INTRODUCTION

At the end of the 1980s, a new discipline named data mining emerged. The introduction of new technologies such as computers, satellites, new mass storage media, and many others have lead to an exponential growth of collected data. Traditional data analysis techniques often fail to process large amounts of, often noisy, data efficiently in an exploratory fashion. The scope of data mining is the knowledge extraction from large data amounts with the help of computers. It is an interdisciplinary area of research that has its roots in databases, machine learning, and statistics and has contributions from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. There are many applications of data mining in the real world. Customer relationship management, fraud detection, market and industry characterization, stock management, medicine, pharmacology, and biology are some examples (Two Crows Corporation, 1999).

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies and the introduction of new ones such as the microarrays. These technological advances have assisted the conduct of large-scale experiments and research programs. An important example is the Human Genome Project that was founded in 1990 by the U.S. Department of Energy and the U.S. National Institutes of Health (NIH) and was completed in 2003 (U.S. Department of Energy Office of Science, 2004). A representative example of the rapid biological data accumulation is the exponential growth of GenBank (Figure 1), the U.S. NIH genetic sequence database (National Center for Biotechnology Information, 2004). The explosive growth in the amount of biological data demands the use of computers for the organization, maintenance, and analysis of these data.

This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology. As Luscombe, Greenbaum, and Gerstein (2001) mention, the aims of bioinformatics are:

- The organization of data in such a way that allows researchers to access existing information and to submit new entries as they are produced.
- The development of tools that help in the analysis of data.
- The use of these tools to analyze the individual systems in detail in order to gain new biological insights.

The field of bioinformatics has many applications in the modern day world, including molecular medicine, industry, agriculture, stock farming, and comparative studies (2can Bioinformatics, 2004).

BACKGROUND

One of the basic characteristics of life is its diversity. Everyone can notice this by just observing the great
differences among living creatures. Despite this diversity, the molecular details underlying living organisms are almost universal. Every living organism depends on the activities of a complex family of molecules called proteins. Proteins are the main structural and functional units of an organism’s cells. A typical example of proteins are the enzymes that catalyze (accelerate) chemical reactions. There are four levels of protein structural arrangement (conformation) as listed in Table 1 (Brazma et al., 2001).

The statement about unity among organisms is strengthened by the observation that similar protein sets, having similar functions, are found in very different organisms (Hunter, 2004). Another common characteristic of all organisms is the presence of a second family of molecules, the nucleic acids. Their role is to carry the information that “codes” life. The force that created both the unity and the diversity of living things is evolution (Hunter, 2004).

Proteins and nucleic acids are both called biological macromolecules, due to their large size compared to other molecules. Important efforts towards understanding life are made by studying the structure and function of biological macromolecules. The branch of biology concerned in this study is called molecular biology.

Both proteins and nucleic acids are linear polymers of smaller molecules called monomers. The term sequence is used to refer to the order of monomers that constitute a macromolecule. A sequence can be represented as a string of different symbols, one for each monomer. There are 20 protein monomers called amino acids. There exist two nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), composed by four different monomers called nucleotides. DNA is the genetic material of almost every living organism. RNA has many functions inside a cell and plays an important role in protein synthesis (Table 2). Moreover, RNA is the genetic material for some viruses such as HIV, which causes AIDS.

The genetic material of an organism is organized in long double-stranded DNA molecules called chromosomes. An organism may contain one or more chromosomes. Gene is a DNA sequence located in a particular chromosome and encodes the information for the synthesis of a protein or RNA molecule. All the genetic material of a particular organism constitutes its genome.

The central dogma of molecular biology, as coined by Francis Crick (1958), describes the flow of genetic information (Figure 2). DNA is transcribed into RNA, and then RNA is translated into proteins. The circular arrow around DNA denotes its replication ability. However, today it is known that in retroviruses RNA is reverse transcribed into DNA. Moreover, in some viruses, RNA is able to replicate itself. The extended statement of central dogma of molecular biology is depicted in Figure 3.

Houle et al. (2000) refer to a classification of three successive levels for the analysis of biological data that is identified on the basis of the central dogma of molecular biology:

1. Genomics is the study of an organism’s genome and deals with the systematic use of genome information to provide new biological knowledge.
2. Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene’s coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription (Houle et al., 2000).

### Table 1. The four levels of protein conformation

- **Primary Structure.** The sequence of amino acids, forming a chain called polypeptide.
- **Secondary Structure.** The structure that forms a polypeptide after folding.
- **Tertiary Structure.** The stable 3D structure that forms a polypeptide.
- **Quaternary Structure.** The final 3D structure of the protein formed by the conjugation of two or more polypeptides.

### Table 2. Some of the basic types of RNA

- **Messenger RNA (mRNA).** Carries information from DNA to protein synthesis site.
- **Ribosomal RNA (rRNA).** The main constituent of ribosomes, the cellular components where the protein synthesis takes place.
- **Transfer RNA (tRNA).** Transfers the amino acids to ribosomes.