Chapter VII
Computer Aided Knowledge Discovery in Biomedicine

Vanathi Gopalakrishnan
University of Pittsburgh, USA

ABSTRACT

This chapter provides a perspective on 3 important collaborative areas in systems biology research. These areas represent biological problems of clinical significance. The first area deals with macromolecular crystallization, which is a crucial step in protein structure determination. The second area deals with proteomic biomarker discovery from high-throughput mass spectral technologies; while the third area is protein structure prediction and complex fold recognition from sequence and prior knowledge of structure properties. For each area, successful case studies are revisited from the perspective of computer-aided knowledge discovery using machine learning and statistical methods. Information about protein sequence, structure, and function is slowly accumulating in standardized forms within databases. Methods are needed to maximize the use of this prior information for prediction and analysis purposes. This chapter provides insights into such methods by which available information in existing databases can be processed and combined with systems biology expertise to expedite biomedical discoveries.

INTRODUCTION

The mission of this chapter is to introduce concepts and terms that form the core of methods devised for important problems in bioinformatics and systems biology applications. Successful case studies are presented that utilize prior knowledge to aid in novel biomedical discoveries. Machine learning techniques are applied to various important biological problems, namely macromolecular crystallization, biomarker discovery from proteomic mass spectra and protein structure prediction via fold recognition. A common theme is the utilization of protein sequence properties and known task-specific information that serve as prior knowledge for guiding knowledge discovery. Much of the task-specific information

is obtained through direct interactions between the bioinformatician and the domain expert in biomedical science.

Systematization of the processes by which biomedical discoveries are made can uncover useful information that can help the bench scientist prioritize and focus efforts. A major goal of this chapter is to describe efforts made toward such systematization in some critical research areas. Recent novel machine learning algorithms have demonstrated some success in identifying and characterizing "interesting relationships" among domain concepts resulting in discovery of explanations for well-known observed scientific phenomena. The domain expert plays a very important role in studying output generated by computer programs and providing input to bioinformaticians on how to focus their subsequent efforts. Thus, communication between multi-disciplines is crucial to successful computer-aided biomedical discovery. For example, modeling protein sequence-structure relationships is a challenging bioinformatics task. Prior knowledge about protein fold can be used to better model protein families containing remote homologs that have very few sequence characters in common between members of the same family. Such knowledge is obtained typically from study of the literature combined with communication with a domain expert.

BACKGROUND

Knowledge discovery in biomedicine in the current world is very often the result of computational analyses combined with interpretation by domain experts. Langley (1998) states that artificial intelligence researchers have tried to develop intelligent artifacts that replicate the act of discovery. There are distinct steps in the scientific discovery process discussed therein (Langley, 1998) during which developers or users can influence the behavior of a computational discovery system. Furthermore, Langley (1998) suggests that such intervention is the preferred approach for using discovery software. In this chapter, we present an approach to data modeling and discovery that is consistent with this viewpoint.

Jurisica and Wigle (2006) define knowledge discovery (KD) as the process of extracting novel, useful, understandable and usable information from large data sets. The authors review knowledge discovery in proteomics and present examples of such algorithms in the literature that aid protein crystallization. The case studies presented in this chapter reflect state-of-the-art challenges in proteomics along with computer-aided solutions. Quantitative and qualitative discoveries are described along with the methods by which they are arrived at. The KD process in complex real-world domains requires multi-disciplinary methods involving both artificial intelligence and statistics applied to databases (Jurisica & Wigle, 2006).

Proteomics can be defined simply as the study of protein composition in a protein complex, organelle, cell or entire organism (Russell, Old, Resing, & Hunter, 2004). Current high-throughput proteomic technologies require robotics and computational techniques to decipher signals within multitudes of data. It is becoming clear that the high dimensionality poses a serious challenge to existing artificial intelligence tools for knowledge discovery and reasoning (Jurisica & Wigle, 2006). The unavailability of large numbers of samples combined with the high dimensionality of the feature space limits the usefulness of models obtained from such data. Moreover, uncertain and missing values in the data combined with evolving knowledge of the underlying mechanisms requires an intelligent information system to be flexible and scalable (Jurisica & Wigle, 2006).