Módulo 1

1.1 Bioinformatics: what, why and how?

*By **Coral del Val Muñoz***
Associate Professor at the University of Granada, Department of Computer Science and Artificial Intelligence (DECSAI).

By ***Carlos Cano Gutiérrez***
Associate Professor at the University of Granada, Department of Computer Science and Artificial Intelligence (DECSAI).

1. THE CONTEXT: A TSUNAMI OF DATA

Since its origins, research in biology and medicine has aimed to unite data and evidence to try to understand the functioning of biological systems and the causes of diseases. A biological system is a network of interacting biological entities and, depending on the scale or resolution of the study underway, the research target may be a single cell, organs and tissues of an organism, assemblages of organisms, or even entire ecosystems. Thus, these fields present research questions of high interest related to disciplines encompassing health, welfare, ecology, and energy, among others. To illustrate these research disciplines, some of the questions they address include:

• How burned soils can be treated to accelerate their recovery.

• The possibility of whether bacteria can be used to generate and store energy.

• Whether the microorganism species composition and diversity in the human intestinal flora significantly varies between individuals and how this might be related to disease.

• How viruses such as SARS-Cov-2 spread and how COVID-19 affects infected organisms.

• Examination of whether the early diagnosis or prevention of Alzheimer's disease and colon cancer is possible.

The main difficulty encountered by researchers trying to make progress in these fields in the 20th century was a lack of data. Thus, most of the research effort was devoted to generating data, the volume of data generated was manageable, and data analysis was conducted manually or with basic computer and statistical analysis tools such as spreadsheet software. However, enormous technological advances have taken place over the last two decades that are bringing about a true

revolution in the life and health sciences. Technology is making it possible to observe or measure biological systems with a precision never seen before, all at affordable costs that are still decreasing.

These technological advances allow us, for example, to measure the abundance of different molecules within a single cell, identify the species of microbes present in a person's intestinal flora, or monitor the spread of an invasive plant species in an ecosystem by using drones. To illustrate this drastic cost reduction, in 2001 the sequencing of the first human genome was completed after an international effort that cost more than $300 million; today, sequencing a human genome costs around $1,000.

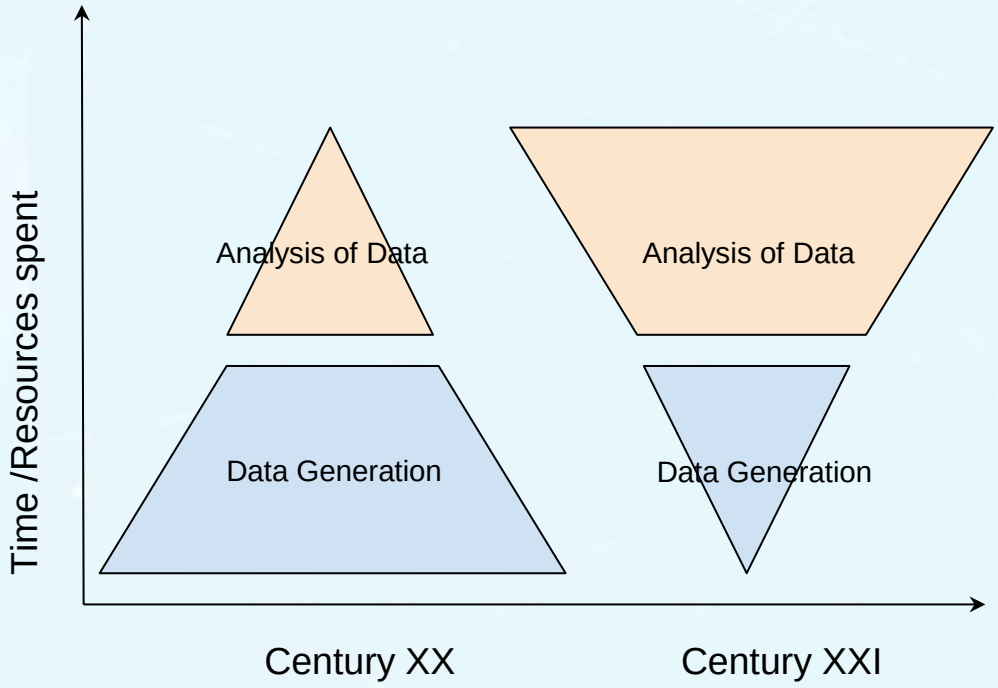

Figure 1. Illustration of the distribution of time and resources in biosciences and biomedicine research in the 20th and 21st centuries.

