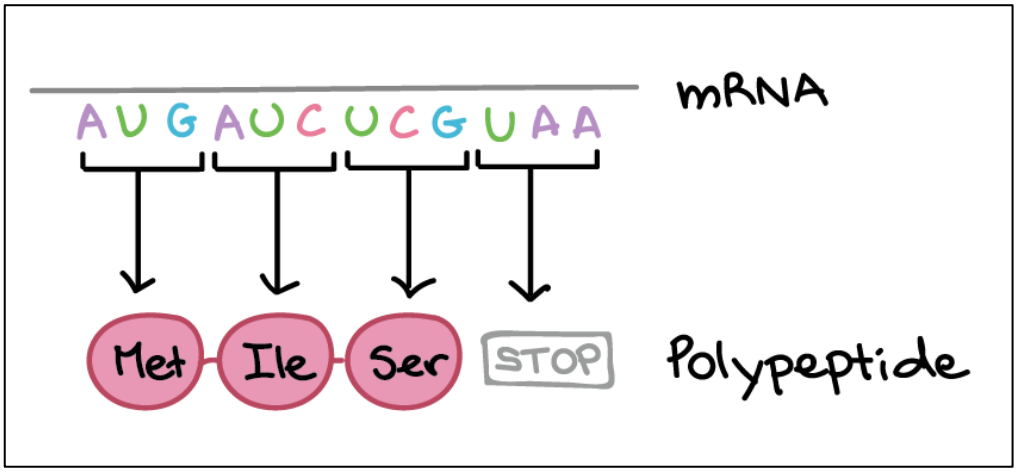


Figure 4. Translation of sequences.
(https://www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/the-stages-of-translation).

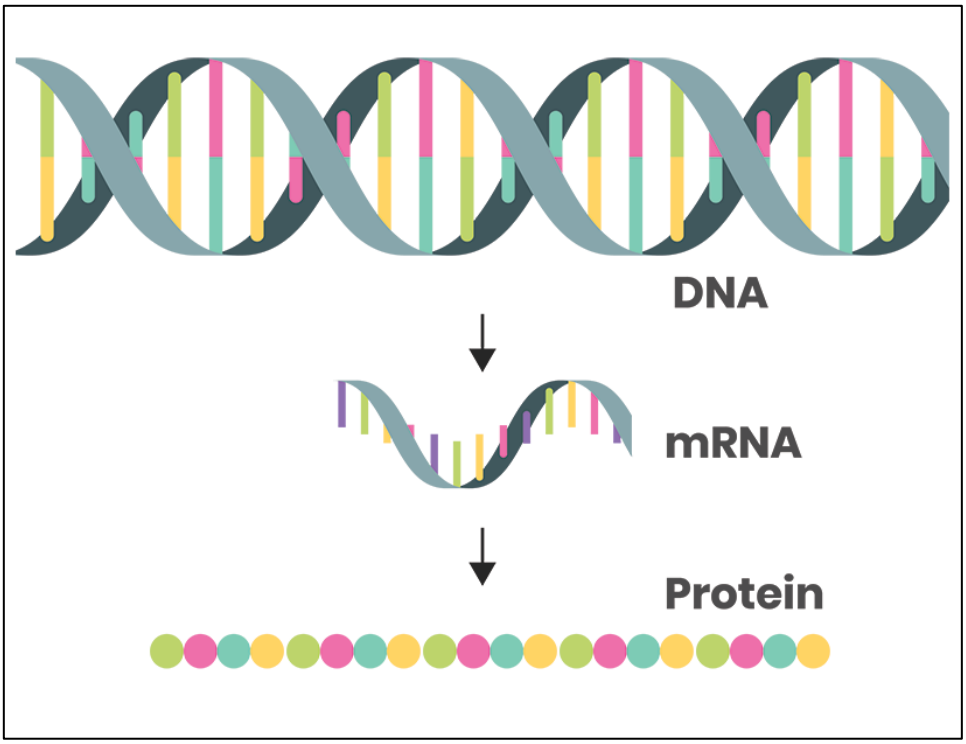


Figure 5. From DNA to cellular function.
(https://ni.vwr.com/cms/function-of-mRNA).

4. Evolution of biological sequences

Very difficult to accurately define an adequate model of sequence evolution. Biological problems are generally too complex to be solved by an exact algorithm in a reasonable time.

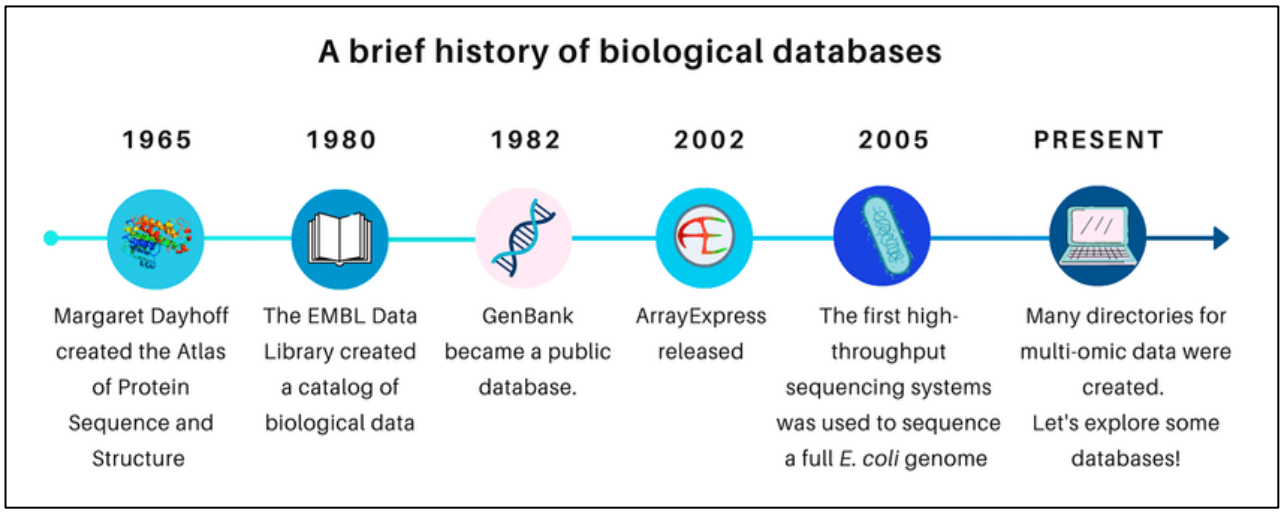


Figure 6. Evolution of biological sequences. (https://www.researchgate.net/publication/350591817_Fantastic_databases_and_where_to_find_them_Web_applications_for_researchers_in_a_rush/figures?lo=1&utm_source=google&utm_medium=organic).

5. Introduction to Sequencing

The genome: All of the genetic material, that is to say the DNA molecules, of a cell. The genome is contained in the nucleus and mitochondria (mithochondrial genome) and in the chloroplasts (chloroplast genome).

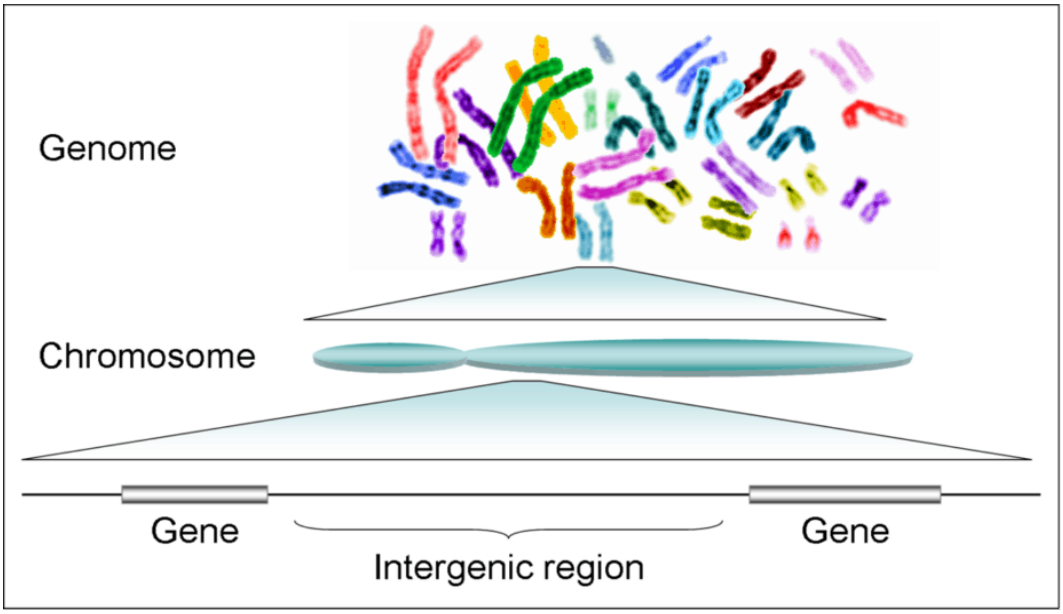


Figure 7. Genome. (https://nebula.org/blog/fr/genome/).

The genome is responsible for the characteristics of a living being.

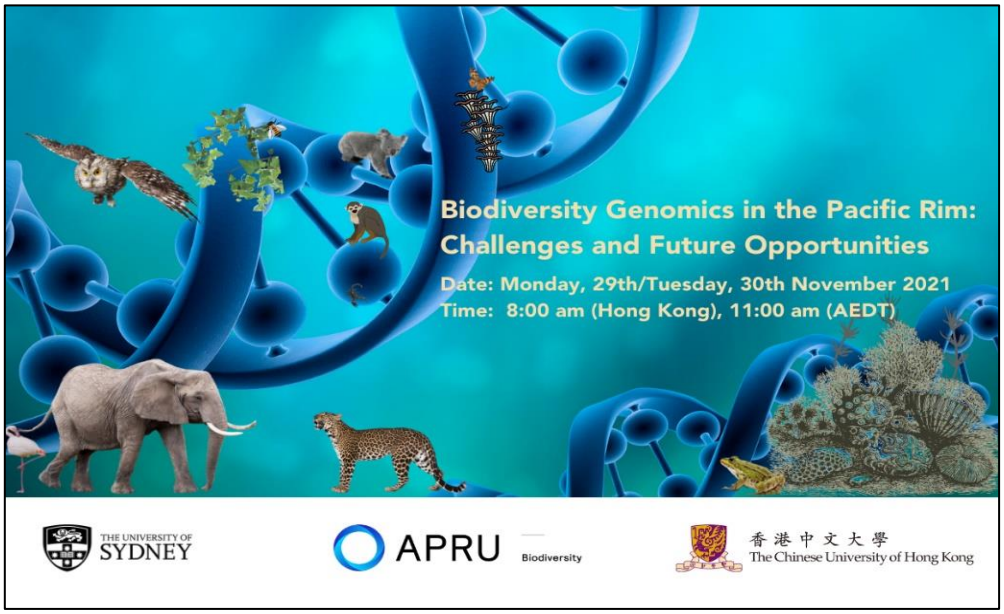


Figure 8. The genome and Biodiversity.
(https://global.sjtu.edu.cn/en/announcement/view/846).

The sequence: Sequences constitute one of the main types of bioinformation analyzed in bioinformatics.

Definition of DNA sequencing: is the determination of the succession of nucleotides component :

- Sequencing of all coding genes: exome sequencing.

- Whole genome sequencing: genome sequencing.

6. History of bioinformatics

For a discipline that has only been around for around twenty years, bioinformatics gets a lot of attention. Its development followed the exponential increase in the quantity of data coming, among other things, from systematic genome sequencing programs. If, at first, the priority was to store this flow of information, the role of bioinformatics quickly evolved. It is now a matter of transforming this raw data into knowledge. And that's no small feat!

A field of research that analyzes and interprets biological data, using computational methods, in order to create new knowledge in biology; here is the definition of bioinformatics recognized by

specialists in this discipline. The main research projects in bioinformatics and genomics are [1-5] :

- Cancer Genome Atlas: Mapping the genome for more than 25 cancer types has generated 1 petabyte of data (to date), representing 7,000 cancer cases. Scientists expect no less than 2.5 petabytes (1petabyte = 1015 bytes = 1000 terabytes).

- Encyclopedia of DNA Elements (ENCODE): The catalog of functional elements of the human genome : 15 terabytes of raw data (1 terabytes = 1000 gigabytes).

- Human Microbiome Project: one of the projects aimed at characterizing the microbiome at different locations in the body: 18 terabytes - approximately 5,000 times more data than the first "human genome" project.

- Earth Microbiome Project: Characterization of microbial communities on earth: 340 gigabytes (17109 sequences, ~ 20,000 samples, 42 biomes). 15 terabytes expected.

- Genome10K: Volume of raw data for the sequencing project of 10,000 vertebrate species should reach 1 petabyte (1 petabyte = 1024 Terabytes).

1951	First protein sequence (Insulin, Sanger)
1960	Link between sequence & structure (Globines, Perutz)
1965	1-First IBM/360 Computers; 2-Evolutionary divergence and convergence in Proteins (Zuckerkandl & Pauling)
1967	1-Construction of Phylogenetic Trees" Fitch & Margoliash
1986	1-Atlas of Protein Sequences (M. Dayhoff, Georgetown; 2-DEC PDP-8 minicomputer
1970	A general method applicable to the search for similarities in sequences of two proteins (Needleman & Wunsch).
1971	First work on RNA folding (J. Ninio)
1972	First Intel 8008 microprocessor
1973	“Genetic Engineering” (Cohen et al.)
1974	“Prediction of Protein Conformation” (Chou & Fasman)
1975	Intel 8080, kit Altair
1977	1-DEC-VAX mini-computer; 2-Microcomputers (Apple, Commodore, Radioshack: 3-DNA sequencing (Sanger, Maxam, Gilbert); 4-First Bioinformatics “package” (Staden)
1978	Databases: EMBL, GenBank, ACNUC, PIR
1980	Telephone access to the PIR database

1981	1-IBM-PC (8088), 16-32kb; 2-Los Alamos-GenBank: 270 sequences, 370,000 nucleotides; 3-Local alignment program (Smith-Waterman)
1983	IBM-XT Hard Drive (10 Mbytes)
1984	MacIntosh: graphic interface & mouse
1985 -1988	"Fasta" program (Pearson-Lipman)
1989	INTERNET succeeds ARPANET and BITNET
1990	1-"Blast" program (Altschul et al.); 2- Positional cloning and sequencing of NF-1
1991	1-Grail, a powerful program for locating genes (Mural et al.); 2- "EST" cDNA tags (Venter et al., Matsubara et al.)
1992-1996	1-Complete sequencing of yeast chromosome III; 2-First complete sequence of a microorganism (Venter et al.; H. influenza); 3-Complete yeast sequence (European consortium)
1997	1- "Gapped Blast" program (Alschul et al.); 2-11 bacterial genomes available
1998	2 Mbase/day of new public sequences
2001	Complete ("first draft") sequence of the human genome.

Table 1. Key milestones in the history of bioinformatics.

7. Bioinformatics and software

It is now easy and common to perform certain more or less complex operations using software rather than manually. However, these practices are not always systematic because it is often difficult for certain users to know which program to use based on a specific biological situation or to exploit the results provided by a method.

This is why this course contains the presentation of a certain number of tools or methods commonly used and recognized in the computer analysis of sequences :

- Command line tools: These tools can be difficult to use for most biologists, but almost always offer more options for running programs. They are more appropriate for analyzing large-scale datasets that are currently encountered in bioinformatics.

- Web Tools (Web-Based Software): Web tools, sometimes called "point-and-click", do not require programming knowledge and are immediately accessible to the scientific community.

The field of bioinformatics relies heavily on the Internet to access sequence data, software useful for analyzing molecular data, and to integrate different types of biology-related resources and information. We will describe a variety of websites.

Initially, we will focus on the major publicly available databases that serve as repositories for DNA and protein data. These include:

- The National Center for Biotechnology Information (NCBI), which hosts GenBank and other resources;

- The European Bioinformatics Institute (EBI);

- Ensemble, which includes a genome browser and resources to study dozens of genomes;

- The University of California Santa Cruz (UCSC) Genome Bioinformatics site, including a web browser and table browser for various species.

Through practical and guided work, we present several websites relating to bioinformatics. The main advantages offered by websites are easy access, rapid updates, good visibility for the scientific community and ease of use (since algorithmic and programming skills are not required).

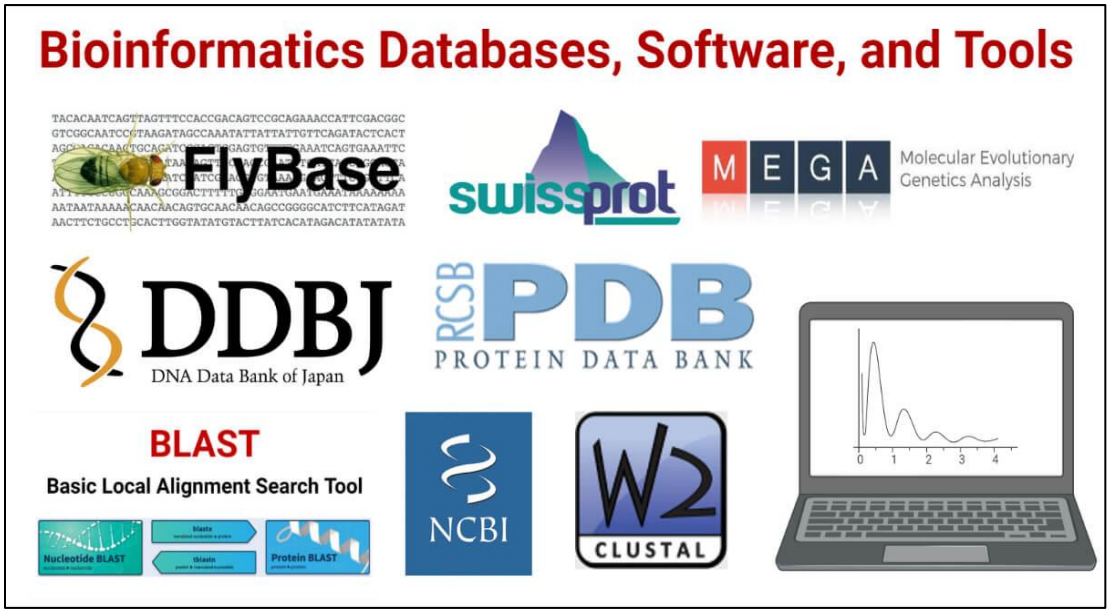


Figure 9. Bioinformatics Databases, Software, and Tools with Uses. (https://microbenotes.com/bioinformatics-databases-software-tools/).

