Development of Computational Methods for Multi-omics Data Analysis

A Thesis submitted by
Madhumita

in partial fulfillment of the requirements for the award of the degree of
Doctor of Philosophy

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
Department of Bioscience and Bioengineering
July 2022

Declaration

I hereby declare that the work presented in this Thesis titled *Development of Computational Methods for Multi-omics Data Analysis* submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of Dr. Sushmita Paul. The contents of this thesis in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Madhumita
P17BS001

Certificate

This is to certify that the thesis titled *Development of Computational Methods for Multi-omics Data Analysis*, submitted by *Madhumita (P17BS001)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy*, is a bonafide record of the research work done by her under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Sushmita Paul

Dr. Sushmita Paul
Ph.D.Thesis Supervisor

Acknowledgements

More than a thesis, a PhD is a journey. As I look back to the years that led to the completion of this thesis, I realize how fortunate I am to find such kind of people around me whose consistent support, supervision, and confidence in me made this journey possible.

First and foremost, I express my deepest gratitude to my supervisor Dr. Sushmita Paul, for her endless support, guidance, training, and teaching throughout my PhD tenure. I thank her for giving me the opportunity to carry out my thesis work in her Computational Biology Laboratory and introducing me to the world of pattern recognition and machine learning. She is the one who made me realize how puissant computational tools can be for a biologist. The thesis would not have been completed without her vision and technical support. Customary thanks always stand inadequate against the priceless effort she has put in to develop my technical knowledge and inspire me to become an independent, disciplined, and creative researcher. Dr. Sushmita Paul has been a guide and guardian throughout my PhD and I feel incredibly proud in saying that I am her student.

I would like to record my sincere gratitude and appreciation to the Department of Bioscience and Bioengineering, IIT Jodhpur, and the HOD of the department, Dr. Mitali Mukerji, for allowing me to work in a great environment during this period. I am grateful to my committee members Dr. Sushmita Jha, Dr. Shankar Manoharan, and Dr. Samanwita Pal, for their timely evaluation and valuable suggestions that facilitated my research progress and allowed me to grow as a research scientist. I am very thankful to Prof. Sankar K. Pal, Scholar-in-Residence, IIT Jodhpur, for sharing his valuable thoughts on my research problem. I would also like to thank Dr. Sukhendu Ghosh, Department of Mathematics, IIT Jodhpur, for his fruitful discussions regarding one of my works. My very special thanks go to all the staff members of the Department of Bioscience and Bioengineering for making all possible facilities available to me.

I am thankful to my lab members, especially, Sonu Bansal, Namrata Pant, Kuldeep Gujar, Animesh Kumar Singh, Amiyangshu De, Jyoti Sharma, Archit Dwivedi, and Khadija Sana Hafeez for creating such a pleasant atmosphere for research and collaboration within the lab. I am delighted to acknowledge my friends Jittu and Diksha for providing me unconditional support on both academic and personal levels. Your presence gives me strength and happiness.

I want to take this opportunity to express my gratitude to my beloved parents and little sister for being the most substantial support not only during my PhD career but also throughout my life. Ma and Papa, you have always inspired me to get out of my comfort zone and work honestly towards my goal. I sincerely acknowledge the hard work and sacrifices you have made for the smooth completions of this thesis work. I am incredibly blessed to have a family which has always supported me in all my endeavors.

Much of what I have learned over the years has come from my friends and well-wishers, both in academic and non-academic spheres. To all such people, I convey my sincere thanks.

I express all my gratitude towards Goddess Saraswati without whose blessings whatever I have done or will do is not imaginable.