These lowering costs and increased technological performance is creating a veritable tsunami of data given that a single experimental study can generate petabytes or even terabytes of information. Given this enormous volume of data and its complexity, its analysis requires the use of dedicated computers and, more specifically, necessitates analysis techniques based on machine learning and big data (precisely, the content covered in this course). Hence, bioinformatics has arisen within this context of technological revolution because of the need to adequately analyze massive amounts of data.

2. WHAT IS BIOINFORMATICS?

There are many definitions of bioinformatics.

On the one hand, Wikipedia defines it as "the interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, chemistry, physics, computer science, information engineering, mathematics, and statistics to analyze and interpret the biological data." Thus, we see that it emphasizes the multidisciplinary nature of bioinformatics, even listing the different areas that converge in this new field of science. This definition also states that one of the main goals in bioinformatics research is the development of methods and software to help scientists interpret biological data.

On the other hand, a slightly broader definition was proposed by the journal *Nature*, which described the discipline as a "field of study that uses computation to extract knowledge from biological data. It includes acquisition, storage, retrieval, and modeling for analysis, visualization, or prediction through the development of algorithms or software." This definition emphasizes the different computational processes involved in bioinformatics: obtaining information and its storage, retrieval, modeling, analysis, visualization, and prediction.

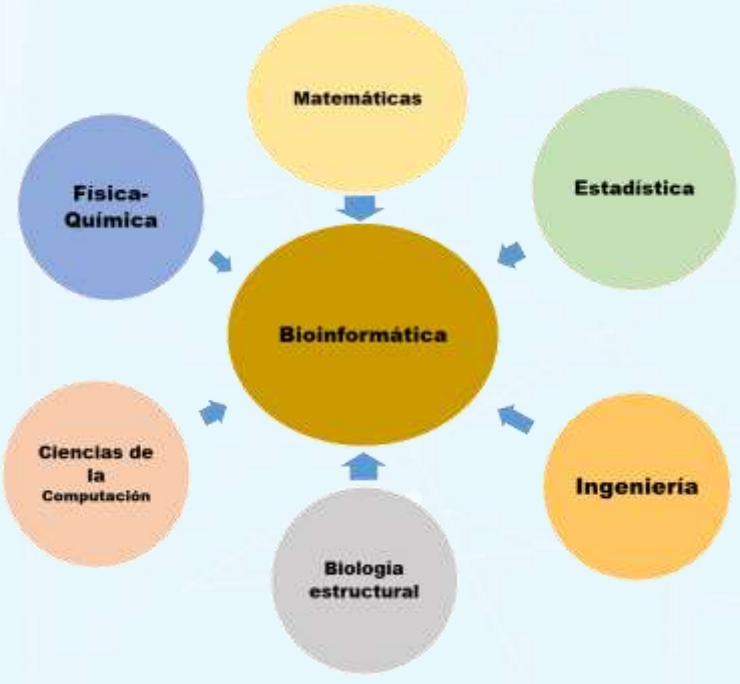

Figure 2. Bioinformatics as an interdisciplinary field.

In any case, just as in any other interdisciplinary field, bioinformatics involves several factors, depending on the prism through which one chooses to consider it. Hence, from the biology perspective, bioinformatics is 'computational biology.' In other words, a discipline that involves the use of computers to analyze any type of biological information (e.g., sequences, X-ray images,

clinical measurements, etc.). However, from the perspective of computer science, bioinformatics is 'biological informatics,' a concept that places much more emphasis on the acquisition, storage, and processing of information and the analysis of these large volumes of data.

Broadly speaking, the five major objectives of bioinformatics can be formulated as described below.

1. **To efficiently organize the copious amounts of data generated in the fields of biosciences and biohealth**. This objective includes the creation of databases and information repositories such as Ensembl, UCSC, or GenBank (for DNA sequences and genomes); UniProt (for proteins); GEO (transcriptomes); KEGG (metabolic networks); and multifactorial databases describing specific diseases (e.g., *The Cancer Genome Atlas*); etcetera.

2. **The design and development of tools and algorithms for data analysis**. In this sense, we can cite thousands of successful examples; one of the most popular algorithms is BLAST, bioinformatics software that allows sequences similar to the user's input sequence to be found very quickly in genomes or among proteins. Thousands of algorithms have been developed for data analysis in the general field of biology, including tools for applications in molecular biology, agriculture, pharmacological processes, biomass biotechnology, renewable energy, vaccine development, immunology, microbiology, food biotechnology, plant breeding, environmental impact, animal production improvement, and forestry, among others. Many instruments are also available for uses in the ambit of chemistry (e.g., chemical biotechnology, toxicology, and pesticide development) and medicine (with uses in neurosciences, psychiatry, nutrition, biomedical computing, risk calculation, disease diagnosis, and cancer, etc.).