8. Work in bioinformatics

Let's take a look at our top 15 bioinformatics degree jobs:

- Bioinformatician
- Bioinformatics Scientist
- Bioinformatics Analyst

- Bioinformatics Consultant
- Bioinformatics Programmer
- Bioinformatics Engineer
- Bioinformatics Technician
- Biostatistician
- Molecular Biologist
- Research Scientist
- Microbiologist
- Zoologist or Wildlife Biologist
- Computational Biologist
- Clinical Database Specialist
- Genomics Scientist

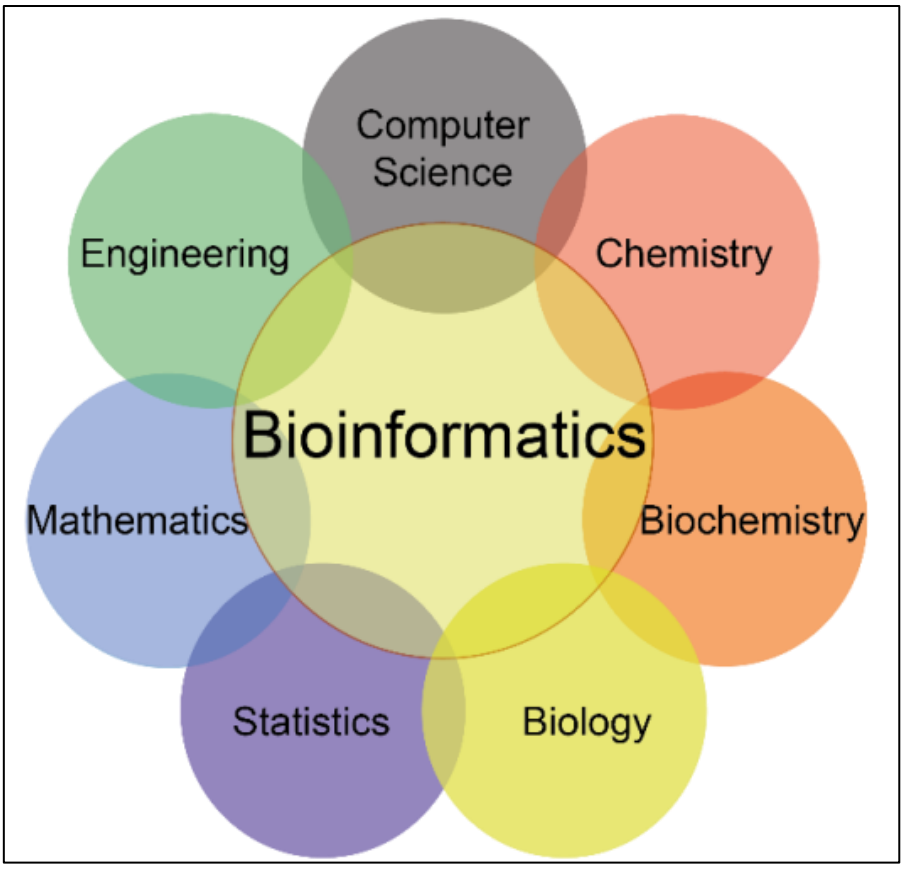


Figure 10. Work in bioinformatics.
(https://careersidekick.com/top-15-bioinformatics-degree-jobs/).

III. Data acquisitions techniques

III. Data acquisition technique

1. The first sequencing techniques

The first sequencing techniques bear the name of their inventors: the Maxam and Gilbert technique and the Sanger technique. Developed at the end of the 1970s, these two techniques use a common principle. The DNA molecule is gradually cut into smaller fragments. The DNA sequence is reconstituted following the separation of single-stranded DNA fragments by polyacrylamide gel electrophoresis. These techniques, which would revolutionize biology at the end of the 20th century, earned Gilbert and Sanger the Nobel Prize in Chemistry in 1980. We distinguish [2,3] :

- The technique of Maxam and Gilbert :

Maxam and Gilbert's technique relies on a chemical process that cuts a radioactively labeled DNA molecule at its 5' or 3' end at a specific base or base family. The conditions used are adapted to lead to a partial cutoff. Therefore, the length of the labeled fragments identifies the position of the base. The chemical reactions carried out preferentially cleave DNA to guanines, adenines, cytosines and thymines and cytosines. The products of the four reactions are resolved by electrophoresis. The DNA sequence can be read directly from the profile of the radioactive bands. The technique of Maxam and Gilbert has not developed well because it requires toxic chemical reagents, moreover it is not easy to automate and remains limited in terms of the size of the DNA fragments that it can analyze (< 250 nucleotides).

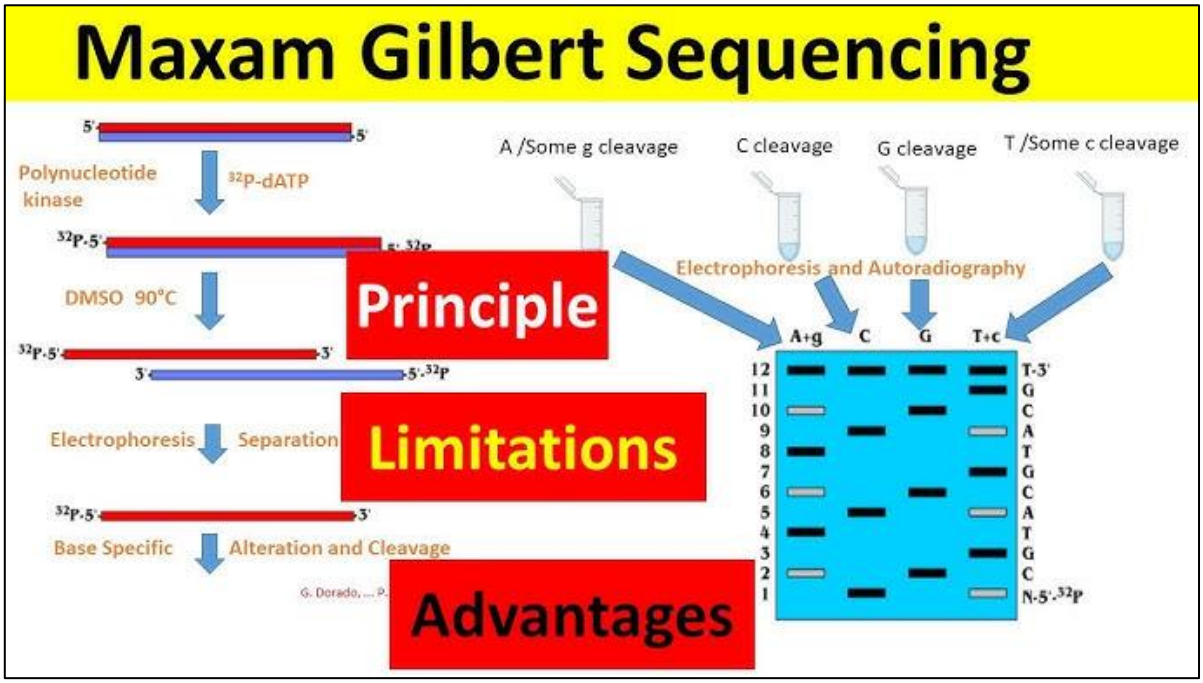


Figure 11. Maxam & Gilbert's method (chemical cleavage).
(https://slideplayer.com/slide/8525757/).

- The Sanger technique :

The Sanger technique uses the principle of enzymatic synthesis of DNA to be sequenced in the presence of DNA polymerase elongation inhibitors, dideoxynucleotides (ddNTPs). The original technique will allow us to describe the principle of this method.

The sequencing reaction is carried out using four enzymatic reactions carried out in parallel. In each tube, are placed [1-3] :

1. Template DNA (DNA to be sequenced): If the DNA to be sequenced is of small quantity, it can be previously amplified by a (PCR) reaction. Otherwise, amplification of the matrix by clonal multiplication, most frequently, in a bacterial vector, is necessary;

2. A primer capable of hybridizing to one of the strands of the matrix and which allows the start of DNA polymerization;

3. An equimolar mixture of dCTP, dGTP and dTTP;

4. Radioactively labeled dATP;

5. A ddNTP corresponding to one of the four bases.

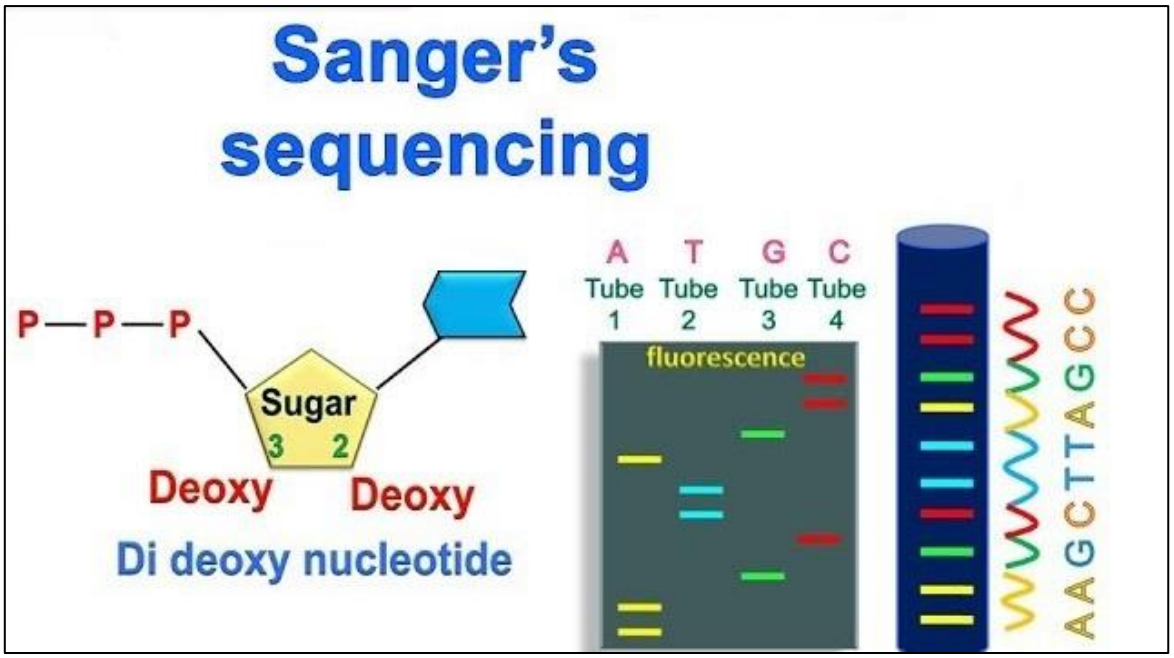


Figure 12. Sanger Sequencing.
(https://parlonssciences.ca/ressources-pedagogiques/documents-dinformation/sequencage-de-sanger).

2. New sequencing techniques (NGS)

Since 2004, new sequencing techniques have been available on the market. In contrast to traditional techniques, they have been developed by manufacturers who market automated platforms allowing these techniques to be used. Another very important point common to all these new technologies is that the amplification of template DNA libraries no longer involves clonal multiplication, but via PCR reactions.

The two types of PCR reactions used are introns:

1. Emulsion PCR (Emulsion PCR or EmPCR); DNA fragments are bound to agarose beads;

2. PCR by bridging (bridge amplification); the DNA fragments are fixed on a glass slide called flow-cell. These PCR amplifications avoid any bias in representation of the fragments.

Indeed, during bacterial cloning, certain DNA fragments can be toxic to bacterial cells. These new techniques are based on [6] :

1. DNA synthesis (Pyrosequencing, Solexa/Illumina and Ion Torrent).

2. Hybridization on DNA chips (SOLiD developed by Applied Biosystems);

3. Real-time detection of molecules (not yet commercially implemented).

Pyrosequencing and the Solexa technique are commonly used in combination with the Sanger technique for de novo sequencing. Solexa and SOLiD techniques are used for resequencing. As these are the most commonly used, we will describe here the techniques based on DNA synthesis.

- The pyrosequencing technique:

The template DNA library is amplified by emulsion PCR. Each agarose bead carrying a copy of DNA from the bank is placed in the well of a cell slide that can be analyzed by optical fiber (Picotiter plate). In each cell the sequence reaction takes place. The dNTPs are added in successive flows (unlike the Sanger technique). When the added dNTP is complementary to the nucleotide of the template strand, it is incorporated into the strand being synthesized and an inorganic pyrophosphate (PPi) is released.

The release of PPi is followed by chemiluminescence following a cascade of enzymatic reactions: in the presence of adenosine 5'-phosphosulfate (APS), ATP sulfurylase transforms PPi into ATP which is used by a luciferase to transform luciferin into oxyluciferin , molecule generating a light signal in the visible. Then, apyrase degrades unincorporated nucleotides and excess ATP before new dNTP is added.

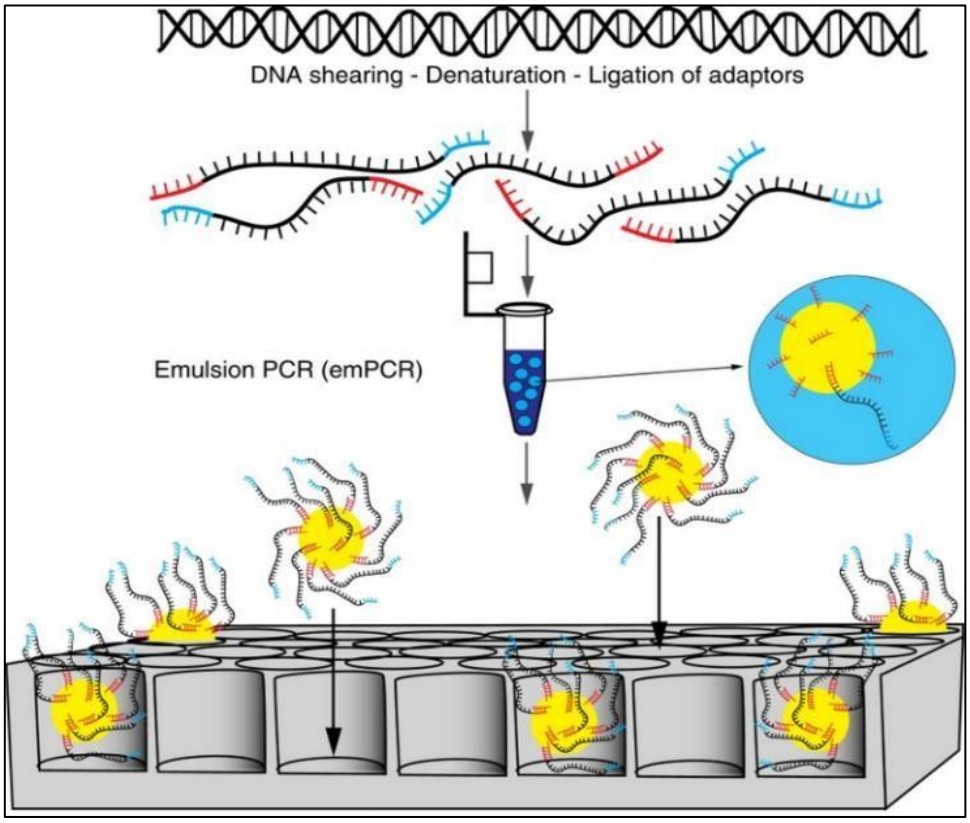


Figure 13. The release of PPi.

