Dedicated to

Maa

Late Lilavati Patowary

&

Deuta

Mr. Kumud Chandra Patowary

Declaration

I, Pallabi Patowary, hereby declare that the thesis entitled *"Disease Biomarker Identification for Esophageal Squamous Cell Carcinoma using Gene Expression Data Analysis"* submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy is based on bonafide work carried out by me. The results embodied in this thesis have not been submitted in part or in full, to any other university or institute for award of any degree or diploma.

(Pallabi Patowary)

Tezpur University

Certificate

This is to certify that the thesis entitled *"Disease Biomarker Identification for Esophageal Squamous Cell Carcinoma using Gene Expression Data Analysis"* submitted to the Department of Computer Science and Engineering under the School of Engineering, Tezpur University in partial fulfilment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by *Mrs. Pallabi Patowary* under my personal supervision and guidance.

All helps received by her from various sources have been duly acknowledged. No part of this thesis has been submitted elsewhere for award of any other degree.

Signature of Supervisor

(Dhruba Kumar Bhattacharyya)
Designation: Professor
School: Engineering
Department: Computer Science and Engineering
Tezpur University
Assam, India-784028

Certificate

This is to certify that the thesis entitled *"Disease Biomarker Identification for Esophageal Squamous Cell Carcinoma using Gene Expression Data Analysis"* submitted by *Mrs. Pallabi Patowary* to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering has been examined by us on .... and found to be satisfactory.

The Committee recommends for award of the degree of Doctor of Philosophy.

Signature of Principal Supervisor
Date:

Signature of External Examiner
Date:

Acknowledgment

"Our greatest glory is not in never falling, but in rising every time we fall."

It will be injustice to the success of this thesis if due acknowledgement is not expressed to the people who stood by me through the thick and thin of this journey. First and Foremost, I would like to thank my supervisor Prof. Dhruba Kumar Bhattacharrya for his constant support, trust, valuable feedback, encouragement and innumerable advices. He gave freedom to pursue my ideas and work at my own pace, and was always available to discuss various problems on the way. His encouragement and guidance have provided a good basis for completion of my research work.

With great pleasure I would like to express my gratitude to Prof. Sarat Saharia, Prof. Utpal Sharma, Dr. Rosy Sarmah, and Dr. Pankaj Barah who were part of my doctoral committee at Tezpur University for their valuable suggestions and guidance. I would also like to thank Prof. Jugal Kumar Kalita, University of Colorado, USA, for all his help, support and valuable guidance. I extend my sincere thanks to other faculty members specially Prof. Bhogeswar Borah, Prof. Smriti Kumar Sinha. Dr. Sanjib Deka and Dr. Arindam Karmakar and the non-teaching staff of the Department, for their generous help in various ways towards the completion of my research work.

My hard work alone would not have resulted this success without the support from my husband (Dr. Madhurjya Pratim Das) and son (Rutvik Taksheel Das).

Guidance and support from seniors, friends and other Departments of Tezpur University helped me travelled this painful journey comfortably. I would like to acknowledge my sincere gratitude to all of them specially to Prathana, Munmi, Trishna, Satya, Hussain Da, Partha, Upasana, Koyel Di, Manaswita Ba, and Nabonita.

Special thanks to all my family members for their blessings, love, and wishes that sustained me this far. A very special thanks to Dada, Bow, Dangor Baa, Biju Baa and Saru Baa. Finally, I would like to thank all those who have directly or indirectly helped me in different capacities to complete my research work.

Last but not the least, I thank the almighty for everything.

Pallabi Patowary

List of Figures

1-1 Structures of DNA and RNA (Credit: Technology Networks, https://www.technologynetworks.com/genomics/lists/, accessed on 21/07/2023) 2

1-2 Central Dogma 3

1-3 An example of Gene Expression Dataset 3

1-4 Screenshot of a micoarray dataset 4

1-5 Screenshot of a RNA-seq count data 6

1-6 In different types of cells within the same tissue, a gene can be expressed at different levels and regulated by different control mechanisms. 6

1-7 Screenshot of a scRNA-seq count data 7

2-1 Schematic representation of a Gene. Source: File:Gene.png. In Wikipedia. https://commons.wikimedia.org/wiki/File:Gene.png, accessed on 21/07/2023 16

2-2 Schematic representation of Gene-expression. Source: https://www.genome.gov/genetics-glossary/, accessed on 21/07/2023 17

2-3 Schematic representation of DNA microarray technology. 19

2-4 RNA-seq data generation steps 20

2-5 Schematic representation of RNA-seq technology. Source: https://en.wikipedia.org/wiki/RNA-Seq, accessed on 21/07/2023 . . 21

