

# **Computational and Statistical methods to analyze multi-omic data**

**Thesis submitted by  
Ashmita Dey**

**DOCTOR OF PHILOSOPHY (Engineering)**

**Department of Computer Science and Engineering,  
Faculty Council of Engineering & Technology,  
Jadavpur University  
Kolkata, India  
2023**



**JADAVPUR UNIVERSITY**  
**KOLKATA-700032, INDIA**

INDEX NO. 286/18/E

1. Title of the Thesis:

**Computational and Statistical methods to analyze multi-omic data**

2. Name, Designation & Institution of the Supervisor/s:

(a) Prof. Ujjwal Maulik

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032, India





# List of Publications

## Papers in Journals

---

1. A. Dey, S. Sen and U. Maulik, "Study of transcription factor druggability for prostate cancer using structure information, gene regulatory networks and protein moonlighting", *Briefings in Bioinformatics*, 23(1), 2022.
2. S. Sen, A. Dey, S. Bandhyopadhyay, V. N. Uversky and U. Maulik, "Understanding structural malleability of the SARS-CoV-2 proteins and relation to the comorbidities", *Briefings in Bioinformatics*, 22(6), 2021.
3. A. Dey, S. Sen and U. Maulik, "Unveiling COVID-19-associated organ-specific cell types and cell-specific pathway cascade" *Briefings in Bioinformatics*, 22(2), 2021.
4. S. Sen, A. Dey, D. Sanyal, U. Maulik, and K. Chattopadhyay, "IMMUND: an Early Stage Diagnostic and Therapeutic Frame for Neurodegenerative Diseases and Multiple Sclerosis based on Immunological Markers." *bioRxiv*, 2021.
5. A. Dey, S. Sen, V.N. Uversky, and U. Maulik, "Structural facets of POU2F1 in light of the functional annotations and sequence-structure patterns." *Journal of Biomolecular Structure and Dynamics*, 39(3), 2020.
6. S. Sen, A. Dey, and Ujjwal Maulik. "Studying the effect of alpha-synuclein and Parkinson's disease linked mutants on inter pathway connectivities.", *Scientific Reports*, 11(16365), 2021.
7. S. Sen, A. Dey, S. Chowdhury, K. Chattopadhyay, and U. Maulik, "Understanding the Evolutionary Trend of Intrinsic Structural Disorders in Cancer Relevant Proteins as probed by Shannon Entropy Scoring and Structure Network Analysis", *BMC Bioinformatics*, pp. 231-242, 19(13), 2019.
8. A. Dey, Sk. Shanwaz and U. Maulik, "scPCN: Revealing Pathway Connectivity among Cell Types using Single-cell Data", *Journal of Genetics and Genomics* (Communicated).

---

## Papers in Conference Proceedings

---

1. A. Dey and U. Maulik. 2022. Bioinformatics pipeline to unveil the heterogeneity of Glioblastoma Multiforme. Oral presentation: 2022 IEEE Calcutta Conference (CALCON), Kolkata, West Bengal.
2. A. Dey and U. Maulik. 2020. Identification of Cell-types based on the Pathway of Markers using Single-cell data. Oral presentation: 2020 IEEE Calcutta Conference (CALCON), Kolkata, West Bengal.
3. S. Sen, A. Dey, U. Maulik. 2018. Identifying potential hubs for kidney renal clear cell carcinoma from tf-mirna-gene regulatory networks. Oral presentation: 2018 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, West Bengal.

---

## List of Presentations in National/International Conference:

---

1. A. Dey and U. Maulik. 2022. Bioinformatics pipeline to unveil the heterogeneity of Glioblastoma Multiforme. Oral presentation: 2022 IEEE Calcutta Conference (CALCON), Kolkata, West Bengal.
2. A. Dey and U. Maulik. 2020. Identification of Cell-types based on the Pathway of Markers using Single-cell data. Oral presentation: 2020 IEEE Calcutta Conference (CALCON), Kolkata, West Bengal.
3. S. Sen, A. Dey, U. Maulik. 2018. Identifying potential hubs for kidney renal clear cell carcinoma from tf-mirna-gene regulatory networks. Oral presentation: 2018 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, West Bengal.
4. S. Sen, A. Dey, U. Maulik. Poster: Unravelling the Dynamicity of POU2F1 based on Evolutionary Conservation and Structure Network Analysis Presented At Function COSI, ISMB/ECCB, Congress Center Basel, Switzerland.
5. A. Dey, R. Das, S. Sen, U. Maulik. Poster: moonPRED: A Deep Neural Network Based Tool to Predict Protein Moonlighting moonPred Presented At MLCSB COSI ISMB/ECCB, Congress Center Basel, Switzerland.
6. A. Dey, U. Maulik. Poster: Cell type-specific pathways associated with diffuse large b-cell lymphoma metastasis related to neuro-diseases Presented At 4th health sciences and innovation congress, Baku, Azerbaijan.
7. S. Sen, A. Dey, S. Chowdhury, U. Maulik, K. Chattopadhyay. Poster: Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis Presented At 17th International Conference on Bioinformatics (InCoB 2018), New Delhi.



# STATEMENT OF ORIGINALITY

I Ashmita Dey registered on 28/05/2018 do hereby declare that this thesis entitled "Computational and Statistical methods to analyze multi-omic data" contains a literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis has been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 9%.

Ashmita Dey

Signature of the Candidate:

Date:

Ujjwal Haulin

Certified by Supervisor:

(Signature with date, seal)

20/5/2024

Professor

Computer Sc. & Engg. Department

Jadavpur University

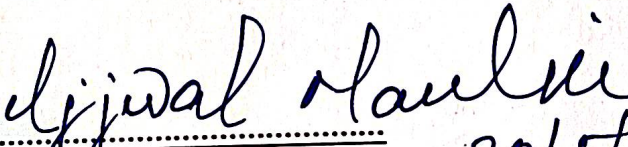
Kolkata-700032





# CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **Computational and Statistical methods to analyze multi-omic data** submitted by Ashmita Dey, who got her name registered on 28.05.2018 for the award of Ph.D. (Engineering) degree of Jadavpur University, is absolutely based upon her own work under the supervision of Prof. Ujjwal Maulik, Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India, and that neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

  
.....  
Signature of the Supervisor  
and date with Office Seal

20/5/2024  
Prof. Ujjwal Maulik  
Computer Sc. & Engg. Department  
Jadavpur University  
Kolkata-700032





---

## *Dedication*

*To my family and supervisor*



# Acknowledgements

Finally, it is time for me to acknowledge all those who inspired me, supported me and helped me to get to the place where I am today.

I take this opportunity to express a deep sense of gratitude to Prof. (Dr.) Ujjwal Maulik for his supervision and invaluable cooperation. This thesis would not have been possible without his constant inspiration and unbelievable support of him over the last few years.

I have had an amazing group of Labmates. Each of them deserves my gratitude: Sagnik Sen, Rangan Das, Sourav Chowdhury. Working with them was a great experience that brought together several good and fruitful ideas. Thank you to all of you for the unselfish help, insights, and feedback, and for making the science a collaborative effort. With that, a special thanks to Prof. Sanghamitra Bandyopadhyay, Dr Indrajit Saha, and Prof. Krishnananda Chattopadhyay. The Journey would remain incomplete without their suggestions, unconditional help, and encouragement.

I would also like to thank all of my colleagues from the Computer Science and Engineering Department, at Jadavpur University for providing me with a friendly research environment.

Finally, most important of all, I would like to dedicate the thesis to my parents Mr Ashim Kr. Dey and Mrs Susmita Dey, my beloved younger sister Archita Dey, to honour their love, patience, encouragement, and support during my research. I would also express my appreciation to my husband, Mr Akash Roy and my mother-in-law Mrs Sarbani Roy for their unwavering support and love.

Date: 20/05/2024

Ashmita Dey

(Ashmita Dey)



# Abstract

The advent of high-throughput technologies has enabled the simultaneous measurement of multiple types of omics data, such as genomics, transcriptomics, proteomics, and metabolomics. This multi-omic approach has revolutionized biological research by providing a comprehensive view of complex biological systems. However, the analysis of such data poses significant challenges due to its high dimensionality, heterogeneity, and interdependencies. In this context, computational and statistical methods play a vital role in extracting meaningful insights from multi-omic data. By incorporating this idea, we intend to address some key questions in this thesis, including (i) interactions and influences between different omics layers, (ii) identification of potential therapeutic targets for guiding drug development efforts, (iii) the role of cellular heterogeneity in disease programs, and (iv) mechanisms of cell communication within tissues or across the cellular system. While various methods are available, these questions still require appropriate answers. The traditional reductionist approaches fail to establish the bridge between biological significance and designed computational methods. This observation motivates us to design novel bioinformatics frameworks by using computational and statistical techniques to decipher unanswered questions regarding multi-omics data. The first chapter of this thesis begins with a general introduction of the multi-omics data and provides background information on each omics and establishes the significance of the research. It involves an exploration of existing knowledge gaps, practical implications, and potential theoretical contributions. Additionally, the chapter delineates the scope of the studies, encompassing any limitations or constraints that may influence the findings' applicability. Furthermore, an overview of the research questions stated earlier is provided, elucidating how they are addressed, along with an outline of the methodology employed in each study. Finally, the chapter concludes by providing a summary of the thesis' overall structure, offering a brief preview of the subsequent chapters' content.

In the subsequent chapter, the main objective is to answer the interaction among the different omics data and their impact on each other. The existing approaches of regulatory networks play a pivotal role in unraveling the complex mechanisms underlying various biological processes and diseases. However, traditional methods fail to achieve a system-level understanding of gene regulation and association with other regulators. In this regard, we proposed a pipeline by incorporating network biology concepts to identify the regulators responsible for disease-specific gene dysregulation. This proposed method helps to decipher the relationships and also the key regulatory genes and pathways that play critical roles in disease development or involves in normal physiological processes.

Chapter 3 proposes models for the identification of potential therapeutic targets to guide drug development efforts. In the preceding chapter, transcription

## Abstract

---

factors (TFs) are identified as promising drug targets due to their ability to modulate the expression of protein-coding genes and miRNAs at the transcriptional and post-transcriptional levels, respectively. However, targeting TFs for drug development is challenging due to their highly dynamic nature, which makes it difficult to identify stable drug-binding sites. To address this challenge, a novel concept called protein moonlighting is employed to determine suitable druggable sites on TFs. Additionally, the structure-function paradigm of proteins is utilized to gain valuable insights into their dynamic properties, which can guide the development of drugs and optimize their binding efficacy. To explore the dynamicity of TFs, an evolutionary study is conducted on the structure-function space of these proteins. Two entropy models, based on Shannon entropy and direct coupling analysis, are proposed to analyze the information content and coevolutionary patterns within the TFs. Moreover, centrality-based clustering techniques are employed to establish the structural network of the TFs. This approach helps identify interconnected regions within the network, known as the structural core. These structural cores represent specific protein conformations or dynamic properties that drugs can target to achieve optimal binding and therapeutic efficacy. By incorporating these models, this chapter provides a comprehensive framework for identifying potential drug targets within the highly dynamic landscape of TFs.

The studies discussed in the previous two chapters of this thesis are performed on bulk sequencing data. However, this sequencing technique does not capture the cellular heterogeneity within the population, potentially masking important subpopulations of cells with distinct gene expression profiles. In contrast, single-cell RNA-sequencing (scRNA-seq) enables the profiling of gene expression at the single-cell level, providing a comprehensive view of cellular heterogeneity within a tissue or sample. During COVID-19, it is imperative for researchers to contribute scientifically towards controlling the spread of the virus. The aim of our study is to investigate the impact of COVID-19 on specific organ cell types and elucidate the associated cell-specific pathway cascades. Through comprehensive case studies, it is found that some specific organs such as lungs, kidneys, liver, ileum, and bladder are affected due to COVID-19. Additionally, it was discovered that the virus exhibits a strong binding affinity with ACE2 and TMPRSS2 proteins. To gain insights into the organ-specific cell types and their differentially expressed biomarkers, clustering techniques were employed. Moreover, protein-protein interaction (PPI) analyses were performed to identify highly influential hubs within these networks, utilizing the K-means clustering algorithm. Furthermore, a pathway semantic similarity network was established to uncover the interactions between pathways responsible for regulating cell function. In another study, the maximal clique centrality (MCC) technique was incorporated to identify essential genes from local interaction networks. By identifying these

essential genes, their critical roles in cellular processes and potential as therapeutic targets can be elucidated. Overall, the proposed bioinformatics methods have the potential to contribute to the development of targeted therapeutic strategies and enhance our understanding of the disease.

Many diseases like cancer, neurodegenerative disorders, and immune-related conditions involve dysregulated cell communication and crosstalk. Simply identifying biomarkers associated with a disease may provide diagnostic or prognostic information, but it does not capture the underlying mechanisms driving the disease. In this regard, Chapter 5 proposed a new method for computing pathway activity scores of the cells to decipher the cell-to-cell communication in a particular tissue. The output of the proposed method provides information regarding the function of genes performed in biological processes. Finally, by incorporating pathway information in our proposed single-cell clustering method, we are able to align pathways across the three gynaecological diseases and successfully identified disease-specific mechanisms.

As demonstrated in this thesis, the application of statistical and computational techniques has made significant contributions to addressing key problems in biology and holds great promise for future research. These approaches have proven to be valuable tools for analyzing complex biological data, uncovering hidden patterns, and gaining insights into fundamental biological processes.





# Contents

<b>Acknowledgements</b>	<b>xiii</b>
<b>List of Figures</b>	<b>xxiii</b>
<b>List of Tables</b>	<b>xxxi</b>
<b>1 Introduction and Scope of the Thesis</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Multiomics . . . . .	1
1.2.1 Genomics . . . . .	2
1.2.2 Transcriptomics . . . . .	3
1.2.3 Proteomics . . . . .	3
1.2.4 Metabolomics . . . . .	4
1.3 Integration of multi-omics at system-level . . . . .	5
1.3.1 Approaches . . . . .	6
1.4 Computational methods for analysing multi-omics data . . . . .	8
1.4.1 Integrative methods . . . . .	8
1.4.2 Clustering and classification methods . . . . .	8
1.4.3 Pathway analysis . . . . .	8
1.4.4 Dimensionality reduction . . . . .	9
1.4.5 Network analysis . . . . .	9
1.5 Approaches for integrating bulk multi-omics data . . . . .	10
1.5.1 Multi-step/multi-stage approaches . . . . .	10
1.5.2 Bayesian methods . . . . .	11
1.5.3 Network based approaches . . . . .	11
1.5.4 Matrix factorization . . . . .	11
1.5.5 Multiple kernel learning . . . . .	11
1.6 Bioinformatics techniques for Single-cell data . . . . .	12
1.6.1 Quality Control and Normalization . . . . .	12
1.6.2 Feature Selection and Dimensionality Reduction . . . . .	12
1.6.3 Clustering and Cell Type Identification . . . . .	13
1.6.4 Trajectory Inference . . . . .	13
1.6.5 Differential Expression Analysis . . . . .	13
1.7 Networks of Networks . . . . .	13
1.7.1 Gene Regulatory Network . . . . .	14
1.7.2 Gene coexpression network . . . . .	14
1.7.3 Protein-protein interaction . . . . .	16
1.7.4 Metabolic Interaction . . . . .	16
1.7.5 Cell-signalling network . . . . .	17

## CONTENTS

---

1.7.6	Disease network . . . . .	17
1.8	General network properties . . . . .	18
1.8.1	Centrality . . . . .	18
1.8.2	Modularity . . . . .	19
1.9	Scope of the Thesis . . . . .	20
1.9.1	Studying the association between gene and its regulators using regulatory network . . . . .	20
1.9.2	Understanding the dynamicity of proteins in terms of network	21
1.9.3	Graph-theoretical modeling to unveil the cell-to-cell hetero- geneity . . . . .	21
1.9.4	Computational framework for pathway-based inference of single-cell RNA-seq data . . . . .	22
1.9.5	Conclusions and Future Scope . . . . .	22
<b>2</b>	<b>Studying the association between gene and its regulators using regula- tory network</b>	<b>23</b>
2.1	Introduction . . . . .	23
2.2	Computation approaches to unveil the association between gene and regulators . . . . .	24
2.2.1	Data description for KIRC . . . . .	24
2.2.2	Data selection for Prostate Cancer . . . . .	24
2.2.3	Establishing gene regulatory network . . . . .	25
2.2.4	Detection of the functional hub using modularity . . . . .	25
2.2.5	Biological validation . . . . .	25
2.3	Experimental Results and Discussion . . . . .	26
2.3.1	Computational transcriptomic approach to provide an idea of the mechanism of gene regulation leads to KIRC disease .	26
2.3.2	Experimental results of the prostate cancer-specific gene reg- ulatory network . . . . .	29
2.4	Conclusion . . . . .	34
<b>3</b>	<b>Understanding the dynamicity of proteins in terms of network</b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	Data Acquisition and Data Preparation . . . . .	38
3.3	Evolutionary Trait based on Sequence Complexity . . . . .	38
3.3.1	Shannon's Entropy . . . . .	39
3.3.2	Coupling Analysis . . . . .	40
3.3.3	Graph theoretical modelling and eigenvector community de- tection . . . . .	41
3.3.4	Disorder region of the protein sequence . . . . .	41
3.3.5	Post Transnational Modification . . . . .	41

3.3.6	Hydropathy calculation . . . . .	42
3.4	Identification of the impact of Structural adaptation in protein functions . . . . .	42
3.4.1	Normal mode-based Structure Network Analysis . . . . .	42
3.4.2	Root Mean Square Fluctuation . . . . .	43
3.4.3	Calculate the stability score . . . . .	43
3.4.4	Liquid Liquid Phase Separation . . . . .	44
3.5	Experimental Results for Prostate Cancer study . . . . .	44
3.6	Results of POU2F1 as an important TF in pan-cancer study . . . . .	49
3.7	Results of protein-specific study for SARS-COV-2 . . . . .	56
3.7.1	Sequence Space Analysis . . . . .	58
3.7.2	Structure Space Analysis . . . . .	59
3.7.3	Diseases Related to the SARS-CoV-2 Infection . . . . .	60
3.7.4	Evolutionary Sequence-Structure Space Study of Proteins E, M, and N . . . . .	62
3.7.5	Evolutionary Sequence-Structure Space Study of the Spike S Glycoprotein . . . . .	65
3.7.6	Evolutionary Sequence-Structure Space of the Replicase Polyprotein ORF1ab . . . . .	66
3.7.7	Host Proteins and Corresponding Viral Protein . . . . .	67
3.7.8	Comorbidities and SARS-CoV-2 . . . . .	69
3.7.9	Malignancies and SARS-CoV-2 Infection . . . . .	69
3.7.10	SARS-CoV-2 and its Impact on Neurodegenerative and Neuropsychiatric Diseases . . . . .	70
3.8	Conclusions . . . . .	71
<b>4</b>	<b>Graph-theoretical modeling to unveil the cell-to-cell heterogeneity</b>	<b>73</b>
4.1	Introduction . . . . .	73
4.2	Computational framework for understanding cellular heterogeneity in disease severity . . . . .	74
4.2.1	Bioinformatics pipeline to unveil the heterogeneity of Glioblastoma Multiforme . . . . .	74
4.2.2	Unveiling COVID-19 associated Organ Specific Cell Types and Cell-Specific Pathway Cascade . . . . .	80
4.3	Conclusion . . . . .	93
<b>5</b>	<b>Computational framework for pathway-based inference of single-cell RNA-seq data</b>	<b>95</b>
5.1	Introduction . . . . .	95
5.2	Unveiling cell-to-cell heterogeneity by incorporating pathway information . . . . .	96

## CONTENTS

---

5.2.1	Method . . . . .	96
5.2.2	Validation . . . . .	97
5.3	Revealing Pathway Connectivity among Cell Types using Single-cell Data . . . . .	100
5.3.1	Method . . . . .	100
6	Conclusions and Future Scope of Research	114
	Bibliography	119

# List of Figures

1.1	Integrative analysis of multi-omics data . . . . .	2
1.2	Bulk sequencing and single-cell sequencing: In bulk sequencing, a collection of RNA extracted from a group of cells is sequenced, resulting in expression data that represents the average expression of a specific gene across all cells. On the other hand, scRNA-seq (single-cell RNA sequencing) preserves the unique transcript information of each individual cell. When multiple cells of the same type are identified, averaging the sequence reads across all those cells provides cell-type-specific expression information similar to the profiles obtained through bulk RNA-seq. . . . .	7
1.3	Gene regulatory network: a graphical representation illustrating the interactions and relationships between genes and molecular regulators, showcasing how they influence each other's expression levels. The green, blue and red coloured nodes represent transcription factors, genes and miRNA respectively, while the arrows indicate activation or repression of gene expression . . . . .	15
1.4	Analysis of the interactivity of human POU2F1 by STRING computational platform that produces the network of predicted associations for a particular group of proteins. . . . .	16
1.5	Disease network- a visual depiction highlighting the complex connections and associations between various diseases and related factors. Here, nodes represent different diseases and their associated cell types, while edges depict the relationships and interactions among them, offering insights into the interconnected nature of diseases and potential underlying mechanisms . . . . .	17
1.6	The graphical abstract to summarize the researches covered under the thesis . . . . .	21
2.1	An interactive network between TF and their targeted host genes of KIRC. . . . .	27
2.2	Venn diagram representation of the common genes targeted by both miRNAs and TFs. . . . .	28
2.3	An regulatory network between TFs and their targeted miRNAs those having significant role in KIRC. . . . .	29
2.4	The interaction among genes, miRNAs and TFs (a) the interaction between SPP1 and hsa-miR-146a-5p which is targeted by different TFs and (b) the interaction between CDH6 and hsa-miR-155-5p which shows how TFs regulate the miRNAs. . . . .	30

## LIST OF FIGURES

---

2.5	The circos plot to represent the association of the selected TFs with the shared pathways. Four pathways are found common among the selected molecular regulators including prostate cancer. The color and different size of the ribbon shows the type of relation depending on the p-value of the pathway and its corresponding TFs. The circos plot is used for better visualization of the tabular data. Here, the ribbons are connected between the TFs and their sharing pathway. The extent of the association is represented through the thickness of the ribbon and this thickness is based on the p-values of the pathways for each TFs. Ribbons touch the segment of the inner circle to define the row value whereas the ribbons do not touch the segments. Moreover, segmentation provides the absolute scale of the specific region of interaction between TFs and pathFA. The percentage of the outer circle indicates the overall total of each segment respectively. . . . .	31
2.6	The final established gene regulatory network depends on the shared pathways of the three molecular regulators. Here red, blue and yellow represent TFs, miRNAs and genes. . . . .	33
3.1	A weighted network $G_{DCA}$ and corresponding colour modules based on overall residual covariation from DI score of the TFs in prostate cancer . . . . .	45
3.2	In prostate cancer, the overlapped areas are obtained from PONDR and IUPred databases. Results from PONDR and IUPred are represented with orange and grey colour, respectively. . . . .	45
3.3	Top five diseases, based on the P-values, of each TF of prostate cancer are selected to establish the network. Green colour nodes are diseases, connected with all four TFs (red oval nodes). Green colour nodes represent the diseases targeted by five out of six TFs . . . . .	46
3.4	The representation of the residues having post-transnational modification for each selected TFs in prostate cancer study. . . . .	47
3.5	The stability score of the amino acids of six TFs of prostate cancer along with their bonds occurred in a particular residue. Torsion, Van der Waals, Chi and Cis-bond are represented by colour yellow, green, red and orange, respectively . . . . .	48
3.6	POU2F1 is associated with multiple diseases. (a) Associations are represented through a scatter plot and (b) the Top ten associated diseases are shown through a network diagram. . . . .	52
3.7	The network representation of the top five common functional pathways among POU2F1 and its targeted miRNAs and genes. Different colors represent diverse pathways and their associations. . . . .	52

3.8	Representing the change in evolutionary trend by (a) The Shannon entropy score of Homo sapiens samples and all other species by blue and black lines respectively and (b) The DCA score of POU2F1.	53
3.9	A weighted network $G_{DCA}$ and corresponding color modules based on overall residual co-variation from DI score. . . . .	54
3.10	The PDB structure of (a) model 1 (-0.56), (b) model 2 (-1.16), (c) model 3 (-4.02) (d) model 4 (-4.16) and (e) model 5 (-4.17), the five models and their c-scores performed from I-TASSER of POU2F1 protein. The highlighted red-coloured zone of each five structures denotes the DNA binding site of the respective models. The DNA binding site of the models is identified using the information from UNIPROT database. . . . .	54
3.11	The structure network of the model The structure network diagram of the models (a) model 1, (b) model 2, (c) model 3, (d) model 4 and (e) model 5, are performed from the I-TASSER. Each structure shows that a large portion of the residues are converged in a single module. . . . .	56
3.12	The disorder analysis of POU domain proteins using PONDR® VSL2 and PONDR® VL-XT algorithms. (a) PONDR-based analysis of the full set of sequences of all the members of the POU domain family shows that the consensus sequence generated for this family is mostly ordered in nature. (b) Disorder analysis of the amino acid sequences of only the members of the POU domain family that belong to the advanced species illustrates that the corresponding consensus sequence is predicted to contain very significant levels of disorder. (c) Analysis of human POU2F1 proteins by PONDR denotes a very high level of disorder. . . . .	57
3.13	The two-dimensional hydropathy plot produced for the selected proteins based on the Kyte and Doolittle scale of amino acid hydropathy. . . . .	58
3.14	The sequence space analysis is performed for the selected protein families along with the SE calculation, coupling analysis and community detection techniques. The residue-wise extreme variability of the SE score has been shown through two colors. . . . .	59
3.15	In the structure space analysis of the selected protein families, the structure fluctuations and structure network are performed. . . . .	60
3.16	To understand the involvement of the proteins with the distortion of the organ behavior, a bar plot is performed of the disease shared by the proteins. . . . .	61

## LIST OF FIGURES

---

3.17	Depending on the P-value of the proteins associated with their respective diseases, the line graph is constructed. . . . .	62
3.18	Evaluation of intrinsic disorder predisposition of major SARS-CoV proteins: envelope (A), membrane (B), nucleocapsid (C), spike (D), ORF1a (E) and ORF1ab (F). These profiles were generated using DiSpi web crawler that aggregate the results from a number of well-known disorder predictors: PONDRVLXT (25), PONDRVSL2, PONDRVL3, IUPred short and IUPred long and PONDRFIT. . . . .	63
4.1	The local networks are established for (a) Oligodendrocyte, (b) M2 macrophage, (c) Stem cell, (d) Regulatory T cell, (e) Vascular endothelial cells, (f) B cell, (g) Astrocytes, (h) Endothelial cell, (i) Neural Stem cell, (j) Inhibitory neurons, (k) Macrophage, (l) Pericytes through single-cell study. The blue and red nodes of the graphs represent the transcription factors and hub genes, respectively. . . . .	77
4.2	The framework of the proposed method. . . . .	78
4.3	The violent color nodes in the graph represent the signaling pathways responsible for the GBM and their association with the cell types shown through an orange color in the graph. . . . .	80
4.4	The flowchart of the proposed method. . . . .	82
4.5	Lung single-cell RNAseq data analysis showed that ACE2 is highly expressed in pulmonary alveolar type II cells (PAT2). (A) The diverse cell types present in the lung are categorized into 17 clusters. (B) Violin plot is used to show the expression level distribution of ACE2 across the cell types. Including PAT2, the ACE2 is also expressed in PAT1, Clara, Club and Cilated cell types. . . . .	85
4.6	Bladder single-cell RNAseq data analysis showed that ACE2 is highly expressed in urothelial cells. (A) The diverse cell types present in bladder are categorized into 12 clusters. (B) A violin plot is used to show the expression level distribution of ACE2 across the clusters labelled with corresponding cell types. The plot shows smooth muscle cells 1 and 2 and basal cells 2, and proliferating fibroblast cells also possess significant expression levels of ACE2. . . . .	85



- 4.7 ScRNASeq data analysis of kidney uncovered that proximal tubule cells show the high expression level of ACE2. (A) The diverse cell types present in kidney are categorized into 11 clusters. (B) The violin plot is used to show the expression level distribution of ACE2 across the clusters. The clusters are labelled as corresponding cell types. ACE2 is expressed in multiple cell types such as in glomerular podocyte, smooth muscle cells, proximal straight tubules (STs), proximal convoluted tubules (CTs), distal tubule, principal cell, kidney progenitor cell, monocytes, B cell and T cell types. . . . . 86
  
- 4.8 The ScRNASeq data analysis of ileum revealed that ACE2 is highly expressed in enterocyte progenitor cells. (A) The diverse cell types present in ileum are categorized into 14 clusters. (B) The clusters in the violin plot are labelled with cell types depending on the present biomarkers. The expression level of ACE2 across the cell types depicts that it is expressed in most of the cell types of the ileum. . . 87
  
- 4.9 Liver single-cell RNAseq data analysis revealed that ACE2 is highly expressed in cholangiocytes. (A) The diverse cell types present in the liver are categorized into 8 clusters. The clusters are labelled with the name of cell types. (B) The Violin plot is used to show the expression level distribution of ACE2 across the cell types. Cell types hepatocytes and endothelial also show high expression level of ACE2 . . . . . 88
  
- 4.10 A cell-type-specific pathway semantic similarity graph: (A) mast cell type, (B) PAT2 and (C) plasma cell of the lung are established by considering the biological process associated with each pathway. Here, the nodes represent a particular pathway and their connecting edges define the weight between two pathways in order to apprehend the highest sharing biological process. The red-marked bold edge in the graphs indicates the higher association score between those pathways. . . . . 89
  
- 4.11 A) Urothelial cell type of bladder and two cell types. (B) Proximal tubule cells. (C) Smooth muscles of the kidney are considered for performing pathway semantic similarity graphs. Each color node of the graph represents a particular pathway associated with the cell type and the connection between the pathways is interpreted through the edges. The red-marked bold edge is used to exhibit the highest relationship between the pathways. . . . . 90

## LIST OF FIGURES

---

- 4.12 Considering the biological processes associated with each pathway, a semantic similarity graph is established for the following: (A) ciliated epithelial cells, (B) enterocyte progenitor cell type of ileum, (C) cholangiocytes cell type of liver. The nodes of the graphs represent a particular pathway and their connecting edges define the weight between two pathways in order to apprehend the highest sharing biological process. The red-marked bold edge in the graphs indicates the higher association score between those pathways. . . . 91
- 4.13 To map the results from pathway semantic and PageRank algorithm within the cell environment, in (A) for the PAT2 cell type of lung the top three ranked pathways excluding RAS and PPAR signaling pathway are shown, whereas the flowchart in (B) urothelial cells and (C) proximal tubule cells of organ bladder and kidney, respectively, show three ranked pathways including RAS and PPAR signaling pathway. For ileum (D) erythrocyte progenitor cells and (E) cholangiocytes of the liver, three pathways excluding RAS and PPAR signaling pathways are reported. As these two pathways are missing in the cholangiocytes cell type. These flowcharts of organ-specific cell types help to reveal the information regarding activating other pathways at the time of this epidemic disease. . . . 92
- 5.1 Representation of Cluster ID 1, 2, 3, 4, 5 and 6 with colors red, green, purple, yellow, brown and ash respectively. . . . . 96
- 5.2 The flowchart of the proposed framework to identify the cell types based on the pathway of cell markers from a single-cell perspective. 97
- 5.3 The violin plot of the cell markers in (a) Cluster 1, (b) Cluster 2, (c) Cluster 3, (d) Cluster 4, (e) Cluster 5 and (f) Cluster 6 respectively. . 98
- 5.4 To facilitate cellular heterogeneity and underlying molecular mechanisms of ovarian cancer, scPCN is applied. A. bar graph shows the top-ranked pathways alongside their activated genes. The colours employed represent the presence of pathways within their respective cell types. B. the heatmap is established to represent the association between disease-specific signalling pathways with their respective cell types. G. functional hub of ovarian cancer is identified. Connecting cell types, pathways, and their associated activated genes. The network is constructed by using a maximal clique centrality algorithm among cell types, pathways, and their associated activated genes. The crucial hubs are denoted by green and pink colours. . . . . 106

5.5	To explore the cellular heterogeneity and underlying molecular mechanisms of endometrial cancer, we employed scPCN analysis. The results are presented through various visualizations: (A). A bar graph displays the top-ranked pathways and their associated activated genes. The colors used indicate the presence of pathways within specific cell types. (B). A heatmap illustrates the association between the top-ranked disease-specific signalling pathways and their corresponding cell types. (C). By employing the maximal clique centrality algorithm, we identify the functional hub of endometrial cancer, which connects cell types, pathways, and their associated activated genes. Crucial hubs are highlighted in green and pink colors. . . . .	108
5.6	To investigate the cellular heterogeneity and underlying molecular mechanisms of cervical cancer, we have utilized the scPCN method. (A). A bar graph displays the top-ranked pathways alongside their activated genes. The colors used represent the presence of pathways within their respective cell types. (B). A heatmap is generated to represent the association between the top-ranked disease-specific signalling pathways and their respective cell types. (C). The functional hub of ovarian cancer is identified, connecting cell types, pathways, and associated activated genes. The network is constructed using the maximal clique centrality algorithm among cell types, pathways, and their associated activated genes. The crucial hubs are denoted by green and pink colors. . . . .	109
5.7	We generated a disease-specific pathway semantic similarity graph for three gynaecological cancers. This graph illustrates the interplay and communication between pathways that are ranked highest using scPCN . . . . .	111



# List of Tables

2.1	KEGG pathways of hsa-miR-146a-5p and hsa-miR-155-5p respectively	28
2.2	GO enrichment analysis of genes SPP1 and CDH6 respectively . . .	29
3.1	The data repositories used to fetch the data and utilized in this study are listed in this table. . . . .	38
3.2	Sharing common pathways with the nearest neighbour of the selected TFs from Protein-Protein Interaction having an impact on Prostate Carcinoma from the Reactome database. . . . .	50
3.3	The moonlighting function of the selected TFs of prostate cancer from MoonProt and MoonDB. . . . .	51
3.4	The List of Alternative Cellular Localizations in Terms of Liquid Liquid Phase Separation . . . . .	51
3.5	The 6 TFs of prostate cancer and their corresponding expression along with the drugs are reported from DrugBank. . . . .	51
3.6	Human-specific protein moonlighting activities of the two proteins similar to POU2F1. The evidence is curated from the MoonProt database after performing BLAST of the POU2F1 sequence. . . . .	56
4.1	The potential cell type determination with their hub gene markers identified by applying maximal clique centrality. . . . .	79
5.1	The potential cell type determination with their cell markers. . . . .	98
5.2	The associated pathways of the cell markers of each cluster . . . . .	99
5.3	The description of the datasets utilized in the study . . . . .	103
5.4	Clustering performance using Adjusted Rand Index scores for different methods with respect to state-of-the-art method. . . . .	105
5.5	Comparison of clustering performance among different methods using Normalized Mutual Information scores . . . . .	105
5.6	The common cancer pathways in three malignant diseases with variable gene sets . . . . .	110



# Introduction and Scope of the Thesis

## 1.1 Introduction

---

The availability of multi-omics data enables researchers to analyze complex biological processes at a systems level, rather than focusing on individual genes or proteins [1]. On the other hand, Systems biology is an interdisciplinary field that enables researchers to study biological systems over time, under varying conditions, and develop solutions to pressing health and environmental issues [2]. However, over the last few decades Systems Biology has witnessed significant growth due to the use of computational and mathematical analysis on multi-omics datasets, that provide an insight to reveal the molecular phenotypes. Therefore, the integration of multi-omics data and systems biology approaches has led to new opportunities for exploration and innovation in biology-based technology and computation [3]. Multi-omics allows for the design of predictive, multiscale models to discover potential biomarkers for disease diagnosis, monitor treatment efficacy, and identify drug targets. Moreover, various computational methods, including mathematical modelling, data integration, network analysis, high-performance computing, simulation, and visualization, are used to tackle new biological questions [4]. Network analysis is a particularly successful computational approach in systems biology [5]. Therefore, proper implementation of statistical and computational models can reveal the complex biological system and the intricate molecular mechanisms associated with pathogenic progression and control conditions.

## 1.2 Multiomics

---

When designing a multi-omic study, it is crucial to take into account the type of disorder being studied. Simple diseases caused by a single gene mutation typically have a small number of causative factors that play deterministic roles in the development of the disease. However, the severity or progression of these diseases may also be influenced by other factors such as modifier genes or environmental

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

factors. With the advancement of sequencing technology, biological studies have increasingly relied on data generated at these different levels of omics, collectively referred to as "multi-omics" data [6].

The primary focus of obtaining meaningful insights into cellular functions now lies in the analysis of multi-omics data and clinical information. Integrating multi-omics data, which provides comprehensive information on biomolecules from different layers, shows promise in systematically and comprehensively understanding complex biology. The integration of omics data, whether performed sequentially or simultaneously, enables the exploration of molecular interplay. These methods facilitate the evaluation of the transfer of information from one omics level to another and bridge the gap between genotype and phenotype. Integrative approaches, due to their holistic study of biological phenomena, have the potential to improve the accuracy of prognostics and disease phenotype predictions, thus contributing to better treatment and prevention.

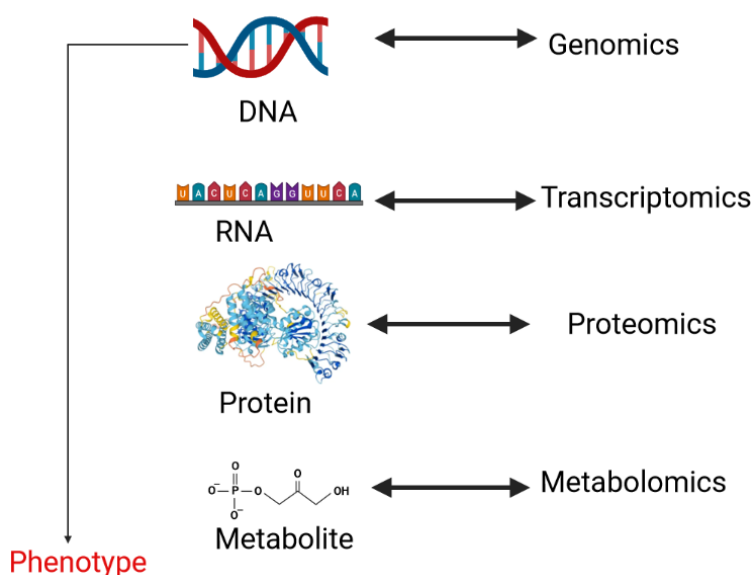


Figure 1.1: Integrative analysis of multi-omics data

### 1.2.1 Genomics

Genomics is the most established of the omics fields and plays a prominent role in medical research by pinpointing genetic variations associated with disease, treatment response, or future patient outcomes. This field investigates the complete set of genetic material of an organism and draws on elements from genetics. To sequence, assemble, and analyze the structure and function of genomes, genomics employs a combination of recombinant DNA, DNA sequencing techniques, and



bioinformatics. Unlike “classical genetics,” which examines one gene or gene product at a time, genomics examines an organism’s entire hereditary material. Additionally, genomics emphasizes interactions between loci and alleles within the genome, as well as other interactions such as epistasis, pleiotropy, and heterosis [7].

One successful approach in this area is Genome-Wide Association Studies (GWAS) [8], which has identified thousands of genetic variants linked to complex diseases across multiple human populations. GWAS studies typically involve genotyping thousands of individuals for over a million genetic markers and using significant differences in minor allele frequencies between cases and controls as evidence of association. Such studies have significantly contributed to our understanding of complex phenotypes. Associated technologies include genotype arrays, whole-genome sequencing using Next-Generation Sequencing (NGS), and exome sequencing.

### 1.2.2 Transcriptomics

Genomics enables the identification of genes present in an organism’s genome, whereas transcriptomics helps to determine the level of gene expression in different types of cells. In multicellular organisms, every cell generally carries the same genome i.e., the same set of genes. However, not every gene is expressed in every cell, leading to variations in gene expression patterns across different cell types. These differences in gene expression patterns are responsible for the wide range of biochemical, physical, and developmental differences seen among various cells and tissues and may also play a role in health and disease [9]. Therefore, by collecting and comparing transcriptomes of different cell types or tissues, scientists can gain a better understanding of the unique characteristics of each cell type and how changes in gene expression may be linked to the development or progression of diseases. For instance, transcriptomics can uncover the genes responsible for conferring unique features such as self-renewal and differentiation potential to stem cells, or identify the specific gene expression changes that accompany the onset and progression of cancer. Moreover, by examining the transcriptome, it is feasible to create a comprehensive map of the genes that are active during various developmental stages [10].

### 1.2.3 Proteomics

Our understanding of the human genome is expanding rapidly. Unraveling the complexities of our genetic blueprint can provide valuable insights into various health conditions and their potential remedies. But the genome is only part of the story. While identifying disease-causing genetic variations has become relatively

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

straightforward, understanding how these genes interact and function together is a more intricate task [11]. When investigating the reasons behind illness, it is not sufficient to merely examine genetic predispositions, which can be obtained through genome studies. Rather, it is crucial to examine the ongoing processes within the body. This necessitates an examination of the proteins produced by our genes. Proteins are produced from genes and collectively form the proteome. The genome acts as a set of instructions for protein formation. Each coding gene is responsible for producing distinct proteins, each with its own specialized role. The transition from genome to proteome occurs through two key steps: transcription and translation. Unlike the genome, the proteome is dynamic and undergoes active changes. Therefore, studying the presence or absence of proteins alone is insufficient for gaining comprehensive understanding [12]. Moreover, certain proteins become functional only in the presence of other proteins, forming complex interconnected networks. Nevertheless, proteins play a vast array of roles within the body, functioning as the active agents within cells. Unlike most other biological molecules (except for RNA), proteins are responsible for the majority of tasks involved in translating genetic information into other cellular components.

Proteomics, the study of the proteome, focuses on investigating how proteins interact with one another and the roles they play within organisms. It provides a holistic perspective on the underlying processes of healthy and diseased cellular functions at the protein level. Since the completion of the Human Genome Project, the study of proteins has emerged as one of the frontiers in biology and personalized medicine. However, scientists are still working towards developing effective methods for measuring our proteomes accurately.