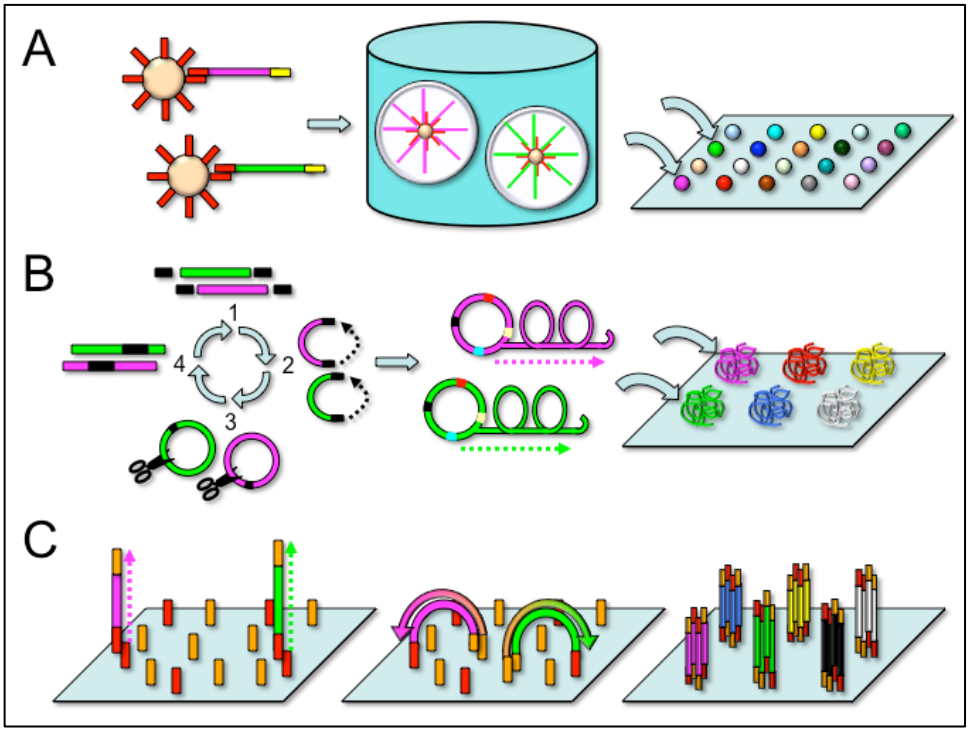


Figure 14. The different steps of emulsion PCR or EmPCR.
(https://www.seqanswers.com/forum/general/15817-emulsion-pcr-in-detail-explained).

A CCD camera captures the light signal emitted and transforms it into an electrical signal which results in a peak on the pyrogram. The height of the peak is proportional to the intensity of the light signal, itself proportional to the number of nucleotides incorporated at the same time. We deduce the sequence from the size of the peaks obtained [1-8].

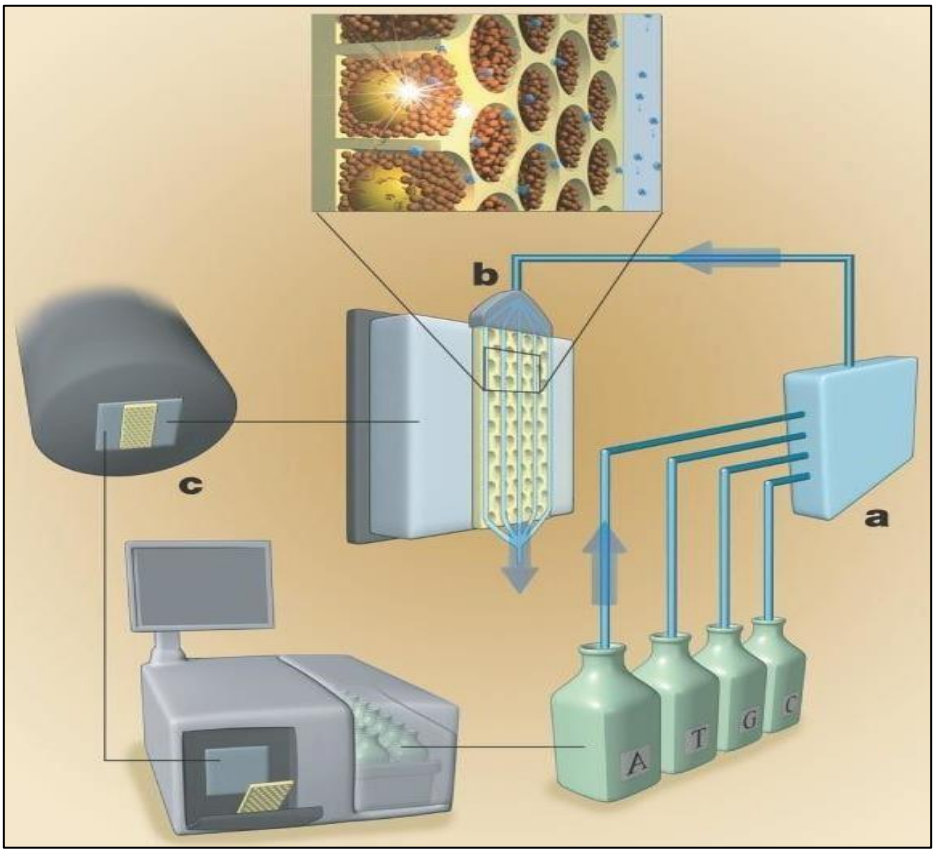


Figure 15. How the Genome Sequencer FLX System works.
(a): The dNTPs are added in successive flow. (b): When the added dNTP is complementary to the nucleotide of the template strand, it is incorporated into the strand being synthesized and an inorganic pyrophosphate (PPi) is released. The release of PPi is monitored by chemiluminescence following a cascade of enzymatic reactions; (C) A CCD camera captures the emitted light signal.

When the added dNTP is complementary to the nucleotide of the template strand, it is incorporated into the strand being synthesized and an inorganic pyrophosphate (PPi) is released.

In the presence of adenosine 5'-phosphosulfate (APS), ATP sulfurylase transforms PPi into ATP which is used by a luciferase to transform luciferin into oxyluciferin, a molecule generating a light signal in the visible.

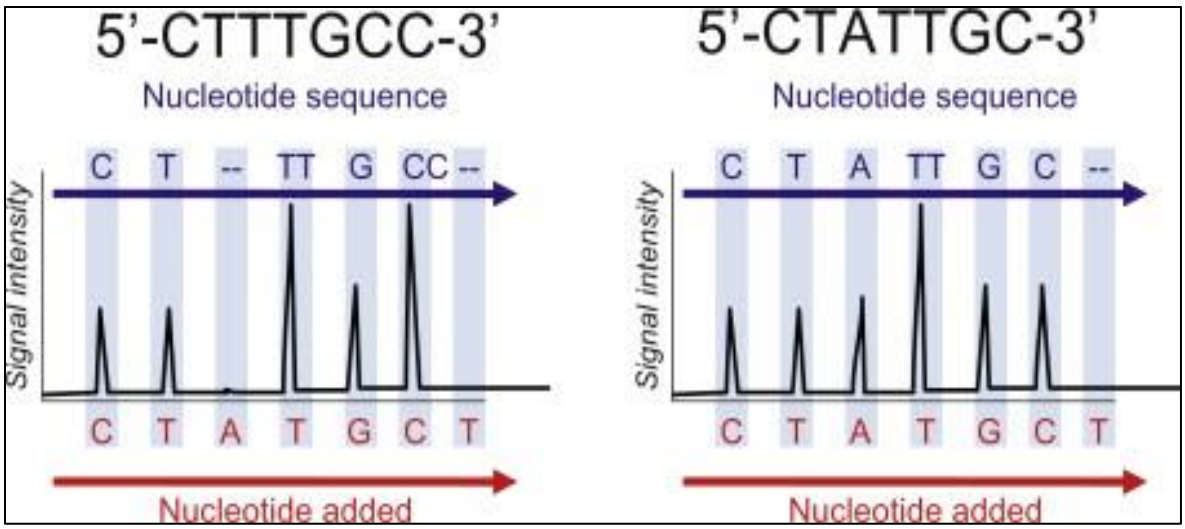


Figure 16. Principle of pyrosequencing (a).

A CCD camera captures the light signal emitted and transforms it into an electrical signal which results in a peak on the pyrogram. The height of the peak is proportional to the intensity of the light signal, itself proportional to the number of nucleotides incorporated at the same time.The sequence is deduced from the size of the peaks obtained.

The pyrosequencing technique was developed by the company 454 Life Science. The corresponding sequencing platforms "Genome Sequencer FLX System" (GS20; first generation), "Genome Sequencer FLX System" (GS FLX and GS FLX Titanium) are marketed by Roche Diagnostic laboratories.

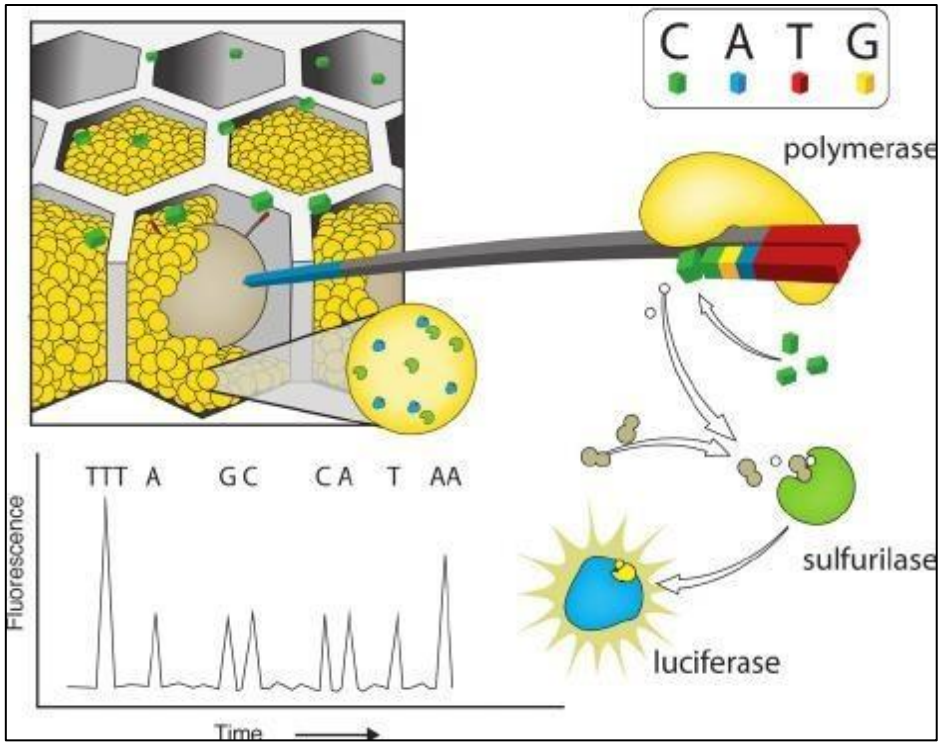


Figure 17. Principle of pyrosequencing (b).

The pyrosequencing technique was developed by the company 454 Life Science. The corresponding sequencing platforms "Genome Sequencer FLX System" (GS20; first generation), "Genome Sequencer FLX System" (GS FLX and GS FLX Titanium) are marketed by Roche® Diagnostic laboratories.

- The Illumina/Solexa sequencing technique :

The fragments of the DNA strand are amplified by "bridge amplification". Each copy of DNA in the library is represented by a group of clonal fragments, called a "cluster". The clusters are attached to the surface of a slide made up of 8 channels and can be analyzed by optical fiber called flow-cell. During a first cycle of synthesis, the DNA polymerase, the primer for initiating DNA synthesis and the 4 dNTPs are added.

dNTPs have two characteristics [7,6] :

1. Each is marked with a different fluorochrome;

2. All are chemically modified so as to block the 3'-OH end and therefore inhibit elongation.

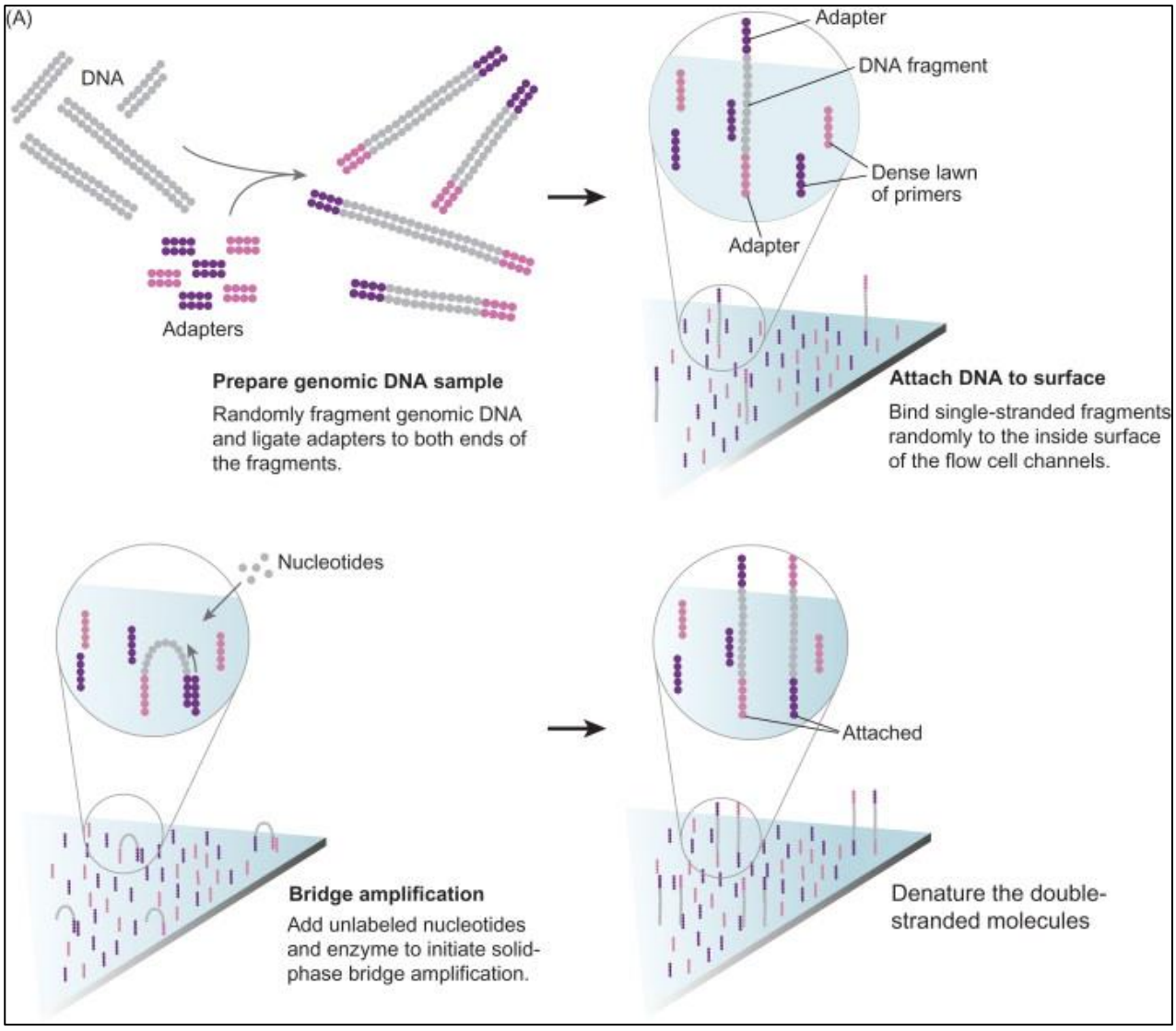


Figure 18. Principle of the Illumina/Solexa sequencing technique.
(https://www.sciencedirect.com/topics/immunology-and-microbiology/illumina-dye-sequencing).

3. 3rd generation sequencing techniques

- SMRT sequencing technology :

This technology uses color fluorescence marking of nucleotides added to DNA strands transcribed by polymerase. Their addition is detected in real time as they are added to the DNA strand to be sequenced.

Its main benefit is that it allows sequences of up to 3000 bases to be read in one go. This helps to reduce the number of errors and reduce the level of coverage rate (the number of reads, i.e. the number of bases to be detected by redundancy / number of bases of the DNA to be sequenced).

- Comparison of techniques :

These new techniques make it possible to carry out rapid and very high throughput sequencing. Moreover. Once the acquisition of the automated platform has been carried out, sequencing is much less expensive than sequencing using the Sanger technique. However, for genome assembly steps and particularly in regions with numerous repetitions, the Sanger technique still sometimes remains necessary.

Technical	**Bank**	**Bank amplification**	**Sequencing principle**	**Length of reading (Nucleotides)**	**Number of nucleotides read by experiment Mb)**	**Approximate price per Mb (euros)**
Sanger technique on an automated sequencer (96 reactions)	DNA fragments double stranded in a replicative vector	Bacterial multiplication	Enzymatic synthesis in the presence of elongation inhibitors, dd NTP and electrophoresis	800	0,096	5000
Pyrosequencing on a FLX platform	DNA fragments double stranded in a replicative vector	Emulsion PCR	Enzymatic synthesis and monitoring of the release of pyrophosphate generated during the incorporation of a nucleotide	200-300	80-120	75
Solexa/Illumina technique	DNA fragments double stranded in a replicative vector	Bridging PCR	Enzymatic synthesis, reversible inhibition of elongation and monitoring of fluorescence of the incorporated nucleotide	30-40	1000	5
Solid technique	DNA fragment	Emulsion PCR	Hybridization / ligation	35	1000-3000	5

	s double stranded in a replicative vector		of primers and monitoring of fluorescence of s hybridized oligonucleotides			

Table 2. Comparison of the 4 main sequencing techniques.

4. Exercise:

The GenBank database

 is a general database of protein sequences

 contains the most exhaustive data possible

 distributed by EBI

 is a specialized database of nucleic sequences

 contains homogeneous data and distributed by theme

5. Exercise:

Among the following databases, which one(s) is/are grouped into general nucleic sequence banks

0 DDBJ

0 UniProt

0 TrEMBL

0 PDB

0 EMBL

IV. Biological banks and databases

IV. Biological banks and databases

1. Definition of a database

When it was created, bioinformatics corresponded to the use of computing to store and analyze molecular biology data. This original definition has now been extended and the term bioinformatics is often associated with the use of computing to solve scientific problems posed by biology as a whole. In all cases, it is a multidisciplinary field of research that brings together computer scientists, mathematicians, physicists and biologists [9].

