CHAPTER I.

INTRODUCTION TO BIOINFORMATICS

I.1. Introduction

The development of bioinformatics follows 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 the flow of information, the role of bioinformatics quickly evolved towards the transformation of this raw data into knowledge.

Bioinformatics is currently defined as a field of research that analyzes and interprets biological data, using computational methods, in order to create new knowledge in biology. This discipline studies the information contained in the sequences of genes and proteins.

Biological systems are very complex and modern techniques for investigating the biological module provide a vast amount of experimental data. The ultimate goal of bioinformatics "is to integrate these data from very diverse origins to model living systems in order to understand and predict their behavior under conditions of normal or pathological operating functions".

Science has experienced an unprecedented (r)evolution with recent technological advances, which have produced a large amount of "omics" data. The increasing production and availability of this information in public databases has been and continues to be a challenge for professionals in various fields (Ritchie et al., 2015). But what is the challenge?

In biology, the greatest challenge is understanding the enormous amount of structural and sequential data generated at multiple levels of biological systems (Pevsner, 2015).

In bioinformatics, it is necessary to develop tools (statistical and computer science) that can contribute to the understanding of the mechanisms underlying the biological questions of the study (Pevsner, 2015). Given the complexity of science, this is also a very reductionist view.

The era of "new biology" is accompanied by the birth/development of other sciences, such as bioinformatics and computational biology, which have an integrated interface with molecular biology. Although considered recently, bioinformatics and genomics have developed interdependently and have had a historical influence on the available knowledge.

Therefore, this report aims to provide a brief overview of these sciences and provide principles supporting bioinformatics by covering the following aspects: i) types of biological

information and databases; ii) sequence analysis and molecular modeling; iii) genome analysis and iv) systems biology. Because these are broad areas, we want to highlight important points when using new techniques and provide tools that can be used in analyzing data and interpreting the results generated by these technologies.

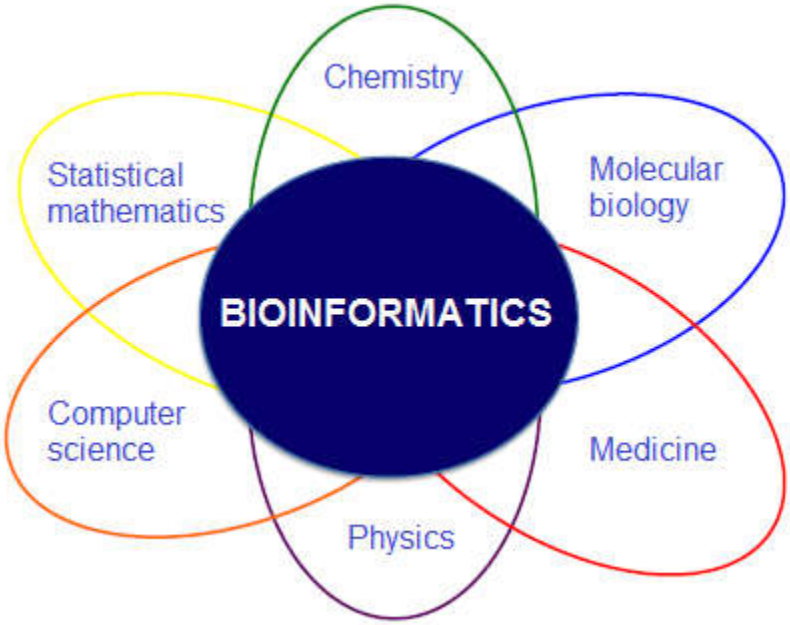
I.2 Definition of bioinformatics

Bioinformatics is a multi-disciplinary field involving biology, computer science, mathematics, and 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. It evolves according to new problems posed by biology.

Bioinformatics is defined as the use of databases and computer algorithms to analyze the genes, proteins, and complete collection of deoxyribonucleic acid (DNA) of a living organism (the genome).

Bioinformatics is a multidisciplinary science. It is located at the crossroads between biology, computer science, chemistry and many other disciplines. It is made up of all the concepts and techniques necessary for the computer interpretation and prediction of information from experimental biology data, in order to establish the links between the structure of biological macromolecules, their functions and their cellular activities. in the body.