### 1.2.4 Metabolomics

Metabolomics is a burgeoning field of omics within systems biology that emerged in the post-genomic era. It encompasses the study of all biochemical reactions regulated by genes and proteins, as dictated by the central dogma of molecular biology [13]. The metabolome refers to the complete collection of low molecular weight compounds involved in biochemical reactions. These compounds, which serve as substrates and by-products of enzymatic reactions, directly influence the cell's phenotype. Metabolites are crucial for maintaining normal physiological function in human cells and organs, and they play a vital role in intercellular signal transduction. Compared to genomics, transcriptomics and proteomics can only tell us what may happen in organisms, metabolomics can directly and accurately reflect the current status of organisms and tell us what has exactly happened in the organisms. In contrast, transcriptomics and proteomics can only offer insights into potential events within organisms. Monitoring cellular function solely based on mRNA or protein levels is inadequate due to factors such as RNA splicing or post-

### 1.3. INTEGRATION OF MULTI-OMICS AT SYSTEM-LEVEL

---

translation, which disrupt the simple relationship between these molecules and metabolism. The smaller scale of the metabolome, with approximately 3,000 commonly used metabolites in key metabolic pathways, in comparison to the proteome and genome, simplifies data analysis. Advancements in testing instruments and equipment have led to the development of next-generation metabolomics, characterized by enhanced sensitivity and accuracy compared to traditional techniques. This emerging discipline is proving to be a powerful tool for cancer diagnosis and treatment. Moreover, metabolomics finds practical applications in medicine, agriculture, and biotechnology. For instance, it enables the identification of potential drug targets, monitoring of disease progression, optimization of crop yields, and improvement of the quality and safety of food products.

### 1.3 Integration of multi-omics at system-level

---

In order to comprehend the regulatory mechanisms of complex biological systems, it is essential to examine all layers of omics, including genomics, epigenomics, transcriptomics, and proteomics. Individual omics layers play a distinct but interconnected role in relation to the others. Although significant mRNA expression does not always result in protein expression, many mRNA outputs are involved in regulating a set of proteins rather than a one-to-one correspondence. Thus, to gain a complete picture of gene regulation, it is necessary to analyze all the omics layers simultaneously at the systems level. This is known as the "Trans-Omics" approach, which reconstructs a global network across multiple omics layers using multi-omics measurements and data integration.

Another important omics layer that can enhance the resolution of the regulatory picture is metabolomics. This layer is different from other omics layers as it focuses on chemical processes involving metabolites from cellular processes, providing a direct physiological state for an active cellular function. Analytical technologies such as Mass spectrometry (MS) or Nuclear Magnetic Resonance (NMR) spectroscopy are used to study metabolomics, which can be difficult to interpret compared to proteomics approaches [14]. However, metabolomics is essential for studying multi-omics as it provides a supplementary layer of information.

In addition to metabolomics, incorporating additional interaction information can enhance the capacity of regulatory networks systematically. For example, HumanNet is a human functional gene network that integrates a series of omics data using Bayesian statistics, allowing for more flexible incorporation of network information into studies. Network analysis can extend or validate the existing network biology. A recent study proposed extending transcriptional drivers using physical and functional interactome networks, successfully identifying known coding drivers in cancer.

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

Therefore, to fully comprehend the underlying mechanism of biological phenotypes, it is necessary to examine a series of omics data and construct a relevant network that is based on the "central dogma" of information flow from transcription to translation. Each step of making mRNAs and proteins is composed of complex networks that interact with each other. Transcriptional regulatory networks, protein interactomes, and functional networks need to be incorporated into the relevant omics data using a systems approach. Furthermore, a transcriptional core regulator-centred approach can be used to identify a core regulator of transcriptional regulatory networks. Perturbing the core regulator can affect both transcription and translation, making it an advantageous approach for multi-omics studies. By using this systematic approach, we can successfully gain insight into the mechanisms underlying complex biological phenotypes.

### 1.3.1 Approaches

Sequencing data provides the foundation for systems biology approaches aimed at modeling and comprehending biological systems as a whole. By combining sequencing data from multiple omics layers, researchers can develop computational models and conduct network-based analyses to explore system-level properties, including emergent behaviors, robustness, and adaptability. These approaches facilitate a holistic understanding of the interactions and contributions of different molecular components to the overall system behavior.

Furthermore, integrating sequencing data across omics layers aids in the discovery of biomarkers with predictive or diagnostic value in disease research. By identifying molecular signatures across diverse sequencing datasets, researchers can pinpoint specific genes, proteins, or metabolites associated with disease phenotypes or treatment responses. These biomarkers are valuable for disease classification, patient stratification, and the development of personalized therapeutic strategies.

#### 1.3.1.1 Bulk analysis approach

With the advancement and decreasing costs of high-throughput bulk profiling technologies, an increasing number of studies have generated large-scale multi-modal data from various samples/individuals to systematically investigate various diseases. The Cancer Genome Atlas (TCGA) project has facilitated the functional interpretation of diverse molecular aberrations and potential mechanisms causing cancers based on multi-omics data. The International Cancer Genome Consortium (ICGC) has identified novel oncogenic mutations and new tumor subtypes for prognosis and therapeutic management through integrative analysis of multi-modal data from various cancer types/subtypes. Combining TCGA and ICGC

### 1.3. INTEGRATION OF MULTI-OMICS AT SYSTEM-LEVEL

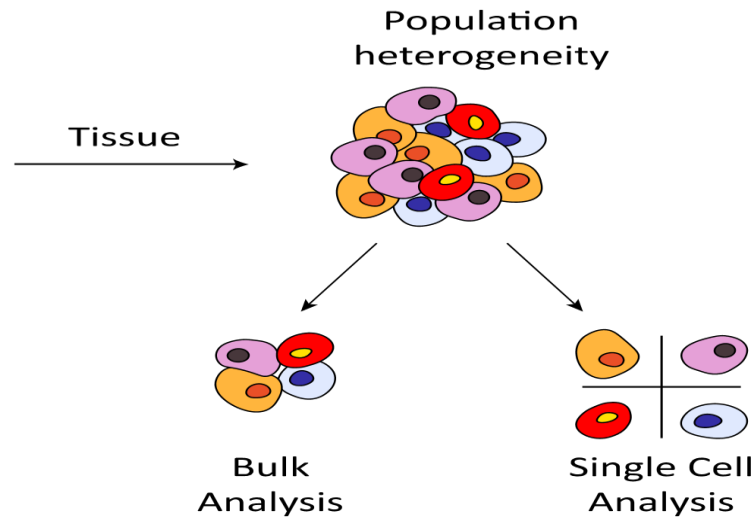


Figure 1.2: Bulk sequencing and single-cell sequencing: In bulk sequencing, a collection of RNA extracted from a group of cells is sequenced, resulting in expression data that represents the average expression of a specific gene across all cells. On the other hand, scRNA-seq (single-cell RNA sequencing) preserves the unique transcript information of each individual cell. When multiple cells of the same type are identified, averaging the sequence reads across all those cells provides cell-type-specific expression information similar to the profiles obtained through bulk RNA-seq.

multi-omics data provides a more comprehensive view of mutational processes and functional effects of mutations in diverse tumors. Bulk multi-omics is essential for systematically elucidating disease pathogenesis and various phenotypes at the individual level. However, bulk technologies mainly capture the averaged signals of a large number of cells for each sample, which ignores the heterogeneity and variations within cell populations.

#### 1.3.1.2 Single-cell analysis approach

Single-cell profiling has provided unprecedented opportunities to explore molecular functions at single-cell resolution and greatly facilitated the delineation of cellular heterogeneity [15]. Single-cell multimodal research has emerged in recent years due to technical limitations and high costs. For example, joint analysis of transcriptomic and epigenetic data of >60,000 cells from the human adult brain revealed cell-type-specific TFs and regulatory elements. Furthermore, single-cell integrative analysis enables the dissection of cellular dynamics and functions, as well as GRNs and cell-to-cell communications at the single-cell level, which cannot be done with traditional bulk strategies. Although continuous innovation in

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

single-cell methods will lead to a gradual shift from bulk multi-omics analysis to single-cell analysis, bulk approaches are still the workhorse for multimodal studies at present, and they are complementary with single-cell strategies to gain a whole-system and cell-based perspectives on health or disease.

### 1.4 Computational methods for analysing multi-omics data

---

Multi-omics data analysis typically involves the use of a range of computational techniques that are tailored to handle the complexity and diversity of the data. Here are some of the commonly used techniques in multi-omics data analysis:

#### 1.4.1 Integrative methods

These methods aim to integrate data from different omics layers to obtain a more comprehensive understanding of biological systems. Network-based approaches, correlation-based approaches, and machine-learning methods are commonly used in the integration.

#### 1.4.2 Clustering and classification methods

Classification and clustering are two distinct learning methods used in data mining to group objects based on their features. While they may seem similar, there are fundamental differences between them. Classification is a supervised learning technique that assigns predefined labels to instances based on their properties. On the other hand, clustering is an unsupervised learning method that groups similar instances together based on their features, without predefined labels [16].

#### 1.4.3 Pathway analysis

A pathway is a graphical depiction of a well-defined segment of the molecular machinery involved in physiological processes [17]. For instance, it can represent a metabolic pathway that describes enzymatic reactions within a cell or tissue, or a signaling pathway model that illustrates regulatory processes leading to downstream metabolic or other regulatory events. Typically, a pathway model begins with an extracellular signaling molecule that activates a specific receptor, initiating a cascade of molecular interactions. The main aim is to understand the implications of gene lists obtained from experiments, particularly omics and functional genomic studies. Mechanistic insights and interpretation of the significance of genes within these lists are gained through pathway analysis. Computational analyses employ specialized pathway representation formats. Such representations are often referred to as gene set enrichment analysis, pathway analysis, gene ontology analysis, or disease ontology analysis. However, in most cases, pathway

---

## 1.4. COMPUTATIONAL METHODS FOR ANALYSING MULTI-OMICS DATA

analysis specifically refers to the initial characterization and interpretation of experimental or pathological conditions studied using omics tools or genome-wide association studies. These studies frequently yield extensive lists of genes that have undergone alterations or variations.

### 1.4.4 Dimensionality reduction

High-dimensional data refers to datasets where the number of variables or features is significantly larger than the number of available samples or observations. This results in a high dimensionality, indicating a substantial number of attributes or measurements recorded for each sample. Dealing with such data presents challenges in terms of visualization, interpretation, and reliability of statistical models due to issues like overfitting and the curse of dimensionality. For instance, in genomics, a high-dimensional dataset might comprise gene expression levels for thousands of genes across multiple samples. To address this, dimension reduction techniques like principal component analysis (PCA), t-SNE (t-Distributed Stochastic Neighbor Embedding), or feature selection methods are employed [18]. These methods aim to extract the most informative features or reduce the dimensionality to a more manageable level while retaining crucial patterns and information. Additionally, they can help eliminate some noise in the data, thereby mitigating the risk of over fitting. Consequently, dimensionality reduction plays a valuable role in managing high-dimensional data, improving interpretability, and enhancing the reliability of statistical modeling.

### 1.4.5 Network analysis

Biological systems are commonly represented as intricate networks, which consist of binary interactions or relationships among various entities. Inherently, every biological entity has interactions with other biological entities, from the molecular to the ecosystem level, providing the opportunity to model biology using many different types of networks such as ecological, neurological, metabolic or molecular interaction networks. The advent of the omics era in biological research has resulted in a tremendous growth of data, prompting the need for more systematic approaches to data analysis that move beyond the traditional focus on single genes or proteins [19]. Network analysis serves as a valuable framework for comprehending these intricate interactions and unravelling the structure and function of complex biological systems. Through network analysis, regulatory relationships, signaling pathways, protein-protein interactions, gene co-expression patterns, and other vital connections within the networks can be revealed. This aids in deciphering the underlying mechanisms and processes at play in biological systems. Additionally, by examining network properties such as centrality mea-

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

asures (e.g., degree, betweenness, closeness), researchers can identify key genes, proteins, or metabolites that act as central hubs or critical mediators within the network. The identification of these key players holds great importance for understanding disease mechanisms, discovering biomarkers, and identifying potential therapeutic targets.

### 1.5 Approaches for integrating bulk multi-omics data

---

Traditional bulk multi-omics data typically involves analyzing a large number of cells from a given sample and assuming that all of these cells are homogeneous. This approach results in averaged signals that fail to capture cellular heterogeneity. Despite advancements in sequencing methods, bulk technologies remain advantageous due to their lower technical noise, simpler experimental procedures, and the ability to analyze the state (living/non-living) of cells. Additionally, generating bulk multi-omics data is generally less expensive than generating single-cell data, making it more affordable for investigating large-scale samples. Bulk approaches are also useful for identifying molecular signatures at the individual or tissue level and for conducting comparative omics analysis. There are several mathematical approaches that can be used for integrating bulk multi-omics data, including:

#### 1.5.1 Multi-step/multi-stage approaches

The ultimate goal of multi-omics data analyses is often to identify the underlying genetic and molecular mechanisms that drive phenotypic variation. To achieve this goal, multi-step strategies can be used in bulk multi-omics data analysis to link genotype to phenotype [20]. One common strategy is to first identify differentially expressed genes or proteins that are associated with a particular phenotype or disease. These genes or proteins can then be further analyzed to identify key pathways and biological processes that are affected by changes in their expression. Next, genomic data such as single nucleotide polymorphisms (SNPs) or copy number variations (CNVs) can be analyzed to identify genetic variants that are associated with changes in the expression of the differentially expressed genes or proteins. This can be done using methods such as genome-wide association studies (GWAS) [8] or quantitative trait locus (QTL) analysis. The identified genetic variants can then be further analyzed to understand how they affect the expression of the differentially expressed genes or proteins and to identify the underlying biological mechanisms that link genotype to phenotype. This can be done using methods such as eQTL analysis, which identifies genetic variants that are associated with changes in gene expression, or by integrating different types of omics data to identify key regulatory networks and pathways that are affected by changes in the genetic variants. Finally, the identified regulatory



---

## 1.5. APPROACHES FOR INTEGRATING BULK MULTI-OMICS DATA

networks and pathways can be further validated using functional experiments, such as gene knockdown or overexpression experiments, to confirm their role in driving phenotypic.

### 1.5.2 Bayesian methods

Bayesian methods are particularly useful in bulk multi-omics data analysis because they allow for the integration of prior knowledge, such as biological pathways and functional annotations, into the analysis [21]. This prior knowledge can be used to inform the model and help to reduce the noise and improve the accuracy of the results. Another useful application of this method is to identify biological markers or biomarkers, indicators of the severity or presence of some disease state. This is also used to estimate the probability of a patient having the disease based on their biomarker profile.

### 1.5.3 Network based approaches

In multiomics bulk sequencing, network-based approaches are employed to explore the intricate connections and interactions between various molecular components, including genes, proteins, and metabolites. By combining information from multiple omics datasets, network-based approaches provide a more holistic understanding of biological systems [19]. They enable the identification of key components, reveal functional relationships, and unravel complex regulatory mechanisms that might be missed when analyzing each omics layer in isolation. These approaches offer valuable insights into the synergistic effects and cross-talk among different molecular entities, facilitating the discovery of novel biomarkers, disease mechanisms, and potential therapeutic targets.

### 1.5.4 Matrix factorization

Matrix factorization is a powerful computational method used in multi-omics bulk sequencing analysis to uncover hidden patterns and extract meaningful information from high-dimensional data [22]. It involves decomposing a multi-omics dataset into lower-dimensional matrices or factors, which capture underlying biological signals and relationships [23]. This approach allows for the integrative analysis of multiple omics layers and facilitates the identification of shared and distinct features across different molecular components.

### 1.5.5 Multiple kernel learning

In the analysis of multi-omics bulk sequencing, diverse types of data, such as gene expression, DNA methylation, and protein abundance, are frequently measured.

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

These data types can be transformed into kernels, which are similarity matrices that capture the relationships or similarities between samples. By utilizing multiple kernels, each representing a distinct omics data type, Multiple kernel learning (MKL) is a technique that enables the integration of these kernels into a unified framework [24]. The main idea behind MKL is to learn the optimal combination of kernels or their weighted contributions to maximize the performance of a prediction or classification model. This can be achieved through various optimization techniques, such as convex optimization or regularization methods.

### 1.6 Bioinformatics techniques for Single-cell data

---

ScRNA-seq provides an opportunity to identify cell-to-cell variations in gene expression, thereby enabling researchers to investigate the molecular mechanisms underlying cell differentiation, development, and disease. The analysis of scRNA-seq data presents several computational challenges, including dealing with high levels of noise, low coverage, and sparsity of data. Therefore, computational methods have become essential for the analysis of scRNA-seq data. These methods aim to identify subpopulations of cells, classify cell types, infer cell trajectories, and detect differentially expressed genes between cell types or conditions.

#### 1.6.1 Quality Control and Normalization

The first step in scRNA-seq data analysis is quality control and normalization. Quality control aims to detect low-quality cells, such as those with low read counts, low expression, or high levels of mitochondrial RNA. Normalization is the process of removing technical variation between cells, such as differences in sequencing depth or library size.

#### 1.6.2 Feature Selection and Dimensionality Reduction

The next step is feature selection, which involves identifying the most informative genes for downstream analysis. Feature selection methods can be based on statistical tests, such as the t-test or Wilcoxon rank-sum test [24], or on machine learning techniques, such as random forests or support vector machines. Dimensionality reduction methods, such as principal component analysis (PCA) [25] or t-distributed stochastic neighbor embedding (t-SNE) [26], are then applied to reduce the high-dimensional data to a lower-dimensional space that can be visualized.

### 1.6.3 Clustering and Cell Type Identification

Clustering methods aim to group cells into subpopulations based on their gene expression profiles. There are several clustering methods available, including k-means, hierarchical clustering, and density-based clustering. Once cells are clustered, cell type identification methods can be used to annotate the clusters with known cell types, based on marker genes or gene ontology enrichment analysis [27].

### 1.6.4 Trajectory Inference

Understanding how cells transition from one state to multiple different states is a common question in biology. This question is relevant not only in developmental biology but also in comprehending cellular transformations during various biological processes like tumor formation, ageing, and immune responses. The advancement of single-cell RNA sequencing (RNA-seq) has provided a means to address such questions. By analyzing and comparing the transcriptomics profiles of individual cells, it is theoretically possible to reconstruct a "trajectory" or path that depicts how cells progress through different states [28]. Cells are positioned along this trajectory based on their "pseudotime" value, which represents their relative position on the trajectory and is unrelated to actual time points. The reconstruction principle is as follows: as "pseudotime" increases, cells are placed further away from the root. The trajectory can take the form of a simple path connecting one cell type to another or a complex tree with multiple branches. The general workflow of single-cell trajectory analysis is depicted in the accompanying figure.

### 1.6.5 Differential Expression Analysis

Finally, differential expression analysis aims to identify genes that are differentially expressed between cell types or conditions [27]. These methods can be based on statistical tests, such as the t-test or Wilcoxon rank-sum test [29], or on machine learning techniques, such as logistic regression or random forests.

## 1.7 Networks of Networks

---

At a biological level, our bodies consist of numerous networks that are interconnected and communicate at various levels. Our genetic material, cells, and organs are all part of a larger network that extends to our environment. Systems biology studies these networks at different scales and integrates their behaviours to develop hypotheses for biological function and gain insights into dynamic biological changes over time and space. To comprehend the complexity of biology, it is

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

insufficient to examine only one component of a system. Therefore, the “Network of Networks” framework provides a valuable perspective for understanding how multi-omics data can aid in unravelling the intricacies of systems biology in a more comprehensible manner.

The first step in this analysis is to use network theory, which involves the use of graphs consisting of nodes and edges, to represent the omics layers. Nodes typically represent molecular entities such as genes, proteins, metabolites, and drugs, or phenotypic entities such as diseases, while edges represent physical, functional, or chemical relationships between pairs of entities. In this regard, network-based approaches have been widely used in the last few decades as mathematical tools for modelling and analysing omics data, which have yielded valuable biological insights. The most significant challenge in network integration is to construct an integrated network representation that captures all gene-gene associations by combining all molecular networks.

### 1.7.1 Gene Regulatory Network

Gene regulatory networks (GRNs) are essential for understanding how genes are regulated and interact in biological systems [30]. They consist of transcription factors (TFs), genes, and microRNAs, with TFs acting as primary regulators. TFs control both genes and microRNAs, while microRNAs can regulate gene transcripts, particularly messenger RNA. GRNs involve the regulation of gene expression, mRNA degradation, translational repression, and more. In GRN models, genes, microRNAs, and TFs are represented as nodes, and the regulatory relationships are depicted as directed edges from TFs to genes and microRNAs, and from microRNAs to genes. The most commonly used model for representing GRNs is Boolean networks, where genes are represented as binary variables (on/off states) and interactions are defined by logical rules. This model simplifies gene regulation by focusing on qualitative behaviors rather than precise quantitative dynamics. However, other models such as Bayesian Networks, Petri Nets, Neural Networks, and Fuzzy Logic Networks are also available. Recently, graph-based algorithms and metrics have been employed to analyze network properties, identify key genes, and infer regulatory relationships within GRNs.

### 1.7.2 Gene coexpression network

A gene co-expression network (GCN) is an undirected graph in which each node represents a gene, and an edge connects a pair of nodes if there is a notable co-expression relationship between them. Gene co-expression networks [31] are of biological importance as co-expressed genes are often controlled by the same transcriptional regulatory program, functionally related, or members of the same

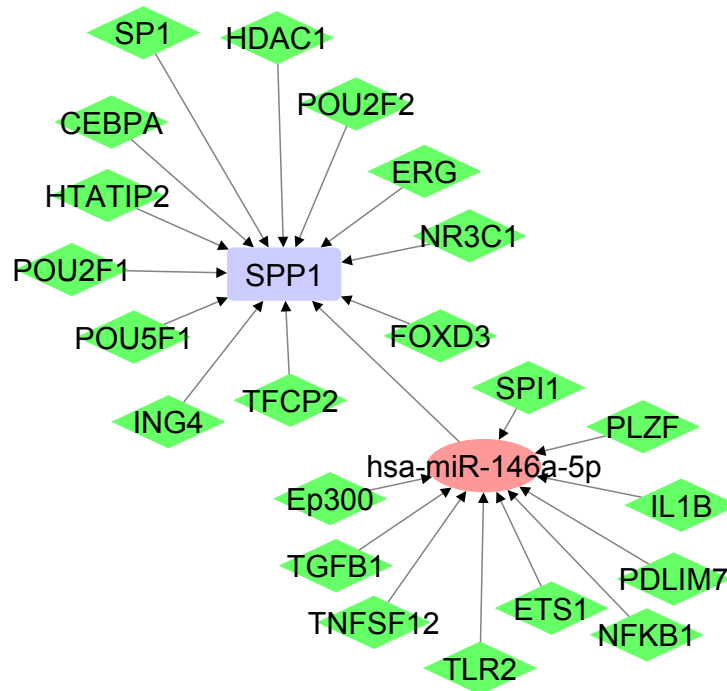


Figure 1.3: Gene regulatory network: a graphical representation illustrating the interactions and relationships between genes and molecular regulators, showcasing how they influence each other's expression levels. The green, blue and red coloured nodes represent transcription factors, genes and miRNA respectively, while the arrows indicate activation or repression of gene expression

pathway or protein complex. Unlike gene regulatory networks (GRNs), which depict directed edges representing biochemical processes such as reactions, transformations, interactions, activations, or inhibitions, gene co-expression networks do not determine the direction or type of co-expression relationships. In a GCN, the edges solely represent correlation or dependency relationships among genes.

Modules or highly connected subgraphs within gene co-expression networks correspond to clusters of genes that share similar functions or participate in common biological processes, leading to numerous interactions among themselves. Several methods have been developed to construct gene co-expression networks, generally following a two-step approach. Firstly, a co-expression measure is chosen, and a similarity score is computed for each gene pair using this measure. Subsequently, a significance threshold is determined, and gene pairs with similarity scores exceeding this threshold are considered to have a significant co-expression relationship, leading to the connection of corresponding nodes by edges in the network.

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

### 1.7.3 Protein-protein interaction

Recent advancements in genomics and proteomics technologies have facilitated extensive studies and generated vast amounts of data. Notably, substantial strides have been made in mapping protein interactions on a large scale, leading to a deeper understanding of both the composition of protein complexes and their organization within broader cellular protein-protein interaction (PPI) networks [32]. These networks often undergo perturbations in disease states, further emphasizing their importance in biological processes. Numerous reviews have focused on the technical developments related to the identification and characterization of PPIs and protein complexes. In Figure 1.4, the general architecture of the PPI network is shown.

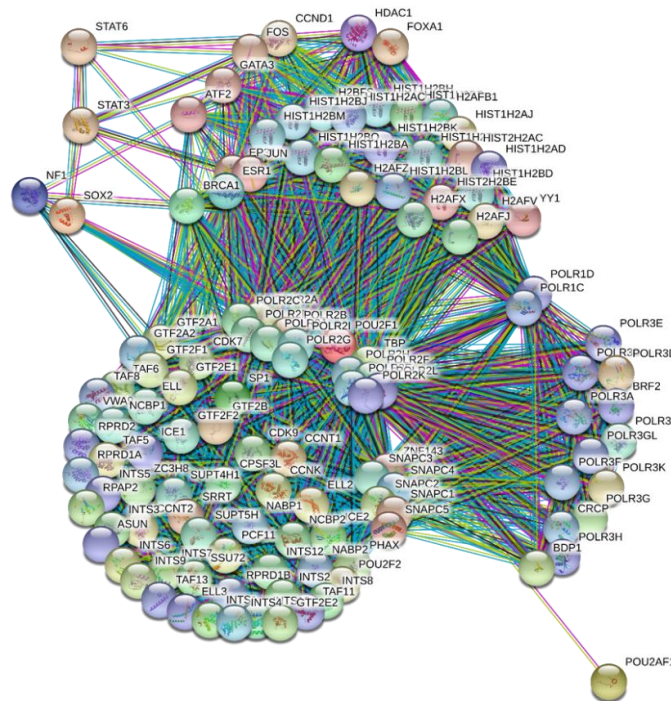


Figure 1.4: Analysis of the interactivity of human POU2F1 by STRING computational platform that produces the network of predicted associations for a particular group of proteins.

### 1.7.4 Metabolic Interaction

These networks represent the biochemical reactions and pathways involved in metabolism. Metabolic networks [33] can be used to simulate metabolic processes, predict the effects of genetic mutations or environmental perturbations, and iden-

tify potential drug targets.

### 1.7.5 Cell-signalling network

These networks represent the interactions between cells and their environment, such as growth factors, hormones, and neurotransmitters. Cell signalling networks [34] can be used to understand how cells respond to different stimuli, to identify key signalling pathways involved in development and disease, and to predict the effects of drug treatments.

### 1.7.6 Disease network

These networks represent the relationships between genes, proteins, and other biological entities that are involved in a particular disease or condition. Disease networks [35] can be used to identify key pathways and interactions that are disrupted in a particular disease, to predict potential drug targets, and to identify biomarkers for diagnosis and prognosis.

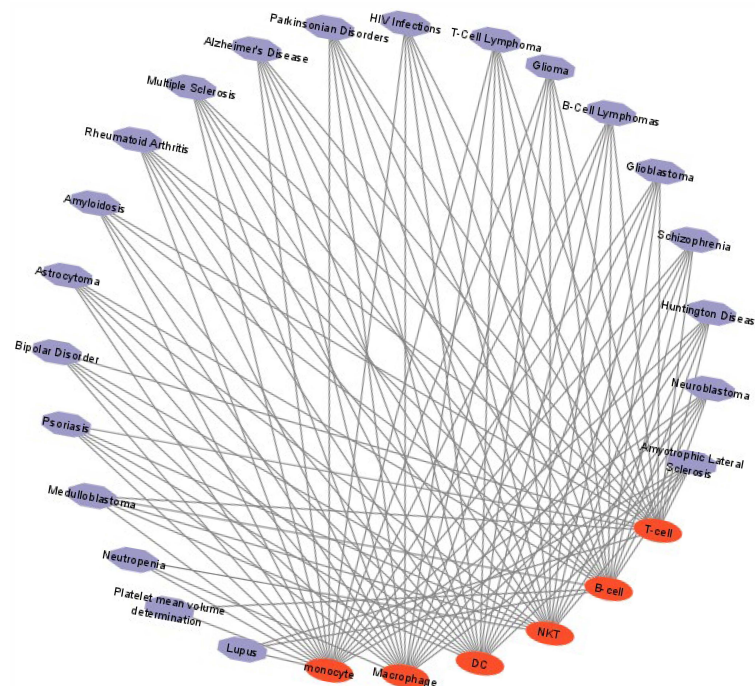


Figure 1.5: Disease network- a visual depiction highlighting the complex connections and associations between various diseases and related factors. Here, nodes represent different diseases and their associated cell types, while edges depict the relationships and interactions among them, offering insights into the interconnected nature of diseases and potential underlying mechanisms

### 1.8 General network properties

---

Networks, in general, can be described by a set of properties that provide insights into their structure and function.

#### 1.8.1 Centrality

Centrality refers to a set of measures used in network analysis to determine the importance or influence of nodes within a network. It quantifies the relative significance of nodes based on their connectivity and their position within the network structure [36]. In biological networks, such as protein-protein interaction networks or gene regulatory networks, centrality measures provide insights into the essential nodes involved in key biological processes. For example, degree centrality measures the number of connections of a node, while betweenness centrality measures how often a node appears on the shortest paths between other nodes in the network.

- **Degree Centrality:** This centrality measuring technique simply counts the number of edges (or connections) that a node has in a network. The degree centrality of node  $i$  is given by equation 1.1:

$$C_d(i) = k_i \quad (1.1)$$

where  $k_i$  is the number of edges that node  $i$  has in the network.

This equation essentially states that the degree centrality of a node is equal to the number of connections it has in the network. Nodes with high degree centrality are often referred to as "hubs", as they are highly connected to other nodes in the network. However, it does not take into account the strength or importance of connections, and it can be biased towards nodes that simply have many connections regardless of their actual importance in the network.

- **Betweenness centrality:** This measure of centrality quantifies how often a node appears on the shortest paths between pairs of other nodes in a network. The betweenness centrality of node  $i$  is given by:

$$C_b(i) = (\text{sum of shortest paths between all pairs of nodes that pass through } i) / (\text{sum of shortest paths between all pairs of nodes})$$

where the "shortest path" refers to the path between two nodes with the fewest number of edges.

This equation essentially calculates the fraction of shortest paths in the network that pass through node  $i$ . Nodes with high betweenness centrality are



---

## 1.8. GENERAL NETWORK PROPERTIES

often referred to as "bridges" or "bottlenecks", as they play a critical role in connecting different parts of the network.

- **Closeness centrality:** This technique quantifies how quickly a node can reach all other nodes in a network. The closeness centrality of node  $i$  is given by equation 1.2:

$$C_c(i) = \frac{N - 1}{\sum_j d(j, i)} \quad (1.2)$$

where,  $d(j, i)$  is the distance between vertices  $j$  and  $i$ . The  $N$  in the  $(N - 1)$  is the total number of nodes in the network.

This equation essentially calculates the average length of the shortest paths between node  $i$  and all other nodes in the network, and then takes the reciprocal to obtain a measure of "closeness". Nodes with high closeness centrality are often referred to as "central" or "well-connected", as they can quickly reach other nodes in the network. However, it may be less relevant in networks where the distance between nodes is less important, or in networks where there are many disconnected components.

- **Eigen vector centrality:** This centrality measure that takes into account the centrality of a node's neighbors. The idea is that a node is important if it is connected to other important nodes. The eigen vector centrality of node  $i$  is given by equation 1.3:

$$C_e(i) = (1/\lambda) * \sum_j (A_{ij} * C_e(j)) \quad (1.3)$$

where  $A$  is the adjacency matrix of the network,  $\lambda$  is the dominant eigenvalue of  $A$ , and  $C_e(j)$  is the eigen vector centrality of node  $j$ .

This equation essentially states that the eigenvector centrality of a node is proportional to the sum of the eigenvector centralities of its neighbours, weighted by the strength of their connections. The constant  $1/\lambda$  is a normalization factor that ensures that the centrality values are bounded and comparable across different networks. The calculation of eigenvector centrality requires the use of linear algebra techniques, such as eigenvalue decomposition.

### 1.8.2 Modularity

Community detection is a popular approach used in the analysis of biological networks to identify groups or communities of nodes that have similar topological or functional properties. In biological networks, such communities often represent groups of genes or proteins that are involved in similar biological processes,

## CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS

---

pathways, or functions. There are several algorithms and methods that have been developed for community detection in biological networks. One of the most popular approaches is the modularity-based algorithm, which aims to maximize the modularity score of a network by partitioning it into communities [20]. It is a measure of the degree to which a network is divided into clearly delineated groups of nodes. Other approaches for community detection in biological networks include hierarchical clustering, k-means clustering, spectral clustering, and stochastic block models. These methods are often combined with biological knowledge or functional annotations to refine the identified communities further and provide insights into the underlying biological processes and mechanisms. Overall, community detection in biological networks is an important tool for understanding the complex relationships and interactions between genes and proteins, and for identifying potential drug targets and therapeutic interventions.

After exploring the various biological and computational aspects related to solving biological questions, a summary of the thesis has been presented. The scope of the thesis and potential future directions are also discussed. Additionally, chapter-wise summaries are provided.

### 1.9 Scope of the Thesis

---

The aim of this thesis is to employ integrative data analysis techniques to gain a comprehensive understanding of cellular processes and molecular mechanisms. To tackle the complexities of biology, various computational methods such as mathematical modeling, network-based architecture, graph theoretical modeling, and unsupervised learning techniques are utilized. The thesis is structured into six chapters, encompassing an introduction and conclusion. The remaining four chapters contribute by presenting a network-based framework that captures interactions among multiple omics layers through graph representations. These graphs aim to accurately depict the molecular connectivity within cells. Each chapter is briefly summarized in the subsequent subsections, providing an overview of their respective titles.

#### 1.9.1 Studying the association between gene and its regulators using regulatory network

In Chapter 2, the computational frameworks used for studying prostate cancer and kidney diseases are discussed. The gene regulatory network approach was utilized to establish the association between genomic and transcriptomic data. Both studies successfully identified biomarkers responsible for the respective diseases, which were found to be regulated by both transcription factors and miRNA. Furthermore, the path leading from genomic variation to disease formation was analyzed.

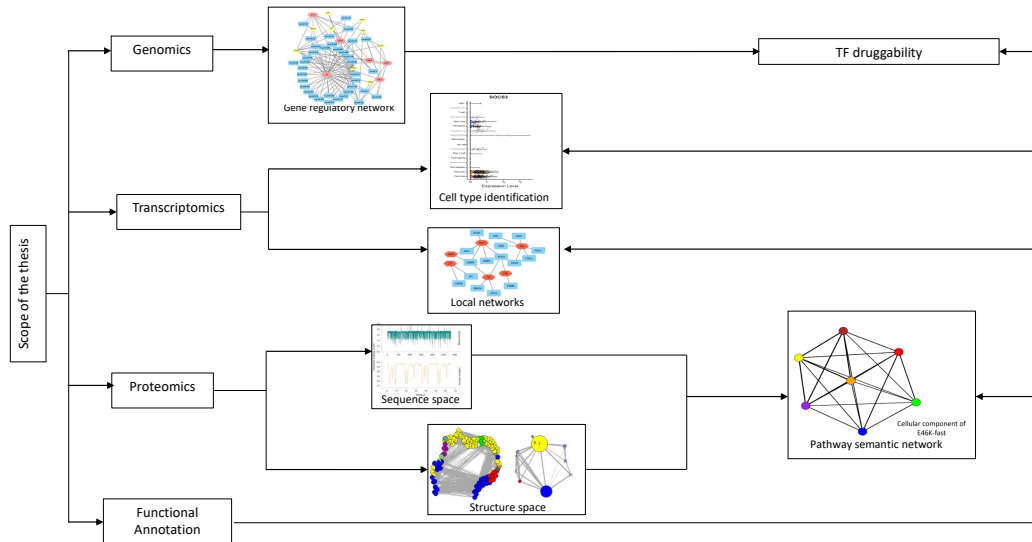


Figure 1.6: The graphical abstract to summarize the researches covered under the thesis

### 1.9.2 Understanding the dynamicity of proteins in terms of network

Chapter 3 delves into the sequence structure-function paradigm of proteins, specifically focusing on their characteristics as transcription factors. Amidst the pandemic of 2020, much research was geared towards finding therapeutic solutions. Most of the attention was given to the druggability of the various SARS-CoV-2 proteins, with very few studies considering their structural malleability and intrinsic disorder potential. However, the framework previously proposed for uncovering the structural malleability of transcription factors is utilized to analyze the evolutionary sequence-structure of these viral proteins.

### 1.9.3 Graph-theoretical modeling to unveil the cell-to-cell heterogeneity

The primary objective of chapter 4 is to develop a computational framework that can uncover the cell-to-cell heterogeneity of multicellular organisms during disease initiation, development, and progression. To achieve this goal, two cases are considered. The first case focuses on glioblastoma multiforme, an aggressive malignancy that affects the brain and spinal cord. Despite advancements in clinical strategies, this disease has a poor prognosis. A bioinformatics pipeline is proposed in this study to comprehend the cell-to-cell heterogeneity and its impact on tissue development. Additionally, the impact of various cell types on a particular tissue is observed. The second case examines single-RNASeq data from organs affected by

## **CHAPTER 1. INTRODUCTION AND SCOPE OF THE THESIS**

---

COVID-19. An organ-specific framework is proposed to identify the involved cell types during disease initiation. Moreover, a pathway semantic model is utilized to decode the contribution of the internal pathway network of each participating cell type towards the infection.

### **1.9.4 Computational framework for pathway-based inference of single-cell RNA-seq data**

Chapter 5 addresses the key role of pathways in cell-to-cell heterogeneity. A framework is proposed on two datasets in order to improve the clustering results by incorporating pathway information. In addition, a novel pathway activity score (PAS) calculation technique is developed. This proposed method resulted in more precise cell clustering and the identification of cell-specific pathways for gynaecological malignancies, which had not been previously studied as per our knowledge.

### **1.9.5 Conclusions and Future Scope**

The final chapter 6 of this thesis provides a summary of its contributions, highlighting key observations from each chapter. Additionally, it presents several areas for future research aligning with the thesis theme.

## Studying the association between gene and its regulators using regulatory network

### 2.1 Introduction

---

Regulatory network analysis involves the construction and analysis of gene regulatory networks (GRNs), which depict the regulatory interactions between genes and their regulatory elements, such as transcription factors (TFs), and miRNAs. These regulatory elements form a dynamic network of interactions, collectively known as GRNs, which control gene expression in a highly coordinated manner. By representing genes as nodes and regulatory interactions as edges, GRNs provide a visual representation of the complex regulatory relationships that govern gene expression. Traditional reductionist approaches to studying gene regulation have largely focused on individual genes or specific regulatory elements. While valuable insights have been gained through such studies, they often fail to capture the complex and interconnected nature of gene regulation. In contrast, network biology further extends the analysis of GRNs by integrating multiomics data, leveraging computational tools, and exploring network properties and dynamics. This approach allows for a systems-level understanding of gene regulation, taking into account the interconnectedness and emergent properties that arise from the collective behavior of genes and their regulators.

The objective of this chapter is to explore the association of GRN and network biology to investigate the association between genes and their regulators in diverse biological contexts. Therefore, we proposed a framework by employing computational and analytical techniques. The aim of this framework is to identify key regulatory elements and elucidate regulatory modules and pathways for two diseases such as kidney renal cell carcinoma and prostate cancer. Furthermore, we seek to explore the implications of these findings in the context of disease development, biomarker discovery, and potential therapeutic targets.

## CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK

---

### 2.2 Computation approaches to unveil the association between gene and regulators

---

From the literature, it is evidenced that the dysregulation of miRNAs and transcription factors (TFs), two crucial mediators of gene expression, influences the prognosis of cancer. In this regard, the regulatory network helps to understand the characteristics of prognostic target genes and their regulators present in the network. During the studies, it is observed that some miRNAs and TFs are very conserved across cancers, despite the fact that the target genes and co-regulatory patterns show cancer-specific characteristics.

#### 2.2.1 Data description for KIRC

The genes associated with KIRC are collected from TissGDB. In this database, 2461 tissue-specific genes are curated for 22 tissue types that match with 22 diverse cancer types. Similarly, miRTarBase [37] is used to identify the interaction between the miRNAs of those targeted to these selected genes. The database contains more than fifty thousand interactions and also, and they are validated experimentally. These interactions are updated and collected manually from a literature survey of research articles related to functional studies of miRNAs. The identified genes are targeted by the number of TFs which are found with the help of the TRRUST database. Finally, the miRNAs and TFs interactions are extracted from the TransmiR database [38]. TransmiR is the database in which regulatory interaction between miRNAs and TFs is found. Currently, the database contains 3,730 literature-curated TF-miRNA interactions among which approximately 623 TFs and 785 miRNAs of 19 organisms from 1,349 publications are found. They also provide many interactions derived from Chip-Seq evidence in five different species.

#### 2.2.2 Data selection for Prostate Cancer

294 prostate tissue-specific genes from the TissGDB database are extracted. The database contains, 2461 tissue-specific genes, curated for 22 tissue types that matched with 22 diverse cancer types. Among the prostate tissue-specific genes, 113 differentially expressed (DE) genes measured by TissGDB are considered to establish the networks. The miRTarBase database [37] includes experimentally verified 502 653 miRNA-target interactions of humans. For the 113 genes, DE in prostate cancer, we identified 1408 experimentally verified miRNA-target pairs from the miRTarBase database. We constructed a network based on the identified interaction pairs. The 113 DE genes are further considered to identify the Transcription factors (TFs) from the experimentally verified TF-gene pairs available in

## 2.2. COMPUTATION APPROACHES TO UNVEIL THE ASSOCIATION BETWEEN GENE AND REGULATORS

the TRRUST database. Moreover, the miRNAs from the established miRNA-gene interactions are pondered for constructing TF-miRNA pairs, where TFs regulate the miRNAs from the TransmiR database.

### 2.2.3 Establishing gene regulatory network

Tissue-specific genes are selected that are responsible for KIRC and among these genes, differentially expressed genes are considered for further analysis. Moreover, two networks are constructed depending on the genes and their targeted miRNAs and genes with their targeted TFs. A third network is also formed by considering those TFs and miRNAs responsible for targeting the genes to find the relation between these TFs and miRNAs and how they interacted with each other as the objective is to analyze the mode of interaction among these three genetic molecules and their contribution towards KIRC. In this regard, from these three networks, a relation is established between gene and miRNAs and genes and TFs and also TFs and miRNAs. These networks are divided depending on the high modularity score. This is an optimizing method to detect the community structure present in the network. It is an NP-hard problem and desired in equation 2.1:

$$Q = \sum_{i=1}^K (e_{ii} - a_i^2) \quad (2.1)$$

Whereas  $e_{ii}$  represents the probability edge is in module  $i$  and  $a_i$  is the probability of a random edge that would fall into module  $i$ .

