Pattern Differentiations and Formulations for Heterogeneous Genomic Data through Hybrid Approaches

Arpad Kelemen, The State University of New York at Buffalo, USA & Niagara University, USA

Yulan Liang, The State University of New York at Buffalo, USA

Abstract

Pattern differentiations and formulations are two main research tracks for heterogeneous genomic data pattern analysis. In this chapter, we develop hybrid methods to tackle the major challenges of power and reproducibility of the dynamic differential gene temporal patterns. The significant differentially expressed genes are selected not only from significant statistical analysis of microarrays but also supergenes resulting from singular value decomposition for extracting the gene components which can maximize the total predictor variability. Furthermore, hybrid clustering methods are developed based on resulting profiles from several clustering methods. We demonstrate the developed hybrid analysis through an application to a time course gene expression data from interferon-β-1a treated multiple sclerosis patients. The resulting integrated-condensed clusters and overrepresented gene lists demonstrate that the hybrid methods can successfully be applied. The post analysis includes function analysis and pathway discovery to validate the findings of the hybrid methods.

Copyright © 2006, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

Progress in mapping the human genome and developments in microarray technologies have provided considerable amount of information for delineating the roles of genes in disease states. Since complex diseases typically involve multiple intercorrelated genetic and environmental factors that interact in a hierarchical fashion and the clinical characteristics of diseases are determined by a network of interrelated biological traits, microarrays hold tremendous latent information but their analysis is still a bottleneck. Pattern analysis can be useful for discovering the knowledge on gene array data related to certain diseases (Neal et al., 2000; Slonim, 2002). The associations between patterns and their causes are the bricks from which the wall of biological knowledge and medical decisions are built.

Pattern differentiations and pattern formulations are two major tracks of patterns analysis. Pattern differentiation of gene expressions is the first step to identify potential relevant genes in biological processes. The coordinated/temporal gene arrays are widely used for pattern formulation in order to study the common functionalities, co-regulations, and pathways that ultimately are responsible for the observed patterns. The identification of groups of genes with “similar” temporal patterns of expression is usually a critical step in the analysis of kinetic data because it provides insights into the gene-gene interactions and thereby facilitates the testing and development of mechanistic models for the regulation of the underlying biological processes. These temporal pattern analyses provide clues for genes that are related in their expression through linkage in a common developmental pathway.

There are several critical challenges in the pattern analyses. One is in the pattern differentiations, the notorious “large p small n” problem (West, 2000). The large number of irrelevant and redundant genes with high level noise measurements and uncertainty severely degrade both classification and prediction accuracy. The solution for the “large p” problem is through affine transformation and feature selection. Affine transformation such as principal component analysis (PCA) or singular value decomposition (SVD) has advantage of simplicity and it may remove non-discriminating and irrelevant features (i.e., genes) by extracting eigenfeatures corresponding to the large eigenvalues (Alter, Brown, & Botstein, 2000; Yeung and Ruzzo, 2001; Wall, Dyck, & Brettin, 2001). Yet it is very difficult to identify important genes with these methods and the inherent linear nature is their prominent disadvantage.

Feature selection consists of two strategies: screening and wrappers. In the screening approaches, all genes are analyzed and tested individually to see whether they have higher expression level in one class than in the other (Baldi & Long 2001; Tusher, Tibshirani, & Chu, 2001; Storey & Tibshirani 2003; Hastiel et al., 2000). The disadvantage of screening processes is that they are non-invertible and can cause multiple testing and model selection problems (Westfall & Young, 1993; Benjamini & Hochberg, 2002).

In wrapper methods, genes are tested not independently, but as ensembles, and according to their performance in the classification model (Golub et al., 1999; Khan et al., 2001; Wuju & Momiao, 2002). Since the number of feature subsets increases exponentially with the dimensions of the feature space, wrappers are computationally intractable for high-dimensional gene data.