In IT, a database is a structured and organized set allowing the storage of large quantities of information in order to facilitate its use (adding, updating, searching for data). This information is related to a given activity and can be used by programs or users. A Database is [10] :

1. A set of data relating to a domain, organized by computer processing, generally accessible online and remotely,

2. Often, data is stored in the form of a formatted text file (respecting a particular layout),

3. Need to develop specific software to query the data contained in these databases.

- Biological databases are electronic, computerized libraries that contain information about the life sciences, collected through scientific experiments, published literature, high-throughput experimental technologies, and computational analyses.

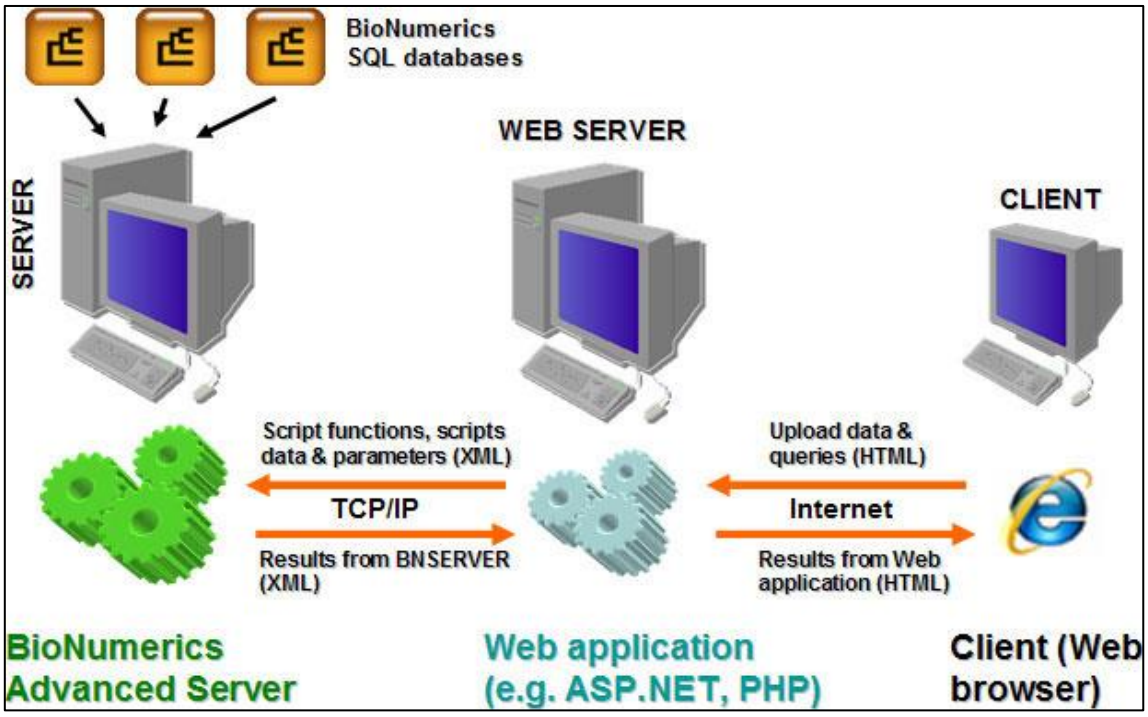


Figure 19. Database server.
(https://www.bionumerics.com/bionumerics-server).

2. Definition of a biological database

Biological databases are computer databases collecting a wide variety of biological data. There are different categories depending on the type of data stored, these are supplemented by annotations. The data is stored there in the form of text files, in relation to each other. These are therefore relational databases. Their design is complex and evolves rapidly with the increase in data and study tools.

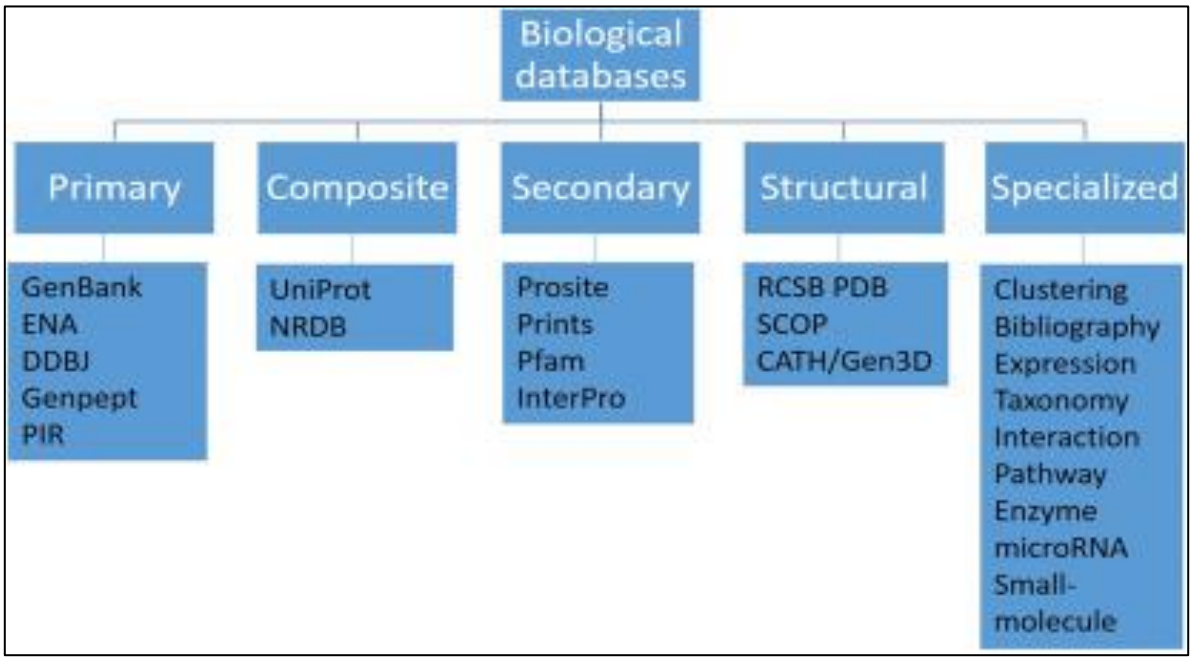


Figure 20. Biological databases and their application.
(https://www.sciencedirect.com/science/article/abs/pii/B9780323897754000213).

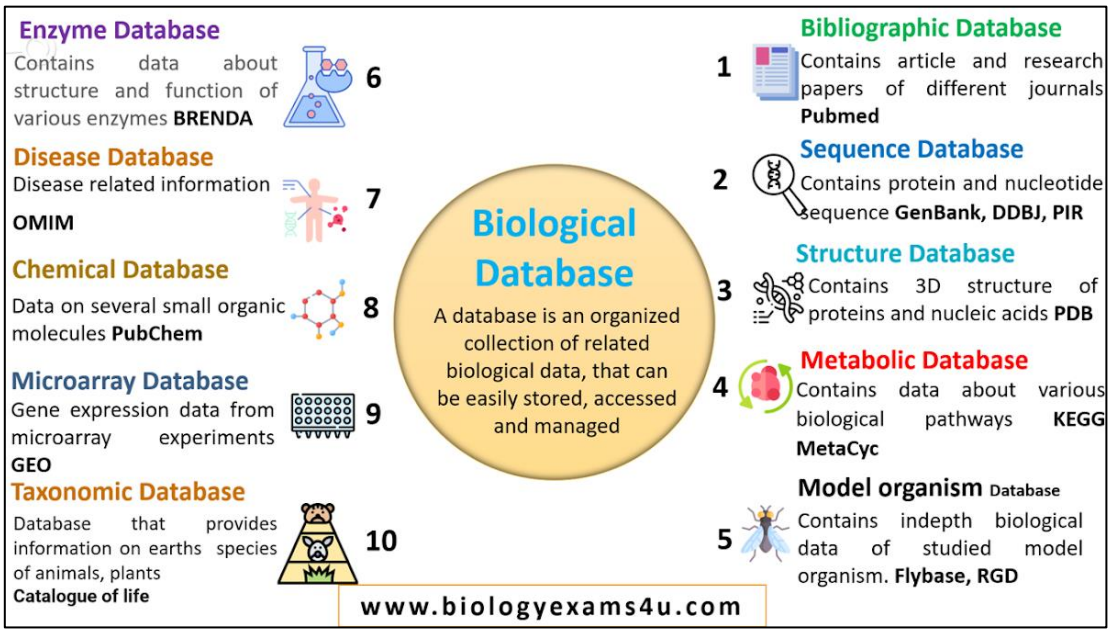


Figure 21. 10 Types of Biological Databases.
(https://www.biologyexams4u.com/2023/02/10-types-of-biological-databases.html#google_vignette).

3. Role of biological databases

The role of biological databases is [1-3]:

1. Collect information from biologists, literature and other databases (Sequences, physical mapping, etc.; Structural, relational data, etc.)

2. Store and organize (Coherent logic)

3. Distribute information (Wide distribution)

4. Facilitate exploitation (User-friendly interface; Definition of search criteria; Data comparison).

4. Contents of biological databases

These databases can contain information: (DNA, proteins, genes and genomes, taxonomy, others, etc.). There is also a bibliography and biological expertise directly linked to the sequences processed.

5. Types of databases

There are a large number of databases of biological interest.

We will distinguish two types of banks, those which correspond to the most exhaustive possible data collection and which ultimately offer a rather heterogeneous set of information (generalist data banks) and those which correspond to more homogeneous data established around a thematic (specialized databases) and which offer added value based on a particular technique or an interest aroused by a group of scientists.

- General databases: they correspond to the most exhaustive collection of data possible and which offer a rather heterogeneous set of information.

General databases contain heterogeneous data:

– Collection as exhaustive as possible,

– Nucleic sequence banks,

– Protein sequence banks,

– 3D structure banks of macromolecules,

– Scientific article banks.

1- Nucleic sequence banks:

EMBL (European Molecular Biology Laboratory): European bank created in 1980 and financed by EMBO

(European Moleculary Biology Organization), it is now distributed by the EBI (European Bioinformatics Institute, Cambridge, UK)

GenBank: created in 1982 by the company IntelliGenetics and now distributed by the NCBI (National Center for Biotechnology Information, Los Alamos, US);

DDBJ (DNA Data Bank of Japan): created in 1986 and distributed by the NIG (National Institute of Genetics, Japan).

2- Protein banks:

PIR-NBRF (Protein Information Resource-National Biomedical Research Foundation): created in 1984 by the NBRF (National Biomedical Research Foundation). It is now a set of data from MIPS (Martinsried Institute for Protein Sequences, Munich, Germany) and the Japanese bank JIPID (Japan International Protein Information Database);

SwissProt: created in 1986 at the University of Geneva and maintained since 1987 as part of a collaboration between this university (via ExPASy, Expert Protein Analysis System) and the EBI. This also brings together annotated sequences from the PIR-NBRF bank as well as coding sequences, translated from EMBL.

▪ Origin of data:

- DNA and RNA sequencing

▪ Stored data: sequences + annotations:

- Genome fragments

- One or more genes, a piece of gene, intergenic sequence, etc.

- Complete genomes

- mRNA, tRNA, rRNA, ... (fragments or whole)

[Note 1]: all sequences (DNA or RNA) are written with Ts [Note 2]: the sequences are always oriented 5' ⟶ 3'. [10-15].

Name	Link	Description
EMBL	http://www.ebi.ac.uk/embl/	European Bank (European Molecular Biology Laboratory) distributed by the EBI (EuropeanBioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	American bank distributed by NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	DNA Data Bank of Japan distributed by the NIG (National Institute of Genetics)

Table 3. Some general databases (general nucleic sequence banks).

UniProt: is a protein sequence database. Its name derives from the contraction of Universal Protein Resource. It is an open, stable and accessible online database, it results from the consolidation of all the data produced by the scientific community.

Name	Link	Description
UniProt	https://www.uniprot.org/	Annotated sequences & coding sequences translated from EMBL

Table 4. Some general databases (general protein sequence banks).

- Specialized databases

Correspond to more homogeneous data established around a theme and which offer added value. Specialized databases contain homogeneous data (a collection established around a particular theme).

- These banks contain homogeneous data
- Collection established around a particular theme
- Advantages: ease of updating data, checking their integrity, offering a suitable interface,
- Disadvantages: does not always target what we want; not all possible banks exist

• Examples: specialized banks for a genome, immunology sequence banks, banks on validated sequences, etc.

Databases	Link	Description
Ensembl	https://www.ensembl.org/index. html	Integrative genomic bank
Prosite	http://prosite.expasy.org	Identifies protein motifs with biological significance
Reactome	https://reactome.org /PathwayBrowser/	Integrative Metabolic Banking
Kegg Pathway	http://www.genome.jp/kegg /pathway.html	Molecular interactions and reactions
PFAM	http://xfam.org/	Protein domains
Interpro	http://www.ebi.ac.uk/interpro/	Brings together several existing banks
PDB	http://www.rcsb.org/pdb/home /home.do	3D structure of proteins, amino acids and biological molecules
PubMed	https://www.ncbi.nlm.nih.gov /pubmed	Citations, abstracts and articles (bibliographic search)

Table 5. Some specialized databases.

6. The most used bioinformatics databases

• The Enter or NCBI portal:

- GenBank: DNA and RNA sequences

- Official BLAST website

- PubMed: Allows searching for scientific articles

- COGs: Orthologous gene families ...

• The EMBL portal: The European Molecular Biology Laboratory

• ExPASy: Expert Protein Analysis System, Proteomics

- UniProt: Protein sequences

- PROSITE: Protein domains and families

- SWISS-MODEL: 3D protein prediction tool

- Different search tools

• PDB: Protein Data Bank

- Database of 3D protein structures

- Visualization and manipulation of structures

• SCOP: Structural Classification of Proteins

7. Structuring and organization

Large general sequence banks such as GenBank or EMBL are international projects that constitute leaders in the field. They have now become essential to the scientific community because they bring together essential data and results, some of which are no longer reproduced in the scientific literature:

- Files and formats:

The sequences are generally stored in the form of text files which can be either personal files (present in a personal space) or public files (bank sequences) accessible by Web tools.

The format corresponds to the set of presentation rules (constraints) to which the sequence(s) in a given file are subject. The format allows [1]:

- Automated formatting

- Homogeneous storage of information

- Subsequent computer processing of the information.

A single piece of information in a database is called an "entry"

So that the user can find their way, all this information is made available to the scientific community according to an organization in sections or fields.

1- The FASTA format:

In this case, the sequence (given in the form of lines of 80 characters maximum) is preceded by a title line (name, definition, etc.) which must begin with the character ">". This allows you to put several sequences in the same file.

Example :

>em|U03177|FL03177 Feline leukemia virus clone FeLV-69TTU3-16.
AGATACAAGGAAGTTAGAGGCTAAAACAGGATATCTGTGGTTAAGCACCTG
TGAGGCCAAGAACAGTTAAACCCCGGATATAGCTGAAACAGCAGAAGTTTC GCCAGCAGTCTCCAGGCTCCCCA

The following lines contain the sequence itself, but with a fixed maximum number of characters per line. This maximum number is generally set at 60, 70 or 80 characters. A sequence of several hundred bases or residues is therefore distributed over several lines.

A file is called multifasta when it contains several sequences in FASTA format, one after the other.

Files containing one or more sequences in FASTA format most often have the .fasta extension but we also find .seq, .fas, .fna or .faa.

FASTA example:

>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens OX=9606 GN=HBB PE=1 SV=2 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH

2- The EMBL format:

Each entry in the EMBL database is made up of lines which begin with a two-character code (field) followed by 3 blanks:

ID Identifier or mnemonic (entry name) XX Blank separator line.

AC Accession number

DT Dates of incorporation into the database and of the last update. DE Description of the sequence

KW Keyword(s) (in alphabetical order). OS Organism from which the sequence comes.

OC Taxonomic classification of the organism

OG Subcellular localization of non-nuclear sequences (chloroplast, kinetoplast, mitochondria, plasmid, etc.)

RN Bibliographic references of the entry. RC Comments on the reference

RX Region for which the bibliographic reference is associated. RP References associated with the different regions of the sequence. RA Article authors

RT Article title

RL Journal references

DR connections with other databases FH Field header FT

FT Characteristics of the sequence (features).

SQ Sequence (60 nucleotides per line in the 5'--->3' direction). CC Comments

// End of entry.

Basic format:

The 1st line contains ID, 3 spaces then the identifier (9 digit characters max).

The 2nd line contains AC, 3 spaces then the accession number (6 digit characters max). The 3rd line contains DE, 3 spaces then the description (6 digit characters max).

The 4th line contains SQ, 3 spaces followed by the sequence size. The following lines contain the sequence, divided into 6 blocks (per line) of 10 characters, separated by a space.

Each entry ends with "//". LINE 1:ID ID_name

LINE 2:AC Accession number

LINE 3 :DE Describe the sequence any way you want LINE 4 :SQ Length BP

LINE 5: ACGTACGTAC GTACGTACGT ACGTACGTAC GTACGTA...

LINE 6: ACGT

LINE 7 ://

3- Stanford / IG format:

The 1st line is a comment line preceded by the character ";". The 2nd contains the identifier (sequence name) in the 10th columns

The following lines contain the sequence (80 characters max/line) ending with the character "1" (for a linear sequence) and "2" (if the sequence is circular).