### 2.2.4 Detection of the functional hub using modularity

In this research, modularity is used to identify the relation among the three genetic molecules and how they regulate each other depending on the cluster they have formed. Through the proposed framework three objectives are possible to compute, those are: (a) the genes are regulated by TFs or miRNAs individually, (b) the genes are targeted by both miRNAs and TFs and regulated by these two molecules and (c) the genes are regulated by miRNAs but also passively targeted by TFs, i.e., TFs targeted the miRNAs and those miRNAs targeted genes and prevent them from protein translation and become the cause of this disease.

### 2.2.5 Biological validation

The observed functional hub molecules of the KIRC disease are validated to understand the biological significance. In this regard, the gene, TF and miRNAs of a network are analyzed by using the KEGG pathway and Gene ontology analysis.

## **CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK**

---

Similarly, in the prostate cancer study to strengthen the established GRN, the pathway of each selected TFs, miRNAs and gene is considered from the KEGG pathway. It is evidenced that they share a number of pathogenic and non-pathogenic pathways. The information on pathways is utilized to construct the final relational network among the three molecular regulators.

### **2.3 Experimental Results and Discussion**

---

This section consists of two sets of experimental outcomes. Also, detailed descriptions of the publicly available data have been provided. The experimental results of the two research have been shown in detail.

#### **2.3.1 Computational transcriptomic approach to provide an idea of the mechanism of gene regulation leads to KIRC disease**

The main aim of this study is to find the impact of TFs and miRNAs in gene regulation that leads to KIRC disease. As it is known that TFs and miRNAs are responsible to regulate gene expression by promoting or suppressing transcription, which has a significant contribution towards tumor formation.

##### **2.3.1.1 Initialization of the networks**

In order to find the mode of regulation and also the pathway through which these genes are targeted, three different networks are constructed. From the gene database, 256 tissue-specific genes are found for KIRC. Among these genes, 56 genes are found differentially expressed and also targeted by 105 TFs, which are identified from TRRUST, and the network is shown in Figure 2.1. Thereafter, these 56 genes are considered for further analysis and 885 interactions are found between genes and miRNAs.

##### **2.3.1.2 Identification of common markers regulated by both TFs and miRNAs**

It is important to consider genes that are regulated by both the TFs and miRNAs, to construct the gene regulatory network. In this regard, a Venn diagram is performed to identify the common genes. Further analysis of these genes helps to unveil the regulation pathway through which gene regulation leads healthy human health to disease initiation. Finally, the result of the Venn diagram, is shown in Figure 2.2. As a result, 48 genes are found.



## 2.3. EXPERIMENTAL RESULTS AND DISCUSSION

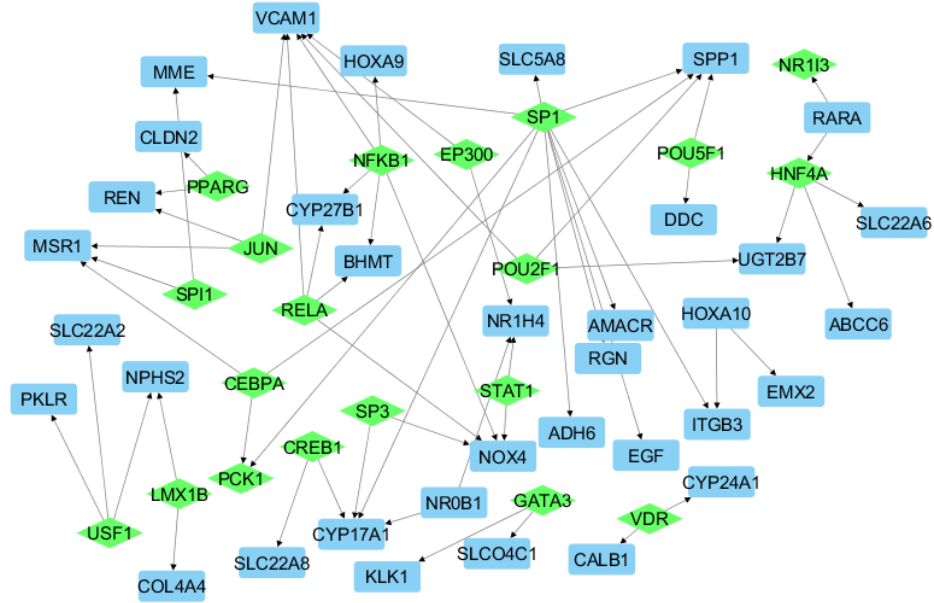


Figure 2.1: An interactive network between TF and their targeted host genes of KIRC.

### 2.3.1.3 Resultant network and hub gene identification

Depending on the results from the Venn diagram, an interaction network is established between 48 genes and 670 miRNAs, which contains 885 combinations. However, these 48 genes are also considered in order to find the TFs, and as a result, 105 TFs are found responsible which is reported through a network diagram in Figure 2.1. However, in Figure 2.3, the 105 TFs are considered to find the targeted miRNA among 670 miRNAs and it is found that only 35 TFs are found responsible for these miRNAs. From these networks, conditions (a) and (b) can be analyzed easily, whereas for the third condition these three networks and merged in order to find the relation between TFs genes and miRNA. In this regard, the networks are also divided depending on the modularity and compared with the merged network. As a result of the proposed framework two genes SPP1, and CDH6, those having a significant impact on KIRC are found as a target of different miRNAs as well as TFs but among all these miRNAs some miRNAs are also regulated by different TFs. The four networks for the relation among gene-miRNA-TF are shown in Figure 2.4. We can conclude that even genes are targeted by miRNAs but from this framework, it is capable to understand the regulation of miRNAs

## CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK

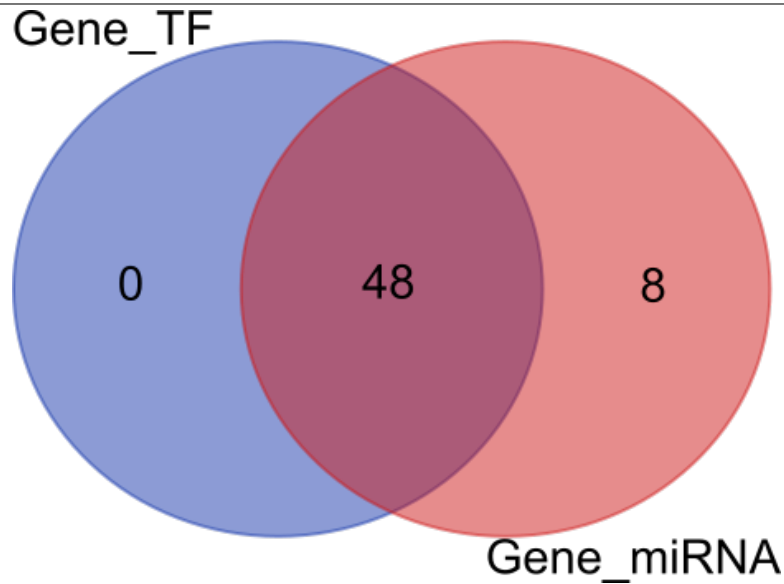


Figure 2.2: Venn diagram representation of the common genes targeted by both miRNAs and TFs.

and also the effect of TFs in gene regulation.

### 2.3.1.4 Biological validation of the hub genes

For biological validation, KEGG pathway analysis is performed shown in Table 2.1 for the two interesting miRNAs and it is found that they are active in different cancer pathways such as *Bladder cancer*, *Prostate cancer*, *NF-kappa B signaling pathway* and *ErbB signaling pathway* etc. Similarly, GO enrichment analysis is performed reported in Table 2.2 for the targeted genes of these miRNAs and it is found some important activities that are having a significant roles in cancer development.

Table 2.1: KEGG pathways of hsa-miR-146a-5p and hsa-miR-155-5p respectively

miRNA-mRNA Pairs	Pathway	Adjusted p-value
hsa-miR-146a-5p	<b>hsa04620</b> Toll-like receptor signaling pathway	5.5E-03
	<b>hsa04064</b> NF-kappa B signaling pathway	9.3E-03
	<b>hsa04012</b> ErbB signaling pathway	2.6E-02
	<b>hsa04110</b> Cell cycle	3.3E-02
	<b>hsa05162</b> Measles	4.7E-02
hsa-miR-155-5p	<b>hsa04064</b> NF-kappa B signaling pathway	2.2E-03
	<b>hsa05212</b> Pancreatic cancer	2.2E-03
	<b>hsa05219</b> Bladder cancer	4.5E-03
	<b>hsa05220</b> Chronic myeloid leukemia	1.5E-02
	<b>hsa05215</b> Prostate cancer	4.1E-02

## 2.3. EXPERIMENTAL RESULTS AND DISCUSSION

Table 2.2: GO enrichment analysis of genes SPP1 and CDH6 respectively

mRNA-mRNA Pairs	Subcategories	GO Term	Adjusted p-value
SPP1	Biological Process	GO:0000902 Cell morphogenesis	4.5E-03
	Cellular Component	GO:0005788 Endoplasmic reticulum lumen	3.1E-02
	Molecular Function	GO:0000983 Transcription factor activity	1.6E-02
CDH6	Biological Process	GO:0008209 Androgen metabolic process	3.0E-03
	Cellular Component	GO:0016342 Catenin complex	1.4E-03
	Molecular Function	GO:0042803 Protein homodimerization activity	3.3E-02

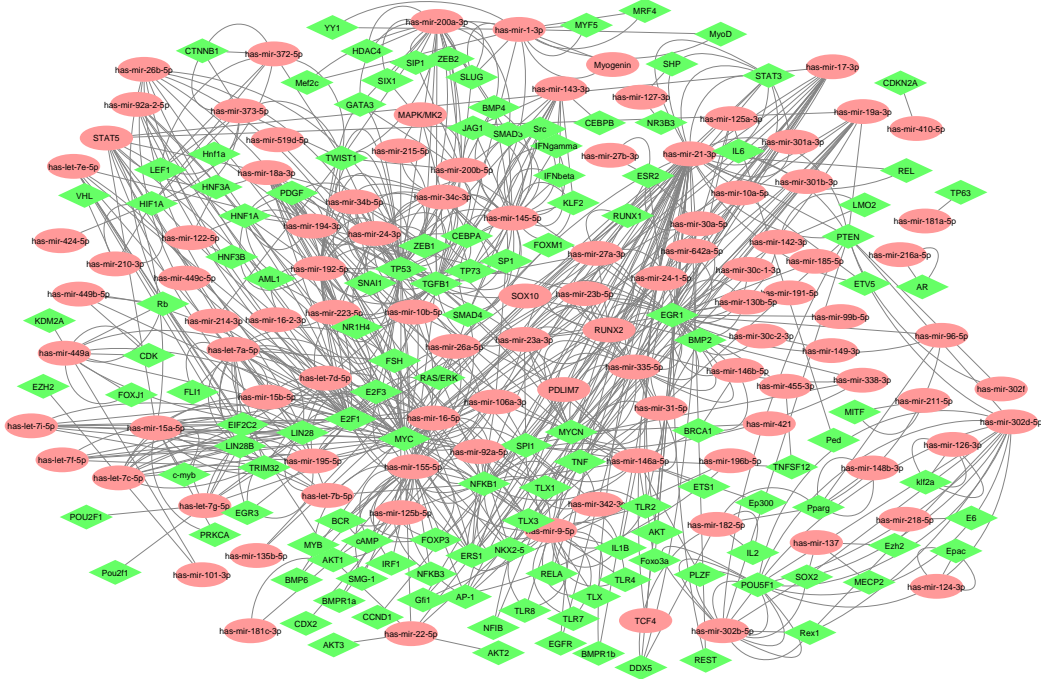


Figure 2.3: An regulatory network between TFs and their targeted miRNAs those having significant role in KIRC.

### 2.3.2 Experimental results of the prostate cancer-specific gene regulatory network

From the literature, it is evidenced that, gene regulatory network plays a vital role in malignancies in this regard, a framework is proposed where the sharing functional annotations of GRN have been considered. The experimental results of prostate cancer have been explained and discussed elaborately.

#### 2.3.2.1 Establishing network between genes and its targets in prostate carcinoma

The objective of the work is to understand the influence of the TF during the regulation of gene expression directly or indirectly (through miRNA). In this regard, a

## CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK

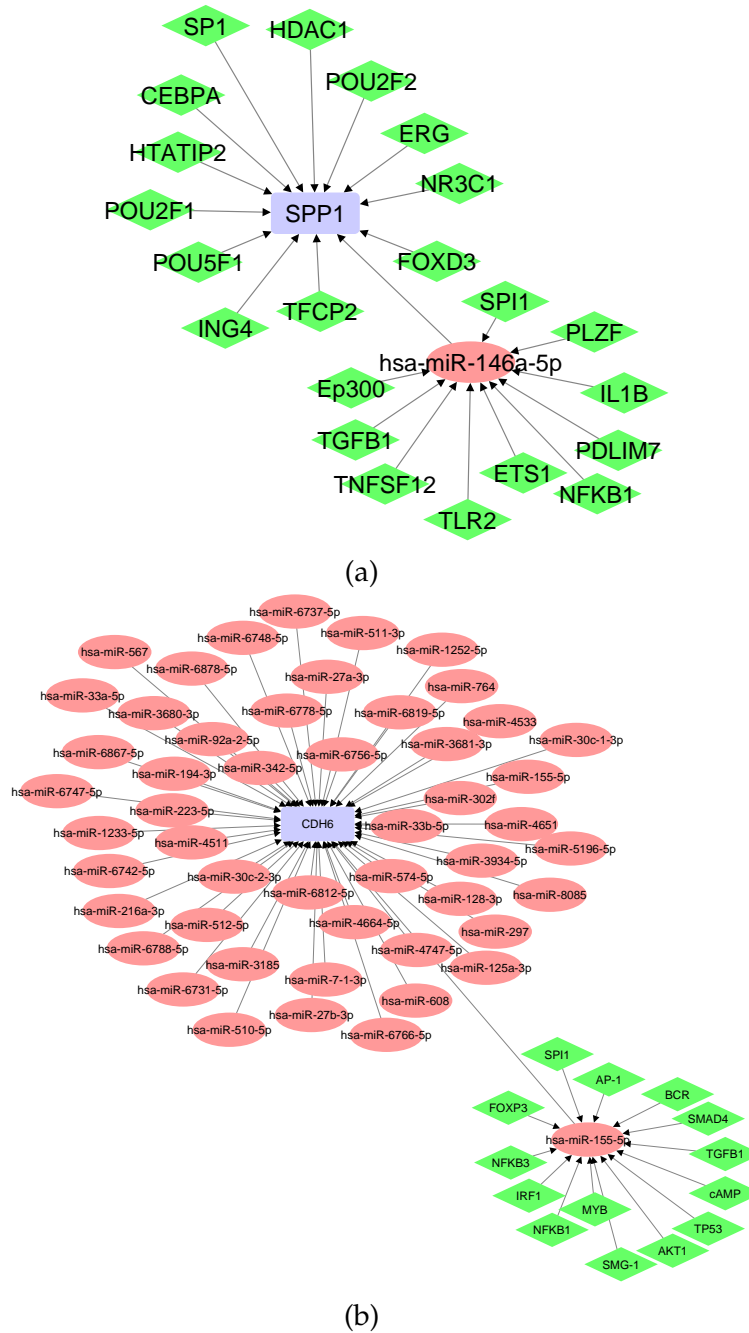


Figure 2.4: The interaction among genes, miRNAs and TFs (a) the interaction between SPP1 and hsa-miR-146a-5p which is targeted by different TFs and (b) the interaction between CDH6 and hsa-miR-155-5p which shows how TFs regulate the miRNAs.

## 2.3. EXPERIMENTAL RESULTS AND DISCUSSION

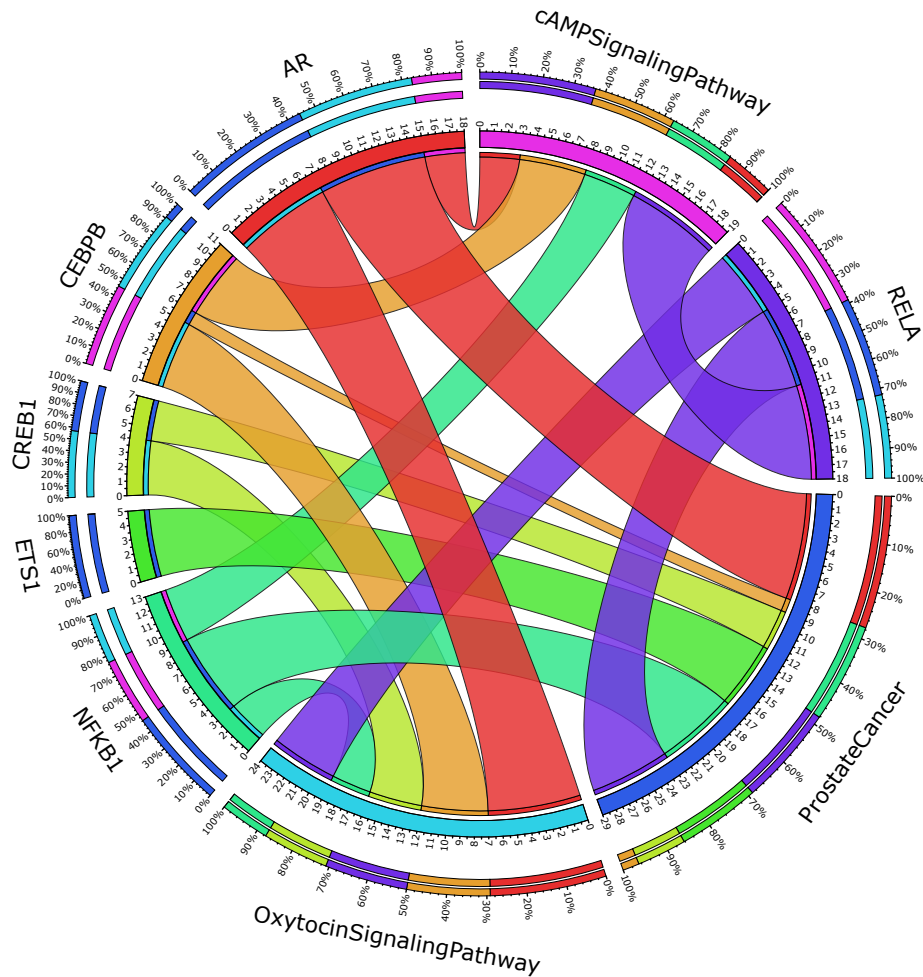


Figure 2.5: The circos plot to represent the association of the selected TFs with the shared pathways. Four pathways are found common among the selected molecular regulators including prostate cancer. The color and different size of the ribbon shows the type of relation depending on the p-value of the pathway and its corresponding TFs. The circos plot is used for better visualization of the tabular data. Here, the ribbons are connected between the TFs and their sharing pathway. The extent of the association is represented through the thickness of the ribbon and this thickness is based on the p-values of the pathways for each TFs. Ribbons touch the segment of the inner circle to define the row value whereas the ribbons do not touch the segments. Moreover, segmentation provides the absolute scale of the specific region of interaction between TFs and pathFA. The percentage of the outer circle indicates the overall total of each segment respectively.

two-fold analysis is performed. First, we established the relationships among the genes, miRNAs, and TFs in the context of PCa. Second, an in-depth analysis of the

## CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK

---

sequence and structure of the selected TFs that have significant associations with the pathology of PCa.

First, we built miRNA-gene and TF-gene networks for the differentially expressed genes responsible for prostate carcinoma. Subsequently, we identified TF-mediated miRNA regulation from the aforementioned TFs and miRNAs. To gain confidence, we have only considered experimentally verified regulator-target associations. Finally, a network is constructed among the genes, miRNAs and TFs. This network heeds only those TFs responsible for both direct and indirect regulation of the FFL. As a result, 183 combinations among 8 genes, 39 miRNAs and 28 TFs are found in the final network. In the network, genes, miRNAs and TFs are represented through a blue rectangle, yellow triangle and red oval respectively.

From 28 TFs we prioritized 6 highly responsible and influential TFs in prostate carcinoma based on their sharing of functional pathways (pathFA). We determined four pathFA that are shared by these 6 TFs. These pathFAs played important roles in cancer including prostate cancer. Furthermore, to understand the path through which the TFs are regulating the miRNAs and genes, the selected four pathFA (CAMP Signaling Pathway, Oxytocin Signaling Pathway, Pathway in Cancer and Prostate Cancer) are highlighted extremely. The association of the TFs with the pathFA are shown in Figure. 2.5 by using a circos plot, but the plot shows only three pathFA as the pathway in cancer is the mother of all the pathways involved in cancer formation.

In the ciros plot, three pathways are shown, those are regulated by most of the 6 selected TFs. Later, these pathways are considered to understand through which path the TFs regulate the miRNA and genes present in the FFL. During the selection of the FFL have an impact on the three pathways 45 combinations are found. Those combinations contain 8 genes, 38 miRNAs and 6 TFs. Moreover, it is evidenced that these 6 TFs have a key role in prostate cancer. In Figure.2.6, the final gene regulatory network is established. As the TFs are selected based on the functional perspective, therefore, it is expected that the final GRN is responsible for regulating the GRN core as well as a functional modification during prostate cancer. Interestingly, the TF AR has also been observed as differentially expressed at the RNA level in prostate cancer which is targeted by two TFs- ETS1 and CREB1 in two distinct FFL.

### 2.3.2.2 Identification of influential TFs for prostate carcinoma

From 28 TFs we prioritized 6 highly responsible and influential TFs in prostate carcinoma based on their sharing of functional pathways (pathFA). We determined four pathFA that are shared by these 6 TFs. These pathFAs played important roles in cancer including prostate cancer. Furthermore, to understand the path through which the TFs are regulating the miRNAs and genes, the selected four pathFA





## CHAPTER 2. STUDYING THE ASSOCIATION BETWEEN GENE AND ITS REGULATORS USING REGULATORY NETWORK

---

is evidenced that these 6 TFs have a key role in prostate cancer. In Figure 2.6, the final gene regulatory network is established. As the TFs are selected based on the functional perspective, therefore, it is expected that the final GRN is responsible for regulating the GRN core as well as a functional modification during prostate cancer. Interestingly, the TF AR has also been observed as differentially expressed at the RNA level in prostate cancer, which is targeted by two TFs- ETS1 and CREB1 in two distinct FFL.

### 2.4 Conclusion

---

In the kidney cancer study, we analyzed the gene regulation for KIRC and also tried to understand the role of TFs and miRNAs that are responsible to target genes and regulate their expressions. In this work, there are actually three objectives that are described in the method section. From the network of miRNA-gene and gene-TF, it is clearly shown which TF and miRNAs played a significant role in targeting the genes and preventing them from protein formation. However, in order to understand the relation between these selected TFs with the miRNAs responsible for targeting KIRC genes a network is formed from which it is concluded that a very small number of miRNAs are targeted by TFs. Thereafter, a relation is built among miRNA, gene and TF. The third objective of this research is to find a relation among TF-miRNA-gene so that a hub can be established through which it can easily predict the cause of regulation of the genes. From our research, we found some genes that are targeted by multiple miRNAs as well as TFs. Among those miRNAs, one miRNA is again targeted by multiple TFs, from this relation we can conclude that the regulation of this gene is not only controlled due to miRNA but TFs are also responsible for this regulation. If the regulation of the miRNA which is targeted by multiple TFs can be controlled then it may provide a better prognosis for genes of KIRC.

In the other study, we established a relationship among genes, TFs and miRNAs that have actively participated in prostate cancer progression. The study mainly focused to identify the influential TFs that are responsible to regulate the expression of genes as well as miRNAs. Additionally, we noticed that under a feed-forward loop, transcription factors can be a good druggable candidate. We have proposed a computational model to study the transcription factors and suggest the appropriate cellular conditions for drug targeting. We have selected feed-forward loops depending on the shared list of the functional annotations among transcription factors, genes, and miRNAs. From the potential feed-forward loop cores, six transcription factors were identified as druggable targets, which include AR, CEBPB, CREB1, ETS1, NFKB1, and RELA. Usually, TFs are not suitable as drug targets due to their lack of druggable pockets at the folding stage. This may be a



## **2.4. CONCLUSION**

---

reason why TFs are not selected as a potential drug target despite their ability to control the pathogenic progression of the malignancies. Very few researches are performed to address this problem.



## Understanding the dynamicity of proteins in terms of network

### 3.1 Introduction

---

In Chapter 2, the regulatory network is studied in detail to understand the underlying mechanism of dysregulation of genes responsible for disease progression and abnormal phenotypical changes. The established feed-forward loops revealed that gene expression is regulated by transcription factors and miRNAs at transcriptional and post-transcriptional levels respectively. Moreover, miRNAs are targeted by TFs that are indirectly responsible for gene dysregulation. So, TFs and miRNAs in the FFLs can control the pathogenic progression of malignancies by affecting the pathways associated with cell proliferation, tumour suppression and cell signalling. Therefore, the evolutionary, structural, and regulatory mechanism study of TFs can aid in the design of small molecules or biologics that can target specific TFs. Such drugs can control the gene expression in various diseases. Despite the ability to control disease-specific pathogenic progression. As per our knowledge, very few information is available regarding TF as a drug target. TFs are proteins in nature. They have been known for their protein moonlighting properties. Protein Moonlighting is a term used for a class of proteins that are associated with completely unrelated functions. Therefore, TFs are naturally extremely uncertain, and they can associate with non-transcription regulation activity at the same or diverse subcellular localization. Usually, many TFs are not suitable as drug targets due to their lack of druggable pockets when folded but this may be a reason why TFs are not selected as a potential drug target. To address this problem and provide a suitable druggable session for the TFs, a comprehensive frame has been proposed. The frame is validated by performing two studies, one is focused on the disease-specific TFs and the other is based on a particular TF and its association with pan-cancer study.

In this chapter, the frame is described in detail. The evolutionary history of transcription factors is studied to reveal the conserved regions and functional do-

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

mains. This information helps in designing molecules that specifically interact with these important residues, disrupting the activity of the transcription factor. In this context, the Shannon Entropy method is applied to the protein sequences. Subsequently, the evolutionary highly co-varying patches of the proteins are observed based on Direct Coupling Analysis (DCA). Following that, the structural space of the protein has been identified by utilizing the structure network models. Finally, the process has been summarized based on pathway analysis which can give an insight into how diverse the association of the protein is due to the mentioned structural malleability. Furthermore, a protein-specific study is performed during the COVID-19 pandemic. The aim of this study is to unveil the structural flexibility of the viral proteins associated with SARS-CoV-2 infection.

### 3.2 Data Acquisition and Data Preparation

During protein sequence-structure space studies, the data are collected from various well-known public databases. In Table 3.1, the list of databases is reported along with their usage. The experiment is divided into two stages: (i) sequence space analysis and (ii) structure space analysis, in order to study the variability of the proteins.

Serial Number	Database	Usage
1	Pfam	The sequences of the selected protein families are curated to perform Multiple Sequence Alignment (MSA)
2	UniProt	15 host cell proteins are considered which are responsible for COVID-19 reported in this database
3	RCSB PDB	The information of each proteins is fetched from this database to execute the structure network analysis
4	DisGeNET	A disease list is prepared to establish the disease network

Table 3.1: The data repositories used to fetch the data and utilized in this study are listed in this table.

### 3.3 Evolutionary Trait based on Sequence Complexity

Protein sequences evolve over time, and certain regions or motifs may be conserved across species. These conserved regions often represent critical functional

### 3.3. EVOLUTIONARY TRAIT BASED ON SEQUENCE COMPLEXITY

domains or residues essential for protein function. By analyzing the sequence complexity within conserved regions, key elements that have been evolutionarily preserved due to their functional importance can be identified. Therefore, by studying the sequence complexity of proteins, we can unveil insights into evolutionary history, functional diversity, and adaptive processes underlying protein evolution. Understanding these traits can help elucidate the relationship between protein sequence and function, as well as guide investigations into the evolution of specific protein families or the design of proteins with desired functionalities. In this context, we performed the below-mentioned methods to understand the evolutionary trait of the protein sequences.

#### 3.3.1 Shannon's Entropy

To understand the evolutionary changes and the associated changes in sequence complexity of the selected protein throughout the evolutionary time, their family sequences are selected from the Pfam database [39]. Some parameters are considered during the collection of the samples, such as, unaligned FASTA format selected from the database, while the sequences include tree ordering and lower-case letters. It is known that utilization of Multiple Sequence Alignment (MSA) [40] plays an important role in comparative functional and structural analyses. In our study, the T-Coffee alignment algorithm [41] is used to align the family sequences, where T-Coffee is a progressive alignment method that produces a list of pairwise alignments to conduct the MSA. Additionally, a consensus sequence [42] is obtained depending on the MSA from a Consensus Maker Tool (<http://www.hiv.lanl.gov/>). Customary parameters are provided to compute the consensus sequence of the family. The resulting consensus represents the logo of the protein family, which depends on the frequency of the amino acids. The high frequency of the amino acids is considered as the conserved region of the respective protein families [43]).

Furthermore, it is important to understand the structural organization of sequences in order to analyze the sequence present in a particular family. High entropy score signifies a higher propensity toward disorderliness [44]. In this regard, Shannon's entropy score is calculated for the consensus sequence to unveil the trait of the family. Shannon's entropy (SE) is defined by equation 3.1:

$$SE(n) = - \sum_{n=1}^N P_n \log_2 P_n \quad (3.1)$$

Here  $P_n$  is the probability and  $N$  is the number of the amino acids present in the sequence. The summation symbol runs over the twenty amino acids that generally present protein sequences.  $P_n$  signifies the probability of the given amino acid in

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

the consensus sequence. Consequently, Shannon entropy score lies between 0 to  $\log_2(20) = 4.32$ . It is known [45] that if the Shannon score of a sequence is more than 2.9 its propensity is more towards disorders and played an important role in disease formation. Similarly, Shannon entropy is computed for each sequence present in the selected protein domain families.

### 3.3.2 Coupling Analysis

The slow changes in a protein sequence are observed throughout the evolutionary time frame maintaining the native structure unaffected [46]. The evolutionary unaltered amino acids are considered conserved. These residues can be implicated in sustaining the structure and function of a particular protein. On the other hand, if the mutations occur in conserved residues, they may lead to structural and functional distortions. To avoid such potentially dangerous alterations, the change in size, shape, and other physicochemical properties of an amino acid by a mutation at a specific position is balanced by a compensatory change in another amino acid located in proximity in the three-dimensional structure of a protein. This signifies that, to restore or preserve protein structure and activity, the co-variation of two amino acids in an evolutionary time frame is immensely important.

To understand the dependencies of the residues and the effect of co-variation, the direct coupling method is used. This is statistical modelling to measure the strength of the direct relation between two residues [47]. Direct Information (DI) is used to perform the DCA. In this regard, the following equation shows how two sites of MSA are directly coupled by :

$$DI_{ij} = \sum_{AB} AB * p_{ij}^{(dir)} * (A, B) * \ln \frac{p_{ij}^{(dir)}(A, B)}{P_i(A)P_j(B)} \quad (3.2)$$

Here  $p_{ij}^{(dir)}$  represents re-weighted frequency counts to introduce two residues for DI.  $P_i(A)$  represents the singular site frequency, i.e., the probability of finding amino acid type A at *i*th position in the sequence.  $P_j(B)$  (for amino acid type B at *j*th position) is equivalent to  $P_i(A)$ .  $p_{ij}(A, B)$  represents the joint probability of observing amino acid type A at position *i* and amino acid type B at *j*th position in the amino acid sequence.

---

### 3.3. EVOLUTIONARY TRAIT BASED ON SEQUENCE COMPLEXITY

#### 3.3.3 Graph theoretical modelling and eigenvector community detection

The depiction of the physicochemical properties from the coupling analysis can be shown through a weighted graph of  $G_{DI}$ , where  $(V_{DI}, E_{DI}) \in G_{DI}$  represented residues and  $E_{DI}$  denoted weighted edges between directly correlated coupled pairs. This study aims to study the evolutionary dynamics of the proteins. In this regard, eigenvector-based community detection has been applied on  $G_{DI}$ . Usually, the eigenvector centrality of a node in a network analyzes the strength of its connectivity with the remaining nodes. Following the concept, a community based on eigenvectors represents the dynamic connectivity among the nodes. Therefore, community detection in  $G_{DI}$  can provide modules with a higher rate of co-varying residues, where the residues are densely connected.

#### 3.3.4 Disorder region of the protein sequence

To understand the disorderness of these TFs, PONDR-vlxt [48] and IUPred-s databases [49] are used. These two web-based servers help to predict the disorder of the amino acids present in intrinsic proteins. However, PONDR is a meta-predictor, that uses a consensus artificial neural network (ANN) prediction method, which was developed by combining the outputs of several individual disorder predictors and predicting intrinsic disorder from an amino acid sequence. Similarly, IUPred presents a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise inter-residue interaction energy, based on the assumption that IUP sequences do not fold due to their inability to form sufficient stabilizing inter-residue interactions. Optional to the prediction are built-in parameter sets optimized for predicting short or long-disordered regions and structured domains. From these two servers, the score of amino acids is identified and the overlapped regions of the sequence are observed to understand the change of score due to disorderness.

#### 3.3.5 Post Transnational Modification

The act of a particular protein in eukaryotic cells is regulated by Post Translational Modification (PTM). This is a collective chemical modification of proteins translated from mRNA before becoming functional in multiple body cells. These modifications are responsible for the heterogeneity of proteins and also help in deploying similar proteins for different functions in different cells. Using the PhosphoSitePlus database, the types of PTM are identified in the selected protein structures.

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

### 3.3.6 Hydropathy calculation

The hydropathy profile of a protein is intermediate between the amino acid sequence and the three-dimensional structure of the protein. This profiling score of amino acids is a measure that reflects their solubility in water. In order to identify the hydrophobic regions in a protein, the well-known Kyte–Doolittle scale [50] is used. During this calculation, a protein sequence is scrutinized with a sliding window. At each position, the mean hydropathy value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window. Expasy (<https://www.expasy.org/>) is utilized with the window size set to 7, the default value, to identify the surface exposed hydrophilic sites.

## 3.4 Identification of the impact of Structural adaptation in protein functions

---

During the study, we tried to map the sequential mutations to the structure space of the proteins. The goal is to use techniques to reveal how proteins function internally. To accomplish this, the Normal mode-based Elastic Network Model was utilized, which considers the independent movement of each residue without simplification. This creates a network model that shows the interactions between residues, with each module displaying a strongly connected core for the protein. Additionally, Alascan was applied to assess the stability pattern of each residue based on alanine mutagenesis, which identifies long, stretched flanking regions. A quality assessment provides interactional cores, and comparing results from different methods sheds light on the structural aspects of proteins. The following section discusses these methods in detail.

### 3.4.1 Normal mode-based Structure Network Analysis

Proteins are dynamic in nature and their fluctuations play vital roles in their functions. To understand the sequential orchestration, a protein structure network was established. The network is constructed based on the Normal mode analysis (NMA), which works better for large structural rearrangements. In this network diagram, amino acids are represented as nodes and their strength of non-covalent interactions are depicted through edges. Equation 3.3 is used to establish the interaction.

$$F_{pq} = \left[ \frac{X_{pq}}{\sqrt{X_p * X_q}} \right] * 100 \geq F_t \quad (3.3)$$

Here,  $X_{pq}$  is the number of side chains, and p and q are atom pairs of residues.



### 3.4. IDENTIFICATION OF THE IMPACT OF STRUCTURAL ADAPTATION IN PROTEIN FUNCTIONS

$X_p$  and  $X_q$  are the normalization factor for residue type  $p$  and  $q$  [51, 52].  $F_t$  is the threshold of interaction strength whereas 4% is the default value.

In this study, a cross-correlation matrix was calculated depending on the correlation matrix of NMA. The hypothesis behind the NMA of protein is the vibrational normal modes manifest the lowest frequencies, which unfold the largest protein movements and the functionally relevant ones. The tertiary structures for the particular proteins of different families are generated from I-TASSER [53]. NMA provides a comprehensive outlook of protein tertiary structure without coarse-graining. Firstly, NMA controls Cartesian coordinates as independent variables. NMA can also represent the chain connectivity of polypeptide chains. Subsequently, the activity of the parameters can easily be controlled by tweaking dihedral bonds. Along with that, NMA considers the individual movement of each of the residues, which perhaps helps to define the external and internal chemistry comprehensively. The value of cross-correlation is represented by the weight connections of a particular node. Simultaneously, a full residue network is established based on the correlation network analysis. Girvan-Newman clustering method is used with a threshold of 0.3 to split it into densely correlated coarse-grained community cluster network [54].

#### 3.4.2 Root Mean Square Fluctuation

Root Mean Square Fluctuation (RMSF) is a measure of particle deviation. In RMSF, a mean over time is considered for a residue  $j$  at the current position and some reference positions. The definition of the RMSF is given in Equation 3.4.

$$RSMF_j = \left( \frac{1}{T} \sum_{t_k=1}^T mod(r_j(t_k) - r_j^{re})^2 \right)^{.05} \quad (3.4)$$

Where  $T$  is the time over which the mean has been taken for reference position of the particle  $j$ ,  $r_j^{re}$ . The RMSF has been observed based on the reference position of the particle  $j$  over time.

#### 3.4.3 Calculate the stability score

The sequence is mapped to the structure of the respective proteins and compares the influential changes in sequence space that affect the structural behaviour. In this regard, the alaskan mutagenesis technique is applied to the PDB structure using FoldX [55]. The alaskan scanning method reveals the contribution of a specific residue to the stability or function of a given protein. Alanine is used due to its non-bulky, chemically inert, methyl functional group that nevertheless mimics the secondary structure preferences that many of the other amino acids possess.

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

Moreover, during this analysis, the quality assessment such as different dihedral bonds, torsion and van der Waal's forces that occurred to particular residues are also identified.

### 3.4.4 Liquid Liquid Phase Separation

Liquid Liquid Phase Separation (LLPS) has been observed to understand the issue mentioned above. LLPS is a method where the condensates help to separate the membrane-less compartments from the homogeneous liquid. Here, two databases i.e., LLPSDB [56] and PhaseDB [57] for the LLPS study are being used. In Table 3.4, the list of such localizations has been reported. Although, the nucleus is a widely known subcellular localization for TFs, nucleus speckles, a type of membrane-less organelles, have been noticed as another localization for most of the TFs. Interestingly, the nucleus speckle is one such membrane-less organelle where the activities at the pre-mRNA phase have been performed. Therefore, the functional existence of TF at nuclear speckles can be avoided. Moreover, when the mRNA is exiting from the nuclear speckles that phase may be considered as an appropriate session for drug targeting. In this regard, further, in vivo validation is required.

## 3.5 Experimental Results for Prostate Cancer study

---

From the final GRN shown in Chapter 1, 6 TFs (AR, RELA, NFKB1, CREB1, CEBPB) are selected. These TFs are analyzed to understand the regions involved in controlling the structural core throughout the natural adaptation. Evolutionary sequence space could detect more stringent domains following the natural adaptation trait of respective families. It is more global and robust than protein-specific information. In this regard, we have designed five DCA-specific networks for six TFs. NFKB1 and RELA are sharing the same protein family. In Figure 3.1, five networks are depending on the coupling propensity. As mentioned earlier, each module of the network can showcase the evolutionarily conserved intra-molecular dependency. Therefore, residual co-variance can contemplate as a key for intra-molecular allostery. This shows how the selected residues can directly or partially control the structural stability while targeted by drug molecules. Besides, the co-variance can also represent the non-covalent residual interaction. So, the sequential frame can provide a set of precise influential residues. Applying the proposed sequential frame, we have identified five probable structural facets for AR, ETS1, CEBPB, CREB1 and RELA. In terms of the prediction algorithm, AR shows the highest rate of disorder at specific residual positions. Additionally, the miRNAs present in the FFL of the final GRN are considered to analyse their evolutionary trait during natural adaptation. Our main aim is to observe the influential TFs in prostate cancer. In this regard, miRNAs that are associated with targeted by at least one

### 3.5. EXPERIMENTAL RESULTS FOR PROSTATE CANCER STUDY

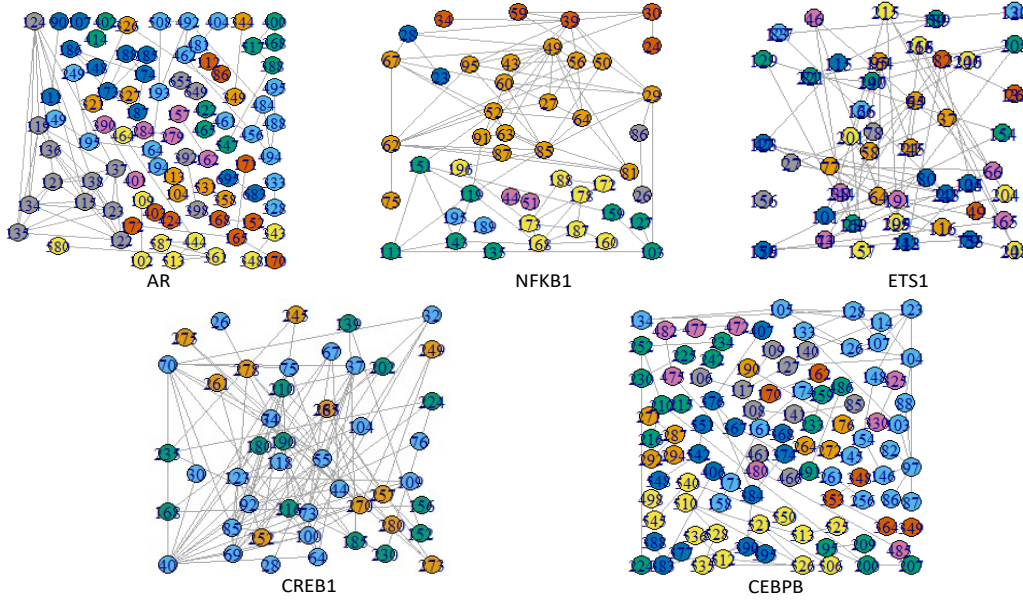


Figure 3.1: A weighted network  $G_{DCA}$  and corresponding colour modules based on overall residual covariation from DI score of the TFs in prostate cancer

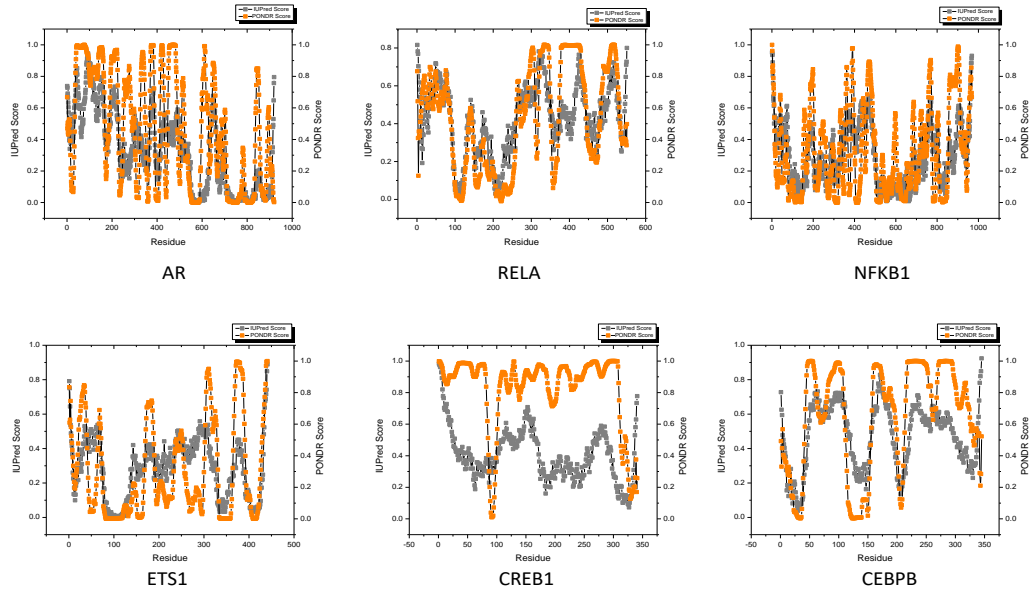


Figure 3.2: In prostate cancer, the overlapped areas are obtained from PONDR and IUPred databases. Results from PONDR and IUPred are represented with orange and grey colour, respectively.

