Abstract

A challenge in bioinformatics is to analyze volumes of gene expression data generated through microarray experiments and yield useful information. Consequently, most microarray studies demands complex data analysis to infer biologically meaningful information from such high throughput data. Selection of informative genes is an important data analysis step to identify a set of genes which can further help in finding the biological information embedded in microarray data, and thus assists in diagnosis, prognosis and treatment of the disease.

The goal of the thesis is to perform gene subset selection, to generate informative signature genes that characterize diseases, and provide insight into underlying biological processes. Our work focuses on extending existing gene selection methodologies by hybridizing statistical, ontology based and network based approaches. The goal is to address the challenges of multiview representation, improved biological interpretation, and study of related biological processes.

The contributions of the thesis are as follows: (i) to obtain multiview clustering of microarray data sets, we suggest an unsupervised feature selection technique, thereby facilitating gene subset selection which are informative genes, not just discriminatory, but biologically enriched genes responsible for concerned biological processes, and may be multi-faceted by nature, (ii) integrate multiview clusters of gene expression data, with the known protein–protein interaction (PPI) knowledge (iii) integrate expression analysis with structural analysis of gene interaction networks, generating dense subgraphs of gene networks, which are functionally associated, and (iv) a data analysis methodology is developed for identification and visualization of co-expressed gene patterns, as emerging clusters, in global transcriptome of epithelial cancer pre-malignant and malignant conditions in comparison to their normal counterparts. It provides an intuitive understanding of molecular course in carcinogenesis and may contribute for combinatorial biomarker discovery.

Experimental results have been furnished to demonstrate the usefulness and efficiency of the proposed techniques based on integration of prior knowledge of the known gene network with transcriptome data for interaction based gene selection. For the proposed graph based multiview gene selection algorithm (GUFS) it is observed that the method can select a small gene subset that provides satisfactory performance in terms of clustering and is able to identify the subset of genes that are biologically significant or correlated. Augmenting the views obtained from GUFS with protein–protein information (PPI) network knowledge improves performance in terms of classification accuracy and is able to identify gene subsets that are biologically relevant and functionally enriched. It is observed that our proposed method for gene selection, combining expression analysis with structural analysis of protein-protein interaction (PPI) networks helps to identify not just differentially expressed genes but also hub genes important in biological processes. This may result in functional prediction of certain genes in disease susceptibility, predicting disease outcome and identifying important therapeutic strategies. The proposed, self-organizing map (SOM) analysis and visualization illustrates genome wide transcriptional changes in specific stage of studied pathobiology and depicts entire process of metabolic changes from normal to cancer progression.

The analysis revealed *six* common biological processes related to pathogenesis of oral and cervical cancers, this information may contribute for combinatorial biomarker discovery. The investigation shows that integration of prior knowledge of the known gene network with transcriptome data for interaction based gene selection or functional module selection provides better biological interpretation and improved statistical analysis.

Keywords

Microarray data, gene selection, clustering, protein-protein network, self organizing map.