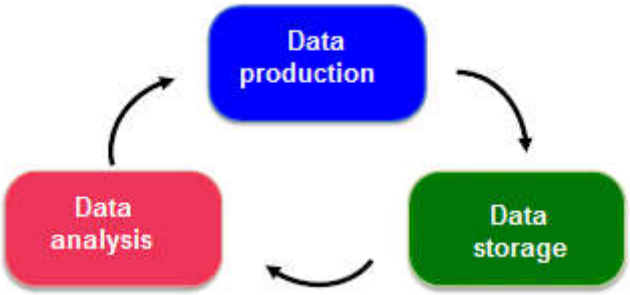
Bioinformatics involves several disciplinary fields:



Bioinformatics is a more pragmatic discipline. Development of practical tools for analyzing and organizing data. Less emphasis on the accuracy or effectiveness of the method. Dedicated to practical applications such as the identification of target proteins for drug design.

Bioinformatics is the "*in silico*" approach to biology which consists of computerized analysis of biological data using a set of means:

- Acquisition and organization of biological data;
- Design of software for data analysis, comparison and modeling;
- Analysis of the results produced by the software.



It is a discipline analogy with the terms:

- In vitro biology in an artificial environment

- In vivo biology of living organisms

- Biology in situ in natural environments.

I.3 History of bioinformatics

Bioinformatics has evolved from basic sequence analysis to a critical field that integrates biology, computer science, and statistics to manage and interpret vast amounts of biological data. The field continues to grow, driven by technological advancements and the increasing complexity of biological research.

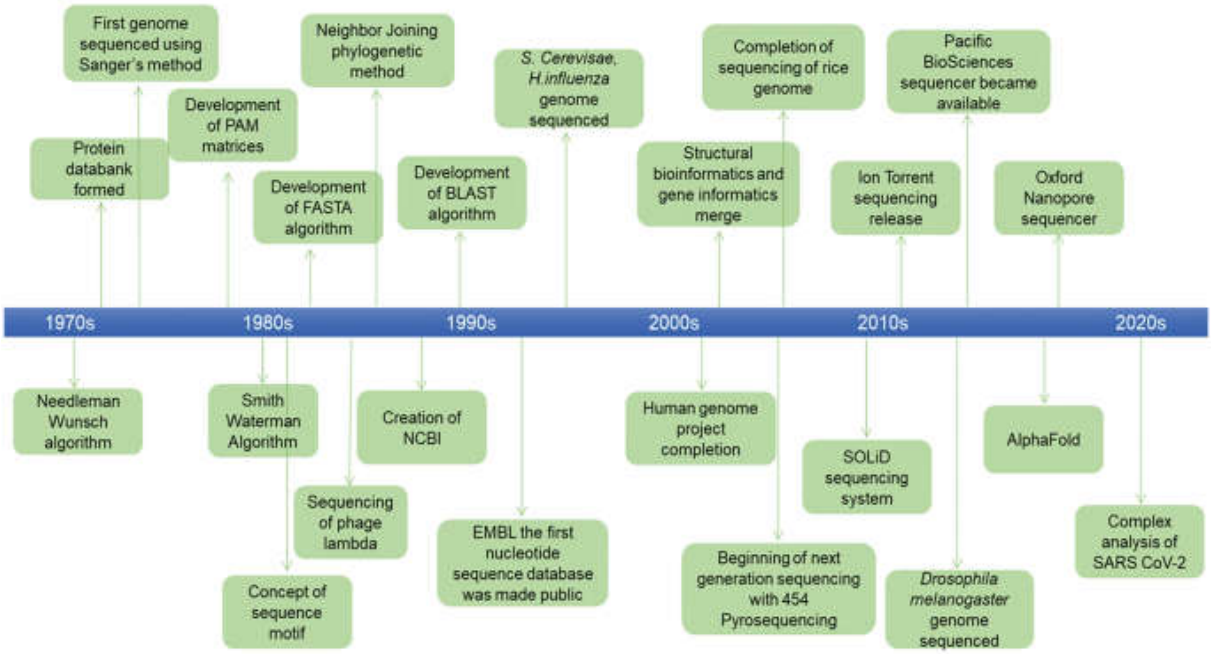


Figure : Introduction to the World of Bioinformatics (Mishra, 2023)

I.4. Objectives of bioinformatics

Bioinformatics applies to all types of biological data, particularly molecular:

- DNA and protein sequences
- Protein structures
- DNA chips (microarrays)
- Interaction networks between proteins
- Metabolic networks
- Phylogeny trees

Among the objectives of bioinformatics, we cite:

- Advance knowledge in biology, human genetics, evolutionary theory, etc.,

- Help with drug design,
- Understand complex diseases,
- Development of software for biology,
- Research in a laboratory,
- Help in the creation of genetically modified organisms (bacteria, plants, etc.).