TF are determined. To observe the evolutionary changes, miRNAs and their tar-

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

geted genes are studied in white mice. As a result, 10 miRNAs are found for six TFs. These miRNAs also significantly regulated the genes in prostate carcinoma. Among the 101 miRNAs, 9 miRNAs are validated through a literature survey. Our results show PTGS2 gene is targeted by all the influential TFs and also targeted by different miRNAs (participated in natural adaptation). This evidence supports the strong impact of the selected TFs in the modulation of the GRN during the disease.

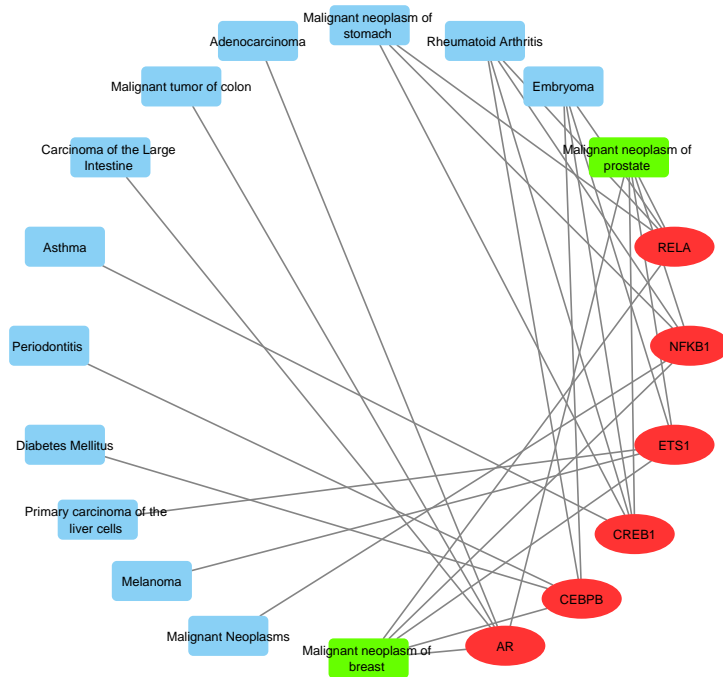


Figure 3.3: Top five diseases, based on the P-values, of each TF of prostate cancer are selected to establish the network. Green colour nodes are diseases, connected with all four TFs (red oval nodes). Green colour nodes represent the diseases targeted by five out of six TFs

PTMs occurring at the amino acids play a crucial role in protein's function, physicochemical properties, conformation, stability and molecular interactions in response to developmental signals or environmental stimuli. In this order, in Figure 3.4 influential subtypes of PTMs for the 6 TFs are shown extensively. The subtypes are represented by using different colours. Red, green yellow and black define acetylation, ubiquitylation, phosphorylation and others respectively Rate of phosphorylation is high in almost every TF except RELA. In AR, three more PTM subtypes viz., acetylation [58], ubiquitination and methylation [59] are well distributed throughout the sequence. Similarly, ETS1 consists of an equal number

### 3.5. EXPERIMENTAL RESULTS FOR PROSTATE CANCER STUDY

of ubiquitination [60] and acetylation sites. Interestingly, CEBPB has dimethylation sites [59]. Unlike other TFs, RELA has O-GlcNAc sites [61, 62] along with a small amount of acetylation and phosphorylation sites. Almost every sub-type has implications on prostate cancer where O-GlcNAc sites are participating in cell proliferation. After PTM site identification, in sequence space, we initially stud-

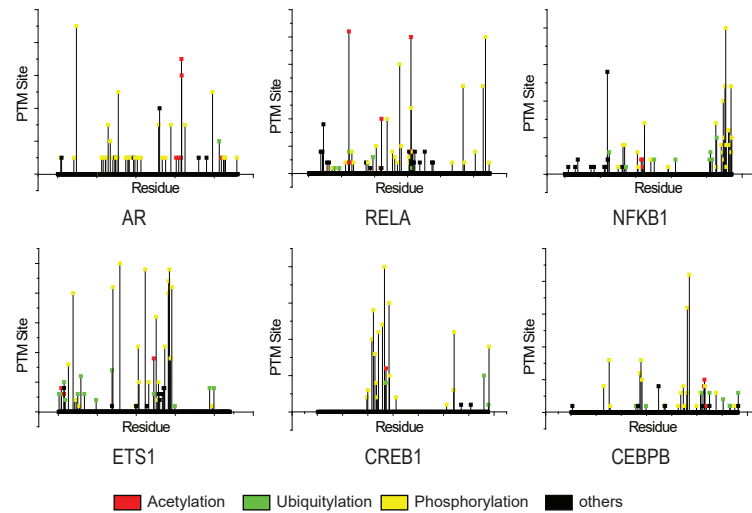


Figure 3.4: The representation of the residues having post-translational modification for each selected TFs in prostate cancer study.

ied the flanking regions by applying the disordered region prediction algorithms shown in Figure 3.3. The prediction has provided protein-specific unstructured as well as expected changing cores. In this regard, intersecting disordered domains from two databases such as IUPred and PONDR are selected as disordered regions. The overlapping regions depict the fluctuations of the amino acids present at that particular region and due to the Ramachandran angles, they are highly responsible for disorderedness of the TFs. Moreover, it is known that the disorder region occurs at the PTM sites mostly and the scores of amino acids support the result of PTM also.

Besides PTM and disorder region identification, Alanine mutagenesis is used to interpret the contribution of a single residue towards the stability of a protein. Due to the non-bulky, chemically inert, methyl functional group that nevertheless mimics the secondary structure preferences alanine is used to check the stability. In Figure 3.5, graphs depict the stability score after interchanging the residues with alanine. Similarly, a quality assessment of selected proteins is performed and the van der Waals clashes occurred to particular residues shown with colours.

Furthermore, the TRRUST database is used to unveil the association of diseases including prostate carcinoma with the six TFs. The top five diseases of each TF are

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

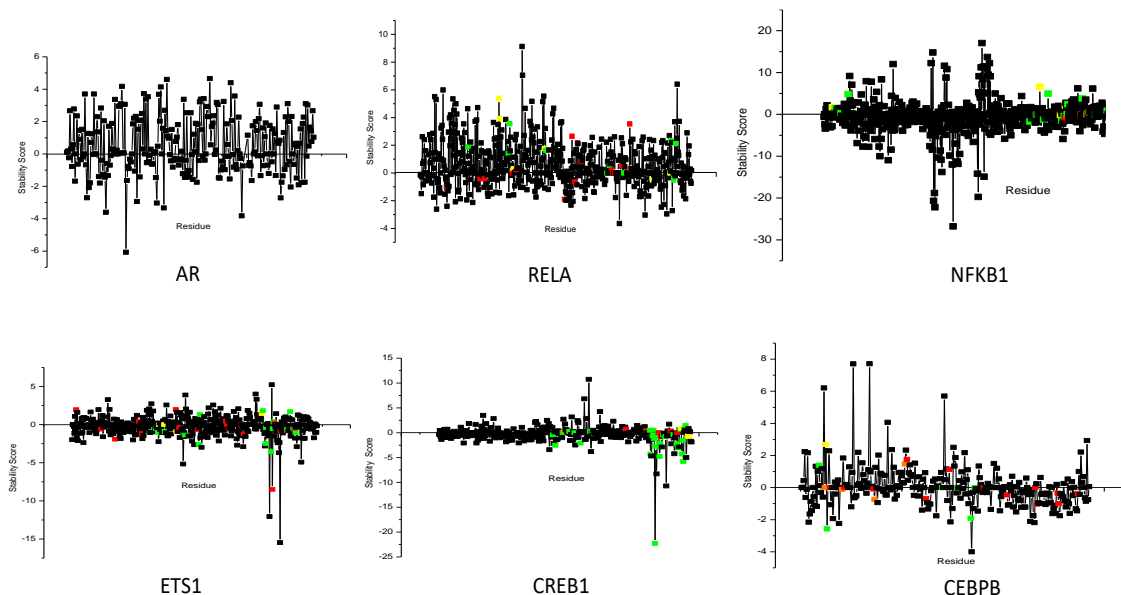


Figure 3.5: The stability score of the amino acids of six TFs of prostate cancer along with their bonds occurred in a particular residue. Torsion, Van der Waals, Chi and Cis-bond are represented by colour yellow, green, red and orange, respectively

sorted depending on their p-values and a network based on that is shown in Figure 3.3. The network diagram depicts the connection between TFs and diseases. Among these two diseases such as malignant neoplasms of the breast and prostate are targeted by five TFs excluding CEBPB with high p-values.

These TFs are found responsible for prostate cancer, but to understand their reaction on the pathways, the protein interaction partners are identified from STRING database (<https://string-db.org/>). Subsequently, the localization of the pathways that are controlled by the TFs and the interaction partners are identified from the Reactome pathway database. The Reactome Pathway (pathreactome) shows that TFs present in a single cellular localization involve two distinct processes. The TFs, and protein interaction partners with their pathreactome for a particular localization are reported in Table 3.2. From this result, it is expected that these TFs may have a moonlighting property as they perform distinct functions in the same localization. Due to this property proteins can execute more than one function. Proteins gather this property through evolution. To validate the moonlighting property of the TFs both MoonProt and MoonDB databases are utilized. We found that AR and ETS1 are present in the database as a moonlighting protein, whereas for the other four TFs Basic Local Alignment Search Tool (BLAST) is used to identify the highest matching with the queried sequence. We found that the highest matching sequence shows moonlighting property and the functions

### 3.6. RESULTS OF POU2F1 AS AN IMPORTANT TF IN PAN-CANCER STUDY

---

are performed by the queried sequence too. Though the searched proteins are not present in the two databases these also show moonlighting properties. This property of TFs along with the curated databases and processes are reported in Table 3.3. In terms of protein moonlighting properties, these can be an appropriate sessions for drug targeting. The main spatial division has shown prime subcellular localization i.e. cytosol and nucleoplasm which are separated through the nuclear membrane. However, a further subdivision within the nucleoplasm has been identified through the LLPS study. In Table 3.4, the list of such localizations has been reported. Although, the nucleus is a widely known subcellular localization for TFs, nucleus speckles, a type of membrane-less organelles, have been noticed as another localization for most of the TFs. Interestingly, the nucleus speckle is one such membrane-less organelle where the activities at the pre-mRNA phase have been performed. Therefore, the functional existence of TF at nucleus speckles can be avoided. Moreover, when the mRNA is exiting from the nucleus speckles that phase may be considered as an appropriate session for drug targeting. In this regard, further, in vivo validation is required.

We have tried to provide suggestive drugs for the selected TFs. In this regard, the respective drugs have been chosen based on the differential activity of the TFs. From the literature, AR is highly associated with Prostate Cancer where the under-expression of the protein is responsible for the disease. Similarly, the two proteins from NFkB family are overexpressed during PCa [63]. NFkB family proteins are regulating the AR gene expression. Also, CEBPB is sharing the same level of expression with NFkB family proteins [64]. CEBPB is modulating the metastatic genes in Prostate Cell. In Table 3.5, we have prepared one such list where TFs and their corresponding expression and drugs are mentioned. In the case of AR, only the top five agonist drugs have been considered. Only for ETS1, we cannot detect the appropriate drugs. All the information has been taken from DrugBank [65].

This chapter consists of three experimental studies. However, some unique computational frameworks are proposed to understand the impact of structural malleability of the protein sequences towards their function annotation and the modification throughout the evolutionary traits. The proposed frameworks are modified according to the need of the studies. Therefore, in the experimental section, the representation of the computed results according to the studies is divided.

### 3.6 Results of POU2F1 as an important TF in pan-cancer study

---

The goal of the study is to develop a complete framework to decipher the structural conservation and co-variation of POU2F1 along with the association of this

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

Table 3.2: Sharing common pathways with the nearest neighbour of the selected TFs from Protein-Protein Interaction having an impact on Prostate Carcinoma from the Reactome database.

TF	Nearest Neighbour	Localization	Pathway
AR	NCOA2	Nucleoplasm	Activated PKN1 stimulates transcription of AR regulated genes KLK2 and KLK3
	KDM1A	Nucleoplasm	Activated PKN1 stimulates transcription of AR regulated genes KLK2 and KLK3
	CCND1	Nucleoplasm	Transcriptional regulation by RUNX2
RELA	CREBBP	nucleoplasm	C-type lectin receptors Cytosolic sensors of pathogen-associated DNA
	NFKBIA	Cytosol	p75 NTR receptor-mediated signalling TCR signaling Signaling by the B Cell Receptor
			TAK1 activates NFkB by phosphorylation and activation of IKKs complex
	IKBKB	Cytosol	p75 NTR receptor-mediated signalling
	IKBKG	Cytosol	TCR signaling
	NFKBIB	Cytosol	Signaling by the B Cell Receptor TAK1 activates NFkB by phosphorylation and activation of IKKs complex
	CHUK	Cytosol	TCR signaling Signaling by the B Cell Receptor TAK1 activates NFkB by phosphorylation and activation of IKKs complex
		Nucleoplasm	Cytosolic sensors of pathogen-associated DNA
NFKB1	NFKBIB	Cytosol	C-type lectin receptors p75 NTR receptor-mediated signalling TCR signaling
			TAK1 activates NFkB by phosphorylation and activation of IKKs complex
			Cytosolic sensors of pathogen-associated DNA
	CHUK	Cytosol	Signaling by the B Cell Receptor TAK1 activates NFkB by phosphorylation and activation of IKKs complex Cytosolic sensors of pathogen-associated DNA
	RELA	Cytosol	TCR signaling Signaling by the B Cell Receptor TAK1 activates NFkB by phosphorylation and activation of IKKs complex
			Cytosolic sensors of pathogen-associated DNA
			p75 NTR receptor-mediated signalling
	IKBKB	Cytosol	p75 NTR receptor-mediated signalling
	IKBKG	Cytosol	TAK1 activates NFkB by phosphorylation and activation of IKKs complex
ETS1	SP1	nucleoplasm	TCR signaling
			Signaling by the B Cell Receptor
CREB1	CAMK4	nucleoplasm	p75 NTR receptor-mediated signalling TCR signaling Signaling by the B Cell Receptor
			TAK1 activates NFkB by phosphorylation and activation of IKKs complex

protein with pan-cancer analysis. It is known that POU2F1 is a transcription factor that actively participates in the pathogenesis of multiple diseases, such as cancer, asthma, dermatitis, periodontal diseases, etc. [66, 67, 68]. The involvement of POU2F1 in diverse diseases is identified using the TRRUST database [69]. This analysis revealed that POU2F1 is responsible for more than 100 diseases. This is illustrated by Figure 3.6(a), which represents the association of POU2F1 with various diseases by a scatter plot. Among the diverse diseases, the top twenty diseases based on their adjusted p-value are shown by a network in Figure 3.6(b).



### 3.6. RESULTS OF POU2F1 AS AN IMPORTANT TF IN PAN-CANCER STUDY

Table 3.3: The moonlighting function of the selected TFs of prostate cancer from MoonProt and MoonDB.

Name of the Protein	Function 1	Function 2	Database Used	Search Process
AR	Signal Transduction	Cellular nitrogen compound metabolic process	moonDB	Data repository
RELA	Inositol phosphate metabolic process	Scaffold, binds protein kinase CK2, TCOF1, and upstream- binding-factor (UBF)	moonProt	BLAST
NFkB1	Carbohydrate metabolic process	Binds to HIF (hypoxia- inducible factor) protein and inhibits nuclear HIF action	moonProt	BLAST
ETS1	Signal Transduction	Positive regulation of transcription from RNA polymerase II promoter	moonDB	Data Repository
CREB1	ATF2 activating transcription factor bZIP family of transcription factors binds DNA as a dimer	Recruiting Mre11 to IR-induced foci (IRIF) in the DNA damage response, this function does not require DNA binding domain	moonProt	BLAST
CEBPB	Not Known	Not Known	Not Known	Not Known

Table 3.4: The List of Alternative Cellular Localizations in Terms of Liquid Liquid Phase Separation

TF	Subcellular Localization1	Subcellular Localization2	Database
AR	Nucleus	Cytoplasm	LLPSDb
RELA	Nucleus	Nucleus Speckles	PhaseDb
NFkB1	Nucleus	Nucleus Speckles	PhaseDb
CREB1	Nucleus	-	-
CEBPB	Nucleus	Nucleus Speckles	PhaseDb

Table 3.5: The 6 TFs of prostate cancer and their corresponding expression along with the drugs are reported from DrugBank.

Transcription Factor	Types of expression	Drug
AR	under expression	Oxandrolone
RELA	overexpression	Dimethyl Fumarate
NFkB1	overexpression	SC-236
ETS1	overexpression	Not known
CREB1	overexpression	Naloxome
CEBPB	underexpression	Quercetin

In order to interpret the regulators through which this TF is regulating the biological pathways that finally lead to disease formation a gene regulatory network is established [70]. Besides constructing a regulatory network based on the literature and experimental evidence, the biological pathways are also considered for each molecule. Each regulator is associated with multiple pathogenic as well as non-pathogenic pathways. Interestingly, 16 pathways are found shared among them. Furthermore, an evolutionary trait of the POU domain family is analyzed by calculating Shannon's entropy (SE) score. During the SE calculations, the

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

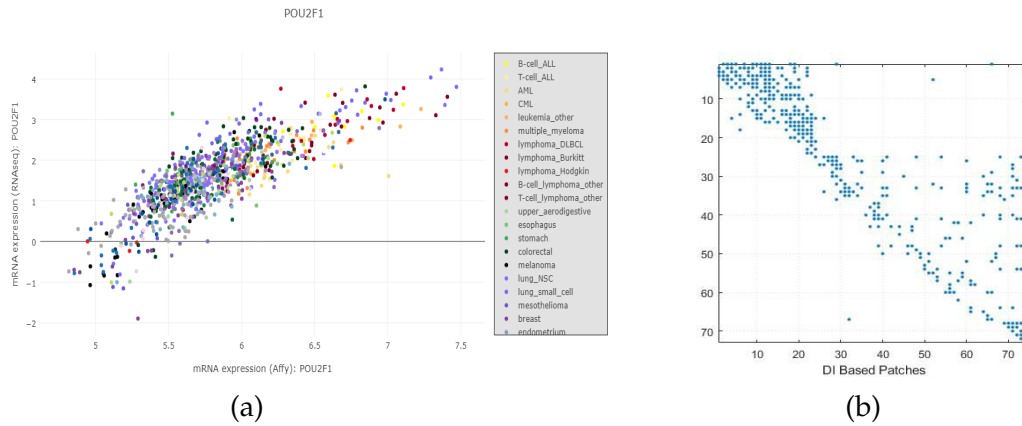


Figure 3.6: POU2F1 is associated with multiple diseases. (a) Associations are represented through a scatter plot and (b) the Top ten associated diseases are shown through a network diagram.

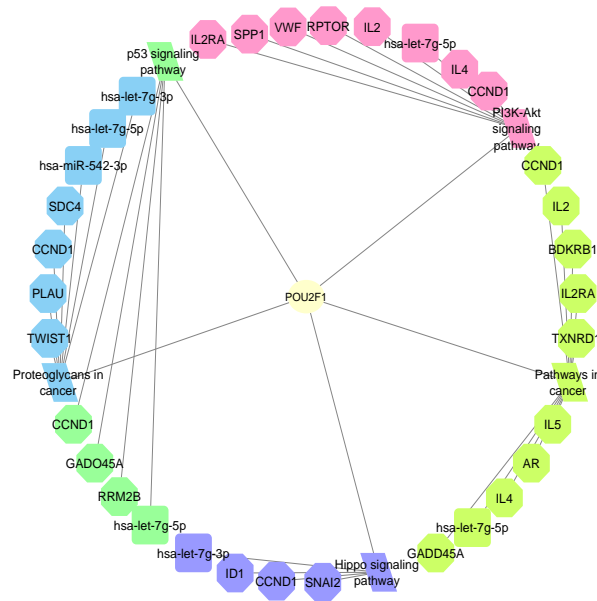


Figure 3.7: The network representation of the top five common functional pathways among POU2F1 and its targeted miRNAs and genes. Different colors represent diverse pathways and their associations.

human and non-human samples are divided into two groups to understand the propensity of order and disorder of humans and other primitive species separately. Figure 3.8(a) depicts the change in entropy scores of groups one and two by green and orange lines respectively. In the next phase of the sequence-spaced study, the co-variation propensity of residue couples at a specific evolutionary conserved

### 3.6. RESULTS OF POU2F1 AS AN IMPORTANT TF IN PAN-CANCER STUDY

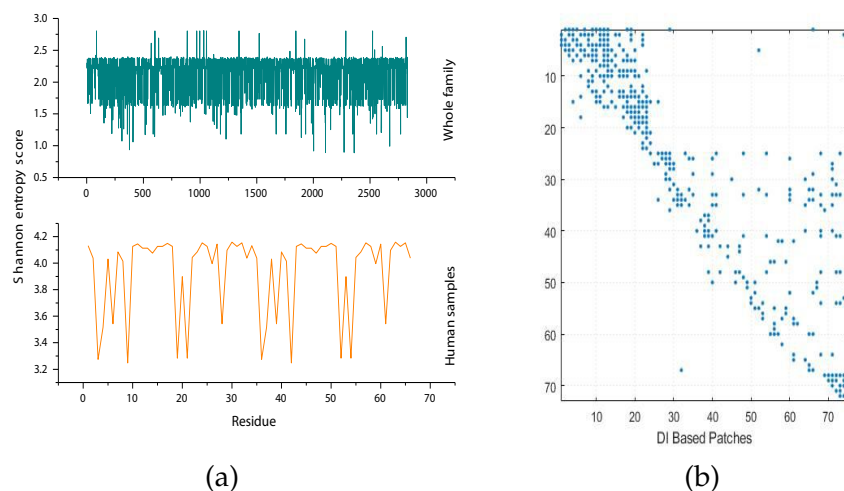


Figure 3.8: Representing the change in evolutionary trend by (a) The Shannon entropy score of Homo sapiens samples and all other species by blue and black lines respectively and (b) The DCA score of POU2F1.

position has been calculated by using DCA. From these calculations, the Direct Information (DI) scores are generated for possible coupling pairs. Among them, 93 coupled pairs are selected as highly evolutionary co-varying patches in terms of DI scores. The detailed score of DCA calculation graph is generated for the visual representation shown in Figure 3.8(b). In parallel, IUPred [71] and PONDR [72] are used to predict the per-residue intrinsic disorder predisposition of POU2F1. The overlapping residues from both prediction models are considered.

These common residues are clustered depending on their DCA score. The residues in the same cluster represent the same rate of evolutionary change shown in Figure 3.9. In parallel, to decipher the residual organization and dependencies at different secondary structural orchestration, structure network analysis is employed.

As the residues are aggregated into one module, which is depicted from the structure network, it is difficult to understand the structural facet of this protein. In this regard, quality assessment is performed using the modelled 3D structure. The non-covalent clashes of residues at disordered regions are identified. These forces are analyzed based on the clusters shown in Figure 3.9. Finally, the structural facet is recognized depending on the cluster distribution of the residues and the corresponding non-covalent forces.

Since POU2F1 is a highly disordered protein, it structurally exists as a dynamic conformational ensemble. The potential 3D structures of several illustrative members of this POU2F1 conformational ensemble are modelled using I-TASSER

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

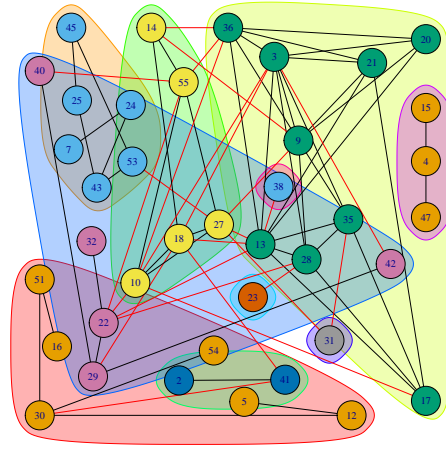


Figure 3.9: A weighted network  $G_{DCA}$  and corresponding color modules based on overall residual co-variation from DI score.

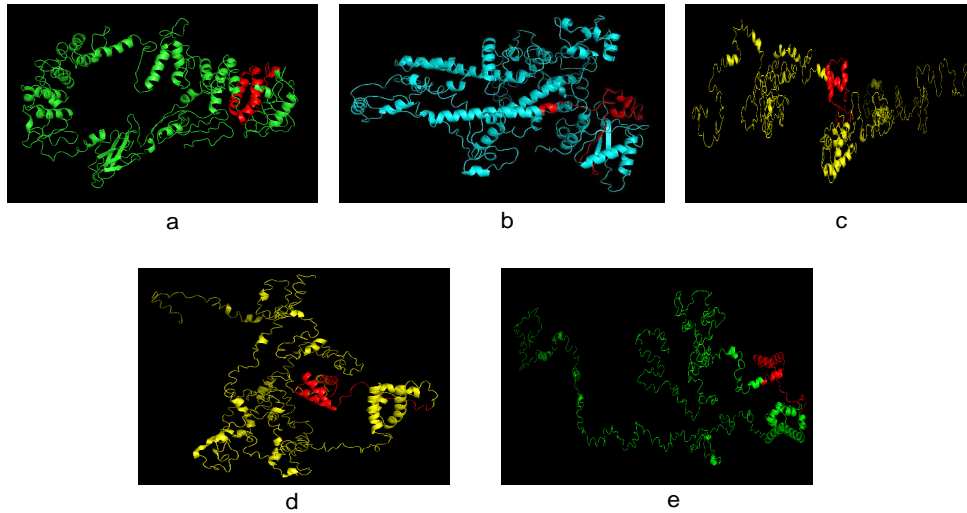


Figure 3.10: The PDB structure of (a) model 1 (-0.56), (b) model 2 (-1.16), (c) model 3 (-4.02) (d) model 4 (-4.16) and (e) model 5 (-4.17), the five models and their c-scores performed from I-TASSER of POU2F1 protein. The highlighted red-coloured zone of each five structures denotes the DNA binding site of the respective models. The DNA binding site of the models is identified using the information from UNIPROT database.

[73, 74]. The top five structures are constructed, and the corresponding models are reported in Figure 3.10 with their respective C-scores. The DNA binding domain

### 3.6. RESULTS OF POU2F1 AS AN IMPORTANT TF IN PAN-CANCER STUDY

in these five models identified from the information in UniProt database (residues 379-438 are shown in red color. Next, the structure network analysis is performed to map the sequence space changes in structure space. The structure network modules of the five models are shown in Figure 3.11. Figure 3.12 shows the PONDR result by overlapping the outcomes of the two algorithms (VSL2 and VL-XT) for three consecutive consensus phases viz., *Consensus<sub>family</sub>*, *Consensus<sub>mammalian</sub>*, and *Consensus<sub>human</sub>*. More elaborately, sequences of POU-domain proteins are considered in three different cases for conserved predicted disorder status (PDS) analysis. In the first case, the sequences of all the 2830 proteins belonging to PF00157 are considered to obtain a consensus sequence that is used for the conserved PDS evaluation. In the subsequent case, amino acid sequences of all the mammalian POU-domain proteins (85 proteins) are aligned to design a single consensus sequence for the PDS analysis. Finally, 33 human members of the POU-domain family are contemplated for a similar purpose. Figure 3.12 shows that when all the sequences of protein from this family are considered, the consensus sequence shows significant levels of order. On the other hand, considering mammalian proteins and especially human POU2F1 proteins only indicates the presence of high levels of predicted disorder which are responsible for the dynamic nature of POU2F1 protein. Furthermore, during this study, it is observed that during the structure prediction, I-TASSER may provide an overview of the protein conformational ensemble (basically generating illustrative examples of this conformational ensemble). For the purpose of the article, we have considered a predicted structure with the least C-score and corresponding structure network. However, the lack of model prediction capability of I-TASSER is largely discussed with a definite set of evidence in the next section.

No evidence has been reported till now regarding the moonlighting potential of POU2F1. Furthermore, from the aforementioned results, it is clear that POU2F1 is responsible for the regulation of diverse biological pathways, that are related to multiple diseases. Strikingly, POU2F1 is found in nucleoplasm only [75]. Due to this, a BLAST is performed and based on the local similarity two proteins are found in the database having moonlighting property [76]. Those proteins along with their human-specific activities (function 1 (F1) and function 2 (F2)) are reported in Table 1.

From the above analysis, we can conclude that POU2F1, a known TF from a pan-cancer study, is associated with diverse functional pathways. During analysis of the associations, multiple targeted molecules are identified as common between pathways. POU2F1 regulates the expression of target genes directly, as well as indirectly, and is also responsible for the impact on targeted miRNAs. As their expression is controlled by POU2F1, it is inferred that POU2F1 is playing an

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

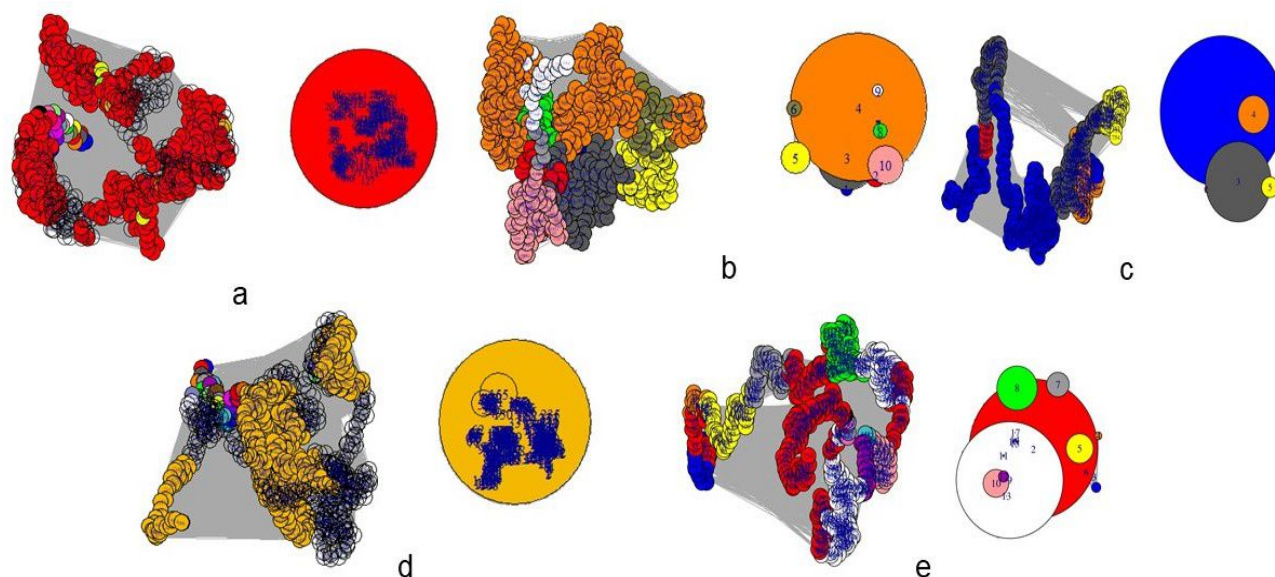


Figure 3.11: The structure network of the model The structure network diagram of the models (a) model 1, (b) model 2, (c) model 3, (d) model 4 and (e) model 5, are performed from the I-TASSER. Each structure shows that a large portion of the residues are converged in a single module.

Table 3.6: Human-specific protein moonlighting activities of the two proteins similar to POU2F1. The evidence is curated from the MoonProt database after performing BLAST of the POU2F1 sequence.

Protein Name	Function 1	Function 2
ATF2	ATF2 activating transcription factor bZIP family of transcription factors binds DNA as a dimer, homodimerization or heterodimerization, sometimes with c-Jun	recruiting Mre11 to IR-induced foci (IRIF) in the DNA damage response, this function does not require DNA binding domain or dimerization with c-Jun
Cdt1	helps load proteins onto chromatin to create the pre-replication complex for initiating DNA replication	role in mitosis localizes to kinetochores through binding to the Hec1 component of the Ndc80 complex

important role in the formation of those reported diseases. Therefore, POU2F1 might not only control the expression of genes directly related to the formation of diseases, but also it is able to control the selected miRNAs responsible for diseases shown in Figure 3.6(a).

### 3.7 Results of protein-specific study for SARS-COV-2

The aim of the study was to reveal the structural flexibility of the viral proteins associated with SARS-CoV-2 infection. Furthermore, the diseases accompanying

### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

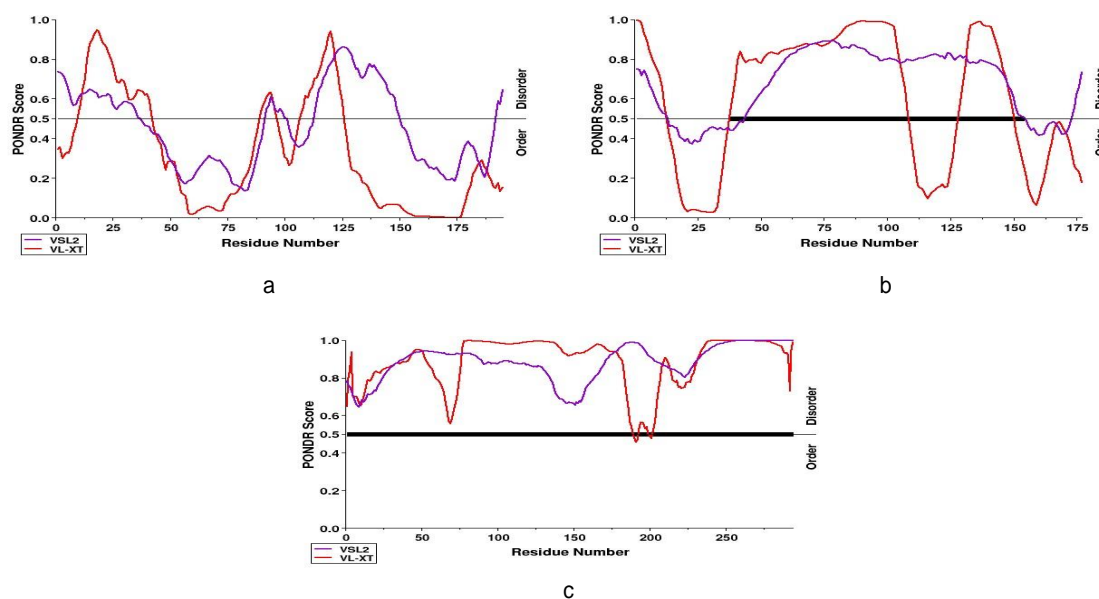


Figure 3.12: The disorder analysis of POU domain proteins using PONDR® VSL2 and PONDR® VL-XT algorithms. (a) PONDR-based analysis of the full set of sequences of all the members of the POU domain family shows that the consensus sequence generated for this family is mostly ordered in nature. (b) Disorder analysis of the amino acid sequences of only the members of the POU domain family that belong to the advanced species illustrates that the corresponding consensus sequence is predicted to contain very significant levels of disorder. (c) Analysis of human POU2F1 proteins by PONDR denotes a very high level of disorder.

COVID-19 were also studied in detail. It is evident from the reported results that the high flexibility of the selected proteins is responsible for the vulnerability of the population to this infection. The study is performed in two different stages. In the first stage, the proteins are analyzed based on their sequence and structure space, whereas in the second stage, the disease associations are established.

A hydropathy plot generated for each protein sequence belonging to 17 protein families is shown in Figure 3.13. Such hydropathy plots allow for the visualization of hydrophobicity over the length of a peptide sequence. A sliding "window" determines the summed hydropathy at each point in the sequence (Y coordinate). These sums are then plotted against their respective positions (X coordinate). Such plots are useful in determining the hydrophobic patterns of globular proteins, as well as determining membrane-spanning regions of membrane proteins.



## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

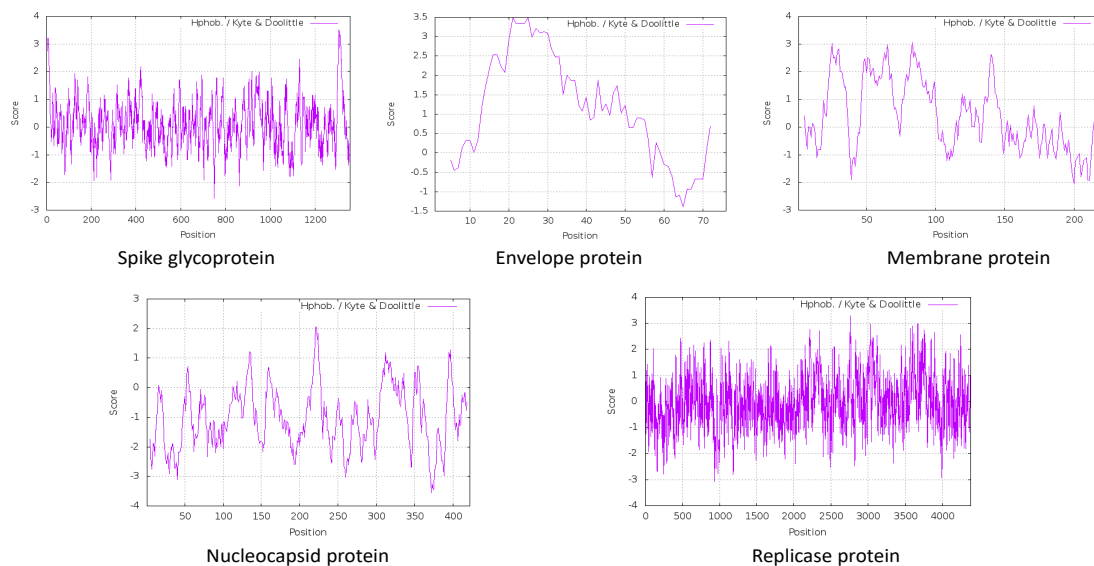


Figure 3.13: The two-dimensional hydropathy plot produced for the selected proteins based on the Kyte and Doolittle scale of amino acid hydropathy.

### 3.7.1 Sequence Space Analysis

From Pfam, we have started with 17 protein families consisting of at least one viral protein. Among them, five proteins, one from each class are reported in the manuscript. During the experiment, we faced one such consequence, where a protein can be a member of more than one protein family (homologically). In such cases, we have considered the most appropriate one, e.g., PF01600 is a more appropriate family for Spike glycoproteins. This has been selected based on the published literature on SARS-CoV [77]. Similarly, NSP1 is also associated with two families based on the sequence similarity at the nucleotide-binding domain and polypeptide conservation [78, 79, 80]. After the selection, SE scorings have been provided for each family. In this case, column-wise sequential occupancy has been reflected in the SE scoring. As aforementioned, we have started with individual protein families. Now, the sequential columns of each family are considered evolutionary stable at the residual position if the score is zero. Higher scores represent more sequential randomness. Therefore, the protein-specific SE plots in Figure 3.14 represent the degree of their evolutionary randomness. Simultaneously, the coupled pairs are studied based on distinct scoring i.e., MI and DI. The co-occurrences of the residues have been shown in the feature map with MI scores. The graphs of each protein show some small white dots which represent the conserved co-varying patches. These also help to portray the coupling strength between the residual position or the co-varying amino acids. In the case



### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

of DI, we have designed weighted networks from the DI scores and corresponding coupled pairs. Hence, the network can represent the evolutionary conserved inner allosteric. The communities from the network are based on the eigenvector centrality. Each community has at least one node with higher eigenvector centrality scores. The connected nodes are involved and responsible for the corresponding scores. Therefore, eigenvector-based communities can represent the internal modifications due to the mutational changes at any residual nodes from the network.

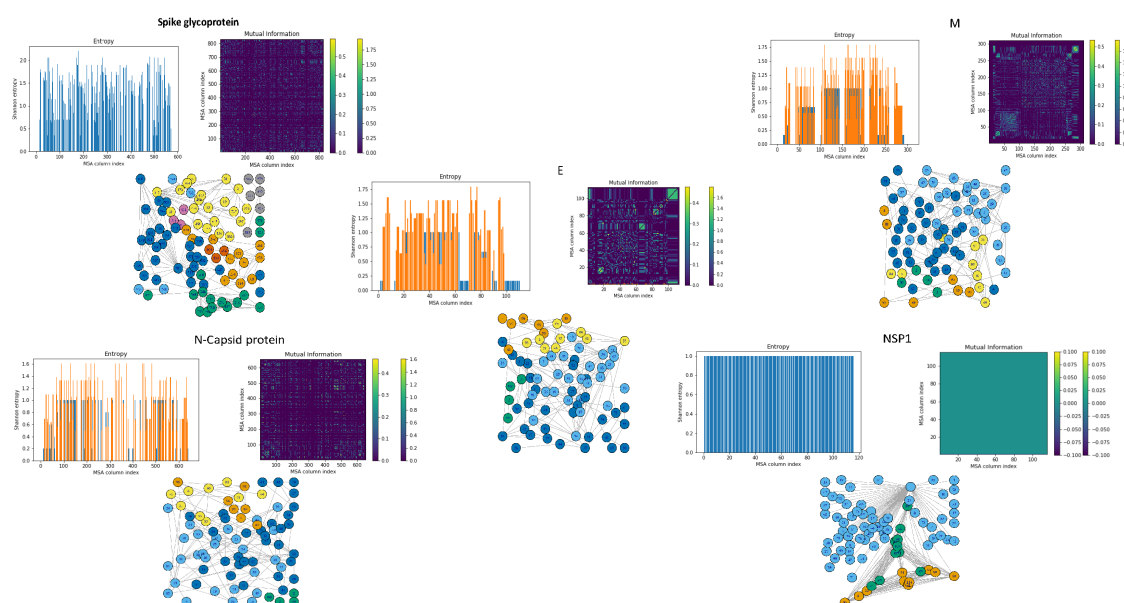


Figure 3.14: The sequence space analysis is performed for the selected protein families along with the SE calculation, coupling analysis and community detection techniques. The residue-wise extreme variability of the SE score has been shown through two colors.