Madhumita

List of Figures

Figure	*Title*	*page*
1.1	A glimpse of multi-omics interactions	2
1.2	Processing workflow for three main high-throughput technologies	4
1.3	Role of computation in multi-omics studies	6
1.4	A simplified illustration of data structure used in for TCGA data storage within GDC	9
1.5	Application of machine learning in multi-omics cancer studies	11
1.6	Unsupervised approaches for multi-omics data integration	11
1.7	Workflow for fair and open sharing of information in a multi-omics study	14
1.8	The overall study presented in this thesis	15
1.9	Outline of the thesis	17
2.1	Schematic portrayal for MMRMs identification using integrative approach	21
2.2	Overview of multi-omics clustering approaches	29
3.1	Diagrammatic representation of *RFCM*^{3} workflow. Given the inputs of the same sampled miRNA and mRNA expression profiles *RFCM*^{3} first generates star-shaped modules containing single miRNA and multiple mRNAs. It selects mRNAs by maximizing relatedness between miRNA and mRNA, also the functional similarity between the mRNAs of the module. *RFCM*^{3} uses MI for identifying relationships between multiple mRNAs and a miRNA. At a later stage, in order to remove insignificant modules and to merge the significant ones, miRNA-miRNA synergistic interaction information is incorporated. Modules are merged rigorously considering their miRNA-miRNA similarity, where most of the modules are functionally enriched. Finally, multiple modules are generated containing more than one miRNA and multiple genes.	39
3.2	Variation in KPES for different values of threshold δ	43
3.3	miRNA interactions across all modules at cut-off $>= 0.7$ generated by *RFCM*^{3}	44
3.4	miRNA-mRNA regulatory interactions of the most significant module generated by *RFCM*^{3}	45
3.5	PPIN generated by STRING for the most significant modules	47
3.6	Comparison of total modules and enriched modules	48
3.7	The human papillomavirus infection pathway (hsa05165) rendered by Pathview	50
3.8	Significant pathways generated by Mirsynergy, SNMNMF, WGCNA, and *RFCM*^{3}	51
3.9	Comparison of Kaplan-Meier curves for patients with cervical cancer plotted for combined expression of 20 genes obtained by the Mirsynergy and 242 genes obtained by the *RFCM*^{3}	51

4.1 Diagrammatic representation of proposed approach showing the pipeline used in this study to identify histological subtypes. miRNA and mRNA expression profiles are first used separately by the proposed method to assign weights to the features. Here, NW_{mi} and NW_m represents normalized feature weights of miRNAs and mRNAs, respectively. These weights are then utilized to calculate distance between pair of samples in each of the expression profiles separately. Two distance matrices are generated for same set of samples using the distance equation. Here, f_m represents feature either miRNA or mRNA. N represents all the features present in either of the expression profile. Two of the distance matrices are then converted into weighted similarity networks. The networks generated by mRNA and miRNA expression data are then fused into a single network. At last clustering of the nodes of this integrated network is performed to identify cancer subtypes. 57
4.2 Variation in average silhouette width with respect to recommended range of the hyper-parameter μ. 63
4.3 Silhouette plots for the two subtypes of each cancer type identified by the proposed approach. 67
4.4 DO plots for the set of differentially expressed genes between cancer subtypes of respective cancers identified by the proposed method. 73
4.5 Kaplan-Meier curves for patients with cervical, breast, and ovarian cancer plotted for combined expression of top 100 differentially expressed genes between respective subtypes obtained by proposed approach of cancer subtypes identification. 74

5.1 Schematic flow diagram of the proposed approach for cancer subtypes identification 79
5.2 Effect of sample size and number of omic-views on the runtime of the proposed algorithm. Values in the parentheses indicate the number of omic-views. 89
5.3 Comparative performance analysis of proposed approach and spectral clustering performed on individual omic-views 93
5.4 Comparative analysis of different integrative sub-spaces on CESC data set 94
5.5 Comparative analysis of different integrative sub-spaces on BRCA data set 95
5.6 Comparative analysis of different integrative sub-spaces on OV data set 96
5.7 Comparative analysis of different integrative sub-spaces on LGG data set 97
5.8 Comparative analysis of different integrative sub-spaces on STAD data set 98
5.9 Method Comparison 106

6.1 Workflow of the proposed cancer subtypes identification framework 111
6.2 Proposed architecture of AE 116
6.3 Varying silhouette score with varying number of neurons in the hidden layers of the undercomplete AE architecture 117
6.4 Varying silhouette score with varying L1 regularization (λ) in the proposed AE architecture 118
6.5 Survival curves for the cancer subtypes identified by the proposed approach 121
6.6 Performance comparison on single-omic and multi-omics datasets using different cluster validity indices 122
6.7 Comparative disease ontology enrichment analysis for DE Genes 125
6.8 Kaplan-Meier survival curves: the log-rank p-value indicates the difference in the survival pattern between Temozolomide treated and not treated GBM patient's subgroups identified by the proposed approach 126

B.1 miRNA-mRNA regulatory interactions of the 22 modules generated by $RFCM^3$ 139
B.2 The human papillomavirus infection pathway (hsa05165) rendered by Pathview for different modules identified by $RFCM^3$ 145

C.1 DO plots for the breast cancer data when differentially expressed genes between two histological groups identified by the method EB, WSNF, and SNF is given as an input. 150

