Abstract

A biomarker is a quantifiable characteristic or substance that may be measured and used to characterize a biological process, a disease state, or a response to a therapeutic intervention. A biomarker's importance in a critical disease based depends on its ability to provide useful information to assist with the diagnosis, prognosis, monitoring, or treatment of the disease. Genes play a crucial role in biological processes and have the ability to provide relevant information about health and diseases, making them highly significant as biomarkers. This dissertation presents a body of works for identification of genes as biomarkers for detection of critical diseases.

Analysis of gene expression data is essential for identifying genes as biomarkers. Biclustering, Differential Expression analysis (DEA), and Differential Co-expression analysis (DCA) are three approaches for analyzing gene expression data. In the first work we propose a biclustering framework, BicGenesis, that identifies topologically significant genes of the biclusters as potential biomarker candidates. To generate these biclusters, BicGenesis employs eight well-known biclustering algorithms and use a selection criteria to identify relevant biclusters for subsequent analysis. In the second work we propose an Integrative DEA framework that comprehensively identifies topologically and biologically relevant Differentially Expressed Genes (DEGs) as potential biomarker candidates. Based on the type of gene expression data the integrative DEA framework employs three chosen microarray or bulk RNA-Seq DEA methods to detect DEGs. To identify relevant DEGs, we implement a consensus-based function that takes into account the DEGs common to the three chosen methods as well as two parameters, *q-value* and *lFDR*, to compensate for information loss. Our third work introduces a DCA framework, CBDCEM that identifies the set of hub-genes of relevant modules extracted from co-expression networks (CENs) constructed under two varying conditions as potential biomarker candidates. As finding the hub gene for relevant differentially co-expressed modules is a key function of DCA we develop a centrality-based hub-gene finding method that identifies hub-genes using seven centrality measures. For validation of our frameworks we chose Esophageal Squamous Cell Carcinoma (ESCC) as the critical disease of interest. Two microarray datasets and one bulk RNA Sequencing dataset for ESCC are chosen. We validate BicGenesis, the Integrative DEA approach and CBDCEM on these three datasets.

The ability to understand the interplay between intrinsic cellular processes and behavioral changes in gene-gene interactions under varying conditions is made possible by single-cell RNA sequencing (scRNA-Seq) technology. However, the scRNA-seq data's high level of sparsity and massive size presents a considerable analytical hurdle. For the purpose of extracting relevant network modules from scRNA-Seq data and identifying subsequent hub-genes as potential biomarker candidates, we develop a comprehensive DCA method, scDiffCoAM, as our final work. By incorporating network appropriate measures, scDiffCoAM implements the hub-gene finding method introduced by CBD-CEM effectively. We validate scDiffCoAM on an scRNA-Seq ESCC dataset.

Experimentation and validation of all the four frameworks on the corresponding datasets are found satisfactory. For each framework, by employing various subsets of methods significant to the framework, we identify topologically significant genes and identify them as biomarker candidate genes (BCGs) and biologically validate them as potential biomarkers. We present a set of biomarker criterion that takes into account the literature evidence that associate the BCGs to ESCC and five other associated diseases, Gene Ontology (GO) enrichment and pathway enrichment of the BCG, and the regulatory behavior exhibited by the BCG in a gene regulatory network (GRN). All four frameworks are efficient in detecting a size-able set of potential biomarkers for ESCC. Seven potential biomarkers , *CAV1, MCM7, KPNA2, DGKA, EPHA2, SEMA3B*, and *FGF5* for ESCC are identified by at least two of our frameworks. By implementing a ranking scheme that takes into consideration the literature evidence as well as the biological evidence presented by each individual biomarker, we rank all potential biomarkers detected by all the four frameworks. We identify eleven high ranked potential biomarkers, namely, *CAV1, MCM7, E2F1, KPNA2, DGKA, EPHA2, EGFR, HIF1A, HSF1, NOTCH3*, and *SEMA3B* as most likely potential biomarkers for ESCC.

Keywords: *Gene Expression Data Analysis, Biclustering, Differential Expression Analysis, Differential Co-expression Analysis, Preservation Analysis, Microarray Data, Bulk RNA Sequencing Data, Single-cell RNA Sequencing Data, Esophageal Squamous Cell Carcinoma, Bioinformatics, Machine Learning.*