Chapter 3.23
Data Mining in Gene Expression Data Analysis:
A Survey

Jilin Han
University of Oklahoma, USA

Le Gruenwald
University of Oklahoma, USA

Tyrrell Conway
University of Oklahoma, USA

ABSTRACT

The study of gene expression levels under defined experimental conditions is an important approach to understand how a living cell works. High-throughput microarray technology is a very powerful tool for simultaneously studying thousands of genes in a single experiment. This revolutionary technology results in an extensive amount of data, which raises an important question: how to extract meaningful biological information from these data? In this chapter, we survey data mining techniques that have been used for clustering, classification and association rules for gene expression data analysis. In addition, we provide a comprehensive list of currently available commercial and academic data mining software together with their features. Lastly, we suggest future research directions.

INTRODUCTION

Recently, bioinformatics has attracted a lot of attention from biologists and computer scientists. One of the most important aspects of bioinformatics is the application of data mining tools to extract meaningful biological information from gene expression data. The study of gene expression levels at a given time or under established conditions is an important approach to understand how a living cell works (Vingron & Hoheisel, 1999). High-throughput microarray technology (Ramsay, 1998; Harrington, Rosenow, & Retief, 2000; Lipshutz et al., 2000; Jordan, 2001) is a powerful tool for simultaneously studying thousands of genes in a single experiment. This revolutionary technology results in an extensive amount of data, which raises a challenging question: how can meaningful biological information be extracted from these data? Important biological information associated with these data may not be discovered or may be misinterpreted due to lack of appropriate and effective data analysis tools and techniques.

Data mining is one of the most important and difficult tasks in gene expression data analysis. Data mining typically includes clustering, classification, and association rule discovery (Lin & Johnson, 2002; Wei, 2002; Johnson & Wichern, 1998; Mirkin, 1996). With extensive microarray data available, clustering can be used to identify genes that are co-regulated in a similar manner under different experimental conditions. Classification provides a way to identify the differences between tissue types such as between normal cells and cancer cells, which facilitates diagnosis of diseases. Discovery of association rules can help biologists to identify genes that govern the expression of other genes in regulatory pathways.

This chapter is organized as follows. First, a background on biology and microarray technology is presented. Then, we discuss the gene expression data characteristics and presentation. Following are reviews of the existing clustering, classification and association rules mining algorithms that have been applied to gene expression analysis. A comprehensive list of available commercial and open source data mining software with their features is then presented. Lastly, we suggest directions for future work.

BIOLOGY BACKGROUND AND MICROARRAY TECHNOLOGY

It is a challenging task for biologists to understand how genes and their products function, interact, and most importantly, cause an organism to function the way it does. Functional genomics plays an important role in accomplishing this task. The goal of functional genomics is to reveal the biological functions of an individual gene and its cooperative roles on a genome-wide scale. Microarray technology, such as the cDNA (Schena et al., 1995) and oligonucleotide microarray (Lipshutz et al., 1999) has emerged as a powerful tool to provide meaningful information about gene expression levels for entire genomes. To help users understand microarray experiments, in this section, we briefly introduce a background of basic molecular biology.

A deoxyribonucleic acid molecule (DNA) is a double-stranded polymer composed of four component "building blocks", called nucleotides. Each nucleotide consists of a phosphate group, a deoxyribose sugar, and a purine or pyrimidine base. The four different bases found in a DNA molecule are the purines, adenine (A) and guanine (G), and the pyrimidines, cytosine (C) and thymine (T). The two DNA strands are linked together by hydrogen bonds between purine and pyrimidine bases, with G always pairing with C, and A always pairing with T. A ribonucleic acid molecule (RNA) has the same general structure as DNA, except that uracil (U) replaces thymine (T), ribose replaces deoxyribose, and the RNA molecule is single stranded with extensive secondary structure, i.e., folds and loops.

A gene is a sequence of DNA containing genetic information that codes for a particular protein. A protein is polymer of twenty different types of amino acids in a sequence that is unique to the particular gene. The construction of the amino acid sequence of a protein comes from the genetic information stored in a DNA molecule (more precisely, a gene) and occurs in three stages (see Figure 1): (1) transcription, in which a section of DNA is transcribed into a single-stranded complementary copy of the DNA termed messenger ribonucleic acid (mRNA); (2) splicing, which occurs only in eukaryotes and removes certain stretches of the mRNA, called introns, leaving the remaining parts, called exons, that are then linked together to form the mature mRNA, and (3) translation, in which the nucleotide sequence of mRNA is translated on intracellular particles called ribosomes to produce a protein. The process