Chapter 16
Complex Biological Data Mining and Knowledge Discovery

Fatima Kabli
Dr. Tahar Moulay University of Saida, Algeria

ABSTRACT

The mass of data available on the Internet is rapidly increasing; the complexity of this data is discussed at the level of the multiplicity of information sources, formats, modals, and versions. Facing the complexity of biological data, such as the DNA sequences, protein sequences, and protein structures, the biologist cannot simply use the traditional techniques to analyze this type of data. The knowledge extraction process with data mining methods for the analysis and processing of biological complex data is considered a real scientific challenge in the search for systematically potential relationships without prior knowledge of the nature of these relationships. In this chapter, the authors discuss the Knowledge Discovery in Databases process (KDD) from the Biological Data. They specifically present a state of the art of the best known and most effective methods of data mining for analysis of the biological data and problems of bioinformatics related to data mining.

INTRODUCTION

In recent years, Bioinformatics has experienced important development, linked the culmination of many works of sequencing, which having led to the arrival of enormous biological quantity of data with different type (DNA, proteins, RNA), all these data are grouped in a variety of databases in their volume and nature. So, it is necessary to implement computer strategies to gain maximum knowledge. The application of data mining techniques in the genomic data is considered as a particular difficult task, represents a real scientific challenge, based on an exploratory analysis of data, to search the systematically potential relationships without prior knowledge of the relationships nature; in order to help the biologist to understand the function of genes and their structures. The biological data mining help the biologist to solve the essential questions for understand the function of genes: what is the role of a gene in which biological process it is involved? How the genes are regulated? What are the genes involved in a particular disease? And many other questions about the structures and functions of genes and proteins.

DOI: 10.4018/978-1-5225-3004-6.ch016

Many data mining techniques have been applied to answer these questions; the classification, clustering, association rules and text mining, this chapter is structured around four stages:

(1) An introduction to biological information with different types of representation format and biological databases available on the internet, (2) A state of the art of bioinformatics and its application fields, (3) the KDD biological process steps with biological data mining methods, (4) the relation among the bioinformatics problems and data mining. Finally, we conclude our chapter.

BIOLOGICAL DATA

Molecular Sequences

To understand the bioinformatics fields, it is necessary to have a rudimentary biology knowledge. This section gives a brief introduction to some basic concepts of molecular biology that are relevant to bioinformatics problems.

Our body consists many organs. Each organism consists of a number of tissues, and each tissue considered as a collection of similar cells that perform a specialized function.

The individual cell is the minimum auto reductive unit in all living species. It performs two different functions:

- Storage and transmission of genetic information to keep life from one generation to another, this information is stored in the form of bi-catenary DNA
- Perform the necessary chemical reactions to keep our life, through proteins that are produced by the transcription of DNA portions to ARN to protein.

The three basic types of molecules are deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins are present in a cell, in this section we discuss these three main molecules.

DNA

Deoxyribonucleic acid (DNA) is the genetic material of all organisms (with the exception of certain viruses), it stores the instructions necessary for the cell to perform the vital functions.

The correct structure of the DNA was deduced by (J.D.Watson and F.H.C.Crick, 1953), they deduced that the DNA consists of two antiparallel strands that are wound around each other to form a double helix. Each strand is a chain of small molecules called nucleotides.

The types of nucleotides depend on the type of the nitrogenous bases, which are adenine (A), guanine (G), cytosine (C), thymine (T).

According to the analysis of E.Charga and colleagues, it is deduced that the concentration of Thymine is always equal to the concentration of adenine and the concentration of cytosine is always equal to the concentration of guanine. This observation strongly suggests that A and T as well as C and G have some fixed relation.