2-6 Schematic representation of scRNA-seq technology. Source:https://www.stephaniehicks.com/2018-bioinfosummer-scrnaseq/introduction-to-single-cell-rna-seq.html, accessed on 21/07/2023 22

2-7 An example of differential expression pattern. The gene expression of G1 varies (increases) significantly in disease conditions. Whereas, gene G2 shows almost steady expression in disease condition. So, gene G1 may be considered as a differentially expressed gene. In this figure, first 5 samples are from normal condition and rest are from disease condition. 29

2-8 An example of co-expressed module. 31

2-9 Types of GCNs and types of co-expression measures and topological properties 31

2-10 Different types of patterns with change. Here, 'a' is a random variable, 'b' and 'c' exhibit shifted pattern, d exhibits a scaling pattern, and e inhibits a shifting-and-scaling pattern. 34

2-11 Clusters 38

2-12 Biclusters 40

3-1 PD_BiBIM framework: This framework consists of five modules - M1, M2, M3, M4, and M5. 58

3-2 The different patterns of gene expression profiles of primary genes across all the normal samples vs tumor samples for dataset GSE20347 (a) RUNX3, (b) CDH1, (c) VIM, (d) WWOX, (e) CTTN, and (f) CCND1. 60

3-3 Variation of connectivities for each secondary gene found in matched Normal (N) vs Tumor (T) samples for the dataset (a) GSE20347, (b) GSE23400. In most cases, secondary gene's connectivities are higher in T sample than that of N sample. 66

3-4 Biological interactions between primary and secondary genes 67

3-5 PPI network among primary and secondary genes 69

4-1 Conceptual Framework for identification of crucial genes for ESCC RNA-seq data 82

4-2 Co-expression network of the suspected eight genes across the states (i.e. Normal and disease). The total number of edge connectivity is 5. 86

4-3 Co-expression network of the suspected eight genes across the disease condition. The total number of edge connectivity is 13. 87

4-4 PPI network of the suspected eight genes 89

4-5 Conceptual framework to find critical genes from ESCC datasets. . 93

4-6 Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs identified by three methods at P-value cutoff .01 for (a) GSE20347 (b) GSE23400 97

4-7 Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs identified by three methods for RNA-seq data SRP064894 P-value cutoff .01 98

4-8 Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs at P-value cutoff .001 identified by three methods for microarray data (a) GSE20347 (b) GSE23400 98

4-9 Venn diagrams to demonstrate the overlapping relationships among the sets of DEGs at P-value cutoff .001 identified by three methods for RNA-seq data SRP064894 99

4-10 GCN analysis of DE genes for microarray data GSE20347 for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM). 100

4-11 GCN analysis of DE genes for microarray data GSE23400 for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM). 101

4-12 GCN analysis of DE genes for RNA-seq data for normal sample [a]-[c] and for tumor sample [d]-[f]. [a],[d] Determination of soft-thresholding power, β by analysing the scale-free fit index for various β. [b],[e] Branches of hierarchical clustering of module eigngenes are grouped together based on positive correlation. [c],[f] Clustering dendogram of genes based on a dissimilarity measure (1-TOM). . . 102

4-13 Preservation analyses: Test-Normal, Reference-Tumor([a], [b], [c]); Test-Tumor, Reference-Normal([d], [e], [f]). [a] GSE20347: Here, the white module is the least preserved module. [b] RNA-seq data: Here, the steelblue module is the least preserved module. [c] GSE23400: Here, the purple module is the least preserved module. [d] GSE23400: Here, the greenyellow module is the least preserved module. [e] GSE20347: Here, the darkgreen module is the least preserved module. [f] RNA-seq data: Here, the dark red module is the least preserved module. 104

4-14 PPI network of low preserved modules constructed in STRING . . . 106

4-15 PPI network of low preserved modules constructed in STRING . . . 107

4-16 PPI network of low preserved modules constructed in STRING . . . 108

5-1 The workflow of module finding using SNMRS measure [DEA: Differential Expression Analysis] 131

5-2 *Synthetic data - D1*: The artificial gene patterns for analysis of different similarity measures. 134

5-3 *Synthetic data - D2*: The artificial gene patterns for analysis of different similarity measures. 134

5-4 Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns ‘b1-b8’ with that of ‘a’ 135

5-5 Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns ‘x1-x7’ with that of ‘x’ 136

5-6 Correlation values obtained from NMRS, NCNMRS, SPEARMAN, KENDALL, MI, and SNMRS while applied on example patterns ‘p1-p7’ with that of ‘p’ 139