#### 3.7.2 Structure Space Analysis

In Figure 3.15, the square-fluctuations of the protein sequences are analyzed by utilizing eigenvector centrality. These fluctuations depict the mutational link between the sequence and structure of a particular protein. The peak of fluctuating regions is mapped with the conserved region of the protein by using sequence space knowledge. This deciphers how the conserved regions are associated with the mutational changes, which may lead to the change in structure-function association. Furthermore, depending on the DI scores, the residues in the same cluster represent the same rate of evolutionary change. In parallel, to decipher the residue-wise organization and dependencies at different secondary structure

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

elements, structure network analysis was employed. The potential 3D structures of these proteins are modelled using I-TASSER [81, 82]. The eigenvector centrality is calculated to unveil the influence of a particular node on the internal dynamics of different protein structures. The structure network analysis is performed to map the sequence space changes in the structure space shown in Figure 3.15.

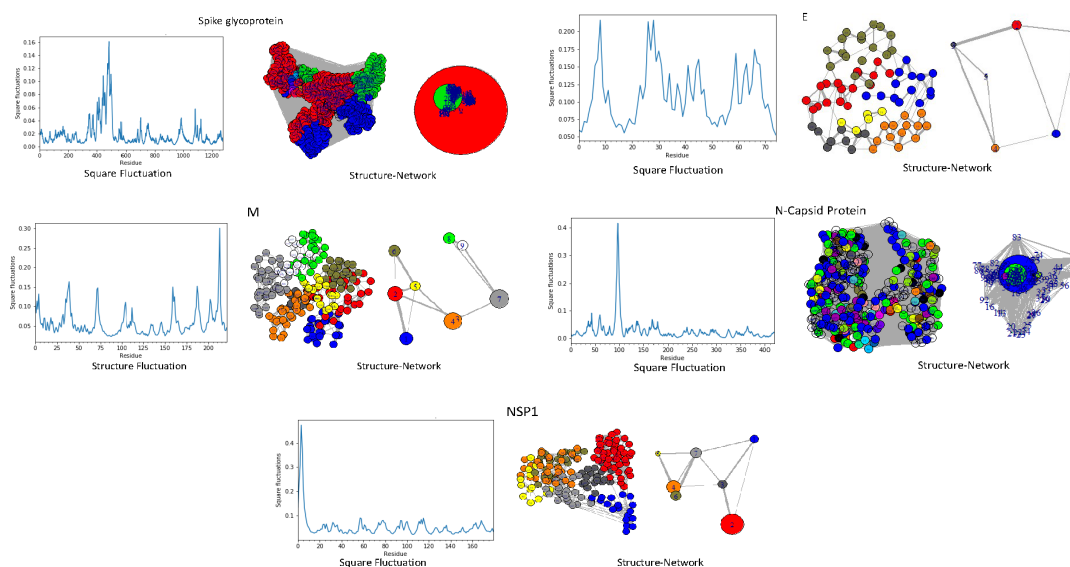


Figure 3.15: In the structure space analysis of the selected protein families, the structure fluctuations and structure network are performed.

### 3.7.3 Diseases Related to the SARS-CoV-2 Infection

Current literature contains a wealth of information supporting the notion that patients suffering from cardiovascular diseases, gastrointestinal disorders, diabetics and cancer show higher vulnerability towards the SARS-CoV-2 infection. Patients in critical condition are more prone to this disease due to multiple organ failure.

Unfortunately, no particular therapeutic means or treatments are found that can cure the affected patients. Furthermore, diverse pathways and biological processes are affected due to the infection, which leads to comorbidity. In order to understand the effect of this infection, in the second stage of this study, the association of diseases with COVID-19 was considered. Firstly, the relations between the SARS-CoV-2 proteins and the host proteins were established. From that relationships, we have identified 15 human proteins. Among the 15 host cell proteins, 14 proteins are associated with a list of human diseases. This list of diseases may help to organize the possibility of comorbidities as well as help to determine diseases at the post-COVID stage. However, one protein i.e., MPP5 which has been excluded from

### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

this list, is not much involved with relevant diseases. Using the STRING database, 14 individual PPI networks were constructed. During this network construction, the neighbouring members of the PPI network are selected based on the sharing same protein homology, experimentally curated and the same co-expression. For each selected protein and its neighbours, all possible diseases are considered from the DisGeNET [35]. The list of diseases was compared with the shortlisted disease list mentioned in the Method section. In Figure 3.16, ten diseases are reported,

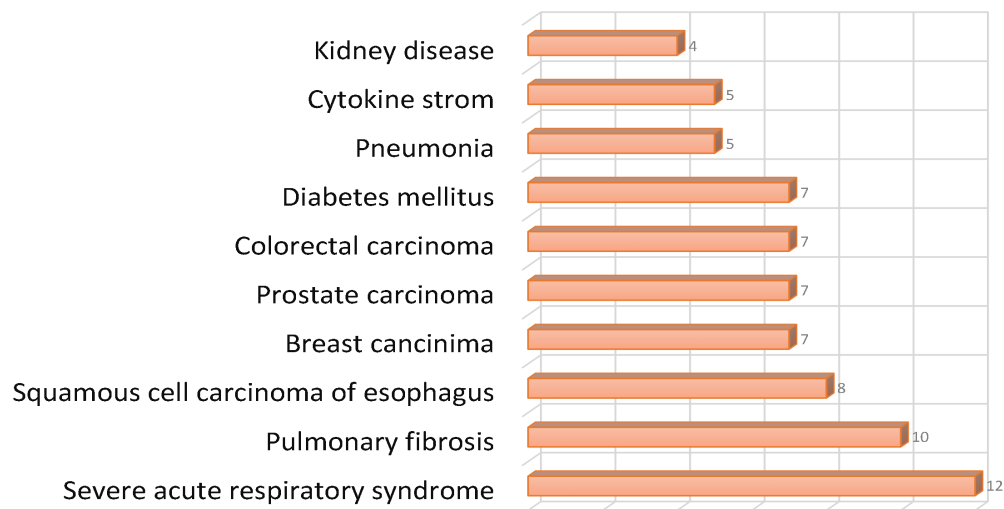


Figure 3.16: To understand the involvement of the proteins with the distortion of the organ behavior, a bar plot is performed of the disease shared by the proteins.

and they are selected based on the number of proteins having an impact on them. During SARS-CoV-2 infection, multiple body organs are affected, such as the lungs, kidneys, liver, cardiovascular system, etc. To understand the involvement of the proteins in the distortion of organ behaviour we constructed a Venn diagram of the diseases shared by the proteins. Interestingly, this analysis revealed that diverse diseases have resulted from the misbehaviour of these proteins. Among multiple diseases, ten diseases are reported in a bar plot along with the number of associated proteins in this paper (see Figure 3.16). Furthermore, the p-values of the diseases with their respective protein are considered, and a corresponding line graph is constructed (see Figure 3.17). Each line of the graph represents a particular disease marked in the figure. The reported diseases include those characterized by a high risk of organ failure and others showing an impact of COVID-19. Those common diseases are selected and a network is built between the proteins and their resulting diseases. As particular organs are highly affected

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

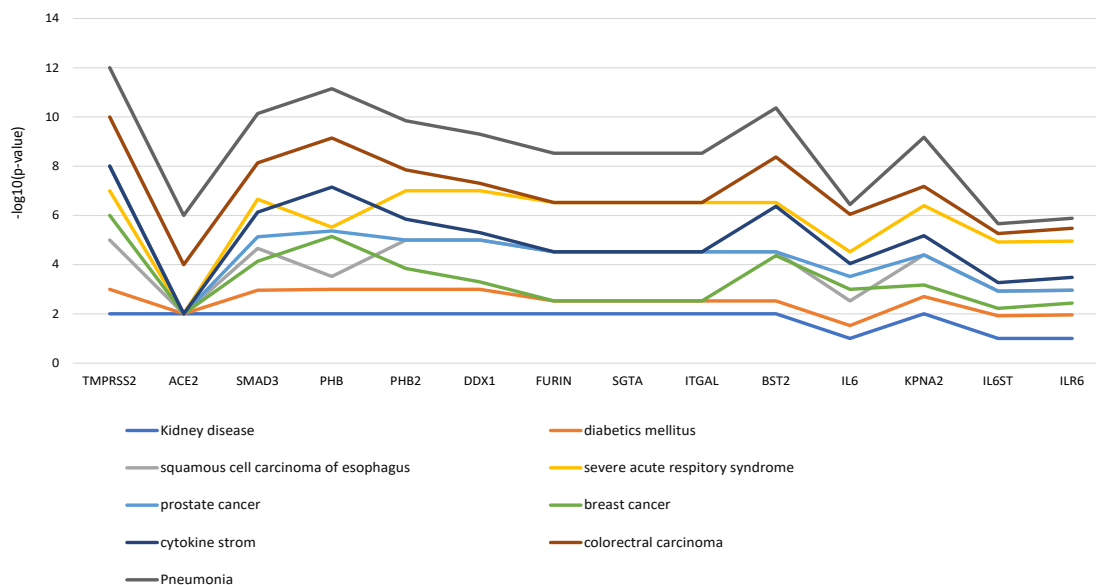


Figure 3.17: Depending on the P-value of the proteins associated with their respective diseases, the line graph is constructed.

due to this disease, we have categorized the diseases according to a specific organ and represented them with diverse colors. The proteins are selected by analyzing the Venn diagram based on the maximum number of common diseases among 14 proteins. In the network, oval and square-shaped nodes represent proteins and diseases, respectively. On the other hand light pink, deep blue, light green, deep green, violet, grey, c-green, orange, and light blue colors of nodes represent different diseases related to COVID-19, such as cardiovascular disease, mental health, immune system diseases, respiratory disease, cancer, diabetics, kidney disease, gastrointestinal disease and others, respectively.

### 3.7.4 Evolutionary Sequence-Structure Space Study of Proteins E, M, and N

Many membrane proteins are known to have biologically important IDRs. The large conformational flexibility of those regions might be related to the multifunctionality of these proteins [83]. Among the SARS-CoV-2 proteins, the M and E proteins are quintessential membrane proteins. Figures 3.18A and 3.18B represent the intrinsic disorder profiles generated for these proteins using several commonly used per-residue disorder predictors and show that although both M and E proteins are mostly ordered, they are expected to have disordered N and C-terminal regions. This is also in line with the results of the hydropathy analysis of these proteins (see Figure 3.13), which shows that these proteins are enriched

### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

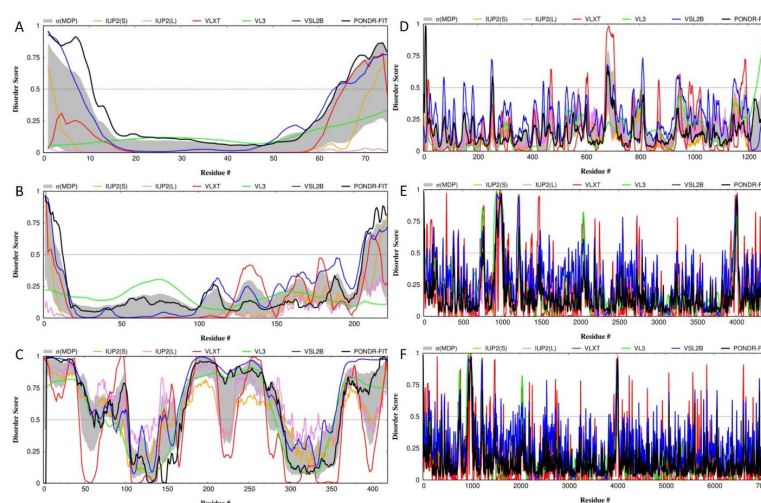


Figure 3.18: Evaluation of intrinsic disorder predisposition of major SARS-CoV proteins: envelope (A), membrane (B), nucleocapsid (C), spike (D), ORF1a (E) and ORF1ab (F). These profiles were generated using DiSpi web crawler that aggregate the results from a number of well-known disorder predictors: PONDRVLXT (25), PONDRVSL2, PONDRVL3, IUPred short and IUPred long and PONDRFIT.

in hydrophobic residues (i.e., residues with positive hydropathy values in the Kyle-Doolittle scale). This enrichment in hydrophobic residues is a typical hallmark of transmembrane proteins. On the other hand, N and C-tails of both M and E proteins are enriched in hydrophilic, polar amino acids. In this analysis, outputs are per residue predictions of the intrinsic disorder propensity in the [0, 1] range. These outputs are then compared to a threshold (we used the default threshold of 0.5), and residues with a prediction value greater than the threshold were predicted to be intrinsically disordered (Shown in Figure 3.18). The local increase in the intrinsic disorder tendency for an ordered protein with overall low disorder propensity scores is expected to correlate with the increased mobility of the studied region [84, 85, 86, 87, 88, 89]. In SARS-CoV, the functions of the M and E proteins are diverse and enigmatic. The higher sequential similarity enhances the possibility of physiological similarity for these proteins in SARS-CoV-2. In the case of the M protein, the assemblies with three distinct classes provide a specific set of functions. Specifically, E and M are involved in intercellular trafficking in Endoplasmic Reticulum (ER), Golgi Apparatus, and ER-Golgi intermediate compartment (ERGIC) after host cell intrusion. Mechanistic advantages have been attributed to the structural flexibility of some of the parts of these proteins, which may provide a higher adaptation ability at diverse conditions (at least more than

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

the soluble proteins) [90]. In Figure 3.14, the residual occupancies have been highlighted for both proteins. The conservation within the N- and C-terminal regions are quite strong. However, the intermediate part is extremely susceptible to mutations in the E protein, whereas, for the M protein, the intermediate part is slightly conserved. On the other hand, the coupling analyses have shown a different aspect based on evolutionary co-variation or co-occurrence. In MI, the distributions of the coupled pairs are analyzed within the sequence space of E and M proteins. However, DI shows a slightly different aspect. Firstly, DI considers only the coupling propensities (or co-occurrence trait) between two residues in a disentangled manner. The coupling score-based weighted network has a minimal number of residues. Subsequently, the application of the eigenvector-based compartment distribution clarifies the evolutionary crosstalk between residues, which are extremely conserved within a few color modules. In E proteins, most of the residues are distributed within two color modules, i.e., cyan and blue, where the residual range is within 100. However, the residual occupancies are evolutionary random at those sequential positions. As mentioned earlier, DI weighted network is quite different for M proteins. Interestingly, the M protein network has distinct modules, which are highly aligned with SE scores. More elaborately, highly random sequential columns belong to the same-coloured module. The longest conserved region as per SE scores has been kept within the blue-coloured module. It emphasizes an important point that the rate of flexibility may enhance the rate of adaptability. It is quite evident, as E and M proteins are basically sub-classes of membrane-bound proteins. The amino acid sequences of the M proteins are extremely variable. In homologically similar viruses, the disorder rate of M proteins is within 4%-14% at least. However, not much information about E proteins is known. In the structure space study, the square-fluctuations of E and M proteins are almost similar. However, the residues at the structure network of E and M proteins are distributed in such a way that even a small percentage of the disorder can affect the whole structure. In the case of M proteins, the C-terminal residues are conserved with transmembrane residues with two structure network modules. In the case of E proteins, the two bigger modules include a maximum number of residues, which covers almost the whole structure (all shown in Figure 3.15). Therefore, the sequence vulnerability can provide structural adaptability in diverse conditions. Similarly, N proteins, known for their RNA binding property, are also being studied in a similar manner. In [91], it has been shown that the disorder at the binding zone for nucleotide-binding protein is quite obvious. In fact, a systematic computational analysis of the nucleosomes (~548 000 nucleic acid binding proteins) of 1121 species from Archaea, Bacteria and Eukaryota revealed the prevalence of the intrinsic disorder in these proteins [92]. Based on the analysis of 5658 dissimilar (below 50% sequence similarity) proteins with known

### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

3D-structures that bind to proteins, DNA or RNAs it was also pointed out that disorder is crucial for the formation of many protein–protein, protein–DNA and protein–RNA complexes, where IDRs undergo the binding-induced disorder-to-order transitions [93]. Also, the significant rate of disorder has been predicted for N proteins in [94]. Interestingly, N and C-terminal regions of N proteins are IDRs (see Figure 3.18C). However, we are expecting that IDRs at N proteins might have a strong impact on the receptor binding domain (RBD). Initially, the N proteins bind with the viral RNA genome. It is also involved in the regulation of the replication cycle and is involved in the host cellular response to viral infection, which finally leads to the formation of virus-like particles (VLPs). Usually, DNA and RNA binding proteins are known for their lower hydrophobicity. In Figure 3.13, the hydropathy plot provides clear support for this statement. In Figure 3.14, the residual occupancy of N proteins shows a random distribution of the SE score throughout the sequence space, where the positions of some of the residue are locally conserved. Interestingly, the coupling propensities of the N proteins are almost similar to those of the M proteins. In the MI map, MI patches are less distributed than in E and M Proteins. However, the communities of the weighted network based on DI scores are equally distributed within the residues. Interestingly, some of the residues from each module are from the conserved portion of the SE score distributions. Therefore, the sequence space is extremely vulnerable and random. In [95], the structurally disordered rate of the homological similar viruses for N proteins is around 48%-57%. Therefore, IDRs at the N and C-terminal regions of the N protein can affect the RBD and nucleotide-binding domains. These also reflect in the structure network shown in Figure 3.15. In the structure network, almost all the residues are conserved within blue modules. The previously analyzed information in terms of sequence space and structure space clarifies the trait of unfolding scenario at the monomeric stage.

#### 3.7.5 Evolutionary Sequence-Structure Space Study of the Spike S Glycoprotein

For S proteins, not much information has been provided by previous researchers. As per some of the information from previous research, spike proteins are expected to have a lesser disorder. However, structurally, spike glycoproteins are full of  $\alpha$ -helices. Therefore, it raises questions on the possibility of the high prevalence of the structural disorder in this protein. In agreement with this notion, Figure 3.18D shows that the S protein is predicted to be mostly ordered, but still contains several short IDRs. In Figure 3.13, the distribution of the hydrophobic index is shown. The lower levels of hydrophobicity can enhance the chances of the local unfolding of a protein at changes in the environmental condition. Figure 3.14 shows that the residual occupancy is extremely random throughout the entire space, with

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

conserved regions being extremely small. However, N and C-terminal regions are sequentially conserved. In the MI map, the distribution of the coupling patches is low mainly at the tail of the sequential distribution. Interestingly, the DI-based residues and cross-talking through the communities consist mostly of the tail-end residues. Figure 3.15 shows that the square fluctuation is comparatively low. It may possibly be due to the length of the sequence because we have considered the normalized square fluctuations here. However, the community distribution in the structure network clarifies its inability to achieve the stability stage at the monomeric phase. Almost 620 (which is more than 50% of the sequence length) residues including the RBD are grouped within the red module. Therefore, the spike glycoprotein can also be an IDP with greater flexibility.

### 3.7.6 Evolutionary Sequence-Structure Space of the Replicase Polyprotein ORF1ab

Replicase polyproteins include a set of proteins, which are firmly associated with the replication of the viral RNA. Individually, they are performing distinct functions. After viral uptake, replication of nucleotide is a vital role, where a large number of the host cell proteins, such as Prohibitin and Prohibitin 2, are engaged. During this process, the vital cellular organelles, such as mitochondria and ribosomes are participating through the viral interactor counterparts. Basically, the process is initiated by the host cell mRNA degradation. The complex of Nsp1 and 40S ribosome creates endonucleolytic cleavage near the 5' UTR of the host cell mRNAs, which assists the mRNA degradation process. However, the leader sequence at 5' UTR of the viral mRNA protects it from the action of the Nsp1-40S complex [96]. Likewise, Nsp2 is interacting with Prohibitin and Prohibitin 2 to initiate the cell survival signals [97]. Each viral protein derived from the polyproteins is performing a distinct set of functions, and such multifunctionality can be attributed to the structural flexibility of these proteins. Figures 3.13, 3.14 and 3.15 represent the results of the detailed sequential and structural analysis of this protein. We have only shown the results for Nsp1. As per the hydropathy index, the distribution of the hydrophobic and hydrophilic residues within this is rather random. Therefore, the presence of a continuous disordered region cannot be observed throughout the sequences. Figures 9E and 9F are in line with the results of the previously published studies, where the sequences of proteins derived from the polyproteins ORF1a and ORF1ab were shown to have large sets of structurally ordered patches. However, all these proteins are expected to have some small sequence patches of disorder [98]. Interestingly, IDRs are effective enough to participate in functional and dysfunctional protein interactions. Therefore, we have aimed to observe the fluctuating hubs of the replicase polyproteins. As discussed earlier, the hydropathy distribution is more variable. Similarly, the SE score dis-



### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

tribution over the MSAs shows high entropy through the evolutionary sequence space. It shows that the position of each residue contains a higher rate of substitution frequencies. Interestingly, sequential co-variation study has shown that the long sequential dependencies are conserved within the one color module; i.e., the cyan module. However, square fluctuation shows a higher rate at the N-terminal end of the structure, whereas the structure network shows localized structural communities. From these results, the dependency of each residue position is clear, which also shows that disorder at any residual point can have a strong impact on the rest of the structure.

#### 3.7.7 Host Proteins and Corresponding Viral Protein

At least 15 host cell proteins were shown to be activated during the virus intrusion. All of these are directly or indirectly affected by viral proteins and are involved in direct or indirect interactions with them. ACE2 and TMPRSS2 [99, 100] are two well-known host proteins, which govern the initial virus entry through spike glycoproteins. In this case, we have considered host cell proteins, that are interacting with or affected by SARS-CoV-2 proteins, and information about which is available in UniProt. In previous focused studies, these selected human proteins were categorized as per their involvement in viral infection. After viral uptake into the host cell, the viral proteins first attenuate or stop the regular protein functions. Subsequently, the cell cycle oscillation is used to grow and created multiple copies. Interestingly, not all human proteins have been performed to enhance the efficiency of infection and increase the rate of virus production in the host cells. Some of these proteins attenuate the proteins, which are involved in the pathogenic progression. RNA helicase-DDX1 complex, clearly observed in SARS-CoV, is one such example, which is related to the defence responses against the virus infection [101]. Basically, a large number of host cell proteins are involved in the viral genome replication through subparts of the replicase polyproteins. ORF1ab replicase polyprotein is a long polypeptide (7,096 residues; UniProt ID: P0DTD1) that includes multiple proteins with various activities. Specific enzymatic cleavages of this polyprotein in vivo by its own proteases, 3CLPRO and PLPRO proteinases, which are autocatalytically processed, yield mature proteins. From the N to C-terminus of the ORF1ab polyprotein, these functional proteins are Host translation inhibitor Nsp1 (residues 1 – 180), Non-structural protein 2 (Nsp2, residues 181 – 818), Nsp3 (residues 819 – 2763), Nsp4 (residues 2764 – 3263), Nsp5 or 3C-like proteinase (3CLPRO, residues 3264 – 3569), Nsp6 (residues 3570 – 3859), Nsp7 (residues 3860 – 3942), Nsp8 (residues 3943 – 4140), Nsp9 (residues 4141 – 4253), Nsp10 (residues 4254 – 4392), RNA-directed RNA polymerase (residues 4393 – 5324), Helicase (residues 5325 – 5925), Proofreading exoribonuclease (residues 5926 – 6452), Uridylate-specific endoribonuclease (residues 6453 – 6798), and 2'-

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

O-methyltransferase (residues 6799 – 7096). Replicase polyprotein ORF1a is a 4,405-residue-long polypeptide (UniProt ID: P0DTC1) that is produced by the ribosomal frameshifting and includes the same set of proteins found within the first half of ORF1ab (residues 1-4392); i.e., Nsp1-4, 3CLPRO, and Nsp6-10, but instead of a set of replication-related enzymes found in the C-terminal part of ORF1ab, ORF1a contains just one protein Nsp10 (residues 4393 – 4405). Starting with Nsp2 [102], each of the proteins excised from the viral polyproteins is performing its specific functions and is playing its specific roles in assisting the viral genome replication, modulating the cell growth, and/or controlling the pathogenic intrusion into the nucleus. As already pointed out, Nsp1 forms a complex with 40S ribosome, and this complex initiates the process of viral replication by inhibiting the host cell mRNAs. This complex binds at the 5' UTR of the host mRNAs and promotes their degradation. However, the viral mRNAs are not degraded, due to the presence of specific 5'-leader sequences. Subsequently, Nsp2 interacts with host PHB and PHB2 proteins and sends the survival signal, which helps keep the normal cell machinery and metabolism active. Likewise, each other Nsps has a specific set of roles associated with interaction with the particular host cell proteins. An interesting example is given by FURIN, which shows an almost similar type of proteolytic behavior as TMPRSS2 [103]. In the case of TMPRSS2, the viral entry mechanism has been facilitated in two ways, e.g., proteolytic cleavage of the host receptor ACE2 and proteolytic cleavage at the s1/s2 domain boundary of spike glycoprotein [99]. However, furin utilizes a proteolytic cleavage site at the spike glycoproteins [104]. SARS-CoV papain-like protease (PLpro) induces Egr-1 dependent up-regulation of the transforming growth factor- $\beta$ 1 (TGF- $\beta$ 1) through the ROS/p38/MAPK/STAT3 pathway, thereby up-regulating the pro-fibrotic responses in vitro and in vivo [105]. SARS-CoV PLpro also caused the change in the ubiquitination profile of Rho GTPase family proteins and plays a role in the TGF- $\beta$ 1-dependent expression of Type I collagen via activating STAT6 pathway [106]. Also, SARS-CoV nucleocapsid (N) protein potentiates TGF- $\beta$ 1-induced expression of plasminogen activator inhibitor-1 and attenuates the Smad3/Smad4-mediated apoptosis of human peripheral lung epithelial cells [107]. Many other host cell proteins either work to assist the viral uptake or are affected due to viral uptake and corresponding pathogenesis. Therefore, viral proteins are capable of modulation of numerous host cell proteins. Therefore, it is possible that such virus-induced distortion in the normal proteostasis within the host cells can be associated with certain comorbidities found in recovered patients in the post-COVID scenario. The list of host cell proteins associated with specific diseases has been provided.

### 3.7. RESULTS OF PROTEIN-SPECIFIC STUDY FOR SARS-COV-2

---

#### 3.7.8 Comorbidities and SARS-CoV-2

In continuation of the previous part, we have analyzed each of the host cell proteins for their association with diseases. In some of the previous studies, researchers have created a distinct classification in terms of associated diseases, where they have mostly studied the diseases during the infection and associated comorbidities. Apart from a regular set of comorbidities (e.g., ARDS, Diabetes, chronic liver or kidney diseases etc.), several more diseases, such as malignancies and neurological disorders have been observed in infected patients. In [108], a group of patients has been observed of raising the risk of schizophrenia. Another paper showed a set of infected male patients, having prostate carcinoma. We are expecting that the developing pathogenic conditions might have a strong connection with dysfunctional host cell proteins. The disordered nature of viral proteins makes them extremely sensitive to subtle changes in environmental conditions but also prioritizes them as promiscuous interaction partners. Therefore, the different classes of viral proteins may modulate the corresponding host cell proteins (some of the examples are given in the previous section). In this regard, we have searched all the diseases associated with dysfunctional proteins. Among such associated diseases (or comorbidities) are the possibilities for co-infection, e.g., with the Influenza A virus. However, the main effect of SARS-CoV-2 pathogenesis is related to the 'cytokine storm' theory. In [109], the importance of IL6 has clearly been described. Following the 'cytokine storm' theory [98], two possibilities have been observed. First, acute inflammatory activities are attributed to the increased levels of many cytokines and chemokines, specially IL6, with the level of IL6 being used as a predictor of the infection severity. Second, this 'cytokine storm' in Central Nervous System (CNS) can trigger neuroinflammation. Apart from acute lung or cardiovascular diseases, two distinct large classes of diseases can be initiated through 'cytokine storm', especially due to the increased IL6 levels. Associated diseases are more elaborately described below.

#### 3.7.9 Malignancies and SARS-CoV-2 Infection

Along with other common diseases, such as hypertension, diabetes type I and type II, malignancies can be among the vital comorbidities of COVID-19. It has been shown that malignant patients are more vulnerable to infection than non-malignant patients. Furthermore, during infection, patients with lung cancer are more prone to die than patients with other cancer types. The study also showed a higher possibility of cancer survivors having severe conditions than normal patients [110]. The outcomes of a multi-cancer study support the previous findings. However, the pro-inflammatory conditions due to the cytokine storm can raise the possibility of malignancies apart from the diseases like ARDS. Pro-inflammatory

## CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

mediators, such as IL6, TLRs, and TNF- $\alpha$ , help in tumorigenesis. Specifically, IL6 works to protect the cancer cell from DNA damage, apoptosis, etc. Attenuation of IL6 is considered a therapeutic technique for malignancies. From multiple studies, we have observed that the upregulation of IL6 can be one of the roads towards malignancies from SARS-CoV-2 infection. In fact, SMAD3 protein, which is known as a prime tumor suppressor, has been attenuated during cytokine storm [111]. This also enhances the possibility of multiple malignancies. More elaborately, SMADs, and especially SMAD3, are responsible for TGF-mediated immune suppression, which would later promote favourable conditions for metastasis. Therefore, systematic expression of SMAD3 proteins can maintain the activity of the TGF- $\beta$  signaling. So far it has been observed that the cross-talk between TGF- $\beta$  signaling and different pathways is responsible for different cancers. For example, the cross-talk between IRS-1 signaling pathway and TGF- $\beta$  signaling in colon cancer or a cross-talk between the BCL signaling pathway and TGF- $\beta$  signaling in hepatocellular carcinoma. Therefore, the attenuated levels of SMAD3 may enhance the chances of malignancy development, especially in the post-COVID situation. Importantly, two proteins (TMPRSS2 and FURIN), which are exclusively associated with uptake, are directly and partially androgen-modulated. Mollica et al. [112] showed that the promoter of the TMPRSS2 can be modulated via androgen receptors (AR). Lucas et al. showed how the androgen-mediated protease TMPRSS2, which plays an important role in viral uptake, might help in creating the proteolytic cascade during prostate carcinoma and prostatic neoplasms. Therefore, the androgen levels can be the determinant factor to understand viral vulnerability, which is two times in males. However, the amount of the androgen hormone varies among the genders. More specifically, males and postmenopausal females, who have elevated androgen levels, are more vulnerable and susceptible to the infection. Among the disease classes, different types of malignancies are associated with a maximal number of host cell proteins affected by SARS-CoV-2. This number is alarming not only during the infection but also in the post-COVID situation.

### 3.7.10 SARS-CoV-2 and its Impact on Neurodegenerative and Neuropsychiatric Diseases

In the case of neurodegenerative and neuropsychiatric diseases, the disease possibilities have been increased due to the neuroinflammation mediated by SARS-CoV-2 infection. From the results, three prominent cases viz., Cytokine Storm, Nervous System Disorder, Chronic Depression, etc. have been identified to be associated with host cell proteins. Some recent research has featured the impact of cytokine storm in the progression of neurological disorders [113]. On the other hand, the interleukins, and chemokines from the host cell have already been ob-

### 3.8. CONCLUSIONS

---

served to be associated with multiple neurological conditions [114]. Therefore, the link between two pathological conditions viz., cytokine storm and neurological disorders. As per previous evidence, neuroinflammatory responses are very common and usual for diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), multiple sclerosis (MS), etc. due to the intervention of the pro-inflammatory mediators, such as chemokines, cytokines, interleukins, etc. Cytokine storm at CNS can affect the microglial cells, a specialized population of macrophages. Usually, these cells show a stable and inactive immunophenotype in a healthy brain. However, functional dysregulation of the microglial cells can occur due to the unusual interactions with cytokines and also due to exposure to soluble Amyloid-beta ( $A\beta$ ) (during the pathogenic progression of dementia in AD). On the other hand, microglial cells are associated with the phagocytosis of extracellular  $A\beta$ . These could be the reasons behind increasing the rate of AD. Though, the relation between immunology and neurodegenerative and neuropsychiatric diseases is not linear. A similar type of neuroinflammation is also observed during the cellular deposition of the  $\alpha$ -synuclein in PD. Basically, the cytokine storm during the intrusion of SARS-CoV-2 may disrupt the blood-brain barrier and may be associated with the early progression of the aforementioned diseases [115]. Also, a few cases of schizophrenia have also been noticed in the post-COVID population [116]. Though the reasons for this increase in schizophrenia incidences are not clear. However, the relation between viral infection and neuropsychiatric diseases is well known. Some of the previous research has shown that chronic viral infection is responsible for losing cognitive senses [117, 118]. SARS-CoV-2 can follow the same path, as Rogers et al. have summarized in their recent review [119]. As per the review report, multiple case studies have shown that neuropsychiatric dysfunctions, including anxiety, insomnia, depression, impaired memory, etc. are associated with 7%-41% of the COVID-19 cases, whereas the overall percentage of affected patients is around 63%. Two possibilities can be discussed in this regard. First, the genetic disorders, where the infection can work as a dominator [110]. Second, neuroinflammation can be another reason behind it.

### 3.8 Conclusions

---

In the TF druggability study, the objective is to define the TF druggability for preventing prostate cancer. Addressing protein moonlighting properties of TFs, we have identified evolutionary covarying patches initially. Along with that, a specific set of non-pathogenic pathways has been selected. Dysregulation of these pathways can initiate prostate carcinoma. A total of six TFs are selected viz., AR, ETS1, CREB1, CEBPB, RELA, and NFKB1. Similarly, all the TFs are strongly associated with two signaling Pathways, viz., the cAMP signaling pathway and

### CHAPTER 3. UNDERSTANDING THE DYNAMICITY OF PROTEINS IN TERMS OF NETWORK

---

the Oxytocin signaling pathway. In the study, we have shown probable drugs for three TFs viz., CEBPB, ETS1, and CREB1 taken from CRAFFT. Therefore, adjusting the structural malleability of the TFs, and targeting the TFs can be much more effective while residing in the aforementioned non-pathogenic pathways. Hence, this frame can be applied to addressing similar kinds of malignancy-related issues in the future.

The sequence-structure-based study unveils the possibilities of the structural malleability of all five viral protein classes. These viral proteins have some pre-determined host protein interaction partners. The structural disorderedness of these viral proteins prioritizes them as interaction partners within the host cell. Here, the viral proteins can strongly associate with host cell proteins and attenuate their usual activity. Such a mechanism increases the possibility of multiple other diseases as comorbidities and post-COVID effects. The host cell proteins and corresponding diseases are divided into two distinct classes Viz., proteins, directly associated with the set of diseases while showing similar activities; cytokine storm-mediated pro-inflammation (e.g., Acute Respiratory Distress Syndrome, malignancies), and neuroinflammation (e.g., neurodegenerative and neuropsychiatric diseases). This list shows the post-COVID disease possibilities, which become one of the leading reasons for death for many COVID recovered patients. Finally, the study also reveals that males and postmenopausal females can be more vulnerable to SARS-CoV-2 infection due to the androgen-mediated protein transmembrane serine protease 2.

# 4

## Graph-theoretical modeling to unveil the cell-to-cell heterogeneity

### 4.1 Introduction

---

In the previous two chapters, experiments are performed based on bulk RNA expression and protein abundance. Traditional bulk omics data captures multiple cells for a given sample and consider all as a homogeneous material. This leads to averaging effect, thus missing important signals on cellular heterogeneity. Recently, technology for sequencing single cells has become available. The advancement of technologies made it easier to understand that cells can show variability at any omics layer under diverse conditions. Insipite of the advantages of the single-cell method, bulk technologies are more affordable in terms of lower technical noise, and experimental procedures, and they do not require living cells. More importantly, bulk methods focus on individual tissues and help to unveil the biological differences among diffident species. However, they fail to address the cell-to-cell disparity of heterogeneous cell populations and are unable to analyze a small number of cells.

With the emerging sequencing technologies, it is evident that cells can possess different transcriptomes, epigenomes and proteomes even though they are derived from the same cell line or present within a single tissue. Cell populations derived from the same tissue of an organism are heterogeneous. Cell-to-cell variability is responsible not only for diversity in cellular states but also affects interactions among numerous other distinct cells. Therefore, studies conducted at the single-cell level are needed to uncover the underlying complexity of biological systems and to gain a greater understanding of biological processes. Since single-cell profiling techniques are still in their early stages of development, the data they produce are typically sparse and have high levels of technical noise. As a result, it makes interpreting results and downstream data analysis more difficult. Therefore, advanced bioinformatics pipelines are required for gaining a deep understanding of biological heterogeneity and contributing to therapeutic decisions.

## **4.2 Computational framework for understanding cellular heterogeneity in disease severity**

---

In this chapter, the main aim is to design a computational framework to unveil the cell-to-cell heterogeneity of multicellular organisms in disease initiation, development, and progression. In this regard two cases are considered, first one is the glioblastoma multiforme cell-specific single-cell data. Glioblastoma is one of the most aggressive malignancies that occurs in the brain and spinal cord area. In spite of advancement in clinical strategies, this disease shows a poor prognosis. In this regard, a bioinformatics pipeline is proposed during this study. This pipeline helps to understand cell-to-cell heterogeneity and its impact on tissue development. Moreover, the impact of the diverse cell types on a particular tissue is also observed.

On the other hand, single-RNASeq data is considered for the COVID-19-affected organs. An organ-specific framework is proposed to identify the cell types involved during the disease initiation. Moreover, a pathway semantic model is applied to decipher the contribution of the internal pathway network of each participated cell type towards the infection. The methods are discussed extensively throughout the chapter.

### **4.2.1 Bioinformatics pipeline to unveil the heterogeneity of Glioblastoma Multiforme**

To understand the cellular heterogeneity in disease state, we performed genetic and molecular studies to enhance one of the most aggressive malignancies i.e., glioblastoma multiforme (GBM). Cell-type astrocyte which supports the nerve cells are mainly involved in GBM [120]. This is the most common central nervous system malignancy that may occur at any age but is mostly found among older people. Since it grows very quickly, it spreads easily to other brain tissue. Even after various advancements in clinical strategies, GBM remains a deadly disease with a poor prognosis [121]. Medical help only may slow down the progression rate of cancer and reduce signs and symptoms, however, the cure is remaining a challenge till now. Patients show a maximum survival rate of 2 years after the first diagnosis. In this regard, genetic and molecular studies are required to enhance the understanding of the etiology of GBM. Depending on this fact, previously, various studies are performed to decipher the regulation of the responsible biomarker of the disease and their impact on it [122]. However, these studies are performed from bulk RNA level, which may result in missing important biological signals of a particular cell. Recently, it is evidenced that numerous cells contain identical genetic information, whereas their functional disparity makes them unique. Moreover, local biological networks are responsible for disease progression and



## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

these networks help in the diversity of the therapeutic response for a particular malignancy [123]. In 2019 a study evidenced that, cell-specific networks were capable to reveal the ‘unstable’ gene expression form to a ‘stable’ gene transformation mechanism. In addition, the dark genes are identified from their generated network, which played important roles at the network level but are generally ignored by traditional differential gene expression analyses [124]. On the other hand, an algorithm locCSN was published to construct the local cell-specific networks for two brain cell samples. Their application showed the evolution of gene networks in fetal brain cells by comparing the cells sampled from case and control subjects [125]. Therefore, it is important to understand the disease heterogeneity at a cellular level by constructing a cell-specific local network, which is the main challenge in biology and medicine.

In this study, we have identified the cell types involved in GBM by utilizing Louvain clustering [126]. The differentially expressed genes of each cell are considered to construct a cell type-specific protein-protein interaction (PPI) network. Additionally, the interactive communities in the network are detected, known as functional hubs, by applying the maximal clique centrality (MCC) method [127], as this is reported as the most efficient method to detect the biologically important functional hubs. Furthermore, these functional hub genes are considered to establish the local networks of the disease cell types. The local networks formed based on the TFs are targeted to the differentially expressed hub genes of each cell type. However, these TFs are responsible to regulate the mode of the biomarkers and become the cause of the poor prognosis of the GBM disease. Interestingly, we found NFKB1, RELA, STAT3, SP1 etc. as crucial TFs responsible to regulate the mode of the biomarkers. However, these TFs can be considered as suitable therapeutic targets to design clinical strategies. Previously, it is evidenced that biological pathways are the set of genetic factors through which cells are communicating with each other in order to perform a particular functional task. Abnormalities in these signaling pathways may lead to disease progression. In this regard, we identified the pathways that played a key role in this disease and established their relationship with the cell types. This study reveals network communities that represent the functional modules of the disease. Finally, these identified modules can be considered to design a potential single-cell network for personalized medicine.

### 4.2.1.1 Data acquisition

The ScRNASeq glioblastoma multiforme data is downloaded from a publicly available dataset, [www.10xgenomics.com](http://www.10xgenomics.com). The dataset is accumulated from 57 years of male donor with, 5604 estimated cells. Furthermore, the transcription factors are identified as those are targeted to the cell-specific genes from the TRRUST database [128]. This database accommodates 8444 human TF-target interactions

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

---

between 800 TF genes and 1975 non-TF genes. Moreover, the mode of regulation of the TFs is also provided in this database.

### 4.2.1.2 Proposed method

In the first phase of the study, the cell types are identified by utilizing Seurat V3.0 [27] method on the raw count matrix. The undesired cells from the dataset are removed by considering 2,500 or less than 200 unique features. Additionally, cells are filtered with  $> 5\%$  mitochondrial counts. This filtered count matrix is further normalized by the "LogNormalize" function. The data has been scaled by multiplying a value by 10,000. Later, principal component analysis (PCA) is performed in order to construct the k-nearest neighbor graph. From the established graph, the cells are clustered with the implementation of the Louvain algorithm. The top differentially expressed genes from each cluster are identified by the "FindMarkers" function, a minimum percentage of this function is set to 0.23. Furthermore, Wilcoxon rank sum statistical test is performed to sort the cluster-specific identified marker genes according to their rank.