3. **Development of models that explain the functioning of complex biological systems**. A model is a representation of a complex reality which is used to facilitate the latter's understanding. The objective of bioinformatics is to propose these representations based on data and algorithms in order to understand the regulatory mechanisms supporting these systems, thereby helping physicians to, for example, prevent, diagnose, or treat a given disease.

4. **Discover new knowledge by leveraging these models and tools.** A good example is the discovery of new biomarkers for the early diagnosis of certain types of cancers.

5. **Facilitate the accurate and meaningful interpretation of results by experts**. This point emphasizes the importance of delivering the knowledge uncovered by implementing these tools to experts in a way that is concise, precise, and easy to interpret and apply.

3. MULTIDISCIPLINARITY, OR, WHAT TO EXPECT FROM THIS COURSE…

The definitions of bioinformatics listed above make the multidisciplinary nature of this field of study clear. That is, when learning to navigate in the world of bioinformatics, it helps to speak different languages. On the one hand, we have the language of the problem domain (biology, medicine, chemistry, pharmacy, nutrition, and agriculture, etc.). But why is it important to speak this language? The reason is because we must be able to understand the nature of the problem and incorporate this knowledge into an appropriate solution. On the other hand, we must be fluent in the language of analysis (computing, machine learning, big data, mathematics, and statistics, among others). Again, why? The answer is so that we can propose techniques that could contribute to solving the problem at hand.

This course will allow you to get into bioinformatics with an emphasis on the language of computer science, especially machine learning and big data, without moving away from the real problems and applications in the worlds of biology, medicine, and other areas. The goal of this course is to allow you to discover what computer science (machine learning and big data techniques) can contribute to bioinformatics, what kind of problems you can solve by applying these techniques, and how to achieve these goals. In addition, we have included some resources in the bibliography that will allow you to further expand your training once you have completed the course.

4. THE ORIGINS OF BIOINFORMATICS: SEQUENCE ANALYSIS

In terms of its origins, bioinformatics arose from the need to understand the genetic code of living beings with the ultimate aim of discovering the molecular mechanisms involved in different developmental processes. Over time, in addition to focusing on the sequences of molecules such as DNA, RNA, or proteins, interest also arose in their structure, interactions with other molecules, and their interrelated mechanisms of regulation. Indeed, the bioinformatics tools developed over the years have become essential in many disciplines (e.g., medicine, agriculture, and nutrition, etc.). The following section provides a brief overview of the origin of this field and some of the most relevant current bioinformatics problems in biology.

As already mentioned, bioinformatics originally arose from the need to understand the genetic code. This implies that molecular biology was the first 'customer' of bioinformatics: the first discipline that generated such a volume of data that it required a new partner for its analysis. In this context, sequence analysis was the first problem that necessitated superior analysis capabilities and therefore, was the first application of bioinformatics.

To understand the origin of bioinformatics and its initial applications, we must first understand some basic concepts referred to in the discipline of molecular biology. For example, it is important to know the so-called central dogma of molecular biology. In a simplified form, this dogma stipulates that genetic information flows from DNA to RNA to protein. The information stored in the genome (DNA molecules) present in the nucleus of each organism's cells is encoded

in nucleotide sequences. These nucleotides are adenine, cytosine, guanine, and thymine (referred to as A, C, G, and T) in a language that uses an alphabet of 4 symbols.

In a biological process called transcription, this information is encoded into another language: that of RNA sequences which are also nucleotides, in this case, adenine, cytosine, guanine, and uracil (A, C, G, and U). RNA molecules leave the cell nucleus and move into the cytoplasm where they interact with each other and with other molecules. Some of these sequences (messenger RNAs, or mRNAs) interact with ribosomes, which translate them into new types of molecules—proteins—that encode information using the language of amino acids, which involves 20 different symbols.

These amino acid sequences acquire a three-dimensional structure which confers them with a function within the cell. Hence, the process of gene transcription and translation into proteins must be understood as a quantitative process subject to constant regulation. Thus, the presence or abundance of certain molecules within the cell will cause or stop the expression of certain genes, halting the production of certain types of proteins. This process is generically called gene expression regulation and will be the focus of some of the problems tackled in this course.

