Disease Biomarker Identification for Esophageal Squamous Cell Carcinoma using Gene Expression Data Analysis

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

by

Pallabi Patowary

Enrollment No. CSP17109

Registration No. TZ166819 of 2016

Department of Computer Science and Engineering

School of Engineering, Tezpur University

Tezpur, Assam, India - 784028

July, 2023

Chapter 7

Conclusions and Future Works

7.1 Concluding Remarks

The amount of messenger RNA (mRNA) produced by a gene in a particular situation is represented by its expression profile. With the advent of newer transcriptomic technologies, it has become possible to measure the gene expressions inside millions of cells in a single experiment. A biologist generates the biological data whereas a computational expert has the job of mining the information hidden inside the high dimensional biological data using efficient algorithms. Gene expression data such as DNA microarray, RNA-seq, and scRNA-seq are high dimensional. Data analysis in biological research is performed in a systematic way and the validation of the final findings is an absolute necessity. Gene expression data analysis is a systematic approach to understand the functions of and interactions among genes. My research deals with identifying potential biomarkers from gene expression data. The obtained biomarkers are found to be associated with ESCC disease progression. Following are the few observations to the area on gene expression data analysis during the course of my PhD work.

1. In Chapter 3, a framework is proposed to identify most crucial genes in ESCC microarray gene expression datasets using an ensemble approach of biclustering methods. For this work, GCNs for each validated biclusters

have been constructed to support topological and other gene-gene associative analysis. Here, prior knowledge of ESCC biomarkers called primary genes has been applied based on which secondary genes are obtained. It is found that secondary genes follow common genetic pathways with the primary genes in the context of ESCC. A ranking scheme is performed to list the secondary genes as the crucial genes and these crucial genes have been validated using GRN analysis, GO enrichment analysis, pathway analysis and literature evidences.

2. Chapter 4 presents two individual methods of integrative analysis which are used to identify a set of most responsible genes that may cause the progression of ESCC disease. In the first method, only RNA-seq ESCC gene expression data is considered and identified a set of crucial genes using DEA supported by gene enrichment analysis. The second method, an ensemble approach is proposed supported by an effective consensus function to find an unbiased set of DEGs by multiple DEG finding tools. These DEGs are considered as input for GCN construction and for subsequent downstream analysis. GCN analysis is carried out to extract modules from the co-expression network to perform module preservation analysis. The low preserved modules are further investigated using topological, pathway, GO enrichment analysis and in light of relevant literature evidence to identify an interesting set of biomarkers for the microarray data and RNA-seq data. Finally, interesting biomarkers are identified.

3. Chapter 5 reports a gene module extraction technique based on an effective similarity measure called SNMRS. This method has been used over ESCC microarray as well as RNA-seq datasets. It is observed that the proposed measure provides better performance than several other measures in terms of cluster-validity indices. Co-expression networks are constructed using SNMRS that handles all types of correlation followed by extraction of network modules from the network applying average linkage clustering algorithm.

This SNMRS based module detection method results in interesting biologically relevant patterns from gene microarray and RNA-seq dataset. Novel biomarkers are identified which are associated with the progression of the disease ESCC and these biomarkers have been validated using GO enrichment analysis, pathway analysis and literature sources.

4. Chapter 6 discusses an effective pipeline to investigate cell heterogeneity and to identify potential biomarkers in single-cell RNA sequencing data using DEA. Several biological networks are constructed and observed that relative to normal (0Gy), 30Gy induced higher variability of genes interms of co-expression, correlation, regulatory network, transcription factors and cancer related genes. The expression of most cancer related genes show high correlations among these three groups. A set of genes involved in ESCC progression are also identified and validated using GO enrichment analysis, pathway analysis and literature sources. An analysis of the hub genes found in modules is also reported here.

7.2 Future Works

In conclusion, bioinformatics and computational biology presents an excellent opportunity for computer engineering background researchers to apply their skills and knowledge to enable new discoveries. As the field continues to grow, the potential for new discoveries and implications is increasing. The work presented in this thesis can be categorised as- data mining techniques in the field of bioinformatics and computational biology and in particulars gene expression data analysis which itself has tremendous potential in future research. During my tenure of PhD work I have come across many challeges which can be converted into a research problems, as follows-

1. My Ph.D work is focused on ESCC disease. A few benchmark microarray, RNA-seq, and scRNA-seq dataset for ESCC are available in public repos-

itories and that too with low dimensions. In future, we will extend our work which can use high or ultra-high dimensional GED specially scRNA-seq datasets. The proposed methods will be implemented in other deadly diseases like gall bladder cancer (GBC) and breast cancer (BC).

2. For a large datasets co-expression network construction using proposed measure SNMRS is time consuming. Parallel computing of SNMRS measure can ease the whole process.

3. Gene expression data analysis towards potential biomarker identification using deep learning technique is another future research direction.

4. A framework will be implemented to handle various sources of data at a time (e.g., micrroarray, miRNA, RNAseq, and scRNA-seq data) towards identification of interesting biomarkers through consensus building for a given disease.

5. We are working on the development of a robust soft computing enabled outlier based DEG finding method to substitute the present ensemble approach.

6. A multi-objective biomarker ranking approach is also a future work.