The differentially expressed genes of each cluster represent the characteristics of the cluster and define the functionality of the particular cell type. Depending on this fact, the top genes of each cluster are considered to identify the respective cell type of each cluster. In this regard, publicly available databases such as CellMarker [129] and Paglaodb [130] are used.

In the next phase of this study, the identified cell type-specific biomarkers are further considered to find the core of interactions from PPI network. The STRING database [32] is utilized in this regard. Once the cell-specific PPI network is established, we have aimed to find the interactive and important biological genes, called hub genes of each network by applying the maximal clique centrality method. This method calculates the score of the nodes present in a network with local rank method. The direct relation between nodes and their neighbors is observed in the local method. Among different local-based methods, we applied the MCC in order to detect the feature node with increased sensitivity and specificity. The logic of MCC believes that biologically important genes are clustered together in a network. Following that, the calculation of MCC is discussed below: Let  $n$  be a given node, then MCC of  $n$  is derived in equation 4.1:

$$MCC(n) = \sum_{C \in S(n)} (|C|-1)! \quad (4.1)$$

Here  $S(n)$  is the set of maximal cliques which contain the node  $n$ , and the product of all the positive integers is denoted as  $(|C|-1)!$ , which is less than  $|C|$ .  $MCC(n)$  becomes equal to its degree when an edge is missing between the selected node and its neighbours.

## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

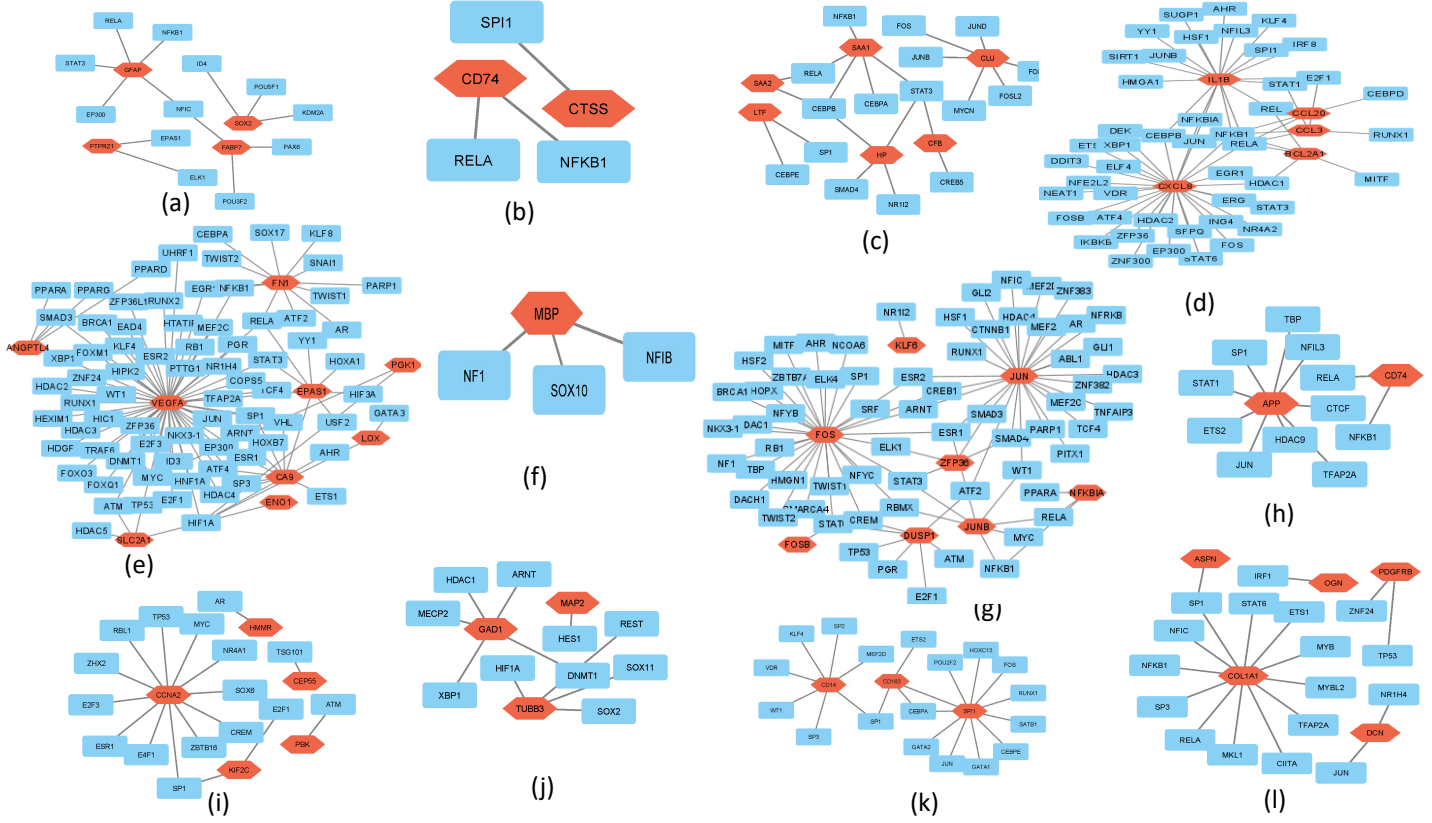


Figure 4.1: The local networks are established for (a) Oligodendrocyte, (b) M2 macrophage, (c) Stem cell, (d) Regulatory T cell, (e) Vascular endothelial cells, (f) B cell, (g) Astrocytes, (h) Endothelial cell, (i) Neural Stem cell, (j) Inhibitory neurons, (k) Macrophage, (l) Pericytes through single-cell study. The blue and red nodes of the graphs represent the transcription factors and hub genes, respectively.

The hub genes are considered to establish the local network biology for each cell type. The TFs are curated from the TRRUST database, those are responsible to regulate the mode of the biomarkers during the disease progression. Moreover, the established TF-gene network is observed to reveal the pathogenicity responsible for the mentioned disease. The framework of the proposed method is shown in Figure 4.2.

### 4.2.1.3 Experimental results and validation of the method

The objective is to identify the cell-specific biological networks that played a key hub during GBM progression. In this regard, the Louvain clustering algorithm is applied to GBM ScRNASeq data in the first phase of the study. The results from clustered biomarkers represent the heterogeneity of the cell types. With the

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

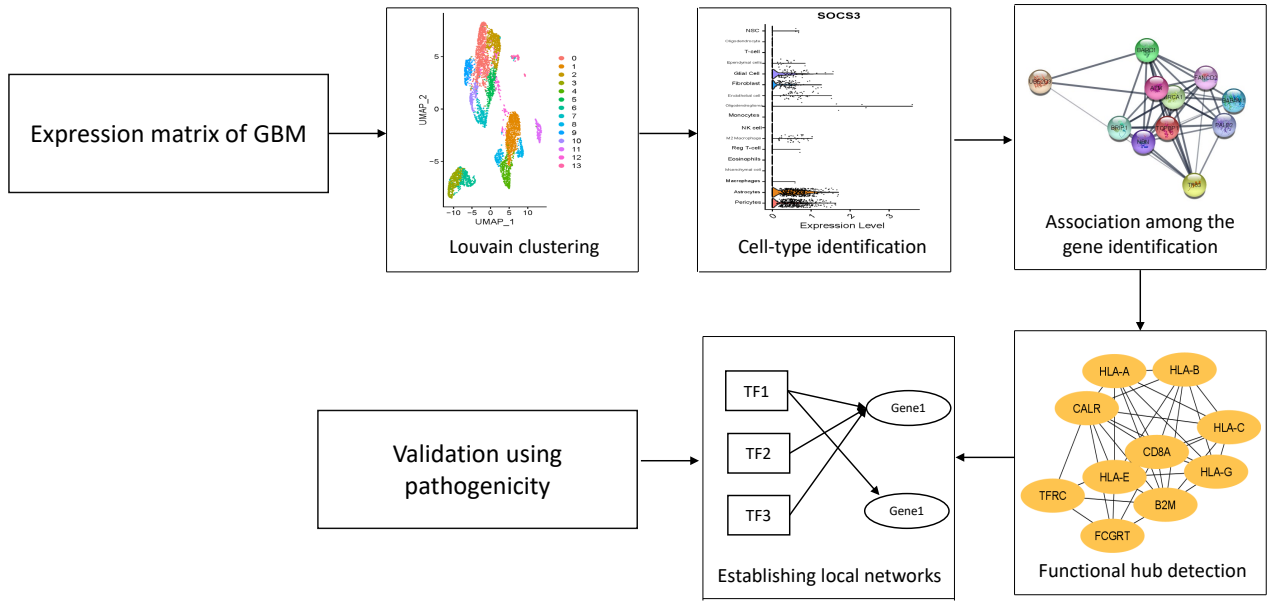


Figure 4.2: The framework of the proposed method.

advancement of technologies, the single-cell study helps to identify the cell types which reveal the group of cells responsible to perform a particular function in the respective disease. This information opens a new path to solving biological complex problems. We have acquired 14 cell types. Publicly available databases are used to identify the cell types depending on the top 50 ranked gene markers of each cluster. It is noticed that three clusters represent the same cell type such as oligodendrocyte, which infer that the clustering algorithm can identify the subtypes also. Interestingly, the identified cell types play a crucial role in GBM disease progression.

As our study is to understand the local biological networks responsible for aggressive malignancies, we performed a cell type-specific gene regulatory network. In this regard, we performed a protein-protein interaction network among the biomarkers of a particular cell type. Once the network is established, we are tried to identify the hub genes using the MCC method. According to the MCC score, among the top ten hub genes, five biomarkers are reported in Table 4.1. along with their respective cell type. From the literature, we observed that identified hub genes have a high impact on the functionality of the cell types [131]. Moreover, abnormalities in the mode of regulation of the biomarkers may lead to disease severity. To understand the regulation of the biomarkers, we established a transcriptional regulatory relationship network locally. The hub genes are considered to identify the transcription factors. During the study, we found that biomarkers of clusters 3 and 11 are not targeted by any TFs and interestingly, these two cell

#### 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

Cluster ID	Cell type	Cell Markers
0,3,11	Oligodendrocyte [132]	GFAP,GAP43,CHL1,FABP7,PTPRZ1
1	M2 macrophage [133]	C1QB,LY86,C1QA,FCGR3A,CTSS
2	Stem cell [134]	ERMN,MOBP,MOG,HAPLN2,CNTN2
4	Regulatory T cell [135]	IL1B,TYROBP,CCL3,BCL2A1,CXCL8
5	Vascular endothelial cells [136]	SLC2A1,VEGFA,LOX,CA9,FN1
6	B cell [137]	MAG,PLP1,MOG,ERMN,MBP
7	Astrocytes [120]	JUN,FOS,DUSP1,ZFP36,FOSB
8	Endothelial cell [138]	CD74,AIF1,TYROBP,C1QB,FCER1G
9	Neural Stem cell [139]	NUF2,CCNA2,CDCA5,PBK,HMMR
10	Inhibitory neurons [140]	DCX,TUBB3,GAD1,MAP2,TUBA1A
12	Macrophage [132]	C1QB,FCGR3A,C1QA,C1QC,CD163
13	Pericytes [141]	COL1A1,COL3A1,LUM,COL6A3,DCN

Table 4.1: The potential cell type determination with their hub gene markers identified by applying maximal clique centrality.

types are oligodendrocytes. The cell-specific regulatory networks are shown in Figure 4.1. The blue and red nodes represent transcription factors and biomarkers, respectively. This is evidenced that TFs can be treated as a successful drug target. Moreover, due to the aggressive nature of GBM, very few clinical strategies are available as a therapeutic solution to this disease. Based on this, we intend to identify the important TFs that are showing an impact on the maximum number of cell types. From a comparison study, we identified TFs such as NFkB-family protein [142], STAT3 [143], RELA, FOS, JUN [144], SP1 etc. as therapeutic targets in GBM disease.

Furthermore, to validate the local networks biologically, we performed pathway analysis. The important pathways responsible for GBM disease are identified from the previous studies. We have also found the pair of TF-genes those are contributed to the pathway abnormalities and lead to the disease progression. The pathways and their corresponding cell types are shown through a network diagram in Figure 4.3. Interestingly, we noticed that the identified pathways indicate some functionality those are dedicated to the cell types. This reveals that local regulatory networks help to understand the biological complexity of a disease and also provides therapeutic solution from a single-cell perspective.

In summary, by integrating the MCC method in single-cell gene expression, our study identified the TFs from the local networks. Moreover, these TFs have the potential for prognosis prediction in GBM, validated by pathway analysis.

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

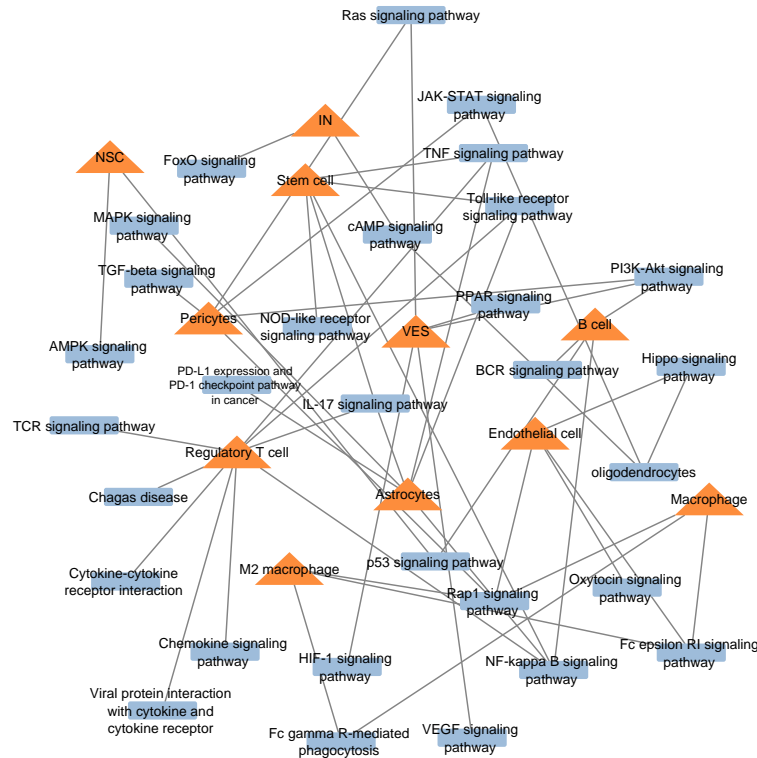


Figure 4.3: The violent color nodes in the graph represent the signaling pathways responsible for the GBM and their association with the cell types shown through an orange color in the graph.

### 4.2.2 Unveiling COVID-19 associated Organ Specific Cell Types and Cell-Specific Pathway Cascade

The current outbreak of the pandemic COVID-19 has been proclaimed as a public health emergency by the World Health Organization (WHO). Interestingly, these viruses have shown a strong binding with a cell receptor, namely Angiotensin-Converting Enzyme type II (ACE2) through the Spike S virulent protein. The study of Zhang et al. [145] has provided the molecular mechanism of the entry of the spike protein. As per the study, Transmembrane proteases play a vital role by creating cleavage in ACE2 and activating ACE2 cell receptors. The significant expression rate of ACE2 and TMPRSS2 has been observed in Pulmonary Alveolar type II (PAT2) cells in Lung [146]. Similarly, each of the vital organs viz., Kidney, Liver, Ileum, and Bladder should have certain cell types where significant RNAseq expression of ACE2 and TMPRSS2 is expected to observe. It is expected that these cell types are participating in nCoV infection. In a cellular condition, the significantly expressed ACE2 can regulate co-interacting functional hubs. Similarly,

## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

the transmission rate of COVID-19 must be engaged with many more molecular members of different cell types. So far, no studies reported the involvement of proteomic samples associated with ACE2. The study of the functional hubs with neighboring proteomic samples can be a key point in terms of therapeutic possibilities by suppressing functional dysregulation during the infection.

In this paper, we have studied the single-cell RNAseq data for Lung, Ileum, Kidney, Bladder, and Liver. The organ-specific cell types and respected markers are extracted dependent on both the ACE2 and TMPRSS2 expressions in PAT2 cells. For each of the defined cell types, significant proteomic markers are shortlisted based on the communities of Protein-protein interaction network (PPIN) [32] of cell-specific significant transcripts. The functional hubs are identified by applying K-Means network clustering [147]. However, members of the functional hubs should be associated with pathways. The pathway-driven systems are responsible for regulating cell functions. In this context, the interconnection of the pathways, associating with the members of the functional hubs, has been revealed through the pathway semantic network.

### 4.2.2.1 Data selection

Publicly available ScRNASeq datasets of different tissues and organs (Bladder [148], Ileum [149], Kidney [150], Liver [151] and Lung [152]) from diverse human bodies are curated from GEO (<https://www.ncbi.nlm.nih.gov/geo/>). The detailed information of the acquired datasets is as follows: Bladder, GEO Accession No. GSE129845 sample GSM3723358; Ileum, GEO Accession No. GSE134809 sample GSM3972018; Kidney, GEO Accession No. GSE131685, 3 healthy kidney tissues; Liver, GEO Accession No. GSE115469, 5 healthy human patients; Lung, GEO Accession No. GSE122960. The proposed framework of this study is described in Figure 4.4.

### 4.2.2.2 Dataset Preparation and Identification of Cell types

Seurat V3.0 [27] is used to process the raw count matrix. To remove the undesired cells from the dataset, unique feature counts over 2,500 or less than 200 are considered. Moreover, cells are filtered with  $> 5\%$  mitochondrial counts. The next step is to normalize the data after deleting undesirable cells from the dataset. The normalization is obtained by the "LogNormalize" function. The data has been scaled by natural log transformation after multiplying with 10,000. Principle Component are used to construct the K-Nearest Neighbors (KNN) graph. Depending on the graph, the cells are clustered using the "FindClusters" function, which implements the Louvain algorithm. These cell clusters are visualized by UMAP techniques. The top differentially expressed gene markers of each cluster are identified with

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

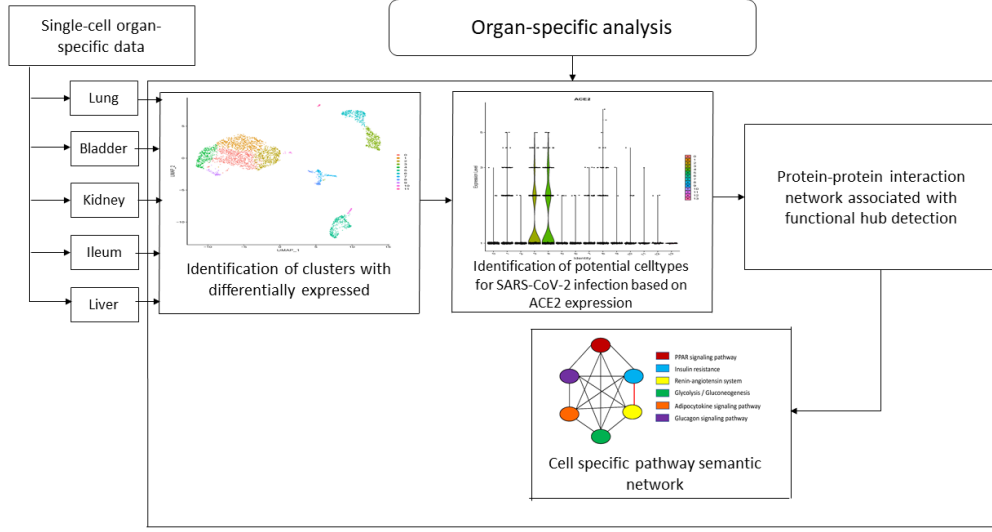


Figure 4.4: The flowchart of the proposed method.

“FindMarkers” function with a minimum percentage set to 0.25 using Wilcoxon Rank Sum statistical test. Finally, the expression level of ACE2 is evaluated from each cluster. It has already been reported in the literature that the SARS-COV-2 virus appears to target lung PAT2 cells via the ACE2 host receptor. Thus, the level of ACE2 expression in PAT2 cells has been used as a reference. Any cell type, consisting proportion of positive ACE2 (UMI count > 0) comparable to or greater than that of PAT2 cells is considered. Consequently, the corresponding organs are reported as high risk.

### 4.2.2.3 Protein-Protein Interaction Network and Functional Hubs

The organ-specific cell types with a significantly high rate of ACE2 expression are selected after initial identification. Subsequently, the interaction cores have been identified from Protein-protein interaction network (PPIN). In this regard, we have utilized STRING database [32]. The top biomarkers, fetched by “FindMarkers” under distinct cellular environments, may be associated with each other. Following that, we observed PPI networks of the selected biomarkers for each cell type (one dedicated PPIN for one specific cell type). Once we have established the connectivity among the biomarkers, we have aimed to detect the interactive communities, called functional hubs, applying K-Means clustering. The number of clusters for the K-Means clustering algorithm is set to 3. The edge weights are calculated using the cumulative scores of edge attributes from STRING database. As these weights help to determine the clusters, we are expecting that the func-



## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

tional hubs having ACE2 as a member show strong connectivity with ACE2. These functional hubs are considered for further study.

### 4.2.2.4 Pathway semantic

The pathways of the potential markers identified from functional hubs are utilized for the pathway semantic networks. For the selected pathways, a set of biological processes are considered to calculate the semantic similarity. Wang [153] defined the semantic value of term  $T$  as the aggregate contribution of all terms in  $DAG_T$  to the semantics of term  $T$ , terms closer to term  $T$  in  $DAG_T$  contribute more to its semantics. GO term  $T$  is defined as  $DAG_T=(T, X_T, E_T)$ ,  $X_T$  and  $E_T$  represent a set of GO terms and a set of GO terms connecting edges respectively. The  $X_T$  includes term  $T$  as well as all its ancestors. Thus, defined the contribution of a GO term  $p$  to the semantic of GO term  $T$  as the  $S$  - value of GO term  $p$  related to term  $T$ . For any of term  $p$  in  $DAG_T$ , its  $S$  - value related to term  $T$ ,  $S_T(p)$  is defined by equation 4.2:

$$\begin{cases} S_T(T) = 1 \\ S_T(p) = \max\{c_e * S_T(p') \mid p' \in \text{children of } (p)\} \text{ if } p \neq T \end{cases} \quad (4.2)$$

Here  $c_e$  is the semantic contribution factor for the edge  $e \in E_T$  linking GO term  $p$  with its child term  $p'$ . After calculating the  $S$  - value for the GO term in  $DAG_T$ , the semantic value of GO term  $T$ ,  $SV(T)$  is defined in equation 4.3:

$$SV(T) = \sum_{p \in X_T} S_T(p) \quad (4.3)$$

For two given GO term,  $T$  and  $Q$ , the semantic similarity between them is defined as:

$$SS_w(T, Q) = \frac{\sum_{p \in X_T \cap X_Q} S_T(p) + S_Q(p)}{SV(T) + SV(Q)} \quad (4.4)$$

In equation 4.4, the method proposed by Wang et al. [153] is used to compute the GO semantic similarity ( $SS_w$ ). Moreover,  $S_T(p)$  is the  $S$  - value of GO term  $p$  related to term  $T$  and  $S_Q(p)$  is the  $S$  - value of GO term  $p$  related to term  $Q$ .  $X_Q$  is the set of GO terms including term  $Q$  as well as all its ancestors.

Based on the semantic similarity of GO terms, the Best-Match Average (BMA) [154] strategy is performed to compute semantic similarity among sets of GO terms associated with the markers associated with a particular pathway, which is defined as equation 4.5:

$$S_{BMA}(G1, G2) = \frac{\sum_{i=1}^i \max_{1 \leq n \leq j} S(go1_m, go2_n) + \sum_{j=1}^j \max_{1 \leq m \leq i} S(go1_m, go2_n)}{i + j} \quad (4.5)$$

here, gene  $G1$  annotated by GO terms set  $GO1 = (go1_1, go1_2 \cdots go1_i)$  and  $G2$  anno-

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

---

tated by  $GO2 = (g_{021}, g_{022} \cdots g_{02j})$ .

### 4.2.2.5 Pathway Ranking Based on PageRank Algorithm

PageRank (PR) [155] Algorithm is introduced by Google to rank the searched pages in their search engine. It has been applied to calculate the rank of the nodes from the graph. The algorithm utilizes probability distribution depending on the occurrence of each node and measures the connection weight among different nodes. The node rank has been defined in equation 4.6:

$$PR(a) = \sum_{b \in B_a} \frac{PR(b)}{E(b)} \quad (4.6)$$

where the rank of node  $a$  relies on the  $PR$  values for each connected node  $b \in B_a$ , divided by  $E(b)$ , edges from node  $a$ . Here the pathways are represented by nodes and the links are weighted edges. The pathway semantic network is a weighted network where the weighted edges signify semantic strength between two pathways. Therefore, the  $PR$ -based ranks signify the dependency and influence of the pathway nodes in the network.

### 4.2.2.6 Experimental results

The single-cell RNA-Seq data of different human organs revealed information regarding the responsible cell types. These vital organs can be subdivided into three classes based on their functions. Interestingly, most of the cell types associated with a significant expression value of *ACE2* also possess a significant expression value of *AGT* and *PPAR* family transcripts e.g., *PPARA* and *PPARG*. However, only those cell types have these significantly expressed samples where *TMPRSS2* shows a higher expression. The highly significant samples are selected based on the functional hub which is strongly associated with *ACE2*. The Protein-protein interaction networks of each selected cell type.

#### 4.2.2.6.1 Cell Specific Functional Hubs from Lung

*PAT2* is detected as potential cells in Lung where *ACE2*, *AGT* and *PPARA* show significant expression. In Figure 4.5 the violin plot of the *ACE2* expression level implies that Lung is highly vulnerable towards viraemia. However, we have also identified two different cells, i.e., the plasma cell and the Mast cell. In the case of plasma cells, all three transcripts are significantly expressed whereas in Mast cells *AGT* and *ACE2* are significantly expressed.

## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

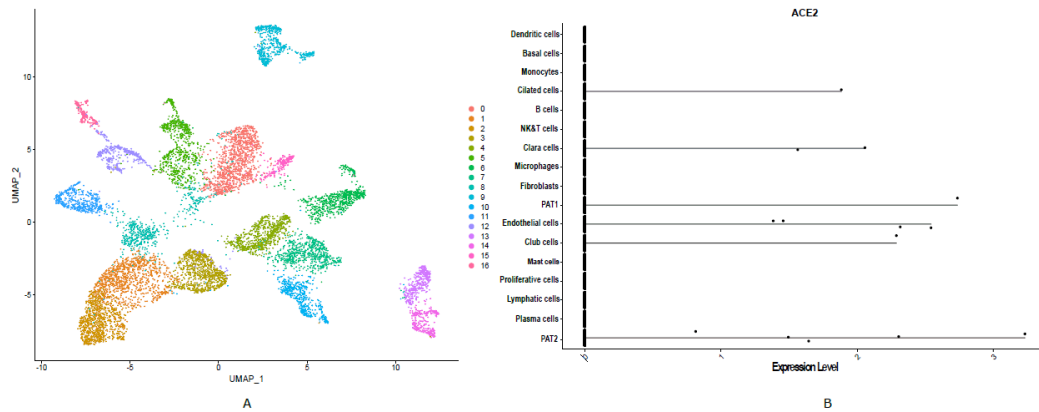


Figure 4.5: Lung single-cell RNAseq data analysis showed that ACE2 is highly expressed in pulmonary alveolar type II cells (PAT2). (A) The diverse cell types present in the lung are categorized into 17 clusters. (B) Violin plot is used to show the expression level distribution of ACE2 across the cell types. Including PAT2, the ACE2 is also expressed in PAT1, Clara, Club and Cilated cell types.

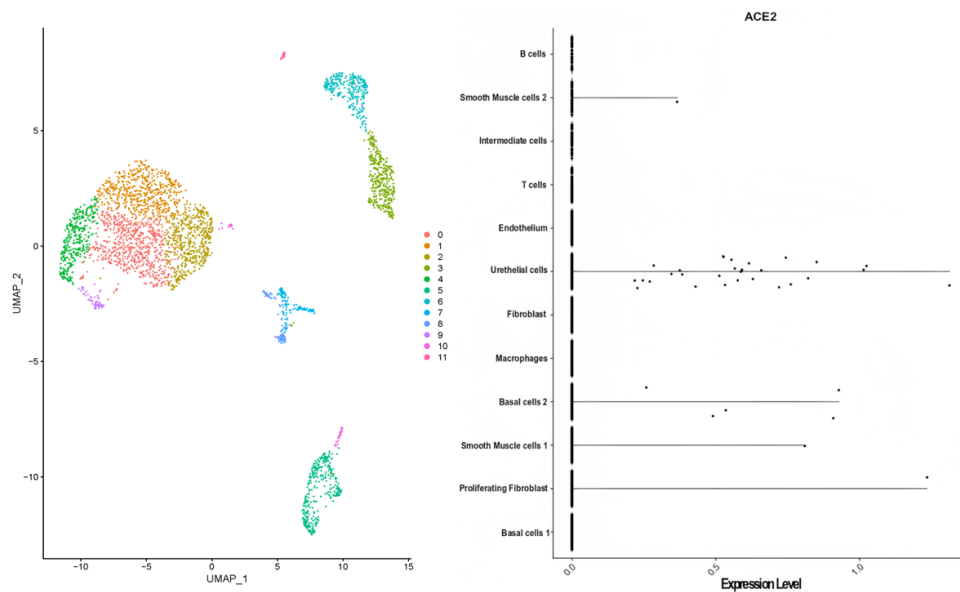


Figure 4.6: Bladder single-cell RNAseq data analysis showed that ACE2 is highly expressed in urothelial cells. (A) The diverse cell types present in bladder are categorized into 12 clusters. (B) A violin plot is used to show the expression level distribution of ACE2 across the clusters labelled with corresponding cell types. The plot shows smooth muscle cells 1 and 2 and basal cells 2, and proliferating fibroblast cells also possess significant expression levels of ACE2.

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

### 4.2.2.6.2 Cell Specific Functional Hubs from Bladder and Kidney

In Bladder, ACE2 shows higher affinity to urothelial cells reported in 4.6. Instead of PPARA, PPARG from PPAR family is showing a significant expression level. From Kidney, ACE2 is showing higher affinity in Proximal Tubule Cells and Smooth Muscle Cells 4.7.

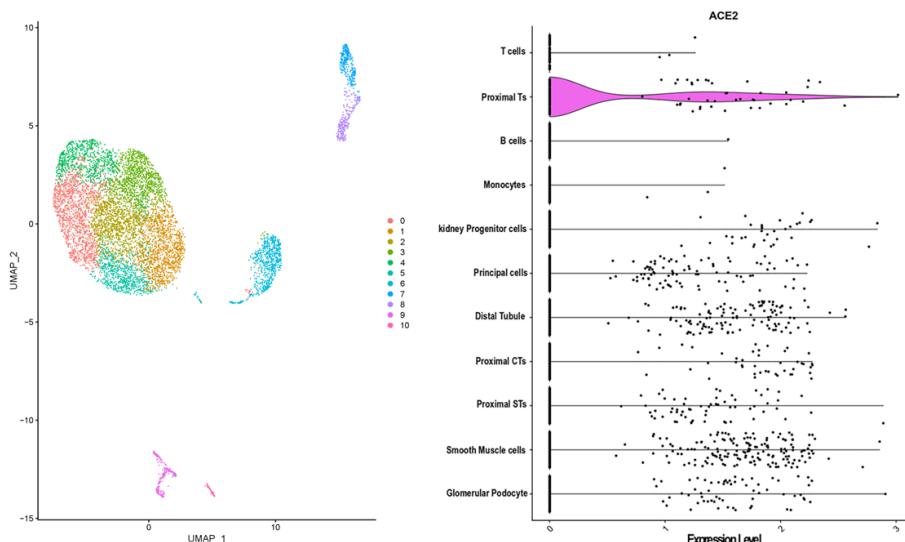


Figure 4.7: ScRNASeq data analysis of kidney uncovered that proximal tubule cells show the high expression level of ACE2. (A) The diverse cell types present in kidney are categorized into 11 clusters. (B) The violin plot is used to show the expression level distribution of ACE2 across the clusters. The clusters are labelled as corresponding cell types. ACE2 is expressed in multiple cell types such as in glomerular podocyte, smooth muscle cells, proximal straight tubules (STs), proximal convoluted tubules (CTs), distal tubule, principal cell, kidney progenitor cell, monocytes, B cell and T cell types.

### 4.2.2.6.3 Cell Specific Functional Hubs from Ileum and Liver

The ileum and Liver is a region of the metabolic system. Both of the organs have ACE2-positive epithelial-like cells. In the case of the Ileum, Enterocyte cells and Ciliated Epithelial cells have a significant expression of ACE2, AGT and PPARA Figure 4.8. However, the figure reveals that the expression level is low compared to other organs. Similarly, cholangiocytes are one of the epithelium cell classes essentially found in Liver tissue. The markers collectively ACE2, AGT and PPARA are detected with higher expression values in these cell types.

## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

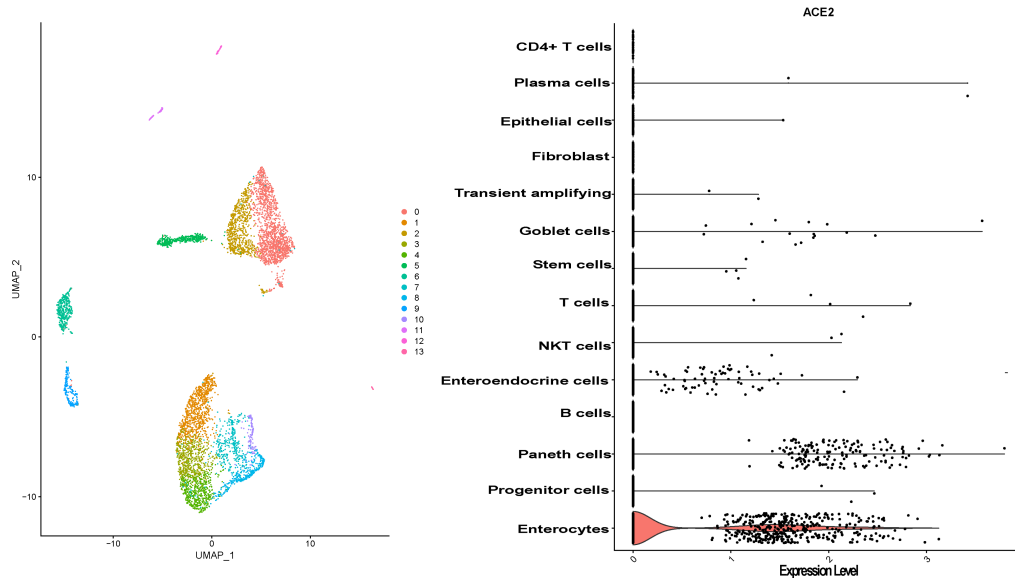


Figure 4.8: The ScRNASeq data analysis of ileum revealed that ACE2 is highly expressed in enterocyte progenitor cells. (A) The diverse cell types present in ileum are categorized into 14 clusters. (B) The clusters in the violin plot are labelled with cell types depending on the present biomarkers. The expression level of ACE2 across the cell types depicts that it is expressed in most of the cell types of the ileum.

### 4.2.2.6.4 Pathway Semantic Network

The influential markers which show an impact on the functional hub obtained from cell type-specific PPIN, are further studied. Pathways that possess the maximum number of potential markers and are also likely to be triggered during COVID-19 infection are curated from the Reactome database [156] and KEGG Pathway database [157]. We observed that the organ-specific cell types shared most of the common pathways but their association with markers differ from each other. Considering the biological process [158] associated with each pathway, the semantic similarity graphs have been constructed. ACE2 shows a high expression rate in three different cell types of lungs. We have provided pathway semantic networks for Mast, PAT2, and Plasma cells in Figure. 4.10(A), 4.10(B), and 4.10(C) respectively. In Bladder, ACE2 is expressed in one particular cell type i.e., Urothelial cells (shown in Figure. 4.11(A)). Whereas in Kidney Proximal tubule cell (shown in Figure. 4.11(B)) and Smooth muscle cell (shown in Figure. 4.11(C)) possess a high expression rate of ACE2. Similar to the Bladder organ, Liver ACE2 is expressed in one particular cell type i.e., Cholangiocytes cells (shown in Figure. 4.12(A)). In Ileum, cell types such as Ciliated Epithelial cells (shown in Figure. 4.12(B)) and

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

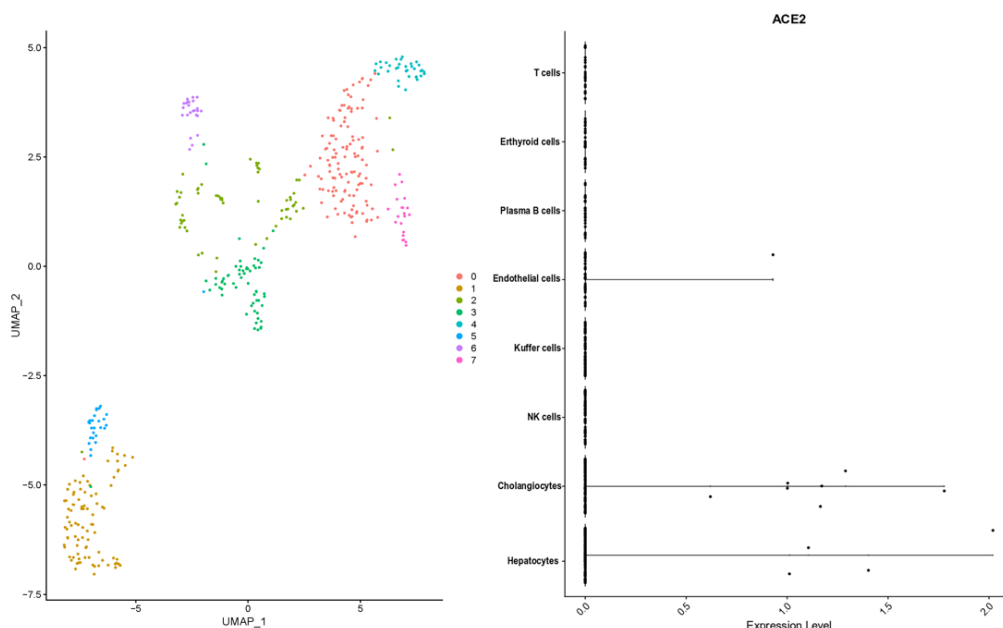


Figure 4.9: Liver single-cell RNAseq data analysis revealed that ACE2 is highly expressed in cholangiocytes. (A) The diverse cell types present in the liver are categorized into 8 clusters. The clusters are labelled with the name of cell types. (B) The Violin plot is used to show the expression level distribution of ACE2 across the cell types. Cell types hepatocytes and endothelial also show high expression level of ACE2

Enterocyte cells (shown in Figure. 4.12(C)), show significant expression levels of ACE2. In pathway semantic graphs, the nodes represent a particular pathway and the weighted edges define the similarity value between two pathways to depict the maximum number of sharing biological processes.

Moreover, during the pathway semantic calculation, a resultant matrix is obtained, this is utilized to perform the PR algorithm for each cell type and ranked them according to the score. We found that the Renin-angiotensin system (RAS) and PPAR signaling pathway secured a high rank in most of the prime cell types. Additionally, Insulin resistance also secures an important position in three organs. As discussed earlier, PR Values and corresponding ranks are calculated to show the influence of each pathway in the network. According to the overall ranking and scoring, the pathways are well segregated in the heatmap. The pathways having AGT and PPAR family proteins are ranked high in Ileum where the Insulin Resistance pathway is missing. Also, PPAR signaling pathway is absent in Liver cell types. Interestingly, Cortisol synthesis and secretion pathways play significant roles in some of the cell types.

#### 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

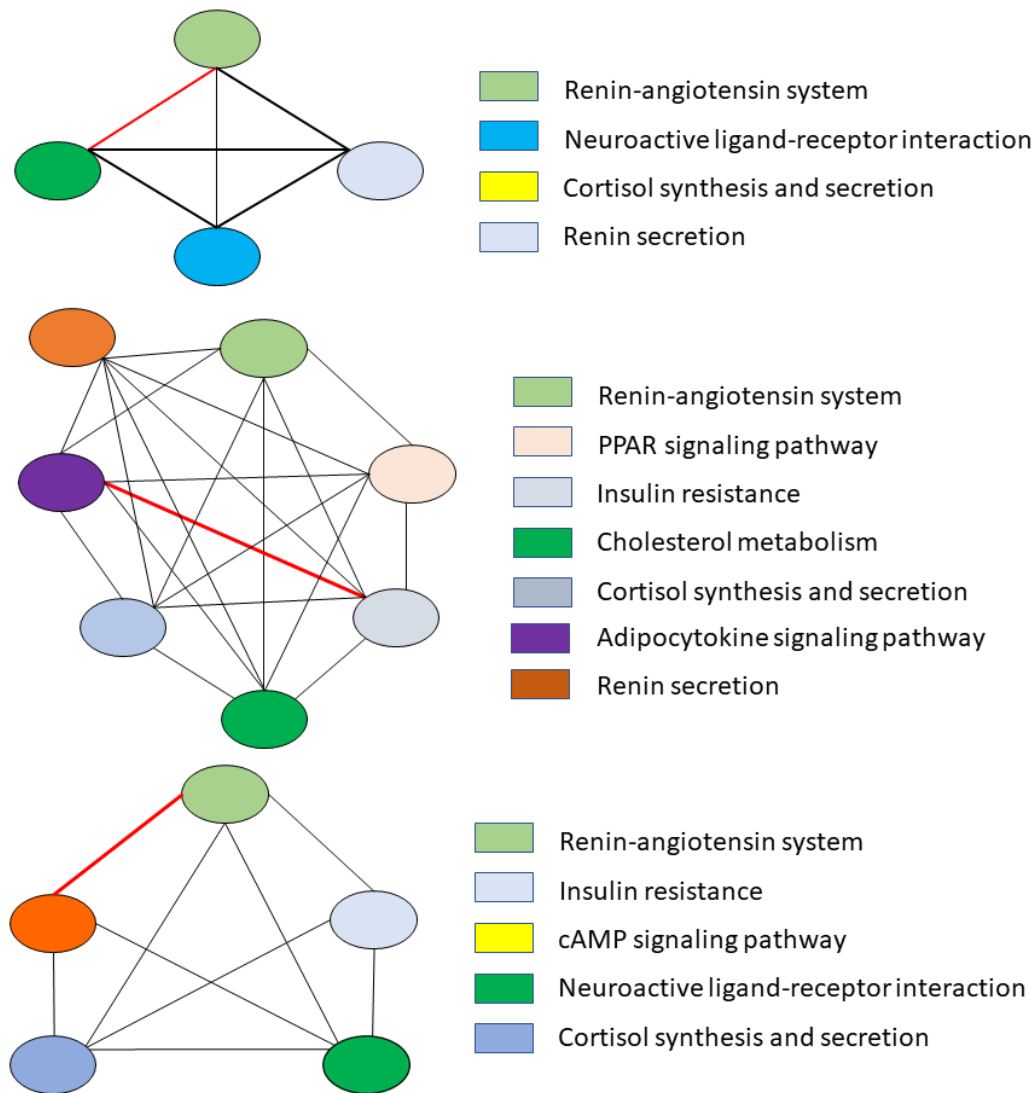


Figure 4.10: A cell-type-specific pathway semantic similarity graph: (A) mast cell type, (B) PAT2 and (C) plasma cell of the lung are established by considering the biological process associated with each pathway. Here, the nodes represent a particular pathway and their connecting edges define the weight between two pathways in order to apprehend the highest sharing biological process. The red-marked bold edge in the graphs indicates the higher association score between those pathways.