C.2 DO plots for the ovarian cancer data when differentially expressed genes between two histological groups identified by the method SFC and WSNF is given as an input. 151
C.3 Kaplan-Meier curves for cervical cancer data when differentially expressed genes between two histological groups identified by different methods are given as an input. 152
C.4 Kaplan-Meier curves for breast cancer data when differentially expressed genes between two histological groups identified by different methods are given as an input. 153
C.5 Kaplan-Meier curves for ovarian cancer data when differentially expressed genes between two histological groups identified by different methods are given as an input. 154

D.1 The plots represent the overall performance of all the methods when comparative analysis based on pair-wise log rank test is performed to evaluate the efficiency of different methods in identifying subtypes with differential survival pattern. In figure (a) success frequency indicates number of times the methods obtained the most significant p-value considering all the datasets. In figure (b) success frequency indicates number of times (out of 25 times), a method obtained a significant difference in the survival pattern between a subtype pair at a confidence interval of 95%. 156
D.2 Feature selection based comparative cluster analysis for GBM dataset 157
D.3 Feature selection based comparative cluster analysis for BIC dataset 158
D.4 Feature selection based comparative cluster analysis for COAD dataset 158
D.5 Feature selection based comparative cluster analysis for LSCC dataset 158
D.6 Feature selection based comparative cluster analysis for KRCCC dataset 159
D.7 Comparative p-value for log-rank test when CPH-based feature selection is applied to the comparative methods 161

List of Tables

Table	*Title*	*page*
1.1	Types of omics-data	4
1.2	Cancer Multi-omics Data Repositories	8
3.1	Modules Description at Different Cut-off Values	43
3.2	Performance Summary of *RFCM*^{3}, Mirsynergy, SNMNMF, WGCNA and MAGIA Web Tool	45
3.3	Comparative Pathway Enrichment of Genes	49
3.4	Comparative Pathway Enrichment of miRNAs	49
3.5	Experimental validation of some of the miRNA-mRNA interactions obtained by *RFCM*^{3}	53
3.6	miRNA Expression in Various Studies	54
4.1	Data set description	58
4.2	Performance of Proposed Approach Versus Other Weighting and Non-Weighting Based Approaches	66
4.3	Comparative Test Statistic Analysis of Different Algorithms	66
4.4	Number of Differentially Expressed Biomarkers	68
4.5	Comparative Gene Enrichment Analysis for Cervical cancer	69
4.6	Comparative Gene Enrichment Analysis for Breast Cancer	70
4.7	Comparative Gene Enrichment Analysis for Ovarian Cancer	70
4.8	Comparative miRNA Enrichment Analysis for Cervical cancer	70
4.9	Comparative miRNA Enrichment Analysis for Breast cancer	71
4.10	Comparative miRNA Enrichment Analysis for Ovarian cancer	71
4.11	Comparative Enrichment Analysis of miRNAs and Genes	71
4.12	Comparative metrices for Kaplan-Meier curves	75
5.1	Data-sets description	88
5.2	Cancer subtypes description: Actual class distribution	90
5.3	Comparative Cluster Analysis of Proposed and Existing Approaches	91
5.4	Comparative Biological Analysis of CESC dataset	101
5.5	Comparative Biological Analysis of BRCA dataset	102
5.6	Comparative Biological Analysis of OV dataset	103
5.7	Comparative Biological Analysis of LGG dataset	103
5.8	Comparative Biological Analysis of STAD dataset	104
5.9	Overlap with experimentally validated gene-list	107
6.1	Description of Cancer Datasets	111
6.2	Comparative Cluster Evaluation	120
6.3	Comparative p-value for Cox Log-rank Test	123
6.4	Comparative Biological Enrichment Scores	124
6.5	Comparative Pathway Enrichment Analysis for DE Genes	124
6.6	Comparative Pathway Enrichment Analysis for DE miRNAs	124
D.1	Comparative p-value for Cox log-rank test for every subtypes pair	156
D.2	Comparative silhouette score when CPH-based feature selection is applied to the comparative methods	160

List of Abbreviations

ACA	Adenocarcinoma
AE	Autoencoder
API	Application Programming Interface
AR	Annotation Ratio
ARI	Adjusted Rand Index
BIC	Breast Invasive Carcinoma
BL	Bottleneck Layer
BP	Biological Processes
BPEA	Biological Processes Enrichment Analysis
BPES	Biological Processes Enrichment Score
BRCA	Breast Cancer
CALM	The Causal Regulatory Modules
CC	Cellular Component
CC	Consensus Clustering
CCA	Canonical Correlation Analysis
CCDB	The Cervical Cancer Gene Database
CCLE	Cancer Cell Line Encyclopedia
CESC	Cervical Cancer
CLASH	Crosslinking, Ligation, and Sequencing of Hybrids
CNMF	Consensus Non-negative Matrix Factorization
CNV	Copy Number Variation
COAD	Colon Adenocarcinoma
CoALa	Convex-combination of Approximate Laplacians
COCA	Cluster-of-Cluster Assignment
Corr-LDA	Correspondence Latent Dirichlet Allocation