The emergence of bioinformatics dates to the 1960s, when Sanger and his collaborators developed a method for sequencing the protein molecules involved in transcription. From hereon in we will usually use the term 'sequencing,' which simply refers to the process of determining the sequence of a certain type of molecule. Sanger and his collaborators developed a method to ascertain the exact sequence of amino acids comprising a given protein. This experimental advance meant that new algorithms had to be developed to analyze and compare the different protein sequences from different organisms. This was because the huge volume of sequences available prevented this work from being done manually. Thus, multiple alignment algorithms designed to establish the level of similarity between two sequences and determine the regions they have in common were created. The main sequence databases, such as GenBank, were also established during this period.

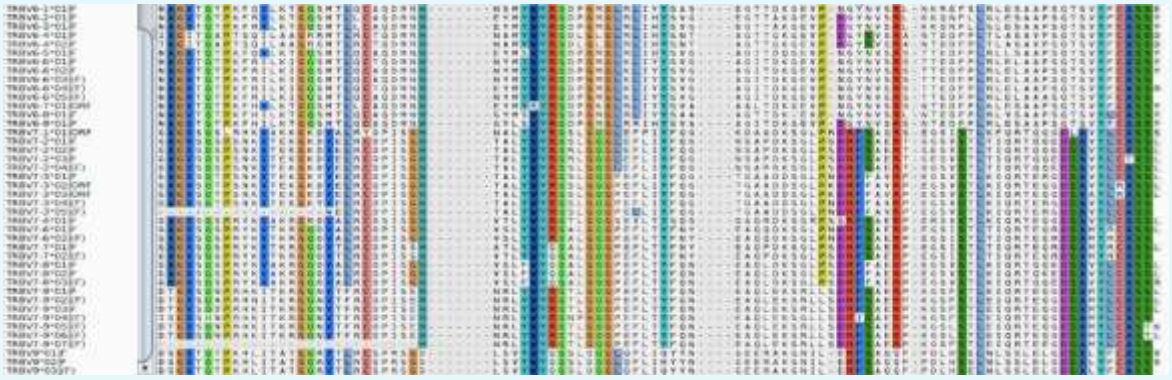

Figure 4. Multiple amino acid sequence alignments. In a multiple alignment, each protein is arranged in a row and its sequence is detailed by adding gaps (-) to highlight the amino acids the sequences have in common. The amino acids with the highest consensus at each position (column) are highlighted with a color code.

The boom in sequence analysis coincided with the popularization of the internet in the 1990s. This made the distribution of new software and discoveries faster and led to the appearance of web pages offering analyses of these data to the scientific community. One example is the website of the US National Center for Biotechnology Information (NCBI).

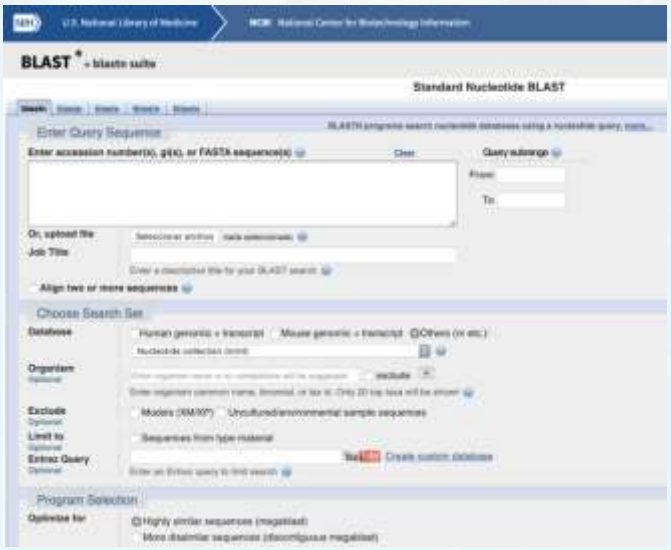

Figure 5. A page from BLAST, an algorithm that analyzes millions of sequences per second to identify the sequence most like another given input sequence.

Since then, there have been countless advances in the methods available for analyzing, comparing, and visualizing molecular sequences. However, the culmination of the *Human Genome Project* in 2001–2003 was an important turning point because it made a complete sequence of a human genome available to the scientific community for the first time. This event marked the beginning of the so-called post-genomic era.

Figure 6. The covers of the journals *Science* and *Nature* upon the publication of the first draft of the human genome in 2001.

