I. INTRODUCTION TO BIOINFORMATICS

Bioinformatics is a recent field of research that started in the late 60's and it combines informatics, theoretical computer science, statistics, and other branches of mathematics to solve molecular biology problems. It has been growing and diversifying extremely rapidly over the last 20 years. It is by nature multidisciplinary and its rapid growth has resulted in the development of many specialized areas.

To understand the needs and problems bioinformatics tries to address it is useful to introduce a simplified model of how genes are represented and expressed in a cell, the so-called Central Dogma of Molecular Biology.

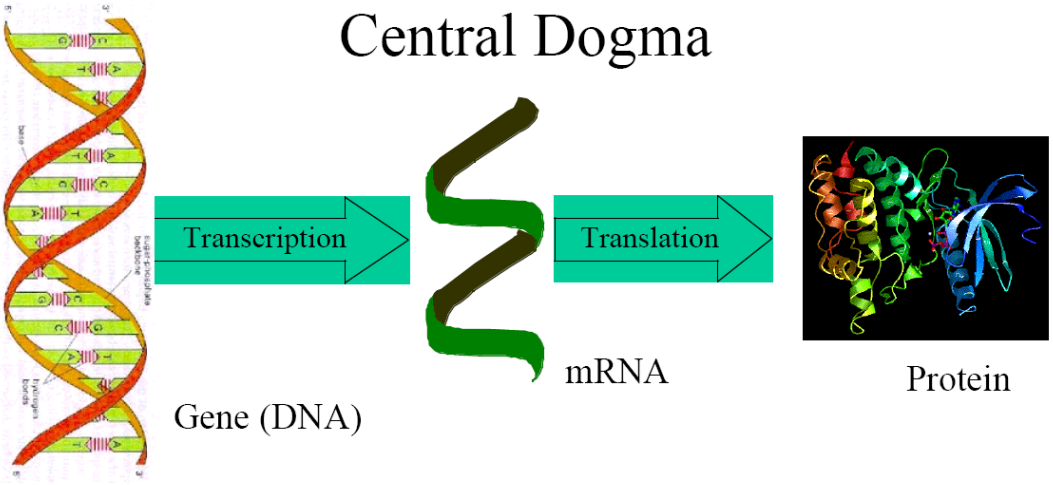

Figure 1: Central Dogma of molecular biology. In a living cell, the chromosomes, made of DNA, code for the genes, which are transcribed as messenger RNAs (mRNAs). mRNAs are finally translated into proteins. Which genes are actually expressed and expression intensity depend on the type of cell and the conditions in which the cell finds itself (environment, cell cycle, etc.). The chromosomes constitute the genome, the mRNAs constitute the transcriptome, and the proteins constitute the proteome. (Source: Esti Yeger-Lotem & Gideon Greenspan.)

As we can see from Figure 1, genetic information is coded by genes, which are present along the chromosomes and constitute the genome; chromosomes are made of deoxyribonucleic acid (DNA). Within the cell, a complicated mechanism of regulation controls the expression of specific genes, which are then used as templates, to produce messenger ribonucleic acids (mRNAs). sRNAs are molecules that are finally translated into proteins. Most of the biochemical reactions occurring in a cell are controlled by and/or involve proteins. They are the „building-blocks of life" and DNA is the "code of life".

This partial and simplified description of the fundamental mechanisms occurring in a cell allows us to present some essential domains of molecular biology research, related data management, and analysis problems bioinformatics has to solve.

A fundamental goal of molecular biology is to know all the genes of a given species or organism. Chromosomes are very large DNA molecules made of two strands of nucleotides: Adenine, Cytosine, Thymine, and Guanine, which can be represented by the four letters A, C, T, and G. The actual sequence of nucleotides in the chromosomes code for genetic information and they contain millions of nucleotides. The whole genome of an organism may contain up to several billions of nucleotides. The gigantic amounts of data, as well as the convenient 4-letter (digital) code, obviously suggest that computers can be used to store, compare, model, and mine genetic information efficiently.

Genetic material is a static template which the cell reads dynamically to respond to its environment and to duplicate itself. A lot of efforts in biology research are oriented towards the understanding of the control mechanisms of gene expression and in comparing gene expression patterns between cells in different conditions. By measuring the concentration of thousands of mRNAs it is possible to obtain a snapshot of the „expression program“ of the cell in a given condition. Comparisons of multiple conditions, following thousands of genes, involve the analysis of large table of numbers by statistical methods mainly. A classical example is the comparison of healthy and diseased cells to discover genes, which expression – or absence of expression – is related to pathology.

Since proteins have a central role in the cell machinery, the analysis of their function is a very important problem in molecular biology. As it is the case for the chromosomes, proteins are made of a finite number of basic molecules, the 20 amino acids. Therefore, proteins also can be conveniently represented by sequences of letters in a computer. The search for patterns in these sequences is named sequence analysis. Specific patterns are related to the function of the proteins and their detection makes predictions possible; this is a very developed field of bioinformatics. In addition, proteins assume a 3-dimentional structure, which conditions their biological function. Thus the analysis and prediction of the 3-dimensional structures of proteins is another – related – very important topic of bioinformatics. Databases to represent protein sequences and patterns as well as algorithms to find patterns and predict or compare 3-dimensional structures are typical bioinformatics tools used to work with proteins.

Besides the essential bioinformatics fields mentioned above, there are many other domains of research such as the analysis of the gene expression control mechanism, the evolution of the genome and speciation, physical interactions between proteins, and the modeling of biochemical reactions within the cell to mention a few.