Abstract

This thesis presents the importance of gene expression data in understanding diseases and the need for computational methods to analyze and interpret large datasets. It highlights the emerging fields of computational biology and bioinformatics and their role in analyzing biological data. The increasing availability of publicly accessible medical databases allows researchers and individuals to extract useful information. Further research is needed to explore the correlations between genes in gene expression datasets and validate hypotheses. The main focus of the thesis is on identifying crucial genes and potential biomarkers for ESCC through the use of gene expression datasets and various data mining analysis techniques.

The method called PD_BiBIM is developed to overcome the limitations of existing biclustering techniques by integrating gene-based analysis to identify potential biomarkers and relationships among genes associated with disease progression. Biclustering results are used to construct biological networks, and further topological, pathway, and causal analysis is conducted on the extracted modules from these networks. Through this comprehensive approach several interesting biomarkers, including IFNGR1, CLIC1, CDK4, and COPS5, have been identified using this method. Recently, apart from microarray data there have been a growing interest in the analysis of Bulk RNA-seq and scRNA-seq transcriptomic data in gene expression research due to its rich and informative content. In the analysis of bulk RNA-seq and scRNA-seq data, potential biomarkers can be identified using several data mining techniques. These include differential expression analysis, co-expression analysis, differential co-expression analysis and topological analysis. These analyses are followed by clustering and validation processes to gain insights into the biological significance of the identified biomarkers. In order to find differentially expressed genes using an ensemble approach to support handling of multiple sources of data, two different frameworks are developed for bulk RNA-seq data and one framework is developed to compare the divergence among sample types towards identification of prognostic biomarkers in ESCC for scRNA-seq data. In order to enhance the analysis of gene expression data, a modified measure called SNMRS has been developed. This measure is capable of

simultaneously identifying positive and negative patterns of shifting, scaling, and shifting-and-scaling, is employed in a module-detection technique. By implementing this technique, the study aims to identify significant and enriched genes within the dataset, leading to the discovery of potential biomarkers. Furthermore, this approach provides valuable insights into the biological relevance and implications of these biomarkers.

Keywords: Gene Expression Data, Biomarker, Differential Expression Analysis, Co-expression Analysis, Differential Co-expression Analysis, Similarity Measure, Clustering, Biclustering