Contents

1 Introduction **1**

1.1 Central Dogma of Molecular Biology 1

1.2 Gene Expression Data (GED) 3

1.2.1 DNA Microarray Data 4

1.2.2 RNA sequencing (RNA-seq) Data 5

1.2.3 Single-cell RNA sequencing (scRNA-seq) Data 5

1.3 Gene Expression Data Analysis 7

1.4 Data Mining 8

1.4.1 Predictive Data Mining Tasks 9

1.4.2 Descriptive Data Mining Tasks 9

1.5 Motivation 10

1.6 Research Objectives 11

1.7 Research Contributions 12

1.7.1 Biclustering-based Biomarker Identification in ESCC Microarray Data 12

1.7.2 Identifying Crucial Genes in ESCC GED using Differential Expression Analysis 12

1.7.3 Identifying Crucial Genes in Esophageal Squamous Cell Carcinoma using an Ensemble Approach 13

1.7.4 An Advanced Measure for Co-expression Network Analysis . 13

1.7.5 Identification of Potential Prognostic Biomarkers for ESCC using Single-cell RNA Sequencing Data Analysis 14

1.8 Organization of the Thesis 14

2 Background 15

2.1 Gene Expression Data Analysis 15

2.1.1 Gene and Gene Expression 15

2.1.2 Gene Expression Data 16

2.1.3 Gene Expression Data Generation 18

2.1.4 Preprocessing of RNA-seq data 24

2.1.5 Preprocessing of scRNA-seq data 25

2.1.6 Gene Expression Data Analysis 28

2.1.7 Degree centrality 35

2.2 Biomarker Identification 37

2.2.1 Cluster Analysis 38

2.2.2 Biclustering or Two-way Clustering Approaches 39

2.3 Gene Expression Data Repositories 41

2.3.1 GEO 41

2.3.2 Recount2 41

2.3.3 NCBI SRA 42

2.3.4 cancerSEA 42

2.4 Datasets Used 42

2.4.1 ESCC disease dataset: GSE20347 42

2.4.2 ESCC disease dataset: GSE23400 43

2.4.3 ESCC disease dataset: GSE32424 43

2.4.4 ESCC disease dataset: SRP064894 43

2.4.5 Yeast Sporulation dataset 43

2.4.6 Iris dataset 44

2.4.7 ESCC disease dataset: GSE81812 44

2.5 Software Tools Used 44

2.5.1 R 45

2.5.2 FuncAssociate 45

2.5.3 DAVID 46

2.5.4 STRING 46

2.5.5 Cytoscape 47

2.5.6 GeneMania 47

2.5.7 GeneMalacard 47

2.5.8 WGCNA 48

2.5.9 Python 48

2.6 Statistical and Biological Evaluation 49

2.6.1 Gene Enrichment Analysis 49

2.6.2 Pathway Analysis 49

2.6.3 Topological Analysis 51

2.6.4 Literature Mining/Evidence 51

2.7 Discussions 51

3 Biclustering-based biomarker identification 53

3.1 Introduction 53

3.2 Related Works 54

3.3 Proposed Method 55

3.3.1 Microarray Data Collection 55

3.3.2 PD_BiBIM framework 55

3.4 Experimental Results 59

3.4.1 Pre-processing of dataset 60

3.4.2 Performance evaluation of PD_BiBIM using GSE20347 and GSE23400 datasets 61

3.4.3 Biological network analysis for Primary and Secondary genes 68

3.4.4 Pathway analysis 68

3.4.5 GO enrichment analysis 69

3.4.6 Gene Regulatory Network (GRN) analysis 70

3.4.7 Literature evidence 71

3.5 Discussion 74

3.6 Conclusion 77

4 Identifying Crucial Genes for ESCC using Differential Expression Analysis 78

4.1 Introduction 78

4.2 Background 79

4.3 Proposed Methods 80

4.3.1 Method I: A Generic DEA Approach: 80

4.3.2 Method II: An Ensemble Approach 90

4.4 Conclusion 118

5 SNMRS : An Effective Measure for Co-expression Network Analysis 121

5.1 Introduction . . . 121

5.2 Background . . . 123

5.2.1 Materials and Method . . . 124

5.2.2 Properties of SNMRS . . . 125

5.2.3 Proposed GCN construction and Module extraction method 129

5.3 Experimental Results . . . 132

5.3.1 Benchmarking of the SNMRS measure using Synthetic and Multivariate data . . . 133

5.3.2 Benchmarking of the SNMRS method using independent RNA-seq and microarray dataset . . . 140

5.3.3 Threshold selection . . . 141

5.3.4 Time Series data . . . 141

5.3.5 Gene Expression data with normal and tumor conditions . . 143

5.4 Discussion . . . 147

5.5 Conclusion . . . 155

6 Identification of Potential Prognostic Biomarkers for ESCC using Single-cell RNA Sequencing Data Analysis 156

6.1 Introduction . . . 156

6.2 Related Work . . . 157

6.3 Materials and Methods . . . 159

6.3.1 Datasets used . . . 159

6.3.2 Proposed framework . . . 159

6.4 Results . . . 162

6.4.1 Preprocessing 162

6.4.2 Cell Clustering and finding DEGs 163

6.4.3 Finding Hub gene 165

6.4.4 Literature Trace 173

6.5 Discussion 174

6.6 Conclusion 175

7 Conclusions and Future Works 176

7.1 Concluding Remarks 176

7.2 Future Works 178

Publications based on the Thesis Works 211