To map the relation between affected cells and pathway regulation inside an individual cell type, a cell-specific diagram is shown in Figure. 4.13. for all organs. Only the highly ACE2-expressed cell type of each organ has been shown. It is revealed from the figure that, PPAR signalling pathway and the Renin-angiotensin

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

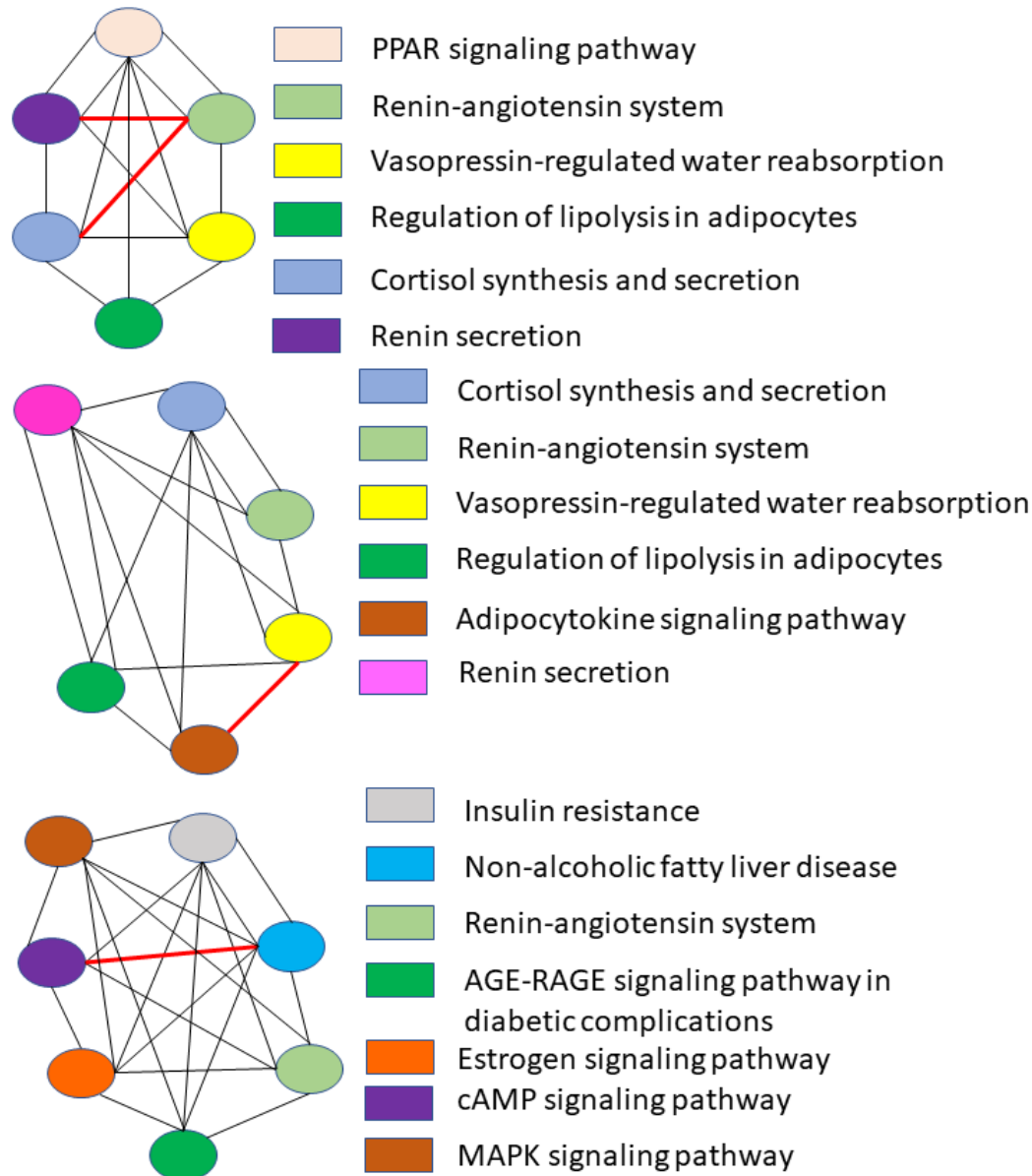


Figure 4.11: A) Urothelial cell type of bladder and two cell types. (B) Proximal tubule cells. (C) Smooth muscles of the kidney are considered for performing pathway semantic similarity graphs. Each color node of the graph represents a particular pathway associated with the cell type and the connection between the pathways is interpreted through the edges. The red-marked bold edge is used to exhibit the highest relationship between the pathways.

system (RAS) are the major pathways even after infection. Other pathways are also reported according to their rank. Similarly, for Bladder and Kidney, this cell



## 4.2. COMPUTATIONAL FRAMEWORK FOR UNDERSTANDING CELLULAR HETEROGENEITY IN DISEASE SEVERITY

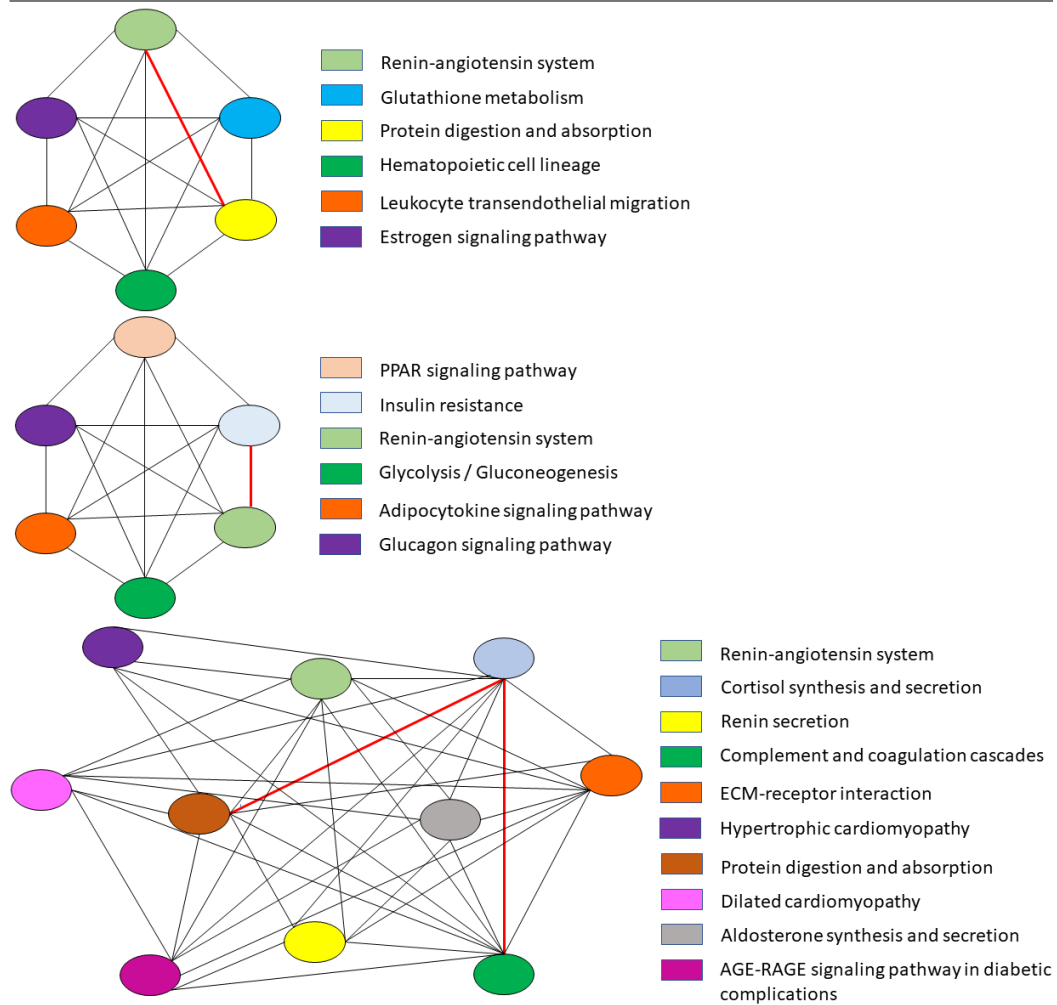


Figure 4.12: Considering the biological processes associated with each pathway, a semantic similarity graph is established for the following: (A) ciliated epithelial cells, (B) enterocyte progenitor cell type of ileum, (C) cholangiocytes cell type of liver. The nodes of the graphs represent a particular pathway and their connecting edges define the weight between two pathways in order to apprehend the highest sharing biological process. The red-marked bold edge in the graphs indicates the higher association score between those pathways.

type specific study is performed and some pathways are found to be common with Lung but the affecting rate is different. Finally, the cell-specific pathway study is shown for Liver and Ileum. Interestingly, the organs share common pathways among them. As mentioned earlier, the PPAR signalling pathway is missing in the cholangiocytes cell type from the liver.

## CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY

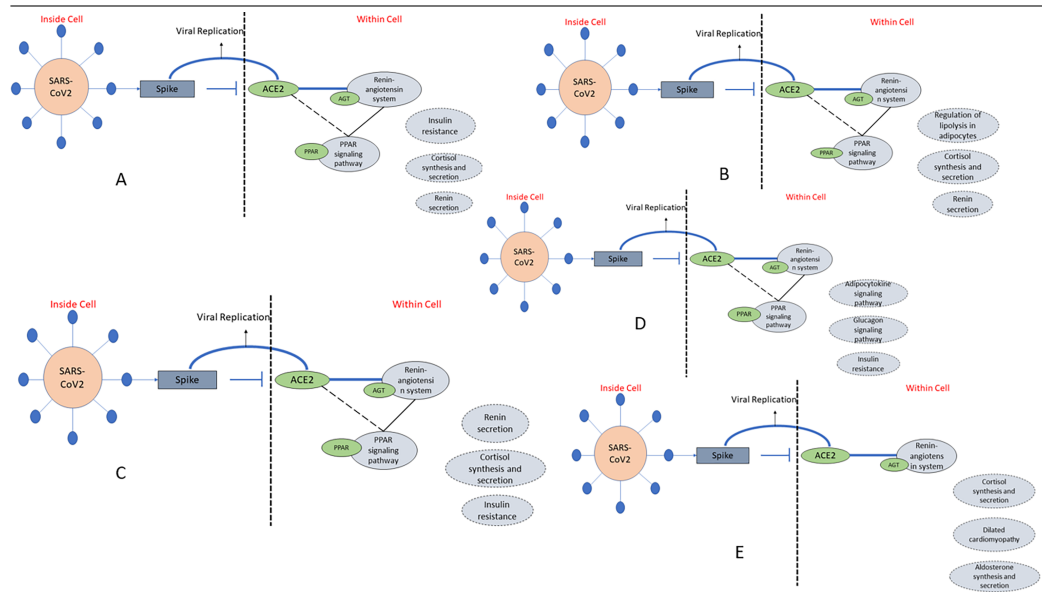


Figure 4.13: To map the results from pathway semantic and PageRank algorithm within the cell environment, in (A) for the PAT2 cell type of lung the top three ranked pathways excluding RAS and PPAR signaling pathway are shown, whereas the flowchart in (B) urothelial cells and (C) proximal tubule cells of organ bladder and kidney, respectively, show three ranked pathways including RAS and PPAR signaling pathway. For ileum (D) erythrocyte progenitor cells and (E) cholangiocytes of the liver, three pathways excluding RAS and PPAR signaling pathways are reported. As these two pathways are missing in the cholangiocytes cell type. These flowcharts of organ-specific cell types help to reveal the information regarding activating other pathways at the time of this epidemic disease.

### 4.2.2.7 Biological validation

The cell-specific findings have provided a trio of significantly expressed samples viz., ACE2, PPAR family samples (in Bladder PPARG) and AGT. As per the potential cell types, if ACE2 and TMPRSS2 are co-expressed [159], ACE2 has a higher propensity of associating with broader functional hubs. Co-expressed markers, AGT and PPAR/PPAG can act as a connection between the functional hub and ACE2. PPAR family proteins are involved in immune cells e.g., macrophages, dendritic cells. A. Erol had reported the importance of PPARG in a pioglitazone study on nCoV infection [160]. Singla et al. [161] have reported a therapeutic strategy, namely statin, for acute lung injury. The study has initiated the protocol of anti-inflammatory effects through transforming growth factor- $\beta$ , and peroxisome proliferator-activated receptor- $\gamma$ . Therefore, the significance of the PPAR family proteins can be observed and this strategy can be re-utilized as a therapeutic protocol for the disease. Similarly, many different articles have reported

the importance of the peroxisome-induced immune response against COVID-19 infection [162, 163]. On the other hand, AGT and ACE2 are directly involved with the renin-angiotensin system (RAS) [164]. Huang et al. have reported the RAS inhibition for H7N9 infection [165]. Also, Kuster et al. report the possibility of RAS inhibition as a therapeutic measure for COVID-19 [166]. In the study, AGT and PPAR family protein have secured vital positions in cellular functional hubs where ACE2 act as an important member. Therefore, there are therapeutic possibilities connected with these two categories of samples as well.

The PPI of the corresponding selected members can only describe the interaction perspective. However, the activities in the inner host cell system remain unveiled. Pathway semantics help to understand internal pathway connectivity. In the last part, we have already discussed the significance of the RAS pathway. In Zhang et al.[145], the functional perspective of the renin-angiotensin pathway has also been described. As per the study, the attenuation of RAS pathway can turn down the rate of viral replication. However, the prolonged effect of knocking down this pathway has not been shown. Similarly, the functionality of PPAR family proteins is previously discussed. These functionalities are completely interdependent. On the other hand, Mehta et al. has considered the infection as a cytokine storm syndrome [167]. In the significant cell-specific marker detection, each of the cell types can produce the TNF- $\alpha$  family proteins. Therefore, the peroxisome proliferator-activated receptor-based statin technique [161] might have worked due to its anti-inflammatory actions. However, we have identified a few potential pathways in most of the prime cell types from each of the organs viz., renin-angiotensin pathway, PPAR-signaling pathways, and adipocytokine signaling pathway. Also, a few basic metabolic pathways are identified. Interestingly, Insulin Resistance pathways are featured in Kidney, Ileum, and Lung specific cell types. These may explain the reason behind the vulnerability of diabetic patients [168]. As per the heatmap, RAS pathways and PPAR signalling (in most of the organ-specific prime cell types) may act as key regulators for most of the cellular systems (as they are securing decent positions in two separate modules). Hence, inhibition of the AGT and PPAR family protein is possibly a critical therapeutic target to attenuate the effect of infection by down-regulating the aforementioned pathways.

### 4.3 Conclusion

---

In the cell-specific study of glioblastoma, single-cell networks are established to understand the heterogeneity of the disease. Firstly, a clustering algorithm is applied to identify the clusters of cells depending on the gene expression. The clustered labelled with the appropriate cell types based on the top-ranked differentially ex-

## **CHAPTER 4. GRAPH-THEORETICAL MODELING TO UNVEIL THE CELL-TO-CELL HETEROGENEITY**

---

pressed gene markers. The top 50 biomarkers of each cell type are considered to establish the connection among the gene markers through a protein-protein interaction network. The functional modules are detected from the cell-specific PPI network in order to construct the local networks. This study reveals network communities that represent the functional modules of the disease. Though TFs are considered as an important druggable target, we identified the TFs that are present among the maximum cell types, to detect the suitable drug target for the disease. Finally, these identified modules can be considered to design a potential single-cell network for personalized medicine.

Subsequently, the single cell-based bioinformatic strategy, applied in COVID-19 data, has aimed to unveil the organ-specific probable infected cell types. The prime objective of the study is to determine the possible pathway connectivity associated with ACE2 dysregulation during COVID-19 infection. Initially, we started with the detection of the cell types from five organs. Also, each of the influential cell types has a list of biomarkers. Markers are sorted based on interacting functional hubs, connected with ACE2. Interestingly, AGT and PPAR family transcripts are common in each of the functional hubs. These two transcripts connect ACE2 and the rest of the samples from the functional communities. As per relevant literature, angiotensin and PPAR family proteins are previously observed to participate in different infections like COVID-19. In this experiment, the impact of PPAR signalling pathways, and RAS systems has been shown in the hub-specific pathway semantic networks. The network shows that the significant regulation of the mentioned pathways can affect the usual functions of the normal metabolic pathways as well as a few other pathways for instance Insulin Resistance. Hence, these samples are important therapeutic candidates. We can conclude that pathways play a key role to differentiate the cells and leading to heterogeneity. In this regard, pathways as well as the impact of genes on those pathways are crucial to identify cell types for a particular tissue. This information may help to understand cell-to-cell heterogeneity more clearly.

# 5

## Computational framework for pathway-based inference of single-cell RNA-seq data

### 5.1 Introduction

---

In the previous chapter, we introduced the pathway semantic similarity mechanism to understand cellular heterogeneity from a molecular level. As a result, it is evidenced from the published work, that pathways changes in each cell type of a particular disease inspite of having a similar set of genes. This is because single-cell clustering involves grouping cells based on similarities in their gene expression profiles. However, interpreting the vast amount of data generated from the single-cell analysis can be challenging, particularly when identifying the functional significance of gene expression changes within specific cell types. Pathway analysis involves identifying sets of genes involved in a specific biological process or pathway. By incorporating pathway information into single-cell analysis, researchers can gain insights into the functional significance of gene expression changes and identify pathways that are dysregulated in specific cell types or disease states. Therefore, integrating pathway data into single-cell clustering can help improve the accuracy and specificity of the clustering results and provide a more comprehensive understanding of the biological processes underlying cellular heterogeneity.

In single-cell studies, pathway activity score (PAS) analysis has been used to translate gene-level data into coherent gene sets representing biological processes or pathways to uncover the potential mechanisms underlying cell heterogeneity. However, no systematic benchmark studies have been conducted to evaluate the performance of unsupervised PAS transformation algorithms, which could enable researchers to analyze scRNA-seq data on PASs instead of individual gene expression. Depending on this fact, we performed two studies to understand the role of pathways in cell-to-cell heterogeneity. The first work is based on improving

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

the clustering by applying imputation. On the other hand, a new PAS calculation technique is developed in the second study. The proposed method clusters cells more accurately and cell-specific pathways are identified.

### 5.2 Unveiling cell-to-cell heterogeneity by incorporating pathway information

Unlike the previous works, this study aims to identify the cell types that are present in a functionally coordinated fashion of PBMC for healthy donors. Therefore, pathway information is considered to decipher the cell variability. The varying expression of a gene in a particular pathway helps to explore the biological causes. The pathway is a molecular biology term which represents an artificial simple framework of a biological process within a cell or tissue. From the recent study, cell types involved in peripheral blood mononuclear cells (PBMC) are identified. Furthermore, the related pathways identified the biological causes of the genes that are expressed differentially in particular cell types.

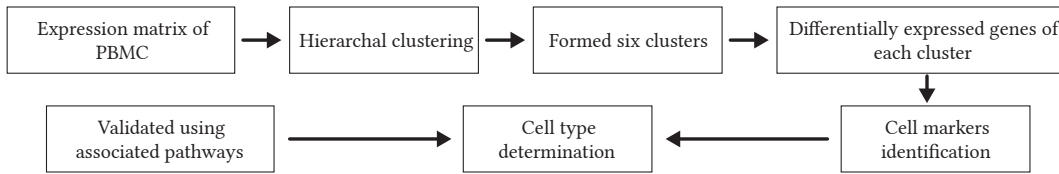


Figure 5.1: Representation of Cluster ID 1, 2, 3, 4, 5 and 6 with colors red, green, purple, yellow, brown and ash respectively.

#### 5.2.1 Method

This study aims to identify the cell types that are present in a functionally coordinated fashion of PBMC for healthy donors. The single-cell RNA-Seq data that has been considered consists of 1,222 cells accumulated from a healthy donor. The expression matrix of the selected dataset is downloaded from <http://www.10xgenomics.com>. The flow of the method is described in Figure.5.1. In this study, the determination of the cell types and also the justification of selecting the clusters as a particular cell type is divided into two phases. In the first phase, dropClust [169] method is applied to the gene expression data. The resultant clusters are further analysed by identifying their cell markers and also validated through their associated pathways. On the other hand, the top 30 ranked markers of each cluster are considered to perform the second phase of the study. Till now several cell types are identified by using a traditional pipeline of single-cell study. But the validation of each cluster with a particular cell type is not yet performed

## 5.2. UNVEILING CELL-TO-CELL HETEROGENEITY BY INCORPORATING PATHWAY INFORMATION

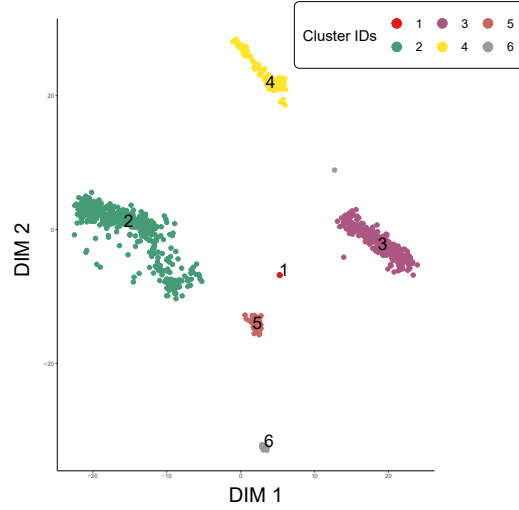


Figure 5.2: The flowchart of the proposed framework to identify the cell types based on the pathway of cell markers from a single-cell perspective.

as per our knowledge. To achieve this, the pathway of the cell marker of each cell type is considered. The cell-type-specific pathways are considered and their tissue-specific impact is curated from the literature. Finally, the results of the two phases are combined. The cluster generated from the first phase, the presences of cell markers in each cell and the pathway of the markers ultimately determine the cell types.

### 5.2.2 Validation

Hierarchical clustering is performed on PBMC gene expression data. The goal is to identify cell types present in the peripheral blood cell. Identification of cell types helps to recognize the group of cells functionally similar in a particular tissue or organ. Using common features cells are divided into groups. This grouping helps to analyse the complex tissues and the species. Finally, this analysis lead to finding the heterogeneity of the organisms in a specific taxonomy. To achieve this, single-cell sequencing contributes largely and rapidly to unzipping a new path. This path follows the unsolved biological questions by uncovering significant cell markers for determining discrete cell types in tissues. We have identified six clusters shown in Figure. 5.2. Multiple colours are used to distinguish the cluster types. Subsequently, genes are ranked according to their expression propensity in each.

Each cell type possesses a particular cell marker, which represents the unique characteristics of the cell type. Five cell markers from each cluster are reported in

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

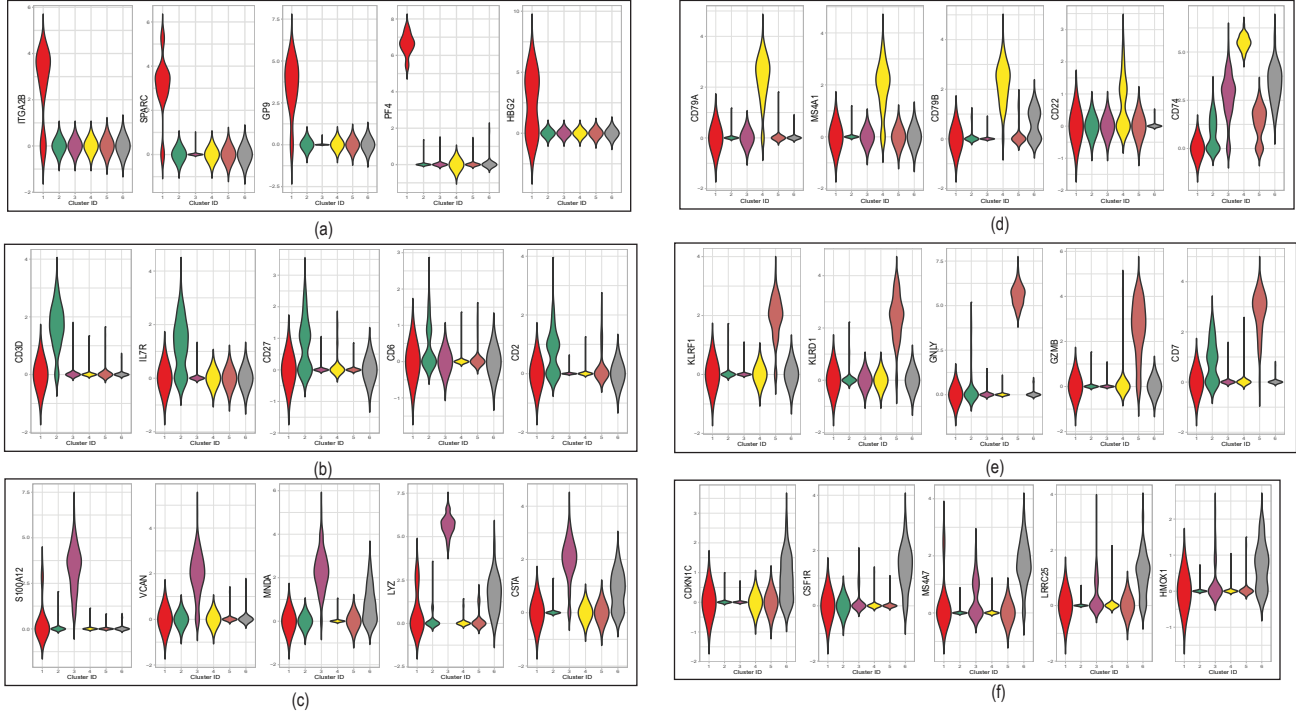


Figure 5.3: The violin plot of the cell markers in (a) Cluster 1, (b) Cluster 2, (c) Cluster 3, (d) Cluster 4, (e) Cluster 5 and (f) Cluster 6 respectively.

Table 5.1: The potential cell type determination with their cell markers.

Cluster ID	Cell type	Cell Markers
1	Megakaryocyte progenitor cell	ITGA2B [170], PPBP [169], GP9 [171], PF4 [169], PLA2G12A [169]
2	T cell	CD3D [172], IL7R [173], CD27 [169], CD6 [174], CD2 [175]
3	Dendritic cell	S100A12 [176], VCAN [176], MND4 [176], LYZ [176], CSTA [176]
4	B cell	CD79A [172], MS4A1 [172], CD79B [172], CD74 [177], CD37 [169]
5	NK cells	KLRF1 [172], KLRD1 [172], GNLY [169], GZMB [169], CD7 [178]
6	CD1C-CD141 dendritic cell	CDKN1C [176], CSF1R [176], MS4A7 [176], LRRC25 [176], HMOX1 [176]

Table 5.1 along with their respective cell types. It is found that, among six clusters, five clusters represent distinct cells whereas, cluster ID 6 represents a cell type CD1C-CD141 dendritic cells which is a subtype of Dendritic cells (cluster ID 3). Interestingly, all the identified cell types are commonly found in the peripheral blood cell. Depending on the p-values, common pathways of the cell markers are reported in Table 5.2. Surprisingly, it is observed that each set of pathways



## 5.2. UNVEILING CELL-TO-CELL HETEROGENEITY BY INCORPORATING PATHWAY INFORMATION

indicates dedicated biological functions to a cell type. Further analysis concludes that cell types determined through markers are similar to the pathways involved in peripheral blood cells. Moreover, a violin plot is shown in Figure. 5.3 which helps to understand the expression of the markers in each cluster. commonly found in the peripheral blood cell.

Table 5.2: The associated pathways of the cell markers of each cluster

Cluster ID	Associated Pathways	Adjusted P-values
1	ECM-receptor interaction	2.46E-04
	Platelet activation	5.63E-04
	Chemokine signaling pathway	5.5e-02
	Rap1 signaling pathway	6.0e-02
	PI3K-Akt signaling pathway	1.0e-01
2	Hematopoietic cell lineage	3.3E-04
	Primary immunodeficiency	5.1E-03
	Cell adhesion molecules (CAMs)	5.2E-02
	Th1 and Th2 cell differentiation	2.0E-01
	T cell receptor signaling pathway	2.0E-01
3	Salivary secretion	2.9e-04
	Cell adhesion molecules (CAMs)	4.2e-02
4	B cell receptor signaling pathway	1.2e-04
	Hematopoietic cell lineage	2.3e-04
	Primary immunodeficiency	9.2e-03
	Antigen processing and presentation	1.9e-02
5	Graft-versus-host disease	4.0e-05
	Natural killer cell-mediated cytotoxicity	4.2e-04
	Allograft rejection	9.3e-03
	Autoimmune thyroid disease	1.3e-02
	Transcriptional misregulation in cancer	4.5e-02
6	Ferroptosis pathway	9.9e-04
	HIF-1 signaling pathway	2.4e-03
	Ras signaling pathway	5.6e-03
	Cytokine-cytokine receptor interaction	7.1e-02
	MAPK signaling pathway	7.0e-01

### 5.3 Revealing Pathway Connectivity among Cell Types using Single-cell Data

---

In the previous research work, pathway information is incorporated to validate the results. Moreover, this process helps to identify rare cells missing during the gene expression-based clustering techniques. This is because genes often work collaboratively, rather than individually. This makes it challenging to identify informative pathways that can enhance our understanding of the functional diversity of cell populations. One approach to address this challenge is to incorporate pathway information into single-cell analysis. In this research, we present the effective utilization of scPCN (single-cell Pathway Correlation Network) in capturing cell-specific pathways within a specific tissue. Our method facilitates the grouping of cells based on their functionality and enables the discovery of novel relationships among enriched pathways. scPCN addresses the critical requirement of elucidating the interplay between pathways across diverse cell types within a specific tissue under various conditions. These relationships establish a comprehensive pathway interaction model, which contributes to a biologically driven understanding of phenotypes, reduces noise, and enhances overall performance. Therefore, scPCN represents a powerful advancement in interpreting results obtained through gene set methods. The python package of scPCN is available at <https://github.com/skshahnawaz/scPCN>

#### 5.3.1 Method

##### 5.3.1.1 Data Processing

The scRNA-Seq data is preprocessed using a standard Scanpy [179] pipeline. After deleting undesirable cells from the dataset, the next step is to normalize the data. This is an important step in scRNA-Seq data analysis, as it can help normalize the data and reduce the effects of technical noise.

##### 5.3.1.2 Pathway selection

One commonly used Equation 5.1 for identifying highly variable genes (HVGs) in scRNA-seq data is the "dispersion" statistic. The dispersion measures the degree of variability of a gene's expression across different cells or samples, taking into account the mean expression level as well.

$$Dispersion_i = \frac{Var_i - Mean_i}{Mean_i} \quad (5.1)$$

where  $Dispersion_i$  is the dispersion of a gene  $i$ ,  $Var_i$  is the variance of gene  $i$  across all cells, and  $Mean_i$  is the mean expression of gene  $i$  across all cells. The dispersion

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

is then scaled by the mean expression and log-transformed by Equation 5.2:

$$Dispersion_{scaled_i} = \log(Mean_i) + \log(Dispersion_i) \quad (5.2)$$

Genes with high values of  $Dispersion_{scaled_i}$  are considered HVGs and can be used for downstream analyses. After calculating and selecting the features, that show high cell-to-cell variation (high expression rate among some cells and low expression rate among other cells) are considered for the pathway selection process. The pathways are selected based on significant enrichment score or p-value ( $\leq 0.05$ ). Moreover, during our study, we only selected the biological pathways and excluded the disease-specific and metabolic pathways.

#### 5.3.1.3 Pathway Co-expression Network

Next, for each selected pathway, the expression values of the variable member genes are gathered. Inferring pathway activity from a pathway data matrix for pathway  $P_k$  is described. For example, a pathway  $p_k$  consists of a set of  $n$  genes  $G = g_1, g_2, g_3, \dots, g_n$  and each gene has  $m$  samples,  $S = s_1, s_2, s_3, \dots, s_m$ . Now, on this gene expression data matrix of pathway  $P_k$ , a gene co-expression network is applied.

To calculate the pathway co-expression network, the Pearson coefficient score is calculated for each gene present in each pathway. The equation for calculating the Pearson correlation coefficient between two genes,  $G_{(p-1)}$  and  $G_n$ , across  $n$  samples by Equation 5.3:

$$r = \frac{\sum (G_{(p-1)_i} - G_{(p-1)(avg)})(G_{p_i} - G_{p(avg)})}{(s-1)\sigma_{G_{(p-1)}}\sigma_{G_p}} \quad (5.3)$$

where,  $r$  is the Pearson correlation coefficient,  $G_{(p-1)_i}$  and  $G_{p_i}$  are the expression values of gene  $G_{(p-1)}$  and gene  $G_p$ , respectively, in the  $i^{th}$  sample  $G_{(p-1)(avg)}$  and  $G_{p(avg)}$  are the average expression values of gene  $G_{(p-1)}$  and gene  $G_{(p-1)}$ , respectively, across all  $p$  samples.  $\sigma_{G_{(p-1)}}$  and  $\sigma_{G_p}$  are the standard deviations of the expression values of gene  $G_{(p-1)}$  and gene  $G_p$ , respectively, across all  $s$  samples. The numerator of the equation calculates the covariance between the expression values, while the denominator normalizes the covariance by the standard deviations of the two genes. Once the coefficient matrix is calculated for all the  $p$  genes present in  $P_k$  pathways, the threshold is applied. The co-expression values  $G_{exp}$  satisfy the range of  $0.2 < G_{exp} < 0.4$  and are considered an active gene of the corresponding pathway.

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

---

### 5.3.1.4 Pathway activity score calculation

Once the co-expression matrix has been calculated, a threshold is employed to identify genes that meet the specified value, indicating their activity within a particular pathway. For example, let us assume,  $P_k$  pathway consists of  $AC_{g_1}$ ,  $AC_{g_2}$  and  $AC_{g_i}$  active genes. This implies that only these three genes will participate in pathway activation calculation. The pathway activity score (PAS) for the pathway  $P_k$  is computed with these three genes by dividing the sum of the sample by the square root of the number of selected genes, as described in Equation [180]. The square root of the number of member genes is used in the denominator to stabilize the variance of the mean. Moreover,  $AC_g$  genes would participate in the pathway  $P_k$  by Equation 5.4.

$$PAS(P_k) = \frac{\sum_{i=1}^q \sum_{j=1,2,p} g_{ij}}{\sqrt{AC_g}} \quad (5.4)$$

### 5.3.1.5 Clustering

The t-score is commonly utilized to assess the capacity to distinguish the accumulated expressions of a pathway's component genes. In previous literature, the t-score has been employed to conduct discriminative power tests. Essentially, it represents the disparity between the central tendency and the dispersion or variability present in the data points, relative to their mean value. Therefore, the t-score is calculated for the  $PAS(p_k)$  as follows by Equation 5.5:

$$t(PAS) = \frac{\mu_x - \mu_y}{\sqrt{(\delta_x/s_x) + (\delta_y/s_y)}} \quad (5.5)$$

In the context of a given pathway, PAS denotes the estimated levels of pathway activity.  $\mu_x$  and  $\delta_x$  indicate the mean and standard deviation of the pathway activity levels for x samples (likewise for group y). Low  $t(PAS)$  indicates the expression level does not exhibit significant differences in values between the two categories, while a high t-score suggests that the expression levels have a wide range of values, and it is anticipated that they differ between the two categories.

To identify informative features for the clustering, we reduced the dimensionality of the PAS matrix. We utilized the Phenograph clustering algorithm [181] to cluster the subpopulation of cells. This algorithm is designed explicitly for high-dimensional single-cell data. It creates a KNN graph based on the Euclidean distance in PCA space and then adjusts the edge weights between any two cells according to the shared overlap in their local neighborhoods. This approach yields a graph that reflects the phenotypic similarities between cells and their highly interconnected communities. Additionally, we used the Leiden method to cal-

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

culate communities within the graph. This algorithm ensures that all subsets of communities are optimally assigned at a local level. Finally, we applied non-linear dimension reduction techniques such as UMAP to visualize and explore the dataset in a low-dimensional space.

The second step of cell type identification is to annotate the clusters to their respective cell types based on the canonical markers. The study aims to reveal the subpopulation of the cell types. In this regard, the cluster-wise markers (pathways) are ranked based on the Wilcoxon rank-sum test. Usually, the top markers are relatively trustworthy, based on those markers the cell types are annotated respectively.

## Results

### Data description

We utilized six publicly available single-cell datasets to evaluate the proposed pathway activity calculation technique for cell type identification. Table 5.3 presents the primary information of these datasets. PBMC 68K consisted of approximately 68,000 PBMCs, collected from a healthy donor, with single-cell expression profiles of 11 purified subpopulations of PBMCs used as references for cell type annotation [182]. This dataset served as a gold standard for performance assessment of the clustering techniques. We also analyzed scRNA-seq data from two batches of peripheral blood mononuclear cells (PBMCs) obtained from a healthy donor (4K PBMCs and 8K PBMCs). Following filtering, we obtained 12,039 cells with 10,310 sampled genes. Litvinukova et al. [183] characterized the transcriptome of 486,134 cells and nuclei from six anatomical cardiac regions.

Dataset	#cells	# population	# genes
PBMC 68k	68,000	1	21932
PBMC	12,039	2	10,310
Heart	486,134	14	26662
OC	13369	8	6939
EC		11	
CC	33,694	2	14220

Table 5.3: The description of the datasets utilized in the study

### Performance of scPCN in single-cell dataset

For the evaluation of the proposed method, we use both intrinsic and extrinsic methods for the quantitative analysis of clustering performance. Three labelled

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

datasets are selected to measure the accuracy of the clustering through extrinsic metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information respectively.

The Adjusted Rand Index (ARI) is a statistical measure used to quantify the similarity between two clustering of data while accounting for chance agreement. The formula gives the ARI in Equation 5.6:

$$ARI(C^a, C^p) = \frac{\sum_{xy} \binom{n_{xy}}{2} - \frac{\sum_x \binom{p_x}{2} \sum_y \binom{q_y}{2}}{\binom{n}{2}}}{\frac{1}{2} \sum_x \binom{p_x}{2} + \sum_y \binom{q_y}{2} - \frac{\sum_x \binom{p_x}{2} \sum_y \binom{q_y}{2}}{\binom{n}{2}}} \quad (5.6)$$

where,  $C^a$  is the anticipated cluster and  $C^p$  is the principle cluster.  $n_{xy}$  is the number of nodes that are present in both cluster  $C_a^x$  and  $C_p^y$ ,  $p_x$  is the summation of all  $n_{xy}$  corresponding to any  $C_p^y$  of  $C_p$  and all  $C_a^x$  of  $C_a$ , and  $q_y$  is the summation of all  $n_{xy}$  corresponding to any  $C_a^x$  of  $C_a$  and all  $C_p^y$  of  $C_p$ . The ARI ranges from  $-1$  to  $1$ , where a value of  $1$  indicates perfect agreement between the two clusters,  $0$  indicates random agreement and negative values indicate disagreement [184]. Normalized Mutual Information (NMI) measures the similarity between two clusterings of a dataset [185]. It is defined as follows in Equation 5.7:

$$NMI(C, T) = \frac{2 * I(C, T)}{H(C) + H(T)} \quad (5.7)$$

where  $C$  and  $T$  are the two clusterings being compared,  $I(C, T)$  is the mutual information between them, and  $H(C)$  and  $H(T)$  are the entropies of the two clusterings. NMI ranges from  $0$  to  $1$ , with a higher value indicating greater similarity between the clusterings. It is a widely used metric in data clustering and information retrieval tasks and is particularly useful in evaluating unsupervised learning algorithms. Therefore, NMI should be carefully considered when assessing the performance of clustering algorithms.

The results demonstrate that scPCN exhibits exceptional performance on the benchmark scRNA-seq datasets, as shown in Table 5.4 and Table 5.5. Particularly, the ARI scores for Heart, PBMC 12k, and PBMC 68k are significantly higher compared to those achieved by state-of-the-art methods.

### Case Study

With the goal of determining the value of our approach in understanding pathway relationships in complex diseases, we consider gynaecological cancers. This group of cancers originate in the female reproductive system. Primarily they affected the organs in reproduction, including the ovaries, uterus, cervix, fallopian tubes, vulva, and vagina. These cancers are among the leading causes of cancer-related deaths in women. The most common types of gynaecological

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

Methods	Heart	PBMC 12k	PBMC 68k
SCANPY	0.4957	0.43721	0.4297
AUCell	0.5773	0.52307	0.4572
scPCN	0.6129	0.6484	0.5013

Table 5.4: Clustering performance using Adjusted Rand Index scores for different methods with respect to state-of-the-art method.

Methods	Heart	PBMC 12k	PBMC 68k
SCANPY	0.7298	0.6058	0.6045
AUcell	0.7523	0.6732	0.6273
scPCN	0.7875	0.6858	0.6785

Table 5.5: Comparison of clustering performance among different methods using Normalized Mutual Information scores

cancer include ovarian cancer, cervical cancer and uterine cancer (endometrial cancer). Early detection, accurate diagnosis, and appropriate treatment are crucial for improving outcomes in gynaecological cancer. With the advancement of technologies, single-cell studies have shown potential revolution of our understanding regarding gynaecological cancer by uncovering the intricate cellular heterogeneity, molecular dynamics, and interactions within the tumor microenvironment. Pathway analysis provides a framework for interpreting and integrating these complex datasets. It helps identify coordinated changes in gene expression, enabling a comprehensive understanding of cellular behaviour and intercellular dynamics. We performed the scPCN methodology to gain insights into the underlying molecular mechanisms of these gynaecological malignancies. Through comprehensive analysis of genomics and transcriptomics data, we sought to uncover critical pathways that are dysregulated in selected cancer and can be identified as a key therapeutic target. In this regard, scPCN is used to uncover the pathological association among these three diseases. Single-cell RNA-seq data of the three cancer types are

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

selected.