I.5. Nature of data in bioinformatics

❖ Nucleic acid sequences: DNA and RNA

- DNA is **the carrier of genetic information**.
- DNA is a **long molecule**, made of **two strands** coiled into a **double helix**.
- The two strands of the double helix suggest a mechanism for DNA replication
- Each strand is the support for a **succession of nucleotides**
- **Four types of nucleotides**: (Adenine, Cytosine, Guanine, Thymine).
- The genomic text is written in an alphabet of 4 letters: A, C, G, T

❖ Nucleic acid sequences: DNA and RNA

DNA:

...AGGAGGATATTCCGAAAACGGTGGAGGTATCGGGATCGGAATTGTGAGTA
CCTGGTCACGTGGTCACATGTGTTTGCCTGGTTGCTAACTATTATTGTTTTT
TTATTCCAGGACCACGGAACCCATGGCCTTCTTGCAGGGATTAAACGTGAG
TTGTGCTTTTAATGTGCAAAGCTATAGCTTACTAACTATTTAATATTATTCC
CCGCGTCCGGGAATCTGATGCAGTTCAGCCAGGTGGGTAACATCGA...

ARN : 'U' replaces 'T'

A: Adenine, C: Cytosine, G: Guanine, T: Thymine, U: Uracil

❖ Protein Sequences: Primary structure

P53 Proteine

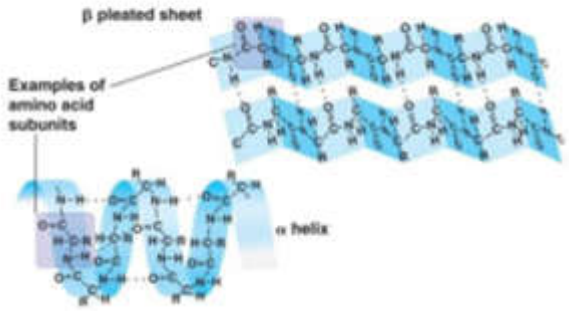
```
  1 MEEPQSDLSI ELPLSQETFS DLWKLLPPNN VLSTLPSSDS IEELFLSENV TGWLEDSGGA
 61 LQGVAAAAAS TAEDPVTETP APVASAPATP WPLSSSVPSY KTFQGDYGFR LGFLHSGTAK
121 SVTCTYSPSL NKLFCQLAKT CPVQLWVNST PPPGTRVRAM AIYKKLQYMT EVVRRCPHHE
181 RSSEGDSLAP PQHLIRVEGN LHAEYLDDKQ TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS
241 SCMGGMNRRP ILTIITLEDP SGNLLGRNSF EVRICACPGR DRRTEEKNFQ KKGEPCPELP
301 PKSAKRALPT NTSSSPPPKK KTLDGEYFTL KIRGHERFKM FQELNEALEL KDAQASKGSE
361 DNGAHSSYLK SKKGQSASRL KKLMIKREGP DSD
```

*Each **letter** represents an **amino acid***

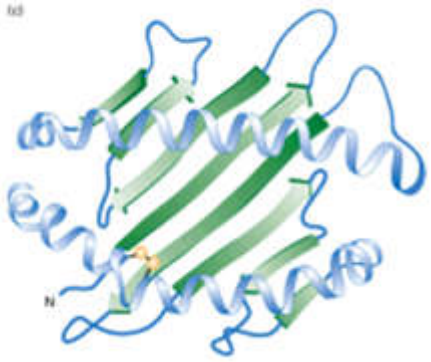
❖ Protein Sequences: Amino Acids Table

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC Ser UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU CUC Leu CUA CUG	CCU CCC Pro CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC Thr ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC Val GUA GUG	GCU GCC Ala GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G

❖ Protein Sequences: Secondary and Tertiary Structure



Secondary structure



Tertiary structure