Chapter XIII
Gene Expression Mining Guided by Background Knowledge

Jiří Kléma
Czech Technical University in Prague, Czech Republic

Filip Železný
Czech Technical University in Prague, Czech Republic

Igor Trajkovski
Jožef Stefan Institute, Slovenia

Filip Karel
Czech Technical University in Prague, Czech Republic

Bruno Crémilleux
Université de Caen, France

Jakub Tolar
University of Minnesota, USA

ABSTRACT

This chapter points out the role of genomic background knowledge in gene expression data mining. The authors demonstrate its application in several tasks such as relational descriptive analysis, constraint-based knowledge discovery, feature selection and construction or quantitative association rule mining. The chapter also accentuates diversity of background knowledge. In genomics, it can be stored in formats such as free texts, ontologies, pathways, links among biological entities, and many others. The authors hope that understanding of automated integration of heterogeneous data sources helps researchers to reach compact and transparent as well as biologically valid and plausible results of their gene-expression data analysis.

INTRODUCTION

High-throughput technologies like microrarrays or SAGE are at the center of a revolution in biotechnology, allowing researchers to simultaneously monitor the expression of tens of thousands of genes. However, gene-expression data analysis represents a difficult task as the data usually show an inconveniently low ratio of samples (biological situations) against variables (genes). Datasets are often noisy and they contain a great part of variables irrelevant in the context under consideration. Independent of the platform and the analysis methods used, the result of a gene-expression experiment should be driven, annotated or at least verified against genomic background knowledge (BK).

As an example, let us consider a list of genes found to be differentially expressed in different types of tissues. A common challenge faced by the researchers is to translate such gene lists into a better understanding of the underlying biological phenomena. Manual or semi-automated analysis of large-scale biological data sets typically requires biological experts with vast knowledge of many genes, to decipher the known biology accounting for genes with correlated experimental patterns. The goal is to identify the relevant "functions", or the global cellular activities, at work in the experiment. Experts routinely scan gene expression clusters to see if any of the clusters are explained by a known biological function. Efficient interpretation of this data is challenging because the number and diversity of genes exceed the ability of any single researcher to track the complex relationships hidden in the data sets. However, much of the information relevant to the data is contained in publicly available gene ontologies and annotations. Including this additional data as a direct knowledge source for any algorithmic strategy may greatly facilitate the analysis.

This chapter gives a summary of our recent experience in mining of transcriptomic data. The chapter accentuates the potential of genomic background knowledge stored in various formats such as free texts, ontologies, pathways, links among biological entities, etc. It shows the ways in which heterogeneous background knowledge can be preprocessed and subsequently applied to improve various learning and data mining techniques. In particular, the chapter demonstrates an application of background knowledge in the following tasks:

- Relational descriptive analysis
- Constraint-based knowledge discovery
- Feature selection and construction (and its impact on classification accuracy)
- Quantitative association rule mining

The chapter starts with an overview of genomic datasets and accompanying background knowledge analyzed in the text. Section on relational descriptive analysis presents a method to identify groups of differentially expressed genes that have functional similarity in background knowledge. Section on genomic classification focuses on methods helping to increase accuracy and understandability of classifiers by incorporation of background knowledge into the learning process. Section on constraint-based knowledge discovery presents and discusses several background knowledge representations enabling effective mining of meaningful over-expression patterns representing intrinsic associations among genes and biological situations. Section on association rule mining briefly introduces a quantitative algorithm suitable for real-valued expression data and demonstrates utilization of background knowledge for pruning of its output ruleset. Conclusion summarizes the chapter content and gives our future plans in further integration of the presented techniques.