Chapter 13
Biological and Medical Big Data Mining

George Tzanis
Aristotle University of Thessaloniki, Greece

ABSTRACT

This chapter discusses the concept of big data mining in the domain of biology and medicine. Biological and medical data are increasing at very rapid rates, which in many cases outpace even Moore's law. This is the result of recent technological development, as well as the exploratory attitude of human beings, that prompts scientists to answer more questions by conducting more experiments. Representative examples are the advances in sequencing and medical imaging technologies. Challenges posed by this data deluge, and the emerging opportunities of their efficient management and analysis are also part of the discussion. The major emphasis is given to the most common biological and medical data mining applications.

INTRODUCTION

Data collection and data analysis were actually taking place from ancient time, even if they were in a primitive form. Many ancient human civilizations had gained important knowledge by observing the planets and the stars. By analyzing these observations they were able to accurately predict the time of the seasonal changes over a year. These predictions were very valuable, especially for agricultural and habitation purposes, providing the means for the survival and development of these civilizations.

Later on in the age of scientific revolution data collection and analysis became a more mature process that guided to a large number of important scientific discoveries. Worth mentioning is the large number of accurate and comprehensive astronomical observations that were collected by the Danish astronomer Tycho Brahe during the early years of scientific revolution in 16th century. After Brahe's death, Johannes Kepler used those astronomical data, a fact that implies a kind of data sharing, and developed his three laws of planetary motion. Another important example of data collection and data analysis was the one of Charles Darwin's in 19th century. Darwin made a voyage that lasted almost five years. During the voyage he investigated geology of the lands he visited and made a lot of natural history collections. The

DOI: 10.4018/978-1-4666-9562-7.ch013

notes and observations he made during his voyage were determinant for the development of natural selection and evolution theories.

In the 20th century the important discoveries concerning DNA, such as the clarification of the correct double-helix model of DNA structure (Watson & Crick, 1953) established molecular biology as one of the most important research fields of biology. These discoveries attracted much attention and changed the direction of research in biology, as well as in medicine. Although the advances in biology during the 20th century were great, the scientific theories and discoveries of physicists are considered even greater. Therefore 20th century is described as the century of physics. However, as it is widely believed we are now living in the century of biology, which promises important advances that will enlighten the constitutive details and rules that characterize and govern life (Venter & Cohen, 2004).

The acquisition of more data has been proceeding through various inventions and technological advancements. For example, the invention and use of telescope made possible the observation of more objects in the sky, whereas the invention and use of the microscope made possible the discovery and study of microscopic organisms such as bacteria. One of the most important recent technological advancements in biology was the development of the polymerase chain reaction (PCR) by Kary Mullis in 1983. The first scientific publication about PCR presented by Mullis et al. three years later (1986). PCR is a biochemical process that amplifies a single or a small number of copies of a piece of DNA sequence across several orders of magnitude. The great importance of PCR is reflected in the fact that PCR was the cornerstone of developing large-scale experiments and sequencing projects making possible to decipher the genetic code of organisms. The representative example is the Human Genome Project, which 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.

After the recent technological advances that made possible the conduction of many large scale experiments, the collection of biological data has been increasing at explosive rates. An important example to perceive the rapidness of this data growth is to consider that the number of transistors on integrated circuits and consequently the processing speed as well as storage capacity of computing hardware doubles approximately every 18 months. This is a very good estimation made by Gordon Moore (1965) and is widely known as Moore's law. However, nowadays Moore's law seems reaching its limits. In contrast, new biological data is doubling approximately every 9 months, and this rate seems to increase dramatically (EMBL, 2013).

BASIC MOLECULAR BIOLOGY CONCEPTS

Diversity is a key property of life and is reflected in the tremendous heterogeneity among living creatures. Surprisingly, the underlying molecular details of organisms are almost universal. All organisms depend on the activities of proteins, a complex family of molecules that comprise the main structural and functional units of cells. The hypothesis of molecular unity of organisms is strengthened by the fact that similar protein sets with similar functions are found in very different organisms. Nucleic acids, namely deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are another family of molecules found in every organism having the role to carry the code of life. Both unity and diversity of living things have been arisen through the force of evolution (Hunter, 2004).

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 these molecules. Because their sequence is usually long, they are both called macromolecules. Sequences of pro-