LINE 1:; Describe the sequence any way you want

LINE 2:ECTRNAGLY2

LINE 3: ACGCACGTAC ACGTACGTAC A C G T C C G T ACG TAC GTA CGT LINE

4: GCTTA GG G C T A1

EX :

; Dro5s-T.Seq Length: 120 April 6, 1989 21:22 Check: 9487 ..

dro5stseq
GCCAACGACCAUACCACGCUGAAUACAUCGGUUCUCGUCCGAUCACCGAAAUUAAGCAGCGUCG
GUUAGUACUUAGAUGGGGGACCGCUUGGGAACACCGCGUGUUGUUGGCCUI

4- GCG format:

The format adopted by the GCG package allows both commenting on the data and verifying the integrity of the sequence by a value (=Ckecksum) calculated on it. The GCG format only allows one sequence per file.

The file is made up of three parts:

- before the "..": comments

- signal line with identifier and "Check ####.."

- after the "..": sequence

EX :

pir:ccho (1-104)

pir:ccho Length: 104 (today) Check: 8847 ..

1 GDVEKGKKIF VQKCAQCHTV EKGGKHKTGP NLHGLFGRKT GQAPGFTYTD

51 ANKNKGITWK EETLMEYLEN PKKYIPGTKM IFAGIKKKTE REDLIAYLKK

101 ATNE

GCG processes three other file formats, which have more specific functions:

FOSN: catalog file containing a list of sequence names (bank mnemonics or personal file names)

MSF (Multiple Sequence File) contains multiple sequences in one file. It comes from a multiple alignment. RSF (GCG version 9.1)

Several GCG programs allow you to perform format conversion

5- Fitch format:

The 1st line contains the name of the sequence.

The following lines contain the sequence, divided into 20 blocks (per line) of 3 characters, separated by a space.

EX :

pir:ccho (1-104) , 104 bases, 7DA79498 checksum.

GDV EKG KKI FVQ KCA QCH TVE KGG KHK TGP NLH GLF GRK TGQ (etc ...) TWK EET LME YLE NPK KYI PGT KMI FAG IKK KTE RED LIA YLK KAT NE

-GENBANK format:

The Japanese bank DDBJ adopted the same format as that of Genbank.

The first 12 columns contain the field name:

LOCUS

DEFINITION

ACCESSION

NID

KEYWORDS

SEGMENT

SOURCE

ORGANISM

REFERENCE

AUTHORS

TITLE

JOURNAL

MEDLINE

COMMENT

FEATURES

BASE COUNT

ORIGIN

Basic format:

The file must contain the header "GENETIC SEQUENCE DATA BANK" and, for each entry, have lines 10 to 16.

LINE 1 : GENETIC SEQUENCE DATA BANK LINE 2 :

LINE 3 :

LINE 4 :

LINE 5 :

LINE 6 :

LINE 7 :

LINE 8 :

LINE 9 :

LINE 10 :LOCUS L_Name Length BP

LINE 11 :DEFINITION Describe the sequence any way you want LINE 12 :ACCESSION Accession Number

LINE 13 :ORIGIN

LINE 14 : 1 acgtacgtac gtacgtacgt acgtacgtac gtacgtacgt a... LINE 15 : 61 acgt...

LINE 16 ://

8. Data quality

It should be noted that the information contained in these databases presents a certain number of gaps. One of the main ones is the lack of systematic verification of data submitted or entered, especially for old sequences.

The authors of the sequences sometimes have difficulty restoring the knowledge they have about their data or have not carried out a certain number of basic checks on their sequences.

For example, we find:

- Segments of cloning vectors in certain sequences or inconsistencies in biological characteristics (coding parts, definition of species or key words, etc.)
- Incomplete or even erroneous biological information.

The organizations responsible for maintaining these banks have become aware of these problems and now numerous checks are carried out systematically upon submission of the sequence.

This does not eliminate all the inaccuracies such as for example the existence of duplicates because these are extremely similar sequences which correspond to different entries in the bank and which it is often difficult to know if they are of polymorphism, duplicated genes or simply duplicated genes of errors established during the determination of the sequences.

There are also electronic mailboxes (e-mail) to inform bank managers of possible errors or corrections that anyone could detect or propose.

Mailboxes

- for EMBL: update@ebi.ac.uk or datasubs@ebi.ac.uk

- for GenBank: update@ncbi.nlm.nih.gov or gb_admin@ncbi.nlm.nih.gov Web pages

- EMBL update http://www3.ebi.ac.uk/Services/webin/update/update.html

- UniProtKB update https://web.expasy.org/docs/swiss-prot_guideline.html

Another important problem is the delay in inserting a new sequence into a bank, often linked to the volume of sequences to be processed which gives rise to priorities or choices. Thus, there can be around ten months of delay between the experimental determination of a sequence and its introduction into a bank.

Despite this, it is necessary to underline the enormous wealth represented by these databases, in particular in the context of sequence analysis.

- The fact that the majority of known sequences is united in a single set is a fundamental element for the search for similarities with a new sequence.

- The great diversity of organisms represented there makes it possible to approach evolutionary type analyses. For example, we can extract the sequences of the same gene from several species.

- Another interest of these bases lies in the information which accompanies the sequences (annotations, expertise, bibliography), even if these are often of unequal quality. These latter can sometimes constitute the rare annotations available on certain sequences.

- Finally, the presence of references to other databases allows access to other unlisted information. Thus we can know the entry into a protein base of the protein which corresponds to the gene that we have identified in a nucleic base.

The UniProt bank, particularly rich in cross-references with other banks and in annotations (for example, the notion of "proven or not experimentally" was recently introduced in the table of biological characteristics) is an example of the quality of the data that the can be found in the various general sequence banks of recent years.

9. Protein Structures Database

- Protein structure:

We can distinguish several levels in the description of the structure of proteins:

• The primary structure: it corresponds to the sequence of the amino acids constituting

the protein. It is a linear assembly of amino acids encoded by messenger RNA.

• Secondary structure: it describes a more complex structural level: the secondary structures which are represented by the local folding of the protein.

It includes the helical structures (α, 310, π, type II) and the sheets (parallel and antiparallel β) and finally the bends (types I, II, III and γ).

• The tertiary structure: describes the three-dimensional structure of the protein or more precisely of a particular shape that the protein of interest can take in space under given experimental conditions and this at a time t.

• The quaternary structure: allows us to describe the interactions between proteins.

- Structural databases:

• E-MSD (European Macromolecular Structure Database): This is the European bank of three-dimensional structures of biological macromolecules, maintained by the EBI. It derives from the PDB but unlike the latter it is a relational bank. There are therefore links for each entry cross-referenced with databases (structural, modular or sequence) with added information. As in the case of the PDB, it is possible to deposit new structures there. One of the major advantages of E-MSD compared to the PDB is to generate a library of ligands (chempdb) from the structures resolved in complexes.

In addition to general and structural information specific to the ligand, it provides fairly precise details of the chemical environment of the latter's binding sites.

• NDB (Nucleic acid structures Database): The structures available in the NDB include structures of RNA and DNA oligonucleotides composed of at

minus two bases. These molecules can be alone or complexed with proteins or small ligands. Archives store primary and derived information from structures. Primary data include atomic coordinates, structure factors for X-ray structures or constraints for NMR structures, and details of experiments (crystallization condition, crystal stacking, data collection, and statistics). refinement).

The derived information corresponds to the analysis of each structure. We find information such as valence geometry, torsion angles and intermolecular contacts. The NDB is partially redundant with the PDB when it comes to complexes with proteins.

• PDB (Protein Data Bank): The protein data bank of the Research Collaboratory for Structural Bioinformatics, more commonly called

PDB, is a global collection of data on the three-dimensional structure (or 3D structure) of biological macromolecules: proteins, mainly, and nucleic acids. These structures are essentially determined by X-ray crystallography or NMR spectroscopy. These experimental data are deposited in the PDB by biologists and biochemists around the world and are in the public domain. Their consultation is free and can be done directly from the bank's websites. The PDB is the main source of structural biology data and in particular provides access to 3D structures of proteins of pharmaceutical interest. As is the case for GenBank, atomic coordinates must be deposited before publication. The structures are checked when they are submitted to ensure that the coordinates submitted comply with established standards.

- PDB file format:

A PDB format file is a text file composed of ASCII characters. It is therefore possible to access the raw information contained in these

files by opening them with a text editor.

PDB files contain the Cartesian coordinates of the atoms that constitute the molecule as well as metadata. This metadata can for example be the primary structure of the molecule, its possible secondary structures, the experimental method which made it possible to obtain the coordinates of the atoms, etc. A PDB file is composed of:

• 1st Part: called header Contains bibliographic information attached to the structure, on the resolution and crystallographic parameters, the sequence and sometimes the secondary structure.

• 2nd part: It contains the atomic coordinates In this part the atoms designated by ATOM are located on the protein chain, while the atoms designated by HETATM (HETeroAToM group) are part of the cofactor molecules, substrates, ions or other groups which are linked by a covalent bond to the protein chain.

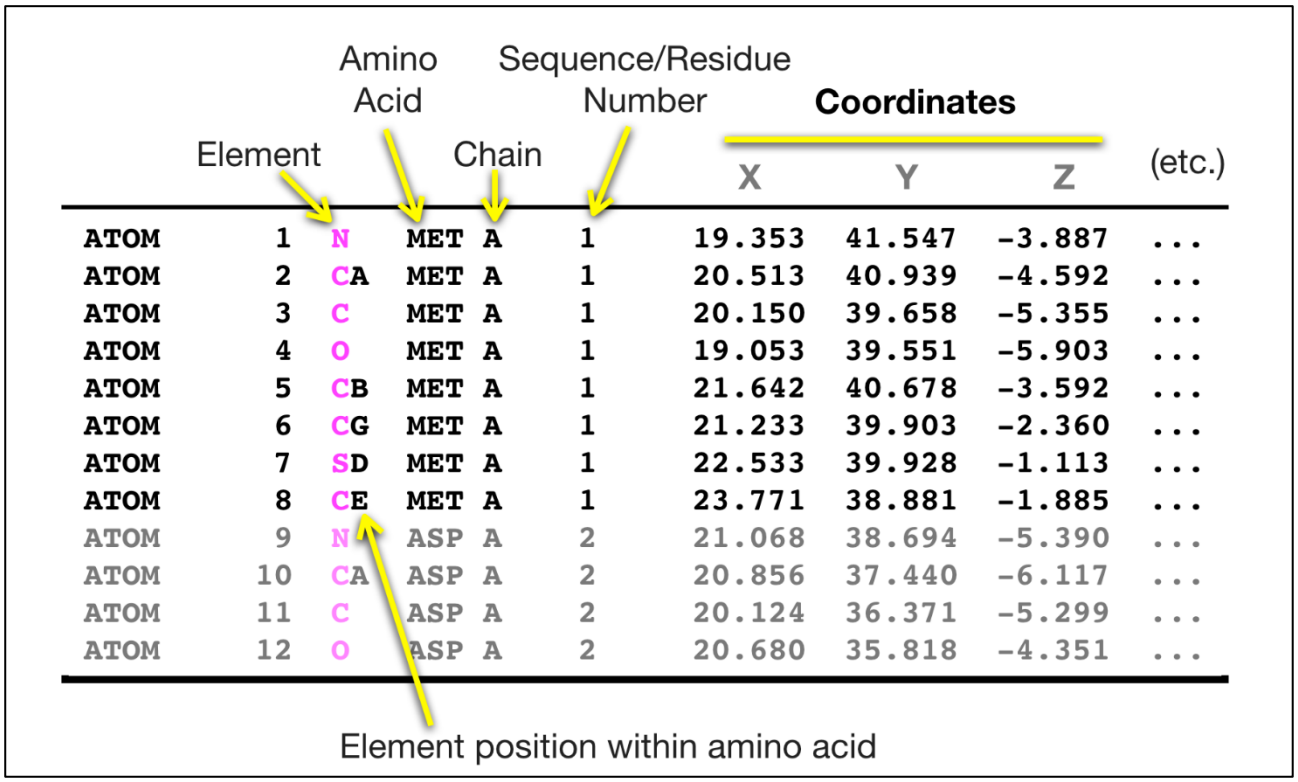


Figure 22. PDB file format (Part I).

Top of 4y2n PDB file

```
HEADER    STRUCTURAL PROTEIN                      10-FEB-15   4Y2N
TITLE     STRUCTURE OF CFA/I PILI MAJOR SUBUNIT CFAB TRIMER
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CFA/I FIMBRIAL SUBUNIT B;
COMPND   3 CHAIN: B, A, C;                                  Chains modeled
COMPND   4 FRAGMENT: UNP RESIDUES 25-170;                   Residues modeled
COMPND   5 SYNONYM: CFA/I ANTIGEN,CFA/I PILIN,COLONIZATION FACTOR ANTIGEN I
COMPND   6 SUBUNIT B;
COMPND   7 ENGINEERED: YES                                  Recombinant form
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI O78:H11 (STRAIN H10407 /
SOURCE   3 ETEC);
SOURCE   4 ORGANISM_TAXID: 316401;                          Native host taxonomy ID
SOURCE   5 STRAIN: H10407 / ETEC;                           Native strain
SOURCE   6 GENE: CFAB, ETEC_P948_0400;                      Gene ID
SOURCE   7 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   8 EXPRESSION_SYSTEM_TAXID: 562                     Expression strain tax. ID
KEYWDS    ENTEROTOXIGENIC ESCHERICHIA COLI, PERIPLASMIC CHAPERONE, MAJOR PILIN,
KEYWDS   2 SELF-ASSEMBLY, FIMBRIAE, STRUCTURAL PROTEIN
EXPDTA    X-RAY DIFFRACTION                                 Experimental method
AUTHOR    R.BAO,D.XIA
REVDAT   2   28-NOV-18 4Y2N    1       JRNL   REMARK
REVDAT   1   10-AUG-16 4Y2N    0
JRNL        AUTH   R.BAO,Y.LIU,S.J.SAVARINO,D.XIA
JRNL        TITL   OFF-PATHWAY ASSEMBLY OF FIMBRIA SUBUNITS IS PREVENTED BY
JRNL        TITL 2 CHAPERONE CFAA OF CFA/I FIMBRIAE FROM ENTEROTOXIGENIC E.
JRNL        TITL 3 COLI.
JRNL        REF    MOL. MICROBIOL.               V. 102   975 2016
JRNL        REFN                   ESSN 1365-2958
JRNL        PMID   27627030
```

Figure 23. An annotated screenshot of the PDB file for entry: 4y2n. (https://bitesizebio.com/61389/protein-data-bank-files/).