5-7 Comparative study of node (Y-axis) and connectivity (Y-axis) at different correlation scores (X-axis) to analyse the threshold value. Effect diagram of the node after threshold 0.8 is increased. Here, (a) Node vs SNMRS for Yeast sporulation dataset, (b) Connectivity vs SNMRS for Yeast sporulation dataset, (c) Node vs SNMRS for ESCC normal dataset, (d) Node vs SNMRS for ESCC normal dataset, (e) Node vs SNMRS for ESCC tumor dataset, and (f) Node vs SNMRS for ESCC tumor dataset. 142

5-8 Visualization of extracted module (salmon) for ESCC normal 146

5-9 Visualization of extracted module (tan) for ESCC tumor 146

5-10 The artificial gene patterns and their scatter plot of Synthetic data - D1: (a) a vs b1, (b) a vs b2, (c) a vs b3, (d) a vs b4, (e) a vs b5, (f) a vs b6, (g) a vs b7, (h) a vs b8 150

5-11 Network obtained from STRING tool for module (Yellow) from ESCC normal 154

6-1 Framework of the proposed method. 160

6-2 Preprocessing of ESCC data. (a) Measurement of library size for genes per cell, (b) Measurement of library size for counts per genes, (c) Quality control, (d) Variable genes in ESCC dataset, (e) DimHeatMap for fifteen principal components, (f) Visualization of strong enrichment of features with low p-values. 164

6-3 (a) UMAP clustering of the dataset, (b) t-SNE clustering of the dataset, (c) Genes distributed based on the percentage of mitochondria, (d) Gene clustering and distribution in all sample-type. . 165

6-4 Venn diagram of differentially expressed genes found using consensus method in sample (a) 0-Gy, (b) 12-Gy, (c) 30-Gy. 166

6-5 Network diagram for KEGG for all the DEGs in (a) O-Gy (b) 12-Gy (c) 30-Gy 168

6-6 Enrichment analysis of DEGs in (a) O-Gy (b) 12-Gy (c) 30-Gy. . . 169

6-7 GRN and hub genes of their subnetworks obtained by cytoHubba for (a) 0-Gy (b) 12-Gy (c) 30-Gy 172

List of Tables

2.1 Differential Gene Expression Analysis Methods/Tools 30

2.2 List of available GO Enrichment Analysis Tools 50

3.1 A brief description of datasets used for evaluation of proposed framework for ESCC disease 55

3.2 Primary Genes associated with ESCC 56

3.3 Descriptions of Primary Genes associated with ESCC based on existing literature 57

3.4 P-values of identified biclusters with primary genes in normal and tumor samples for dataset GSE20347. Note: SN: Serial Number, G: Gene, S: Sample 62

3.5 P-values of identified biclusters with primary genes for normal and tumor samples for dataset GSE23400. Note: SN: Serial Number, G: Gene, S: Sample 63

3.6 Number of Biclusters extracted by different reported biclustering methods for GSE20347 and GSE23400 datasets. Note: T: Tumor sample, N: Adjacent normal sample. 64

3.7 Degree analysis of primary genes and their respective hub genes (secondary genes) for the dataset GSE20347 along with association type (Direct (D) or Indirect (I)) between the primary gene and hub genes are also mentioned. Note: P: Primary Gene, N: Normal sample, T: Tumor sample, NF: Not found, AT(H,T): Association Type between H and T 65

3.8 Degree analysis of primary genes and their respective hub genes (secondary genes) for the dataset GSE23400 along with association type (Direct (D) or Indirect (I)) between the primary gene and hub genes are also mentioned. Note: P: Primary Gene, N: Normal sample, T: Tumor sample, NF: Not found, AT(H,T): Association Type between H and T 65

3.9 Common Pathway shared by Primary and Secondary genes 70

3.10 List of Primary and Secondary genes associated with Cancer Disease 71

3.11 GO enrichment analysis of primary and secondary genes. Note: GC: GOterm Category 72

3.12 GRN network statistics for normal sample 72

3.13 GRN network statistics for tumor sample 73

3.14 Biomarker ranking: Secondary genes are ranked based on their average score. Note: 1 means yes, 0 means No; S: Secondary gene, P: Primary gene. 76