6. THE POST-GENOMIC ERA AND THE 'OMICS' SCIENCES

The fact that, for the first time, a reference sequence (assembly) of the human genome had become available marked the beginning of the post-genomic era, which was characterized by the advent of new, cheaper sequencing technologies. These have made it possible to study genomes of every type (genomics) including plants, animals, microbes, and humans (e.g., *1,000 Genome Project*). Importantly, this technical revolution has led to the generation of new knowledge on the structure and functioning of genomes.

In fact, together with biotechnological advances, this improved understanding has led to the emergence and development of other omics sciences. These new fields are focusing on the characterization and quantification of many types of molecules including lipids, microRNAs (miRNAs), proteins, and metabolites, grouped according to their biological, structural, and/or functional characteristics. Some of these omics sciences and their respective study objectives are briefly described in the following points.

● Genomics: study of the structure and function of genomes. For example, this field includes the study of sequences, mutations, copy number variations, insertions and deletions, and structural variations such as translocations of chromosome pieces, etc.

● Transcriptomics: research into the expression of all the RNA molecules in a cell or collection of cells under specific circumstances. This includes the study of differential gene expression, gene fusion, alternative splicing, RNA editing, and the expression of protein-coding genes and regulatory RNA genes, etc.

● Epigenomics: analysis of chemical changes in DNA and histones (proteins responsible for DNA compaction). For example, this discipline includes the study of DNA methylation, histone modification by de/acetylation, and transcription factor binding, etc. We can say that the epigenome adds reversible marks onto the genome and that these marks affect the activation/deactivation of gene expression. The field of epigenetics focuses on these marks.

● Proteomics: study of all the proteins in a cell, tissue, or organism.

● Metabolomics: study of all the small chemical molecules termed metabolites (e.g., hormones) in a cell, tissue, or organism. Metabolites are intermediates responsible for the functions of signaling, stimulation, enzyme inhibition, and interactions with other organisms (e.g., pigments and pheromones).

● Microbiomics: exploration of the complete collection of microbes present in an organism.

● Metagenomics: study of all the genetic material obtained from an environmental sample (e.g., a pond or snow, or the gut, skin, or oral mucosa).

● **Lipidomics:** investigation of cellular lipids in biological systems. The Lipidome is part of the metabolome and its study uses the same tools as metabolomics but its results are especially important in diseases whose pathogenesis is related to lipid metabolism, such as obesity or hypertension.

● **Glycomics:** analysis of carbohydrate (sugar) compounds generated in extraordinarily complex metabolic pathways; these usually bind to other elements to form glycoproteins (important for cell-to-cell recognition) or glycolipids (vital for cell stability). Of note, different cancers have different glycoprotein profiles.

● **Phenomics:** systematic study of all the observable or measurable characteristics of a given organism. The phenotype is the result of the expression of the genotype in each environment.

Omics and their respective data ('omas') have now provided a more complete and complex picture of gene regulation than that offered by the central dogma of molecular biology. Indeed, advances in these disciplines have shown that not only do proteins interact with each other, RNA, and DNA to regulate transcription, but that RNA also plays a direct role in gene regulation. It has even become clear that the flow of information goes not only from DNA to RNA, but it also moves in the opposite direction. This knowledge has led to the discovery of new key elements in the functioning of gene expression with the discovery of regulatory RNA molecules including long non-coding RNAs (lncRNAs), miRNAs, and cis-activators of transcription such as enhancers.

The omic sciences have allowed the generation of enormous amounts of data. Together with important advances in computational mathematics, machine learning and big data strategies have now been used for the analysis and study of biological systems. The vastness of the results generated by such experiments was unthinkable only 20 years ago and has led to the extension of bioinformatics into other areas of research. In the next capsule, we will describe some of the most relevant areas of the applications of bioinformatics, both related to the omic sciences and those now arising from other disciplines.