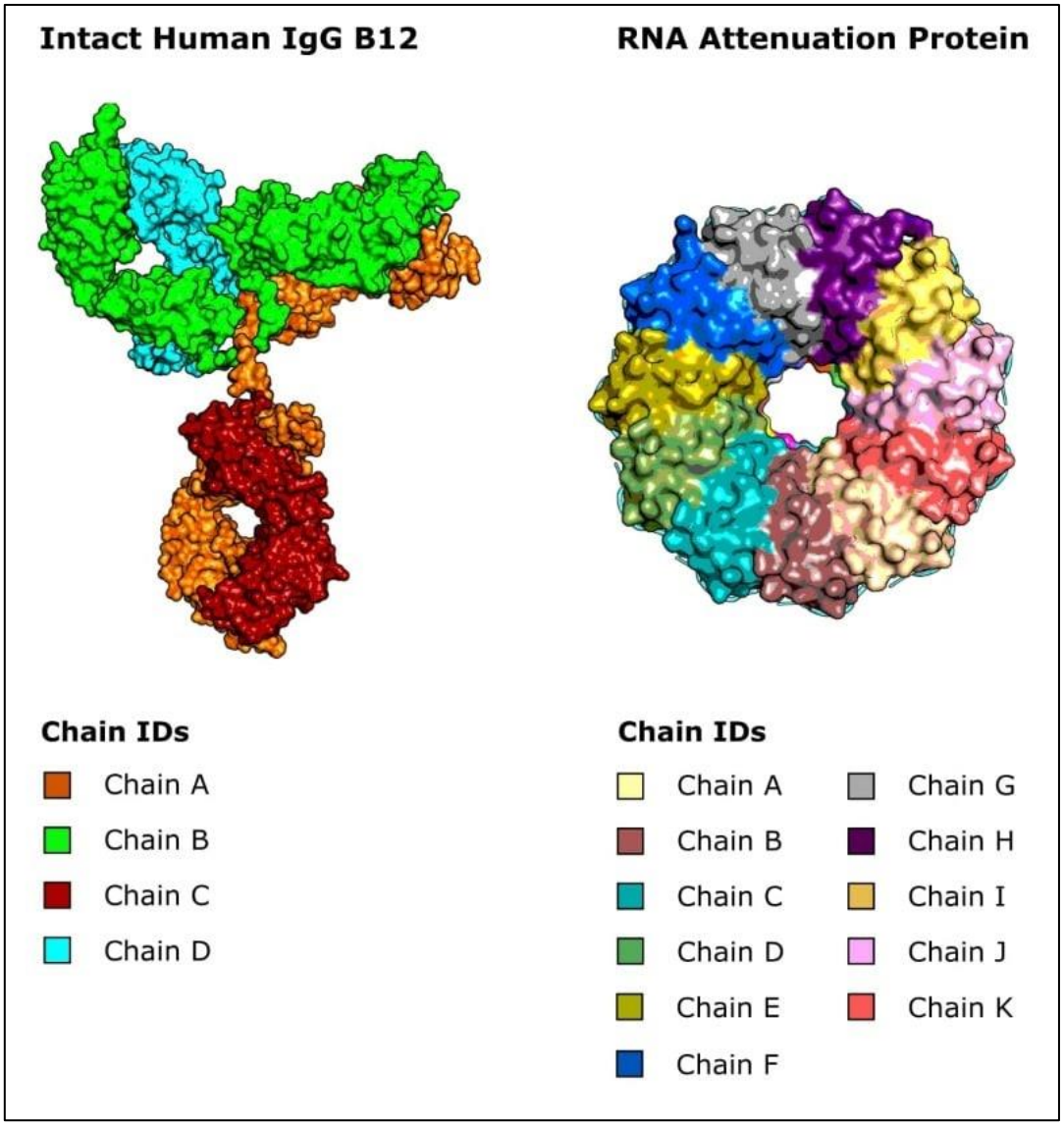
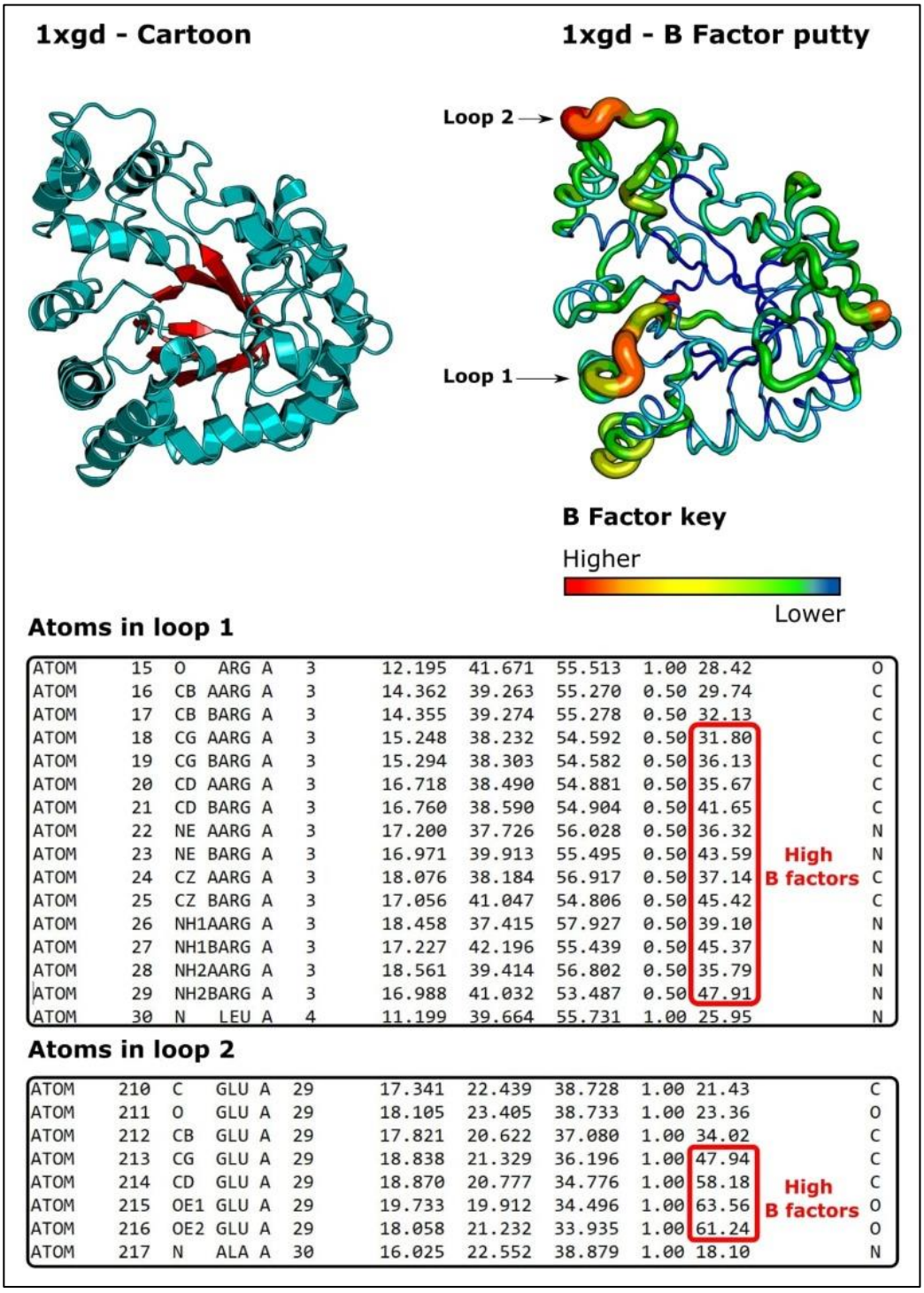


Figure 24a. Two examples of multi-chain proteins. (Left) Human IgG B12 (PDB: 1hzh) comprises 4 unique chains. (Right) The undecameric form of Trp RNA-binding attenuation protein (TRAP) from Geobacillus stearothermophilus (PDB ID: 1c9s) comprises 11 chains with identical amino acid sequences. (https://bitesizebio.com/61389/protein-data-bank-files/).



Atoms in loop 1

ATOM	15	O	ARG	A	3	12.195	41.671	55.513	1.00	28.42	O
ATOM	16	CB	AARG	A	3	14.362	39.263	55.270	0.50	29.74	C
ATOM	17	CB	BARG	A	3	14.355	39.274	55.278	0.50	32.13	C
ATOM	18	CG	AARG	A	3	15.248	38.232	54.592	0.50	31.80	C
ATOM	19	CG	BARG	A	3	15.294	38.303	54.582	0.50	36.13	C
ATOM	20	CD	AARG	A	3	16.718	38.490	54.881	0.50	35.67	C
ATOM	21	CD	BARG	A	3	16.760	38.590	54.904	0.50	41.65	C
ATOM	22	NE	AARG	A	3	17.200	37.726	56.028	0.50	36.32	N
ATOM	23	NE	BARG	A	3	16.971	39.913	55.495	0.50	43.59	N
ATOM	24	CZ	AARG	A	3	18.076	38.184	56.917	0.50	37.14	C
ATOM	25	CZ	BARG	A	3	17.056	41.047	54.806	0.50	45.42	C
ATOM	26	NH1	AARG	A	3	18.458	37.415	57.927	0.50	39.10	N
ATOM	27	NH1	BARG	A	3	17.227	42.196	55.439	0.50	45.37	N
ATOM	28	NH2	AARG	A	3	18.561	39.414	56.802	0.50	35.79	N
ATOM	29	NH2	BARG	A	3	16.988	41.032	53.487	0.50	47.91	N
ATOM	30	N	LEU	A	4	11.199	39.664	55.731	1.00	25.95	N

High B factors

Atoms in loop 2

ATOM	210	C	GLU	A	29	17.341	22.439	38.728	1.00	21.43	C
ATOM	211	O	GLU	A	29	18.105	23.405	38.733	1.00	23.36	O
ATOM	212	CB	GLU	A	29	17.821	20.622	37.080	1.00	34.02	C
ATOM	213	CG	GLU	A	29	18.838	21.329	36.196	1.00	47.94	C
ATOM	214	CD	GLU	A	29	18.870	20.777	34.776	1.00	58.18	C
ATOM	215	OE1	GLU	A	29	19.733	19.912	34.496	1.00	63.56	O
ATOM	216	OE2	GLU	A	29	18.058	21.232	33.935	1.00	61.24	O
ATOM	217	N	ALA	A	30	16.025	22.552	38.879	1.00	18.10	N

High B factors

Figure 24b. Two examples Top left). A cartoon representation of human aldose reductase. (Top right) A B factor putty representation. (Bottom) Atoms belonging to loops 1 and 2 in the corresponding PDB file. (https://bitesizebio.com/61389/protein-data-bank-files/).

V. Algorithms, exploitation and analysis of data (Annotation)

V. Algorithms, exploitation and analysis of data (Annotation)

1. Algorithm, Program, Software, data structures

Genomes can be considered as a long series of letters written in the alphabet A, C, G, T.

How can these texts be interpreted?

This is going to be the subject of bioinformatics using appropriate algorithm.

An algorithm is a finite and unambiguous sequence of operations or instructions used to solve a problem or obtain a result.

The word algorithm comes from the Arabic name 'KHAWRZMIA' from the 9th century Persian mathematician Al-Khwarizmî. The field that studies algorithms is called algorithmics.

Algorithms are found today in many applications such as computer operation, cryptography, planning, image processing, word processing, bioinformatics, etc.

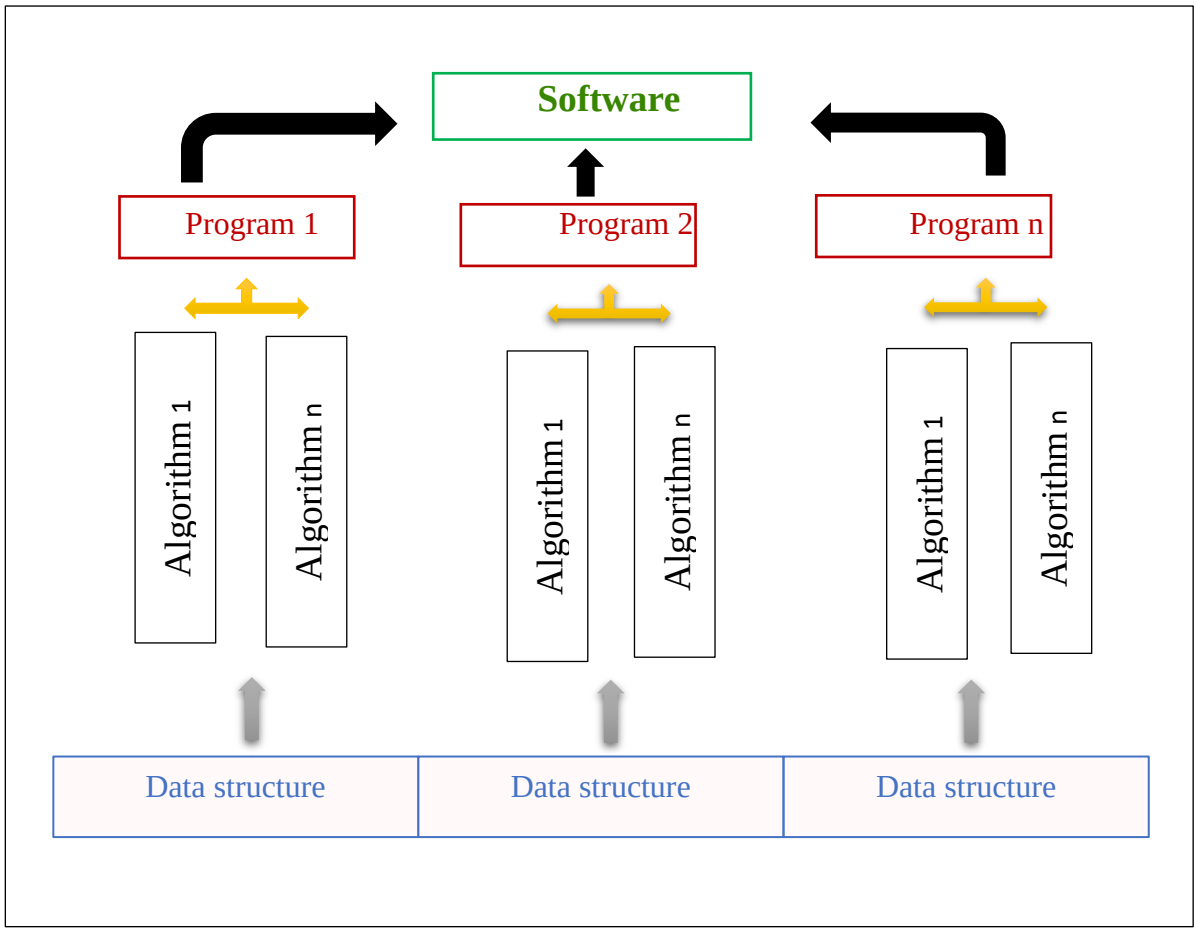


Figure 25. Algorithms + data structures = programs.

A sequence of operations to be executed to solve a problem (or a class of problems), which must be expressed in a formal, explicit and unambiguous manner.

Furthermore, we expect an algorithm to have certain properties:

- Let it end

- That it is relevant (that its execution leads to the resolution of the problem posed, or a diagnosis that this problem cannot be solved)

- That it is effective (that it is executed in a reasonable time)

If it is intended to be executed by a computer, the algorithm must be written in a programming language. There are many programming languages: R, Java, C++, Perl, Python, Matlab, etc.

In this course the algorithms presented will be written in "pseudo-language"

2. Writing an algorithm (The DNA Walk algorithm)

The procedure is as follows:

- We have 4 letters, and there are 4 directions in a plane.

- Up, down, left, right.

- Several choices are possible

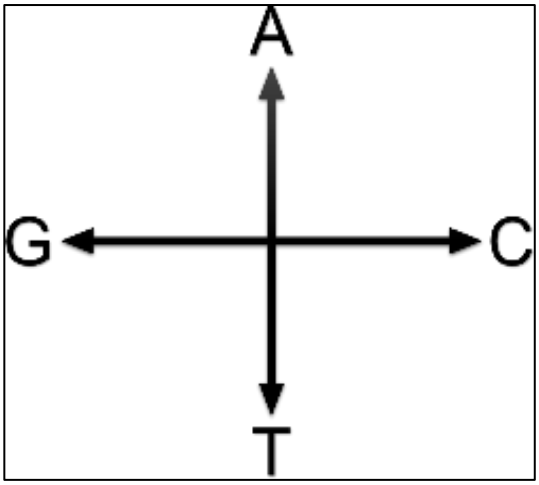


Figure 26. The DNA Walk algorithm (part I).

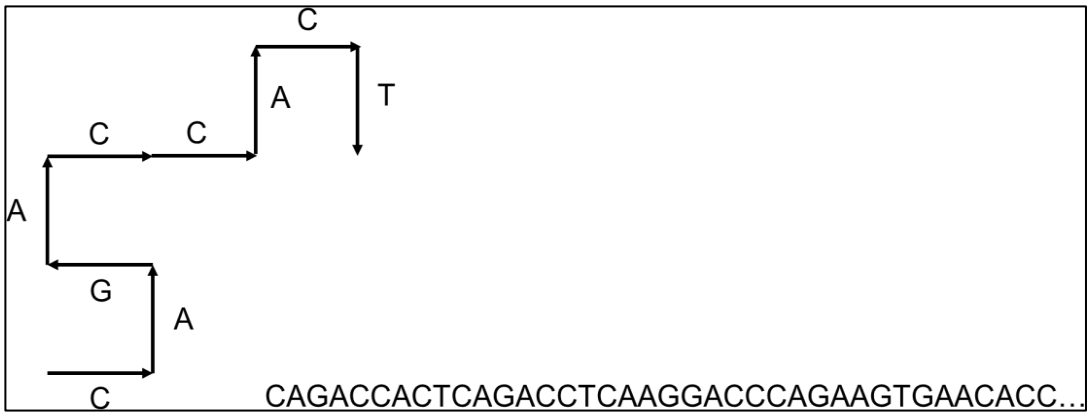


Figure 27. The DNA Walk algorithm (part 2).

```
L 01   index : integer
L 02   sequence: character string [1:*]
L 03   index ←1
L 04   repeat
L 05      case sequence [index] of
L 06       "A": drawUp
L 07       "C": drawRight
L 08       "G": drawLeft
L 09       "T": drawDown
L 10      endcase
L 11      index ← index + 1
L 12   until sequence [index] = "#"
```

Figure 28. Writing an algorithm.

3. Sequence annotation

Classically, there are three main stages in the genome annotation process:

- **Syntactic annotation:** this is the step which makes it possible to identify genetic objects presenting a

biological relevance (coding sequences, RNA, repeated sequences, etc.).

- **Functional annotation:** this is the step which makes it possible to predict the potential functions of previously identified genetic objects (similarities of sequences, motifs, structures, etc.) and to collect possible experimental information (literature, games of large-scale data);

• **Relational annotation:** this is the step which makes it possible to determine the interactions that previously identified biological objects are likely to maintain (gene families, regulatory networks, metabolic networks, etc.).

All of this information will then be stored in databases that can be consulted by the experimenter. Computing plays a vital role in annotation due to the computing power required to perform searches and the enormous amount of information generated. Most of the tools and databases presented here are freely accessible via the Internet. On the other hand, the bioinformatics tools necessary for the different stages of annotation can be grouped into annotation platforms. These platforms have user-friendly interfaces that can be used by non-computational biologists.

4. Syntactic annotation: searching for genetic objects

Principle: The search for genetic objects mainly involves the search for genes in the broad sense, that is, any sequence that, transcribed and/or translated, can play a role in the biological functioning of the cell. This therefore covers coding sequences (Coding Sequence or CDS in English, that is, sequences translated into proteins), untranslated RNAs (transfer RNA or tRNA, ribosomal RNA or rRNA, small RNAs, interfering RNAs, etc.).

The search for coding sequences, although insufficient for a good understanding of the functioning of a genome, is nevertheless the most developed and for which a large number of computer tools exist. This is what we will develop in this part.

Syntactic annotation of prokaryotic genomes is relatively easier than that of eukaryotic genomes for the following reasons:

• Prokaryotic genomes are smaller than eukaryotic genomes and above all have a much higher coding density, of the order of 80-90%, while it can range from 70% in yeast to a few percentages in humans;

• Prokaryotic genes are frequently organized in operons, i.e. a single transcription unit can contain several coding sequences;

• Prokaryotic genes are not fragmentedI unlike those of eukaryotes.

