Preface

With the development of molecular biology during the past decades, we are witnessing an exponential growth of both the volume and the complexity of biological data. For example, the Human Genome Project provided the sequence of the 3 billion DNA bases that constitute the human genome. Consequently, we are also provided with the sequences of about 100,000 proteins. Therefore, we are entering the post-genomic era: after having focused so much effort on the accumulation of data, we now need to focus as much effort—and even more—on the analysis of these data. Analyzing this huge volume of data is a challenging task not only because of its complexity and its multiple and numerous correlated factors, but also because of the continuous evolution of our understanding of the biological mechanisms. Classical approaches of biological data analysis are no longer efficient and produce only a very limited amount of information, compared to the numerous and complex biological mechanisms under study. From here comes the necessity to use computer tools and develop new in silico high-performance approaches to support us in the analysis of biological data and, hence, to help us in our understanding of the correlations that exist between, on one hand, structures and functional patterns of biological sequences and, on the other hand, genetic and biochemical mechanisms. Knowledge discovery and data mining (KDD) are a response to these new trends.

Knowledge discovery is a field where we combine techniques from algorithmics, artificial intelligence, mathematics and statistics to deal with the theoretical and practical issues of extracting knowledge, i.e., new concepts or concept relationships, hidden in volumes of raw data. Knowledge discovery offers the capacity to automate complex search and data analysis tasks. We distinguish two types of knowledge discovery systems: verification systems and discovery ones. Verification systems are limited to verifying the user’s hypothesis, while discovery ones autonomously predict and explain new knowledge. The biological knowledge discovery process should take into account both the characteristics of the biological data and the general requirements of the knowledge discovery process.

Data mining is the main phase in the knowledge discovery process. It consists of extracting nuggets of information, i.e., pertinent patterns, patterns correlations, estimations or rules, hidden in bodies of data. The extracted information will be used in the verification of hypothesis or the prediction and explanation of knowledge. Biological data mining aims at extracting motifs, functional sites, or clustering/classification rules from biological sequences.

Biological KDD are complementary to laboratory experiments and help to speed up and deepen research in modern molecular biology. They promise to bring us new insights into the growing volumes of biological data. This workshop contains very interesting papers that deal with biological KDD.

Mourad Elloumi, LaTICE, University of Tunis, Tunisia
Program Committee Chair