CPH	COX Proportional Hazard
CPTAC	Clinical Proteomics Tumor Analysis Consortium
CV	Coefficient of Variation
DAG	Directed Acyclic Graph
DAVID	Database for Annotation, Visualization, and Integrated Discovery
dbGap	The Database of Genotype and Phenotype
DE	Differentially Expressed
DEA	Differential Expression Analysis
DL	Deep Learning
DLSF	Deep Latent Space Fusion
DNA	Deoxyribonucleic acid
DO	Disease Ontology
DOEA	Disease Ontology Enrichment Analysis
DOSE	Disease Ontology Semantic and Enrichment analysis
DSML	Deep Subspace Mutual Learning
ECMC	Multi-view Clustering Approach with Enhanced Consensus
ELISA	Enzyme-linked Immunosorbent Assay
FDR	False Discovery Rate
FMRMs	Functional miRNA-mRNA Regulatory Modules
GBM	Glioblastoma multiforme
GDC	Genome Data Commons
GEO	Gene Expression Omnibus
GGIs	Gene-gene Interactions
HGP	Human Genome Project
HITS-CLIP	High-throughput Sequencing of RNA Isolated by Crosslinking Immunoprecipitation
HMDD	Human miRNA Disease Database
HPV	Human Papilloma Virus
ICGC	International Cancer Genomics Consortium
IDC	Infiltrating Ductal Carcinoma

Ids	Identifiers
ILC	Infiltrating Lobular Carcinoma
KEGG	Kyoto Encyclopedia of Genes and Genomes
KPCA	Kernel Principal Component Analysis
KPES	Kegg Pathway Enrichment Score
KRCCC	Kidney Renal Clear Cell Carcinoma
LASSO	Least Absolute Shrinkage and Selection Operator
LGG	Lower-grade Glioma
LSCC	Lung Squamous Cell Carcinoma
MAD	Median Absolute Deviation
MAGIA	miRNA and Genes Integrated Analysis
METBARIC	Molecular Taxonomy of Breast Cancer International Consortium
MF	Molecular Function
MI	Mutual Information
MiMIC	Multi-manifold Integrative Clustering
miRNA	microRNA
MISIM	miRNA Similarity Matrix
ML	Machine Learning
MMRMs	miRNA-mRNA Regulatory Modules
MOFA	Multi-omics Factor Analysis
mRNA	messengerRNA
MS	Mass-spectrometry
MSE	Mean Square Error
MTIs	miRNA-Target Interactions
NCBI	National Center for Biotechnology Information
ncRNA	non-codingRNAs
NGS	Next Generation Sequencing
NMF	Non-negative Matrix Factorization
NMI	Normalized Mutual Information

OV	Ovarian Cancer
PAR-CLIP	Photoactivatable Ribonucleoside-enhanced Crosslinking and Immunoprecipitation
PCA	Principal Component Analysis
PEA	Pathway Enrichment Analysis
PLA	Partial Least Square
PPINs	Protein Protein Interaction Networks
PPIs	Protein Protein Interactions
qRT-PCR	Quantitative Real Time Polymerase Chain Reaction
RFCM3	Relevant and Functionally Consistent miRNA-mRNA Modules
RISynG	Recursive Integration of Synergised Graph representations
RNA	Ribonucleic acid
RNA-seq	RNA Sequencing
RPPA	Reverse Phase Protein Array
rRNA	ribosomalRNA
RrSNF	Robust-regression based Similarity Network Fusion
SCC	Squamous Cell Carcinoma
SNF	Similarity Network Fusion
SNMNMF	Sparse Network-Regularized Multiple Negative Matrix Factorization
SNP	Single-nucleotide Polymorphism
SNV	Single-nucleotide Variant
SRA	Sequence Read Archive
SS	Silhouette Score
STAD	Stomach Cancer
SURE	Selective Update of Relevant Eigenspaces
TCGA	The Cancer Genome Atlas
tRNA	transferRNA
tSNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
WGBS	Whole-genome Bisulfite Sequencing

WGCNA	Weighted Gene Correlation Network Analysis
WGS	Whole Genome Sequencing
WMLRR	Weighted Multi-View Low Rank Representation