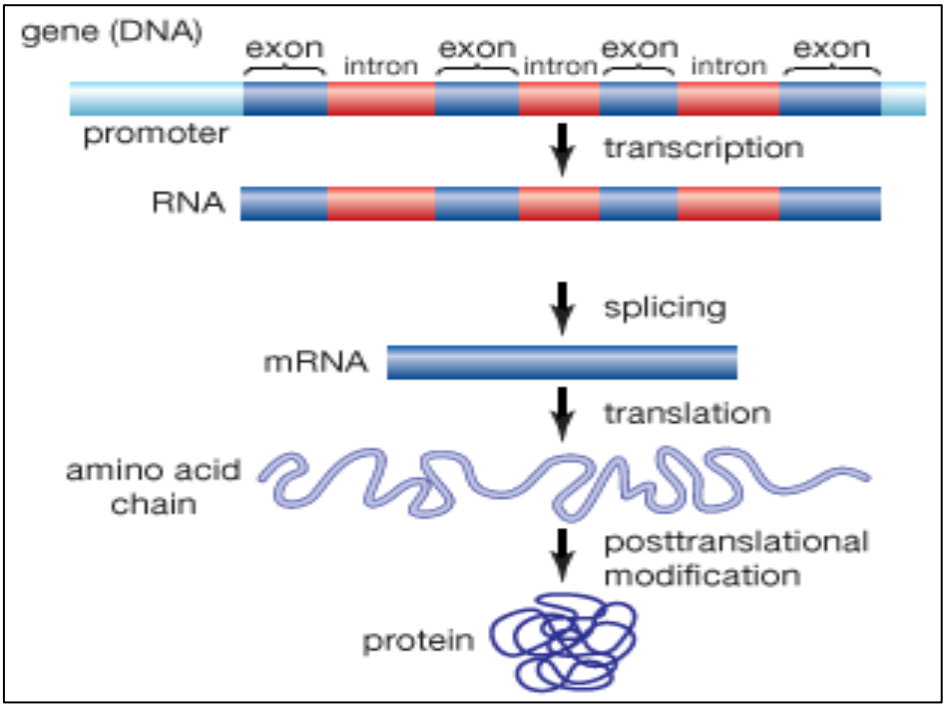


Figure 29. Simplified diagram of the central dogma of molecular biology. Certain DNA sequences undergo transcription to generate a primary messenger RNA. This mRNA undergoes various transformations, including splicing, by which introns are removed, to generate a mature transcript. Finally, ribosomes, with the help of tRNAs and translation factors, translate the coding sequence into protein. The coding sequence (CDS) is shown in blue.

a. ORF and CDS in prokaryotes

The open reading frame (ORF) is the region of DNA that separates two translation termination codons (therefore potentially coding). In this, a coding sequence (CDS) always begins with a translation initiation codon and always ends with a translation termination codon.

The universal translation initiation codon or "Start" codon is the ATG codon. However, in prokaryotes there are rarer "Start" codons such as the GTG and TTG codons. The translation termination codons or "Stop" codons are the TAA, TAG and TGA codons.

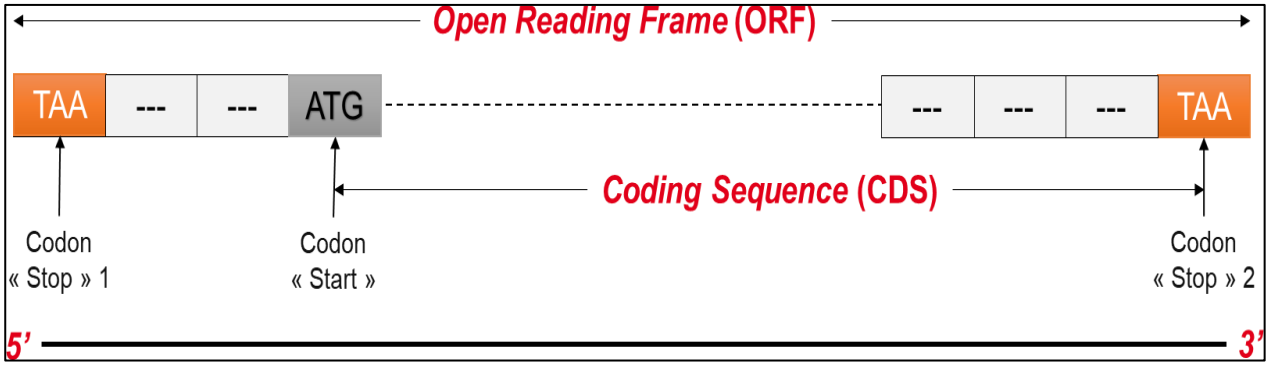


Figure 30. ORFs and CDSs in prokaryotes.

In prokaryotes, each coding sequence is called a cistron. Many prokaryotic messenger RNAs are polycistronic: they contain multiple cistrons or CDSs and therefore code for multiple proteins.

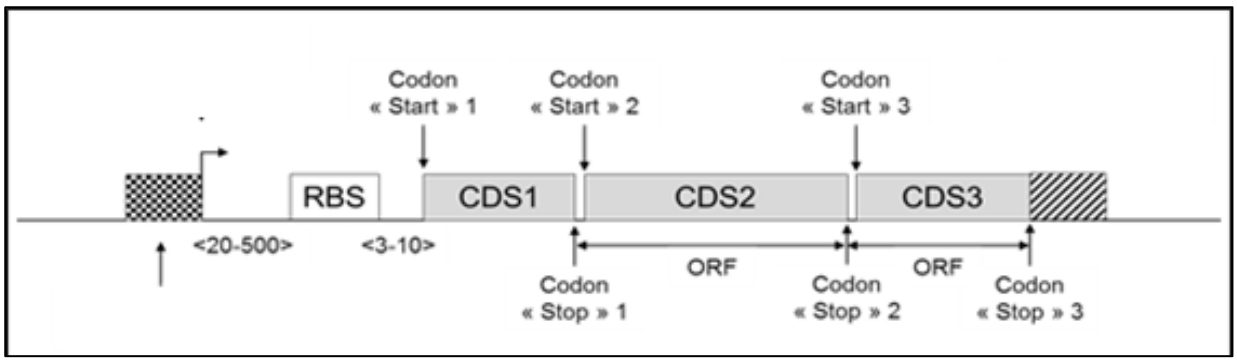


Figure 31. The concept of coding sequence in prokaryotes.

The Shine-Dalgarno sequence or ribosome binding site (RBS) is located 3 to 10 nucleotides upstream of the "Start" codon. It is a purine-rich region of 5-6 nucleotides that allows the ribosome to bind specifically to the AUG corresponding to a true "Start" codon. This signal also allows the annotator to distinguish a true "Start" codon from an ATG codon encoding a methyonine. In Escherichia coli, the consensus sequence of the RBS is: 5' -AGGAGG-3'

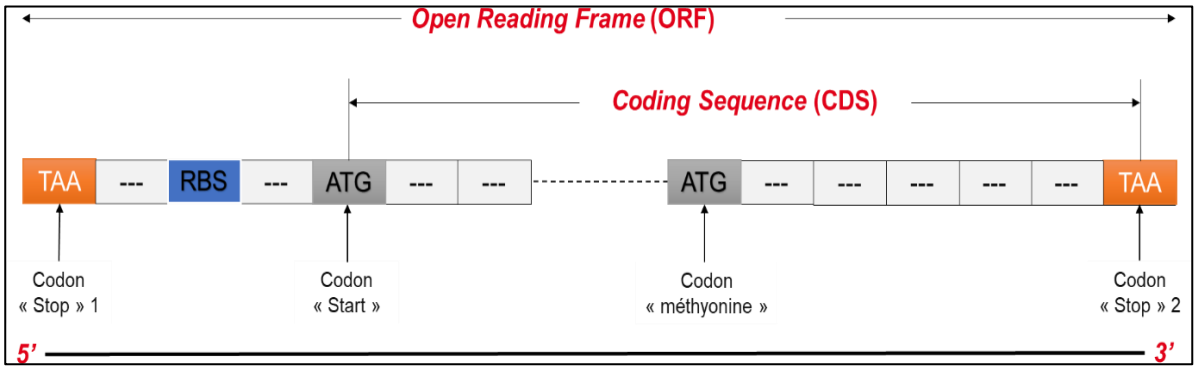


Figure 32. The ribosome binding site (RBS).

The promoter or promoter region is the region specifically recognized by the complex between RNA polymerase (enzyme that ensures the transcription of DNA) and the sigma factor (protein factor that ensures the specificity of transcription initiation).

The promoter consists of two elements: a highly conserved sequence that is located approximately 10 nucleotides before the transcription start site, the TATA or -10 box; and a moderately conserved sequence that is located approximately 35 nucleotides before the transcription start site, the -35 box. In an operon, the promoter is located only upstream of the first CDS, since the operon is a transcription unit.

The transcription terminator is a sequence through which the transcription complex will disassemble and thus terminate transcription. Terminators are palindromic sequences2 rich in GC followed by sequences rich in A (case of Rho-independent terminators) or not (case of Rho-dependent terminators).

The detection of an RBS, a promoter or a transcription terminator can validate the existence of a coding sequence a posteriori. However, their consensus is too weakly conserved for them to constitute reliable signals a priori.

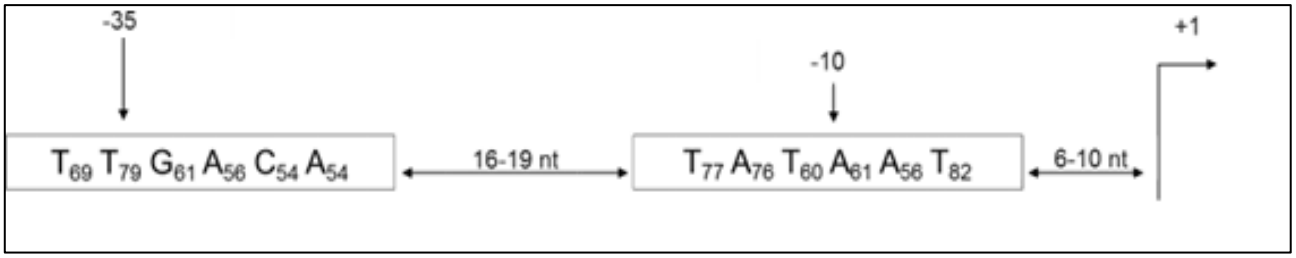


Figure 33. Consensus sequences of promoters recognized by the sigma 70 factor in E. coli. The subscript numbers indicate the percentage of occurrence of each nucleotide in the definition of the consensus sequence. The +1 indicates the transcription initiation site.

Searching for coding sequence signals in eukaryotes:

In eukaryotic genomes, syntactic annotation is significantly more complicated. For the following reasons:

• Eukaryotic genomes have a low coding density. There are therefore large genomic regions without coding sequence;

• Eukaryotic genes are fragmented; they undergo modifications of the nucleotide sequence (splicing) of the pre-messenger RNA. Splicing consists of the excision of one or more sequences (introns). The non-excised sequences (exons) form the "coding sequence" after joining together.
• Finally, splicing can be alternative: different splicing profiles exist for the same pre-messenger RNA and consequently a gene can produce different CDS.

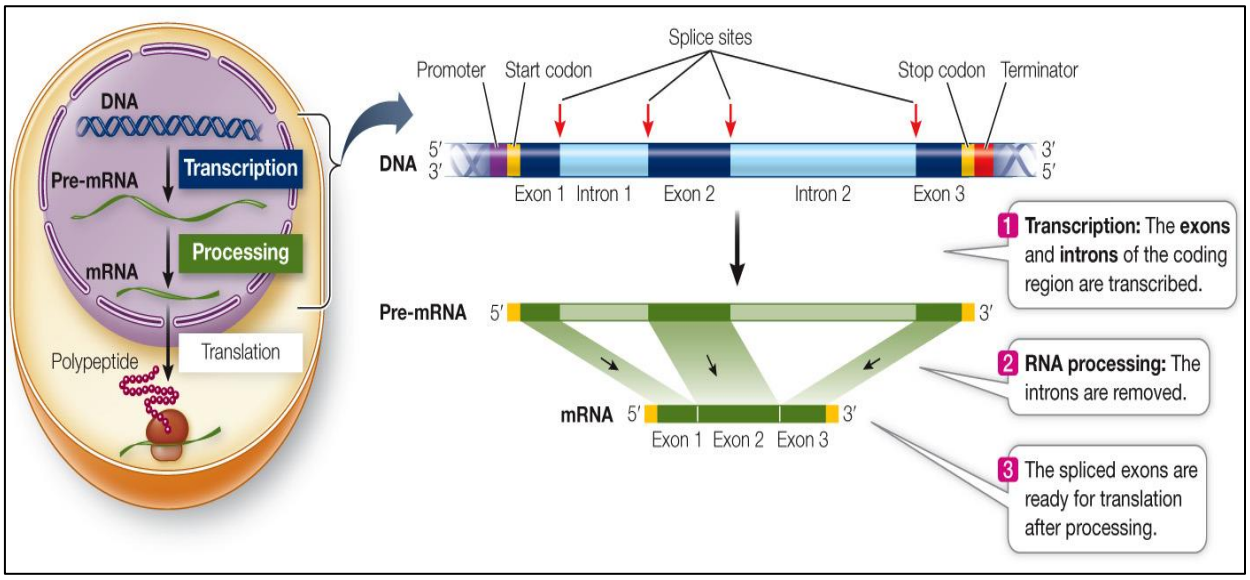


Figure 34. The concept of coding sequence in eukaryotes.

a. Promoters and 5' signals

In addition to "Start" codons and "Stop" codons, different types of signals marking the 5' region of the coding sequence can be searched for. The transcription signals are as follows:

• Promoter sequences recognized by RNA polymerases. In eukaryotes, there are three types of RNA polymerase (RNApolI, RNApolII and RNA polIII). Each RNA polymerase recognizes a type of promoter. PolI promoters are found upstream of the 18S and 28S ribosomal RNA genes. PolII promoters are found upstream of the messenger RNA genes. PolIII promoters are found upstream of the 5S ribosomal RNA and transfer RNA genes. Upstream of the protein coding sequences, we therefore look for a TATA box, a conserved sequence, rich in AT and of 8 nucleotides, which is found in the PolII promoters 25 to 30 nucleotides upstream of the transcription initiation site.

• Transcription factor binding sites;

• The initiator (INR), a weakly conserved sequence located near the transcription start site between positions -3 and +5;

• CpG islands. These are 1-2 kb regions, rich in CG dinucleotides, that are frequently associated with the 5' regions of vertebrate genes and extend over the promoter and the first exon.

• For translation signals, we look in particular for the ribosome binding site or Kozak sequence located upstream of the "Start" codon.

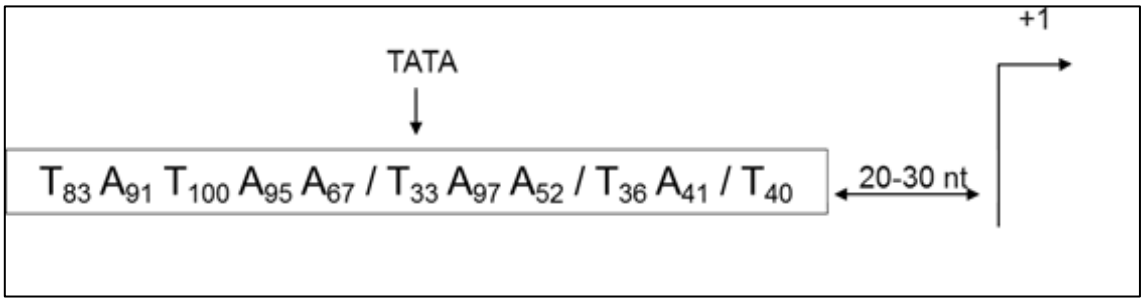


Figure 35. The consensus sequence of the TATA box in eukaryotes. The subscript numbers indicate the percentage of occurrence of each nucleotide in the definition of the consensus sequence. The +1 indicates the transcription initiation site.

b. Exon-intron junctions

Introns have four important signatures, the first two of which indicate exon-intron junctions:

- The donor site /GTRAGT at the 5' end of the intron;
- The dinucleotide GT is systematically excluded from mature mRNAs;
- The acceptor site NYAG/G at the 3' end of the intron;
- The dinucleotide AG is systematically excluded from mature mRNAs;
- The CTRAY branch point with an A that plays a central role in the splicing process;
- A pyrimidine-rich region between the branch point and the receptor site.

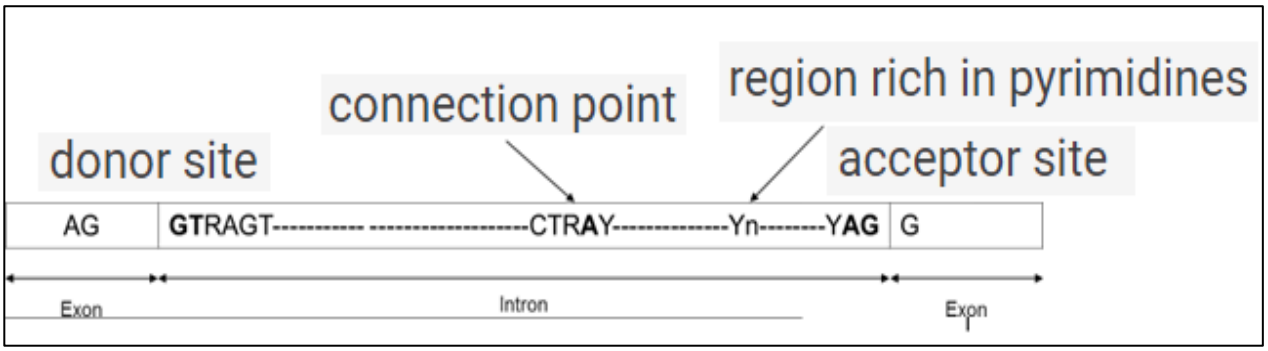


Figure 36. Exon-intron junction signals in vertebrates. Y: pyrimidines; R: purines.

c. 3' Signals

The terminal exon contains a set of signals indicating transcription termination. For messenger RNAs transcribed by RNA polymerase II, these signals are also necessary for polyadenylation because the latter event is coupled to transcription termination:

- The polyadenylation signal: 5'-AAUAAA-3' or 5'-AUUAAA-3';
- A cleavage signal. This is a rather poorly conserved CA dinucleotide, located 10 to 30 bases downstream of the polyadenylation signal.
- At this level, the DNA is cleaved before the addition of the polyA tail;
- A GU-rich region of variable sequence, 20 to 40 bases after the cleavage site.

