Chapter X
Knowledge Discovery in Biomedical Data Facilitated by Domain Ontologies

Amandeep S. Sidhu
Curtin University of Technology, Australia

Paul J. Kennedy
University of Technology Sydney, Australia

Simeon Simoff
University of of Western Sydney, Australia

Tharam S. Dillon
Curtin University of Technology, Australia

Elizabeth Chang
Curtin University of Technology, Australia

ABSTRACT

In some real-world areas, it is important to enrich the data with external background knowledge so as to provide context and to facilitate pattern recognition. These areas may be described as data rich but knowledge poor. There are two challenges to incorporate this biological knowledge into the data mining cycle: (1) generating the ontologies; and (2) adapting the data mining algorithms to make use of the ontologies. This chapter presents the state-of-the-art in bringing the background ontology knowledge into the pattern recognition task for biomedical data.

INTRODUCTION

Data mining is traditionally conducted in areas where data abounds. In these areas, the task of the data mining is to identify patterns within the data, which may eventually become knowledge. To this end, the data mining methods used, such as cluster analysis, link analysis and classifica-

tion and regression, typically aim to reduce the amount of information (or data) to facilitate this pattern recognition. These methods do not tend to contain (or bring to the problem) specific domain specific information. In this way, they may be termed "knowledge-empty." However, in some real-world areas, it is important to enrich the data with external background knowledge so as to provide context and to facilitate pattern recognition. These areas may be described as data rich but knowledge poor. External background information that may be used to enrich data and to add context information, and facilitate data mining is in the form of ontologies, or structured vocabularies. So long as the original data can be linked to terms in the ontology, the ontology may be used to provide the necessary knowledge to explain the results and even generate new knowledge.

In accelerating quest for disease biomarkers, the use of high-throughput technologies, such as DNA microarrays and proteomics experiments, has produced vast datasets identifying thousands of genes whose expression patterns differ in diseased vs. normal samples. Although many of these differences may reach statistical significance, they are not biologically meaningful. For example, reports of mRNA or protein changes of as little as two-fold are not uncommon, and although some changes of this magnitude turn out to be important, most are attributes to disease-independent differences between the samples. Evidence gleaned from other studies linking genes to disease is helpful, but with such large datasets, a manual literature review is often not practical. The power of these emerging technologies—the ability to quickly generate large sets of data—has challenged current means of evaluating and validating these data. Thus, one important example of a data rich but knowledge poor area is biological sequence mining. In this area, there exist massive quantities of data generated by the data acquisition technologies. The bioinformatics solutions addressing these data are a major current challenge. However, domain specific ontologies such as gene ontology (GO Consortium, 2001), MeSH (Nelson & Schopen, 2004) and protein ontology (Sidhu & Dillon, 2005a, 2006a) exist to provide context to this complex real world data.

There are two challenges to incorporate this biological knowledge into the data mining cycle: (1) generating the ontologies; and (2) adapting the data mining algorithms to make use of the ontologies. This chapter presents the state-of-the-art in bringing the background ontology knowledge into the pattern recognition task for biomedical data. These methods are also applicable to other areas where domain ontologies are available, such as text mining and multimedia and complex data mining.

GENERATING OONTOLOGIES: CASE OF PROTEIN ONTOLOGY

This section is devoted to the practical aspects of generating ontologies. It presents the work on building the protein ontology (Sidhu et al., 2006a; Sidhu & Dillon, 2005a, 2006b; Sidhu et al., 2005b) in the section "Protein Ontology (PO)." It then compares the structures of the protein ontology and the well established gene ontology (GO Consortium, 2001) in the section "Comparing PO and GO."

Protein Ontology (PO)

Advances in technology and the growth of life sciences are generating ever increasing amounts of data. High-throughput techniques are regularly used to capture thousands of data points in an experiment. The results of these experiments normally end up in scientific databases and publications. Although there have been concerted efforts to capture more scientific data in specialist databases, it is generally acknowledged that only 20% of biological knowledge and data is available in a structured format. The remaining 80% of biological information is hidden in the unstruc-