4.1 Top 10 Up-regulated DEGs, ranked by *log2FoldChange* value 83

4.2 Top 10 Down-regulated DEGs, ranked by *log2FoldChange* value . . 83

4.3 Top 100 Up and Down-regulated DEGs, ranked by *padj* value, Bold gene: common gene 84

4.4 Topological statistics of GRN among 3 regulatory gene and 8 suspected genes in normal state 87

4.5 Topological statistics of GRN among 3 regulatory gene and 8 suspected genes in disease state 88

4.6 Dataset Used. Note: T: Tumor, N:Normal 92

4.7 DEA statistics for GSE20347 96

4.8 DEA statistics for GSE23400 96

4.9 DEA statistics for SRP064894 97

4.10 Percentage of common gene detected by DEA methods 97

4.11 Top hub genes identified by WGCNA in each non-preserved module 105

4.12 Top hub genes identified from PPI network by STRING tool in each non-preserved module. Note: CG: Central Gene, CS: Confidence score 105

4.13 Enrichment analysis result (ordered) of low-preserved modules for microarray and RNAseq data. Abbreviations: MF-Molecular Function, BP-Biologiocal Process, CC-Cellular Component 109

4.14 GO enrichment analysis for White module 109

4.15 GO enrichment analysis for Darkgreen module 109

4.16 GO enrichment analysis for Steelblue module 110

4.17 GO enrichment analysis for Darkred module 110

4.18 GO enrichment analysis for purple module 110

4.19 GO enrichment analysis for greenyellow module 111

4.20 Pathway analysis for White module. (R: Reactome; N: NCBI Biosystem; K: KEGG; C: Cell signalling technology) 112

4.21 Pathway analysis for Darkgreen module. (R: Reactome; N: NCBI Biosystem; K: KEGG; G: GeneGo; Q: Qiagen) 113

4.22 Pathway analysis for Steelblue module. (R: Reactome; N: NCBI Biosystem; K: KEGG; C: Cell signalling technology; RD: RandD Systems; PH: PharmGKB; G: GeneGo; Q: Qiagen) 114

4.23 Pathway analysis for Darkred module (R: Reactome; N: NCBI Biosystem; K: KEGG) 115

4.24 Pathway analysis for Purple module 115

4.25 Pathway analysis for Greenyellow module 116

4.26 Comparison of existing methods to find critical genes in ESCC . . 119

5.1 Dataset description 129

5.2 *Synthetic data - D3* 135

5.3 Comparisons among several measures including SNMRS in terms of internal cluster validation results for Iris data for average linkage hierarchical clustering. 137

5.4 Continuation of Table 5.3. 137

5.5 Comparisons among several measures including SNMRS in terms of internal cluster validation results for Iris data and for K-means clustering. 138

5.6 Continuation of Table 5.5 138

5.7 Comparisons among measures including SNMRS in terms of internal cluster validation results for Yeast Sporulation Gene expression data for average linkage hierarchical clustering. 144

5.8 Comparative enrichment analysis results obtained for Yeast sporulation dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component. 144

5.9 Comparative biological analysis results obtained by Yeast sporulation dataset 145

5.10 Comparative biological analysis results obtained using ESCC normal dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component. 147

5.11 Comparative biological analysis results obtained using ESCC tumor dataset. Different types of colours are assigned to module names. Abbreviations: BP-Biological Process, MF-Molecular Function, CC-Cellular Component. 148

5.12 P-value biological analysis results obtained by ESCC Normal dataset149

5.13 P-value biological analysis results obtained by ESCC Tumor dataset 151

5.14 Potential biomarker list for ESCC with biological processes and literature evidences 155

6.1 DEG Details 167

6.2 Details of GCN Analysis 170

6.3 Biological analysis of selected hub genes 173

List of Algorithms

1 Ensemble based DEA to identify potential biomarkers. 94

2 Construction of weighted signed GCN using SNMRS and extraction of modules 132

Glossary of Terms

BC	Bicluster
BP	Biological Process
CC	Cellular Component
DCA	Differential co-expression analysis
DEA	Differential expression analysis
DEGs	Differentially Expressed genes
DE	Differentially Expressed
ESCC	Esophageal Squamous Cell Carcinoma
FIR	Fractionated Irradiation
G	Gene
GCN	Gene Co-expression Network
GED	Gene Expression Data
GEO	Gene Expression Omnibus
GO	Gene-Ontology
GRN	Gene regulatory network
H	Hub gene
MF	Molecular Function
N	Normal
NMRS	Normalized mean residue similarity
PCC	Pearson Correlation Coefficient
PPI	Protein-Protein Interaction Network
RNA-seq	RNA sequencing
S	Sample
scRNA-Seq	single-cell RNA sequencing
SNMRS	Scaling-and-Shifting Normalized Mean Residue Similarity
T	Tumor
TF	Transcription Factor