5. Analysis of the base content of coding sequences

The search for signals indicating the presence of coding sequences is insufficient. This is true for all genomes, but even more marked for eukaryotic genomes. In the latter, the signals can be very degenerate (poorly conserved consensus) and the mosaic structure of the genes can be a source of error. A second approach is used to search for coding sequences in genomes: the analysis of the content of the sequences and the biases of this content in the coding regions compared to the non-coding regions.

a. Base composition

There are biases between coding and non-coding sequences in the base composition of dinucleotide, hexanucleotide, etc. sequences. This bias is used for the search for coding sequences and in particular to distinguish introns from exons in eukaryotes.

b. Codon usage bias

The abundance and usage of amino acids varies from one organism to another. This results in different frequencies for each codon within a genome. However, one might expect synonymous codons to be used with the same frequency. However, this is not the case. This is called codon usage bias.

There is a codon usage bias specific to each species. Within a genome, there are also biases specific to certain genomic regions and even to certain genes. It has been observed that the

most expressed genes are the most biased, the most used codons being those for which transfer RNAs are the most numerous. Codon usage bias is used to identify:

• CDS signatures in prokaryotes;

• Coding exon signatures in eukaryotes;

• Signatures of horizontal transfers of genetic material since the code usage bias is often different between different species.

6. Bioinformatics programs for syntactic annotation

The best computer programs for syntax annotation are those that combine gene signal detection and gene content analysis. They use algorithms that, after a training phase on a data set, can differentiate between genic regions and intergenic regions. In most cases, these methods use hidden Markov models (HMMs). HMMs can model and predict local features of molecular sequences (e.g., base composition) while incorporating knowledge from previous research into the analyses (e.g., RBS or promoter sequence).

Programs	Prokaryotes	Eukaryotes
GenMark-P	+	
Glimmer	+	
Genemark-E		+
Grail		+
Genescan		+
Genie		+

Table 6. Examples of coding sequence search programs and their applications.

Particularly in eukaryotes, only functional annotation by comparison with expression databases (see below) will allow to validate the genes predicted during syntactic annotation and to find genes that the computer prediction programs have failed to identify.

7. Functional annotation: Bioinformatics tools for sequence comparison

Sequences can be compared with programs such as FASTA (FAST-ALL) or BLAST. These similarity search tools are based on the concept of local alignment. Local alignment algorithms search for isolated regions in pairs of sequences that have a high degree of similarity.

Here we will describe the usage of the most commonly used program (cited in Google Scholar +70000 times), BLAST . The user provides a query sequence which is then compared to all the sequences in a chosen database. Different subroutines exist depending on the nature of the query sequence and the sequences in the database.

The BLAST result ranks the similarity results according to a significance index called the E-value (expected value), with the most significant result being the first in the list. The E-value is the number of different alignments with the same degree of similarity that one would expect to find by chance, if there were no true similar sequences in the database. So, practically speaking, the closer the E-value is to 0, the less the similarity is due to chance.

Program Name	Nature of the query sequence	Nature of database sequences
Blast ou Blastn	Nucleotides	Nucleotides
Blastp	Amino acids	Amino acids
Blastx	Nucleotides translated in the 6 reading phases	Amino acids
Blastn	Amino acids	Nucleotides translated in the 6 reading phases
Blastx	Nucleotides translated in the 6 reading phases	Nucleotides translated in the 6 reading phases

Table 7. The different BLAST programs.

8. Sequence alignment

Sequence alignment is the process of comparing two (pair-wise alignment) or more (multiple sequence alignment) nucleotide or amino acid sequences to determine their degree of similarity. A marked similarity between two gene or protein sequences to reflect their evolution from a common ancestral sequence. Sequences related in this way are called homologous, and the similarity maintained between these sequences during evolution is called homology.

To compare two sequences, one of the key steps is to match them with each other using an alignment that reveals the existing similarities. Alignments are based on the principle of the presence of a few identical residues at least at corresponding positions. There are two types of alignments: global and local according to (Needleman-Wunsch algorithm), global alignment is the alignment of all the sequences. While local alignment is the alignment of a part of the sequences.

There are different types of alignment:

- **Pairwise alignment:** consists of aligning 2 sequences. It is possible to perform a:

– Global alignment: alignment of two sequences over their entire length, taking into account all residues. If the lengths of the sequences are different, insertions or deletions are introduced to align the two ends of the two sequences. It is used to measure the degree of similarity between 2 known sequences.

– Local alignment: alignment between a sequence and part of the other sequence, i.e. the alignment of two sequences relating to isolated regions and making it possible to find segments that have a high degree of similarity. Efficient and rapid tool for searching databases by comparing an unknown sequence to those in the bank.

- **Multiple alignment:** alignment covering several sequences at once and in their entirety. It requires exponential computing time and storage space depending on the size of the data.

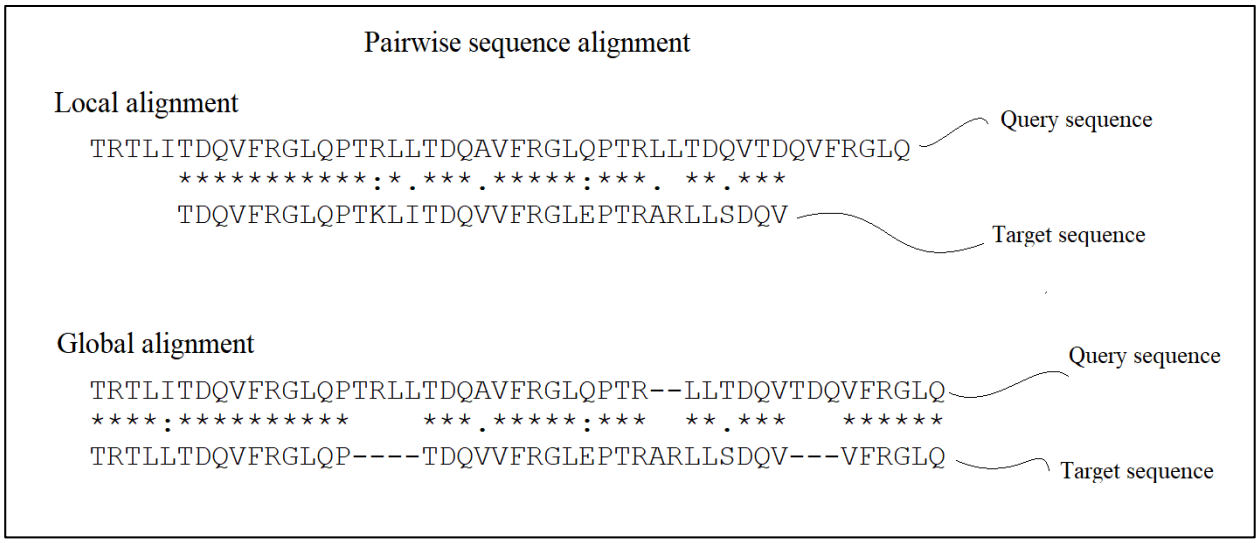


Figure 37. Pairwise alignment.

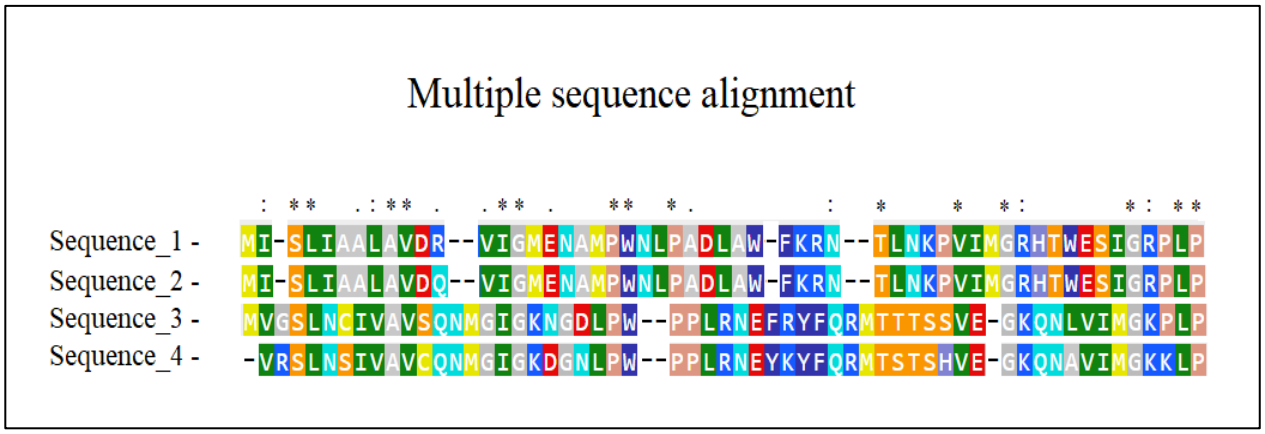


Figure 38. Multiple alignment.

VI. Conclusion

The structural bioinformatics tools described in this module provide a very useful support for the biologist aware of the importance of 3D structure in the functional elucidation of biological macromolecules. Most of them are available as web services or can be downloaded and installed for different operating systems, i.e. Linux, MacOs or Windows. Beyond their ease of use, it is essential to be aware of the limitations of the proposed methodologies by reading, of course, the reference publications but also by taking note of the results of the competitions organized regularly to evaluate the performance of these methods. It provides us with relatively objective data on the quality of protein structural models. This task has made it possible to make spectacular progress in the quality of predictions by encouraging the development of the evaluation tools essential for the use of these methods.

Bibliography

1-References used

[1] Amara Korba, R. Bioinformatique. Université de Bourdj. 2020.

[2] Sad Houari N. Bioinformatique et modélisation. Université des Sciences et de la Technologie Mohamed Boudiaf Oran . 2018-2019.

[3] Ammamra, R. Bio-analyse et Bio-informatique. Université de Bourdj. 2020.

[4] Chelgham, A. Bio-analyse et Bio-informatique. Universite de Chlef. 2022.

[5] Jungfer K, Cameron G, Flores T. EBI : CORBA and the EBI databases in bioinformatics. In : Letovsky, ed. Databases and systems. New York: Kluwer Academic Publishers, 2000 : 245–54.

[6] Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A. An ontology for bioinformatics applications.Bioinformatic s 1999; 15 : 510–20.

[7] Ducournau R, Euzenat J, Masini G, Napoli A. Langages et modèles à objets : état des recherches et perspectives. Collection Didactique. Paris: INRIA, 1998.

[8] Muller PA, Gaertner N. Modélisation objet avec UML. Paris: Éditions Eyrolles, 2000.

[9] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R. TAMBIS : transparent access to multiple bioinformatics information sources. Proc Int Conf Intell Syst Mol Biol 1998; 6 : 25–34

[10] Karp P. An ontology for biological function based on molecular interactions. Bioinformatics 2000; 16 : 269–85.

[11] Karp P, Riley M. Representation of metabolic knowledge. Proc Int Conf Intell Syst Mol Biol 1993; 1 : 207–15.

[12] Karp P. Pathway databases : a case study in computational symbolic theories. Science 2001; 293 : 2040–4.

[13] Kanehisa M. Post-genome informatics. Oxford (GBR) : Oxford University Press, 2000.

[14] Pittard AJ. Biosynthesis of aromatic amino acids. In : Neidhardt FC, et al., eds. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, 2nd ed. Washington DC : ASM, 1996 : 458–84.

[15] Page M, Gensel J, Capponi C, Bruley C, Genoud P, Ziébelin D. Représentation de connaissances au moyen de classes et d'associations : le système AROM. Actes du colloque Langages et Modèles à Objets (LMO), Mont Saint-Hilaire. Canada : Éditions Hermes, 2000 : 91–106.

2-References visited

Aouadj, S. A. (2009). Apport de la télédétection et SIG pour la cartographie des risques d'érosionhydrique du sol dans la région de Saida– (Western Algeria). These of engineering, Saida University,Tlemcen, Algeria.

Aouadj, S. A. (2021). Impact of ecological restoration techniques on the dynamics of degradedecosystems in the Saida mountains: Case of the forests of Doui Thabet – (Western Algeria). PhDtheses, Aboubakr Belkaïd University, Tlemcen, Algeria.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O. & Khatir, H. (2020). Impact of ecological restorationtechniques on the dynamics of degraded ecosystems of the mounts of Saida: Case of the forests ofDoui Thabet (West Algeria). Acta scientifica naturalis, 7 (2): 68-77.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). Impacts of anthropogenic pressureon the degradation of the forest of Doui Thabet (Saida, Western Algeria) in the context of therestoration. Acta scientifica naturalis, 7 (2): 68-78.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). Note on the orchids of mountsof Saida (Saida Western Algeria) in the context of the restoration. Eco. Env. & Cons, 26(2): 37-45.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). The rare, endemic and threatenedflora of the mountains of Saida (Algeria). Agrobiologia, 10(1), 86 – 98.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). Ethnobotanical Approach andFloristic Inventory of Medicinal Plants in the Doui Thabet Region (Saida-Western Algeria).PhytoChem & BioSub Journal, 14 (1): 92-104.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). Regional phytogeographic analysisof the flora of the Mounts of Saida(Algeria): evaluation-restoration report. Biodiversity Journal,11 (1): 25-34.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2020). Ecological characterization andevaluation of the floristic potential of the forest of Doui Thabet (Saida Western Algeria) in thecontext of the restoration. Eco. Env. & Cons, 26 (1): 266-278.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2022). Preliminary study of the pre-germinative treatments of Juniperus oxycedrus L. and Pistacia lentiscus L. Res. Conserv, (67): 13-20, 2022.

Aouadj, S. A; Nasrallah, Y; Hasnaoui, O & Khatir, H. (2023). Floristic and ecologicaldiagnostic of the Mounts of Saida in the context of ecological restoration: Assessment of six yearsof field research (2017-2022). Journal Concepts in Structural Biology & Bioinformatics (JSBB),7 (1) : 1-42.

Aouadj, S. A; Degdag, H; Hasnaoui, O; Nasrallah, Y; Zouidi, M; Allame, A; Nouar, B. & Khatir,H. (2023). Contribution of G.I.S and Remote Sensing for the Risk Mapping of Soil WaterErosion at Saida Province (Western of Algeria). Advanced Research In Life Sciences, 7 (1) :10 – 21
(PDF) The role of rehabilitating and restoration the green dam in managing the impact of invasive forest pests in Algeria. Available from: https://www.researchgate.net/publication/386140662_The_role_of_rehabilitating_and_restoration_the_green_dam_in_managing_the_impact_of_invasive_forest_pests_in_Algeria [accessed May 16 2025].

Aouadj, S. A; Degdag, H; Hasnaoui, O; Nasrallah, Y; Zouidi, M; Allame, A; Nouar, B. & Khatir,H. (2023). New data on Orchid flora (Orchidaceae) in the Tell region of Saida (western of Algeria):Assessment of six years of field research (2017-2022). Advanced Research In Life Sciences, 8(1): 10 – 21.
(PDF) The role of rehabilitating and restoration the green dam in managing the impact of invasive forest pests in Algeria. Available from:https://www.researchgate.net/publication/386140662_The_role_of_rehabilitating_and_restoration_the_green_dam_in_managing_the_impact_of_invasive_forest_pests_in_Algeria [accessed May 16 2025].

3-Webography

- (https://lifesciences.danaher.com/us/en/library/sequencing.html).
- (https://www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/the-stages-of-translation).
- (https://www.researchgate.net/publication/355466069_The_Dynamism_of_Transposon_Methylation_for_Plant_Development_and_Stress_Adaptation/figures?lo=1&utm_source=google&utm_medium=organic).
- (https://www.researchgate.net/publication/358134553_Application_and_Challenge_of_3rd_Generation_Sequencing_for_Clinical_Bacterial_Studies/figures?lo=1).
- (https://nebula.org/blog/fr/genome/).
- (https://global.sjtu.edu.cn/en/announcement/view/846).

- (https://microbenotes.com/bioinformatics-databases-software-tools/).
- (https://careersidekick.com/top-15-bioinformatics-degree-jobs/).
- (https://slideplayer.com/slide/8525757/).
- (https://parlonssciences.ca/ressources-pedagogiques/documents-dinformation/sequencage-de-sanger).
- (https://www.seqanswers.com/forum/general/15817-emulsion-pcr-in-detail-explained).
- (https://www.sciencedirect.com/topics/immunology-and-microbiology/illumina-dye-sequencing).
- (https://www.bionumerics.com/bionumerics-server).
- (https://www.sciencedirect.com/science/article/abs/pii/B9780323897754000213).
- (https://www.biologyexams4u.com/2023/02/10-types-of-biological-databases.html#google_vignette