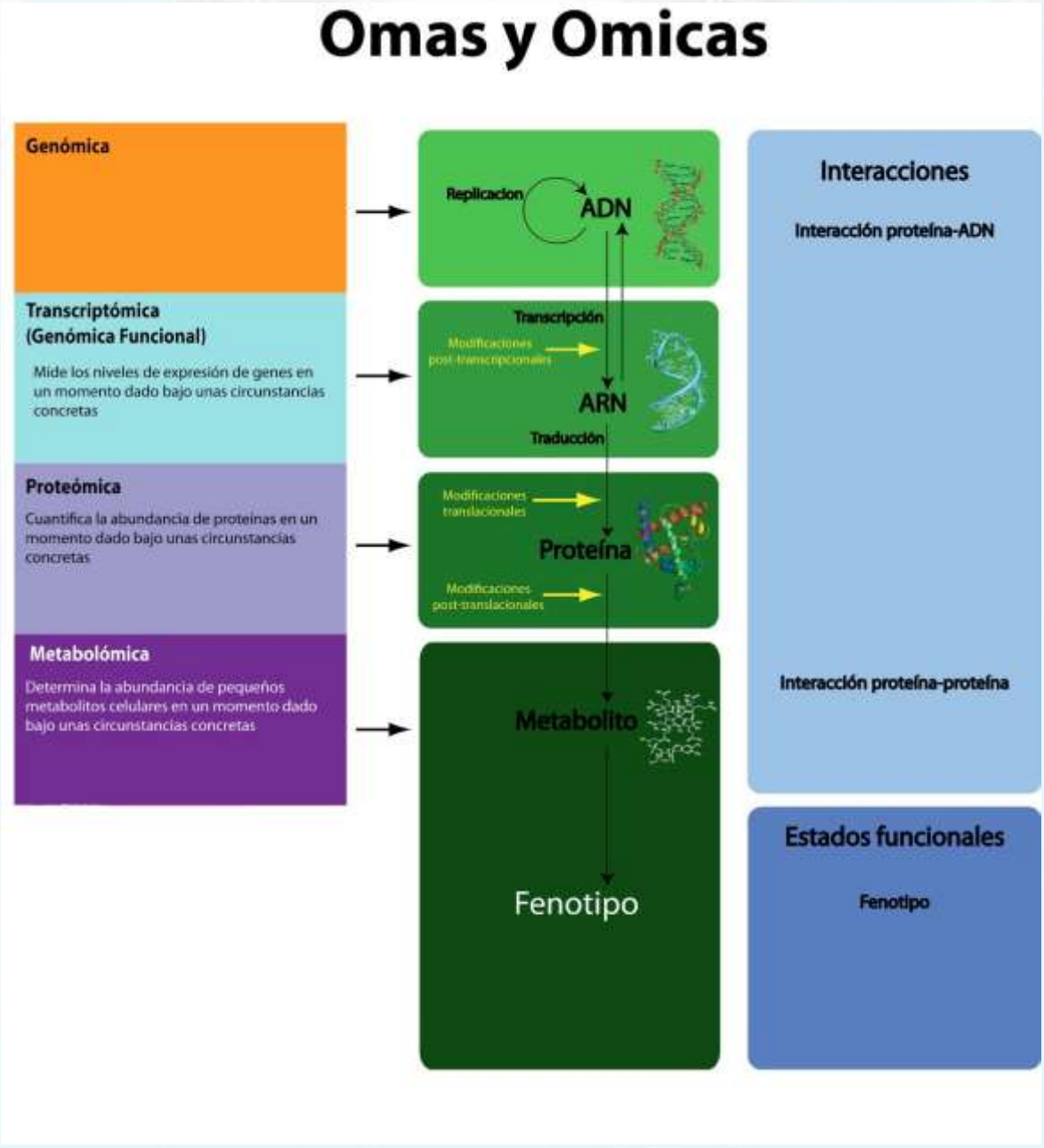

Figure 7. Diagram showing the relationship between different omics sciences and their principal study objectives.

BIBLIOGRAPHYCAL REFERENCES

- **Introduction to Bioinformatics**. A. Lesk. OUP Oxford, 2014. ISBN 0199651566, 9780199651566
- **Recurso web:** The cost of Sequencing a Human Genome. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost
- **WEB RESOURCE:** Bioinformatics for the terrified. C. Brooksbank , A. Cowley. EMBL-EBI. doi: 10.6019/TOL.BioinfTer-c_2016.00001.1
- **WEB RESOURCE:** Proyecto colaborativo LibreTexts: Introducción a la Biología y Biología Molecular (CK-12) https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/Book%3A_Introductory_Biology_(CK-12)/04%3A_Molecular_Biology
- **Bioinformatics Curriculum Guidelines:** Toward a Definition of Core Competencies. L. Welch et al. PLoS Comput Biol 10(3): e1003496. https://doi.org/10.1371/journal.pcbi.1003496
- **Recurso web:** Careers in Bioinformatics. International Society for Computational Biology. https://www.iscb.org/bioinformatics-resources-for-high-schools/careers-in-bioinformatics
- **Searls DB (2014) A New Online Computational Biology Curriculum**. PLoS Comput Biol 10(6): e1003662. https://doi.org/10.1371/journal.pcbi.1003662