Chapter LXVI
Machine Learning and Data Mining in Bioinformatics

George Tzanis
Aristotle University of Thessaloniki, Greece

Christos Berberidis
Aristotle University of Thessaloniki, Greece

Ioannis Vlahavas
Aristotle University of Thessaloniki, Greece

INTRODUCTION

Machine learning is one of the oldest subfields of artificial intelligence and is concerned with the design and development of computational systems that can adapt themselves and learn. The most common machine learning algorithms can be either supervised or unsupervised. Supervised learning algorithms generate a function that maps inputs to desired outputs, based on a set of examples with known output (labeled examples). Unsupervised learning algorithms find patterns and relationships over a given set of inputs (unlabeled examples). Other categories of machine learning are semi-supervised learning, where an algorithm uses both labeled and unlabeled examples, and reinforcement learning, where an algorithm learns a policy of how to act given an observation of the world.

Data mining is a more recently emerged field than machine learning is. Traditional data analysis techniques often fail to process large amounts of -often noisy- data efficiently. The scope of data mining is the knowledge discovery 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. The main difference between machine learning and data mining is that machine learning algorithms focus on their effectiveness, whereas data mining algorithms focus on their efficiency and scalability.

Recently, the collection of biological data has been increasing at explosive rates due to improvements of existing technologies as well as the introduction of new ones that made possible the conduction of many large scale experiments. 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. A representative example of the rapid biological data accumulation is the exponential growth of GenBank (Figure 1), the U.S. NIH genetic sequence database (www.ncbi.nlm.nih.gov). The explosive growth in the amount of biological data demands the use of computers for the organization, the maintenance and the analysis of these data. This led to the evolution of bioinformatics, an interdisciplinary field at the intersection of biology, computer science, and information technology. Luscombe et al. (2001) identify the aims of bioinformatics as follows:

- The organization of data in 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.

There is a strong interest in methods of knowledge discovery and data mining to generate models of biological systems. In order to build knowledge discovery systems that contribute to our understanding of biological systems, biological research requires efficient and scalable data mining systems.

BACKGROUND

One of the basic characteristics of life is diversity, which can be noticed by the great differences among living creatures. Despite this diversity, the molecular details underlying living organisms are

Figure 1. Growth of GenBank (1982-2005)