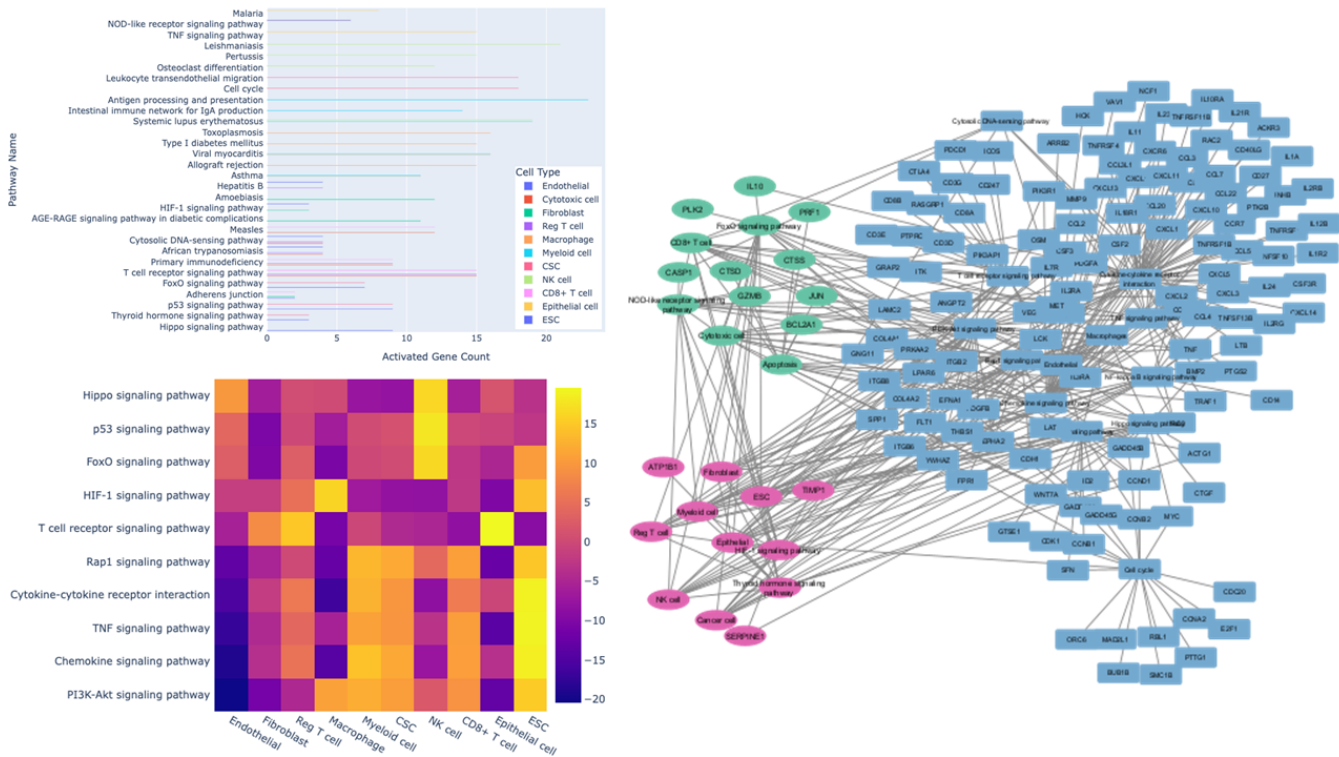


Figure 5.4: To facilitate cellular heterogeneity and underlying molecular mechanisms of ovarian cancer, scPCN is applied. A. bar graph shows the top-ranked pathways alongside their activated genes. The colours employed represent the presence of pathways within their respective cell types. B. the heatmap is established to represent the association between disease-specific signalling pathways with their respective cell types. G. functional hub of ovarian cancer is identified. Connecting cell types, pathways, and their associated activated genes. The network is constructed by using a maximal clique centrality algorithm among cell types, pathways, and their associated activated genes. The crucial hubs are denoted by green and pink colours.

### Biological significance of scPCN in Ovarian cancer study

We analyzed ovarian cancer (OV) single-cell RNA sequencing on 13,369 cells obtained from eight samples, consisting of four primary tumors, two peritoneal metastases, and two relapse tumors [186]. Pathways are selected based on the highly-expressed genes, and pathway co-expression networks are established. For each pathway, a Pathway Activity Score (PAS) is calculated, and clustering is performed using this score. The proposed methodology successfully identifies eleven



### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

distinct clusters of cells. Subsequently, efforts are made to identify key pathways associated with ovarian cancer. Through our proposed method, cell-specific pathways are ranked, with the ranking based on a predetermined threshold criterion. This criterion mandates that each pathway must include more than five activated genes for inclusion. In Figure 5.4A, a bar plot showcases the top-ranked pathways alongside their activated genes. The colours employed represent the presence of pathways within their respective cell types. Our methodology effectively identifies disease-specific key pathways and elucidates their activation patterns in distinct cell types. To gain insights into the temporal dynamics of gene expression changes within each cell type, a heatmap is constructed. The heatmap in Figure 5.4B shows the association between top-ranked disease-specific signalling pathways with their respective cell types. The change in pathway expressions provides valuable insights into the diverse molecular mechanisms underlying cell-to-cell heterogeneity in this disease. Finally, a network is established to depict the functional hub of ovary cancer, connecting cell types, pathways, and associated activated genes. To accomplish this, the maximal clique centrality algorithm is employed, and the resulting subgraphs are displayed in Figure 5.4G, denoted by green and pink colours. These identified subgraphs signify the associations between cell types and their respective pathways, offering potential therapeutic targets for ovary cancer. Furthermore, the associated activated genes are found to play pivotal roles in disease progression.

#### scPCN reveals the cell type-specific pathways in Endometrial cancer

Similar to ovarian cancer, endometrial cancer is a complex disease characterized by the abnormal growth of cells in the lining of the uterus [187]. It is associated with various molecular alterations and dysregulated signalling pathways that contribute to its development and progression. In this context, scPCN is applied to the single-cell RNA-sequencing data of endometrial cancer patients to identify the cell type-specific pathways. The cell-specific pathways of the endometrial cancer dataset are ranked based on the Wilcoxon test. Among the top-ranked pathways of the seven clusters, a boxplot is performed to understand the number of activated genes present in each pathway, shown in Figure 5.5A. However, their associated genes are significantly varied. To understand the diverse association of the pathways, disease-specific important pathways are selected and a heatmap is constructed. The heatmap in Figure 5.5B represents how the association of the same pathways varies in each cell type. Furthermore, the subgraph is established based on the association of cell types and their associated pathways, shown in Figure 5.5C. Moreover, this network shows the geneset of the pathways dysregulated during the disease condition. Interestingly, we found two genes such as ATM and CASP8 are present in both the subgraph of the network. The expression of

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

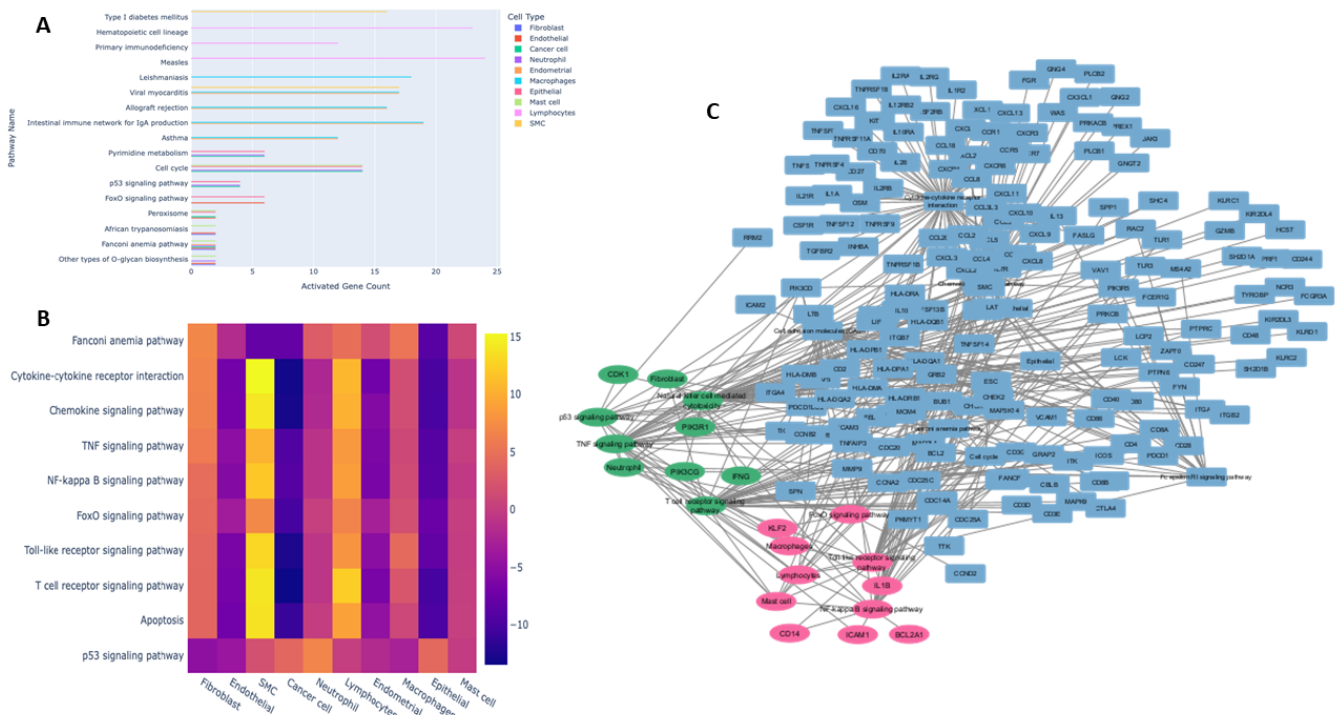


Figure 5.5: To explore the cellular heterogeneity and underlying molecular mechanisms of endometrial cancer, we employed scPCN analysis. The results are presented through various visualizations: (A). A bar graph displays the top-ranked pathways and their associated activated genes. The colors used indicate the presence of pathways within specific cell types. (B). A heatmap illustrates the association between the top-ranked disease-specific signalling pathways and their corresponding cell types. (C). By employing the maximal clique centrality algorithm, we identify the functional hub of endometrial cancer, which connects cell types, pathways, and their associated activated genes. Crucial hubs are highlighted in green and pink colors.

ATM is negatively correlated with the progression of endometrial cancer [188]. On the other hand, mutations in the CASP8 gene have been identified in a subset of endometrial cancers [189]. These mutations can lead to the inactivation or reduced expression of CASP8, impairing its apoptotic function and promoting tumor development. Therefore, we can conclude that the pathway associated with these two key genes can be considered a potential therapeutic target in endometrial cancer studies.

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

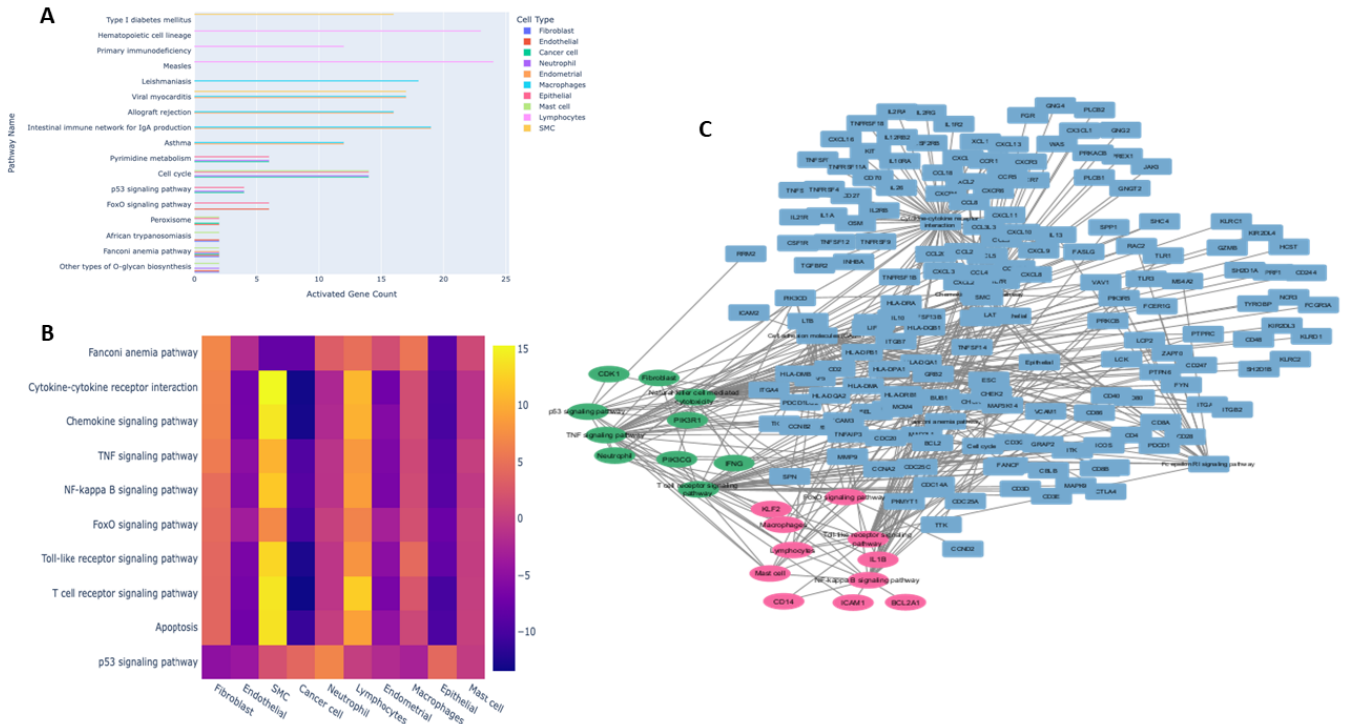


Figure 5.6: To investigate the cellular heterogeneity and underlying molecular mechanisms of cervical cancer, we have utilized the scPCN method. (A). A bar graph displays the top-ranked pathways alongside their activated genes. The colors used represent the presence of pathways within their respective cell types. (B). A heatmap is generated to represent the association between the top-ranked disease-specific signalling pathways and their respective cell types. (C). The functional hub of ovarian cancer is identified, connecting cell types, pathways, and associated activated genes. The network is constructed using the maximal clique centrality algorithm among cell types, pathways, and their associated activated genes. The crucial hubs are denoted by green and pink colors.

#### Application of scPCN on Cervical cancer

Subsequently, the cervical cancer dataset [190] is utilized to provide valuable insights into cellular heterogeneity and underlying molecular mechanisms. By utilizing scPCN the cell-specific pathways are ranked and a bar plot is established shown in Figure 5.6E. Moreover, to understand the rate of change of expression in each cell type, a heatmap (Figure 5.6F) is constructed. Finally, in figure 5.6G, a network is established among cell type, pathways and their associated activated genes to identify the functional hub.

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

### Pathway crosstalk analysis using scPCN

To investigate how the pathways interact with each other, a pathway crosstalk analysis is conducted. The approach is based on the assumption that two pathways can be considered if they are common among all three malignancies and possess a certain number of genes. A total of 24 pathways are found significantly common, of which six pathways met the criterion for crosstalk analysis.

The network of crosstalk, which includes these six significant common pathways of each disease, is presented in Figure 5.7. The bold red edge represents the strength of the association between them, which is measured by a semantic similarity network. From the three diseases, we have noticed that the NF- $\kappa$ B signalling pathway plays a crucial role in each cancer type. However, the gene set varies in each disease for a particular pathway. The common six pathways along with their gene set and cell types are reported in Table 5.6, respectively. The selected pathways are interconnected to form a complex disease network, which indicates the complexity of the pathogenesis of OC, EC and CC.

Disease	pathway	geneset	cell type
EC	T cell receptor signalling pathway	CTLA4, ICOS, HRAS, PTPN6, IFNG	T-cell
	p53 signalling pathway	GADD45G, IGF1, SESN3, SESN1, IGFBP3	smooth muscle cell
	NF- $\kappa$ B signalling pathway	CD40LG, CCL4, TNFSF13B, CCL19, MALT1	Macrophage
	Chemokine signalling pathway	CXCR2, GRK5, GNG2, CXCL16, HRAS	T cell
	TNF signalling pathway	SELE, CXCL1, CXCL3, SOCS3, LIF	Fibroblast
OC	T cell receptor signalling pathway	PIK3R1, CTLA4, CD8B, CD8A, PTPRC	cy-T cell, reg-T cell, CD8+ T cell
	p53 signalling pathway	CCND1, CCNB2, GADD45B, GTSE1, SFN	Endothelial
	NF- $\kappa$ B signalling pathway	TNF, LAT, BCL2A1, LTB, CD14	Epithelial cell
	Chemokine signalling pathway	CXCL3, NCF1, CCL3, CCL4L2, CCL4	Epithelial cell
	TNF signalling pathway	CXCL5, MMP9, IL6, TNFRSF1B, TNF	Epithelial cell
CC	T cell receptor signalling pathway	LCK, LAT, PIK3R5, CBLB, LCP2	Lymphocytes
	p53 signalling pathway	CDK1, CCNB2, RRM2, CHEK2	Cancer cell
	NF- $\kappa$ B signalling pathway	CHUK, CD14, TNFSF13B, BCL2, IL1B	Fibroblast
	Chemokine signalling pathway	CX3CL1, CCR1, XCL2, XCL1, PIK3R5	Cancer cell
	TNF signalling pathway	CREB5, CCL5, TNFAIP3, CSF1, LIF	Fibroblast

Table 5.6: The common cancer pathways in three malignant diseases with variable gene sets

### Discussion

In recent years, the emergence of scRNA-Seq has highlighted the need for corresponding computational analytic approaches and user-friendly packages. While numerous machine learning-based unsupervised methods have been proposed for scRNA-Seq data, assigning heterogeneous biological functions to determined cell populations remains a challenging issue. Additionally, scRNA-Seq data often suffer from technical noise and variability. Recognizing the importance of these challenges, our study introduces a pathway-based clustering approach to identify the heterogeneous functional pathways that differentiate cell populations using a co-expression method. By leveraging prior knowledge of pathways, we

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

address these issues and gain a better understanding of the underlying biology, enabling us to identify cell subpopulations from single-cell data. scPCN integrates

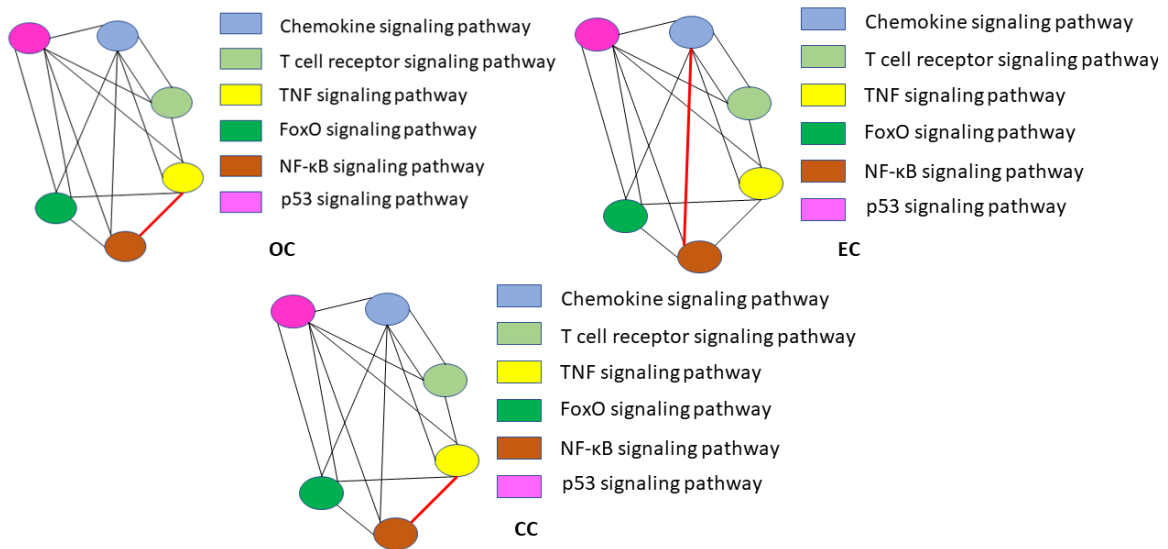


Figure 5.7: We generated a disease-specific pathway semantic similarity graph for three gynaecological cancers. This graph illustrates the interplay and communication between pathways that are ranked highest using scPCN

highly variable genes (HVGs) from the scRNA-Seq dataset. The significant genes are then used to identify pathways from three public databases. We specifically focus on biological pathways, eliminating disease-specific and metabolic pathways. An active gene of the corresponding pathway is defined to compute the co-expression network. Based on the pathway co-expression network, we perform a pathway activity score (PAS) calculation. Unlike other established PAS algorithms [191, 192, 193, 194, 195], scPCN incorporates both the pathway and the corresponding expression values of its sample-wise active genes. The results of our analysis indicate a significant correlation between two pathways with related gene sets. It is worth noting that redundant annotations across pathway databases are often overlooked. Pathway databases sometimes include pathways that share genes to varying degrees, either due to closely related functions or redundant annotations from different sources. Ignoring such redundancies during pathway analysis can lead to identifying pathway relationships based on content similarity rather than truly related biological mechanisms. ScPCN adjusts the correlation between pathways by considering only active genes to address this. Finally, we perform clustering to identify cell-cell heterogeneity using pathway information. The performance and computational time of our method are compared with Scanpy (gene expression-based clustering) and AUcell [196] (pathway-based

## CHAPTER 5. COMPUTATIONAL FRAMEWORK FOR PATHWAY-BASED INFERENCE OF SINGLE-CELL RNA-SEQ DATA

---

clustering). The method identifies crucial cell-type-specific pathways that have a significant impact on the respective tissue type.

To investigate how the identified pathways interact with each other, we conduct a pathway crosstalk analysis using pathway semantic similarity. Studying pathway crosstalk at the single-cell level has far-reaching implications. It provides insights into the heterogeneity of cellular responses within a population and emphasizes the importance of context-dependent interactions between pathways. Furthermore, it offers a deeper understanding of disease mechanisms, as alterations in pathway crosstalk can contribute to pathological conditions.

In our study, among the six common significant pathways, NF- $\kappa$ B pathway signalling exhibits a significant contribution to three malignancies. Dysregulation of this pathway has been implicated in tumor initiation, progression, and therapy resistance in these gynaecological cancers [197]. In ovarian cancer, aberrant NF- $\kappa$ B signalling is observed in both tumor cells and the tumor microenvironment [198]. Activation of the NF- $\kappa$ B pathway promotes cell survival, proliferation, and resistance to chemotherapy. NF- $\kappa$ B also induces the production of cytokines, chemokines, and growth factors that foster an inflammatory microenvironment, supporting tumor growth, angiogenesis, and metastasis [199]. Additionally, NF- $\kappa$ B signalling has been linked to the acquisition of cancer stem cell-like properties, which are associated with therapy resistance and disease recurrence in ovarian cancer. Similarly, in endometrial cancer, the NF- $\kappa$ B pathway has been implicated in promoting tumor growth, invasion, and metastasis [200]. Furthermore, NF- $\kappa$ B activation is linked to hormonal therapy resistance in endometrial cancer, highlighting its role in therapeutic resistance mechanisms. The NF- $\kappa$ B pathway is also activated by human papillomavirus (HPV) infection [201], a major risk factor for cervical cancer development [202]. Given the significance of the NF- $\kappa$ B signalling pathway in cervical cancer, ovarian cancer, and endometrial cancer, targeting this pathway has emerged as a potential therapeutic strategy. Several preclinical and clinical studies have investigated the efficacy of NF- $\kappa$ B inhibitors, both alone and in combination with standard therapies, to overcome treatment resistance and improve patient outcomes in these cancers. However, further research is needed to fully understand the complex interplay between NF- $\kappa$ B signalling and the tumor microenvironment, as well as its specific role in different subtypes and stages of these gynaecological cancers. The results indicate that gynaecological cancer types may exhibit homologous mechanisms with tumor types in other systems. Therefore, scPCN accurately quantifies the level of activity of each pathway in single-cell datasets and demonstrates good performance in diagnosis and prognosis, offering potential clinical value in the early detection of disease states. Our study provides novel insights into understanding the pathological mechanisms of gynaecological cancer, ultimately paving the way for personalized medicine in the treatment of

### 5.3. REVEALING PATHWAY CONNECTIVITY AMONG CELL TYPES USING SINGLE-CELL DATA

---

these diseases.

## Conclusion

---

In the previous chapter, we observed that single-cell clustering has revolutionized our understanding of cellular heterogeneity by grouping cells based on gene expression profiles. However, identifying rare cell types with unique gene expression patterns remains challenging using conventional clustering methods. In these two research articles, we propose the incorporation of pathway information to improve the identification of rare cell types. By leveraging the knowledge of pathway interactions and functions, we enhance the functional characterization of cells and their involvement in specific biological processes. We proposed a method for integrating pathway information into single-cell clustering by using a pathway co-expression network. This approach enables the identification of key regulators, markers, and pathway-level behaviours specific to cell types, offering deeper insights into their biological significance. Moreover, it provides a common pathway-level therapeutic target for gynaecological cancers. Overall, the integration of pathway information in single-cell clustering provides a powerful framework for unravelling the complex cellular composition and uncovering the underlying regulatory mechanisms of cell populations.

## Conclusions and Future Scope of Research

The fundamental objective of this thesis is to establish computational and statistical solutions for handling complex biological problems using multi-omics datasets. This thesis has demonstrated the power and significance of integrating diverse omics data, which comprises genomics, transcriptomics, proteomics, and epigenomics, to unravel the complexities of biological processes. Computational and statistical approaches, such as the Gaussian network model, modularity, clustering, and pathway semantic network, have been utilized to analyze the multiomics datasets. The studies discussed throughout the thesis have provided valuable insights into the understanding of regulatory networks, co-expression networks, metabolic interactions, cell-to-cell communication, disease interactions, and so on. Moreover, the multiomics data are studied from both bulk sequencing and single-cell sequencing perspectives.

This thesis introduces a discussion on the relationship among different omics layers. In systems biology, the connection among multiple omics layers is responsible for the progression of tumorigenesis or other diseases. In Chapter 2, we proposed a computational framework by establishing a regulatory network among a set of macromolecules (RNAs and proteins). Their interactions provide insights into the level of gene expression regulation in the genome of kidney and prostate cancer. In the kidney cancer study, we identify the regulatory molecules and their relationships responsible for regulating gene expression. Furthermore, we report transcription factors (TFs) that regulate both miRNAs and mRNAs in a regulatory network. Additionally, the analysis of TFs provides insight into disease prognosis and reveals the biological pathways responsible for the changes occurring during the disease state. On the other hand, the main goal of the prostate cancer study described in this chapter is to establish a bioinformatics pipeline to study the nature of transcription factors and propose suitable cellular conditions for drug targeting. Here, we find that transcription factors can act as effective druggable targets under a feed-forward loop. Despite the ability of TFs to control the pathogenic progression of malignancies, they are not considered potential drug targets due to the lack of druggable pockets at the folding stage. A detailed

---



---

description of the proposed frameworks is provided in the chapter.

The following contributory chapter of the thesis helps address the unsolved question of the previous chapter. This chapter focuses on understanding the dynamicity of the proteins in terms of networks. Chapter 3, reveals the possibility of the TFs as a potential drug target. However, a handful of research works are considering TFs as a druggable target. Hence, we tried to propose a framework to address this lacuna by providing a solution to this problem. Two studies are conducted, and a comprehensive framework is proposed to understand the structural conservation and co-variation of a protein, as well as its significance as a pan-cancer candidate and protein moonlighting candidate. The first study examined the protein's association with pathogenic annotations by analyzing its relationship to the gene regulatory network (GRN). The study focused on identifying the regions of transcription factors involved in controlling the protein's structural core during natural adaptation. By applying the Shannon Entropy method to the protein sequences, we could detect more stringent domains following the natural adaptation trait of respective families. Next, Direct Coupling Analysis (DCA) is used to observe the evolutionary highly co-varying patches of the proteins, and the protein's structural space is identified using structure network models. Finally, pathway analysis is done to summarize the process and gain insight into the diversity of the protein's association due to its structural malleability. The second study is conducted on SARS-CoV-2 proteins to identify potential drug targets, with Spike Protein and Replicases. The proposed framework for revealing the structural malleability of TFs is used to uncover the evolutionary sequence-structure analysis of viral proteins in the study.

In the previous two chapters, experiments are conducted using bulk RNA expression and protein abundance, which provided a limited understanding of the variability of cells under different conditions at any omics layer. However, technological advancements allow for a better understanding of cell-to-cell heterogeneity at any omics layer. Therefore, Chapter 4 focuses on designing a computational framework to uncover the cell-to-cell heterogeneity of multicellular organisms during disease initiation, development, and progression. Two cases are considered in this regard. The first case involves glioblastoma multiforme, which is one of the most aggressive malignancies in the brain and spinal cord area. Despite advancements in clinical strategies, this disease has a poor prognosis. Therefore, a bioinformatics pipeline is proposed in this study to understand cell-to-cell heterogeneity and its impact on tissue development, as well as the impact of diverse cell types on a particular tissue. In the second case, single-RNASeq data from COVID-19-affected organs is considered. An organ-specific framework is proposed to identify the cell types involved in disease initiation, and a pathway semantic model is applied to decipher the contribution of the internal pathway net-

## CHAPTER 6. CONCLUSIONS AND FUTURE SCOPE OF RESEARCH

---

work of each participating cell type towards the infection. The chapter extensively discusses the methods used in these cases.

In chapter 5, the main objective is to establish a bioinformatics pipeline to understand cellular heterogeneity by incorporating pathway information. In the aforementioned chapters, we proposed computational frameworks in order to unveil the biological complexities. The studies described in this chapter considered the previously mentioned frames to enrich the functional study. Two studies have been performed on single-cell data depending on the discussed frames. The main objective of this study is to incorporate pathway information to identify rare cell types. Here, we developed a novel method for calculating pathway activity scores. The key points of this method are described below:

- In this study a framework is proposed to cluster the cell-specific signaling pathways
- Our novel single-cell clustering framework clusters the cells and highlights the important cell-specific pathways.
- It helps to decipher the responsible genes and pathways that participated in cross-talk.
- We evaluate the performance of the proposed framework through a thorough biological analysis.
- A network is established to explain the path through which cells communicate with each other and how disruptions in those pathways lead to tissue-specific diseases.

In conclusion, this thesis has made notable contributions to the field of multi-omics data analysis by studying it from the perspectives of both bulk and single-cell RNA sequencing. By harnessing the strengths of each approach, this work has deepened our understanding of complex biological systems and paved the way for future fundamental research endeavors. However, both data generated from bulk and single-cell sequencing techniques are expected to undergo advancements in several areas. The future of research based on bulk sequencing and single-cell sequencing techniques is intertwined and complementary. Integrating spatial information with gene expression profiles, using techniques like spatial transcriptomics, allows researchers to map gene expression within tissues. Therefore, advanced research strategies will focus on refining and expanding spatial transcriptomics techniques to provide a detailed spatial context and explore cellular interactions, spatial heterogeneity, and tissue organization at a high resolution. Furthermore, with the increasing complexity and volume of sequencing data, advanced computational methods and data analysis tools are highly required. These tools can facilitate the integration, interpretation, and modeling

---

of bulk and single-cell sequencing data, enabling the extraction of meaningful insights from large-scale genomic datasets. The continued exploration of multi-omics data from both bulk and single-cell perspectives holds immense promise for uncovering novel biomarkers and developing personalized medicine approaches based on individual cellular characteristics and gene expression profiles. This can also decipher the intricate gene regulatory networks, cellular interactions, and dynamic cellular transitions during development and tissue repair in the future.



# Bibliography

- [1] F. R. Pinu and et.al, "Systems Biology and Multi-Omics Integration: View-points from the Metabolomics Research Community," *Metabolites*, vol. 9, no. 4, pp. 1662–1664, 2019.
- [2] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.
- [3] I. Thiele and et.al, "A community-driven global reconstruction of human metabolism," *Nature Biotechnology*, vol. 31, no. 5, pp. 419–425, 2013.
- [4] R. Nussinov, "Advancements and Challenges in Computational Biology," *PLOS Computational Biology*, vol. 11, no. 1, pp. 1–2, 2015.
- [5] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin, "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1370–1381, 2019.
- [6] M. Krassowski, V. Das, S. K. Sahu, and B. B. Misr, "State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing," *Briefings in Bioinformatics*, vol. 11, no. 610798, pp. 1–17, 2020.
- [7] T. Pang, "The Impact of Genomics on Global Health," *American Journal of Public Health*, vol. 92, no. 7, pp. 1077–1079, 2002.
- [8] E. Uffelmann, Q. Q. Huang, N. S. Munung, and et al., "Genome-wide association studies," *Nat Rev Methods*, vol. 1, no. 59, pp. 1–21, 2021.
- [9] L. Atta and J. Fan, "Computational challenges and opportunities in spatially resolved transcriptomic data analysis," *Nature Communications*, vol. 12, no. 5283, pp. 1–5, 2021.
- [10] S. Djebali, V. Wucher, S. Foissac, C. Hitte, E. Corre, and T. Derrien, "Bioinformatics Pipeline for Transcriptome Sequencing Analysis," *Methods in Molecular Biology*, vol. 1468, pp. 201–219, 2017.
- [11] M. Cannataro, "Computational proteomics: management and analysis of proteomics data," *Briefings in Bioinformatics*, vol. 9, no. 2, p. 97–101, 2009.
- [12] A. Schmidt, I. Forne, and A. Imhof, "Bioinformatic analysis of proteomics data," *BMC Systems Biology*, vol. 8, no. 2, pp. 1–17, 2014.
- [13] S. Qiu, Y. Cai, H. Yao, and et al., "Small molecule metabolites: discovery of biomarkers and therapeutic targets," *Signal Transduction and Targeted Therapy*, vol. 8, no. 132, pp. 1–37, 2023.

## BIBLIOGRAPHY

---

- [14] X. Dai and L. Shen, "Advances and Trends in Omics Technology Development," *Frontiers in Medicine*, vol. 9, pp. 1–25, 2022.
- [15] D. Lahnemann, J. Koster, E. Szczurek, and et al, "Eleven grand challenges in single-cell data science," *Genome Biology*, vol. 21, no. 31, pp. 1–35, 2020.
- [16] H. A. AL-kuhali, M. Shan, M. A. Haeland, and et al, "Multiview clustering of multi-omics data integration by using a penalty model," *BMC Bioinformatics*, vol. 23, no. 288, pp. 1–19, 2022.
- [17] S. Canzler and J. Hackermuller, "multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data," *BMC Bioinformatics*, vol. 21, no. 561, pp. 1–13, 2020.
- [18] C. Meng, O. A. Zeleznik, G. G. Thallinger, and et al, "Dimension reduction techniques for the integrative analysis of multi-omics data," *Briefings in Bioinformatics*, vol. 17, no. 4, pp. 628–641, 2016.
- [19] G. Zhou, S. Li, and J. Xia, "Network-Based Approaches for Multi-omics Integration," *Methods in Molecular Biology*, vol. 2104, pp. 469–487, 2016.
- [20] S. Sass, F. Buettner, N. S. Mueller, and F. J. Theis, "A modular framework for gene set analysis integrating multilevel omics data," *Nucleic Acids Research*, vol. 41, no. 21, p. 9622–9633, 2013.
- [21] S. Imoto, T. Higuchi, T. Goto, and et al., "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," *Journal of Bioinformatics and Computational Biology*, vol. 2, p. 77–98, 2004.
- [22] S. Zhang, Q. Li, J. Liu, and X. J. Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules," *Bioinformatics*, vol. 27, pp. 1–9, 2011.
- [23] S. Zhang, C. C. Liu, W. Li, and et al, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, p. 9379–9391, 2012.
- [24] J. A. Seoane, I. N. M. Day, T. R. Gaunt, and C. Campbell, "A pathway-based data integration framework for prediction of disease progression," *Bioinformatics*, vol. 30, p. 838–845, 2014.
- [25] F. W. Townes, S. C. Hicks, and M. J. A. and. et al., "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model," *Genome Biology*, vol. 20, no. 295, pp. 1–19, 2019.

- [26] D. Kobak and P. Berens, "The art of using t-SNE for single-cell transcriptomics," *Nature Communications*, vol. 10, no. 5416, pp. 1–14, 2019.
- [27] T. Stuart and et.al, "Comprehensive Integration of Single-Cell Data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.
- [28] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature Biotechnology*, vol. 37, p. 547–554, 2019.
- [29] T. Mou, W. Deng, F. Gu, Y. Pawitan, and T. N. Vu, "Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing," *Frontiers in Genetics*, vol. 10, no. 1331, pp. 1–12, 2020.
- [30] A. Dey, S. Sen, and U. Maulik, "Study of transcription factor druggability for prostate cancer using structure information, gene regulatory networks and protein moonlighting," *Briefings in Bioinformatics*, vol. 23, no. 1, 2022.
- [31] S. V. Dam, U. Vosa, A. V. Graaf, L. Franke, and J. P. Magalhaes, "Gene co-expression analysis for functional classification and gene-disease predictions," *Briefings in Bioinformatics*, vol. 19, no. 4, p. 575–592, 2018.
- [32] D. Szklarczyk and et.al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D561–D568, 2011.
- [33] R. Levy and E. Borenstein, "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules," *Briefings in Bioinformatics*, vol. 110, no. 31, pp. 12 804–12 809, 2022.
- [34] R. E. Chen and J. Thorner, "Systems biology approaches in cell signaling research," *Genome Biology volume*, vol. 235, no. 6, pp. 1–5, 2005.
- [35] J. Pinero, A. Bravo, N. Q. Rosinach, A. G. Sacristan, J. D. Pons, E. Centeno, J. G. Garcia, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, no. D1, pp. D833–D839, 2017.
- [36] G. Brysbaert, K. Lorgouilloux, W. F. Vranken, and M. F. Lensink, "RINSpector: a Cytoscape app for centrality analyses and DynaMine flexibility prediction," *Bioinformatics*, vol. 34, no. 2, p. 294–296, 2018.
- [37] C. H. Chou, S. Shrestha, and C. D. Yang, "miRTarBase update 2018: a resource for experimentally validated microRNA- target interactions," *Nucleic Acids Research*, vol. 46, no. D1, pp. D296–D302, 2018.

## BIBLIOGRAPHY

---

- [38] J. Wang, M. Lu, C. Qiu, and Q. Cui, "Transmir: a transcription factor-microRNA regulation database," *Nucleic Acids Research*, vol. 38, no. D1, pp. 119–122, 2010.
- [39] S. E. Gebali and et. al., "The pfam protein families database in 2019," *Nucleic Acids Research*, vol. 47, no. D1, p. D427–D432, 2019.
- [40] M. Chatzou, C. Magis, J. M. Chang, C. Kemena, G. Bussotti, I. Erb, and C. Notredame, "Multiple sequence alignment modeling: methods and applications," *Briefings in Bioinformatics*, vol. 17, pp. 1009–1023, 2016.
- [41] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: A novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, pp. 205–217, 2000.
- [42] T. D. Schneider, "Consensus sequence zen," *Applied Bioinformatics*, vol. 1, no. 3, pp. 111–119, 2002.
- [43] T. Dogan and B. Karacali, "Automatic identification of highly conserved family regions and relationships in genome wide datasets including remote protein sequences," *Proteins*, vol. 42, no. 1, pp. 38–48, 2000.
- [44] S. Sen, A. Dey, S. Chowdhury, U. Maulik, and K. Chattopadhyay, "Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis," *BMC Bioinformatics*, vol. 19, no. 13, pp. 231–242, 2019.
- [45] P. Romero, Z. Obradovic, X. Li, E. Garner, C. Brown, and A. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, no. 1, pp. 38–48, 2000.
- [46] H. Maity and et. al., "Protein folding: The stepwise assembly of foldon units," *Proceedings of the National Academy of Sciences of the United States of America*, no. 13, p. 4741–4746, 2005.
- [47] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, pp. 1293–1301, 2011.
- [48] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky, "Pondr-fit: A meta-predictor of intrinsically disordered amino acids," *Biochim Biophys Acta*, vol. 1804, no. 4, p. 996–1010, 2010.



- 
- [49] Z. Dosztanyi, "Prediction of protein disorder based on iupred," *Protein Science*, vol. 27, no. 1, p. 331–340, 2018.
- [50] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.
- [51] K. V. Brinda and S. Vishveshwara, "A Network Representation of Protein Structures: Implications for Protein Stability," *Biophysical Journal*, vol. 89, no. 6, pp. 4159–4170, 2005.
- [52] N. Kannan and S. Vishveshwara, "Identification of side-chain clusters in protein structures by a graph spectral method," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 441–464, 1999.
- [53] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010.
- [54] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.
- [55] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The foldx web server: an online force field," *Nucleic Acids Research*, vol. 33, p. W382–W388, 2005.
- [56] Q. Li, X. Peng, Y. Li, W. Tang, J. Zhu, J. Huang, Y. Qi, and Z. Zhang, "Llpsdb: a database of proteins undergoing liquid–liquid phase separation in vitro," *Nucleic Acids Research*, vol. 48, no. D1, p. D320–D327, 2020.
- [57] K. You, Q. Huang, C. Yu, and et al, "Phasepdb: a database of liquid–liquid phase separation related proteins," *Nucleic Acids Research*, vol. 48, no. D1, p. D354–D359, 2020.
- [58] M. Fu, M. Rao, and C. Wang, "Acetylation of androgen receptor enhances coactivator binding and promotes prostate cancer cell growth," *Molecular and Cellular Biology*, vol. 23, no. 23, p. 8563–8575, 2003.
- [59] H. Kinoshita, Y. Shi, C. Sandefur, L. F. Meisner, C. Chang, A. Choon, C. R. Reznikoff, G. S. Bova, A. Friedl, and D. F. Jarrard, "Methylation of the androgen receptor minimal promoter silences transcription in human prostate cancer," *Cancer Research*, vol. 160, no. 13, pp. 3623–3630, 2000.
-

## BIBLIOGRAPHY

---

- [60] F. Morcos, A. Pagnani, and B. Lunt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 49, pp. E1293–E1301, 2011.
- [61] J. Pang, Y. W. Yang, and Y. Huang, "P110 $\beta$  inhibition reduces histone h3k4 di-methylation in prostate cancer," *Prostate*, vol. 77, no. 3, pp. 299–308, 2017.
- [62] H. M. Itkonen, S. S. Gorad, and D. Y. Duveau, "Inhibition of o-glcnac transferase activity reprograms prostate cancer cell metabolism," *Oncotarget*, vol. 7, no. 11, pp. 12 464–12 476, 2016.
- [63] L. Zhang, S. Altuwaijri, F. Deng, and et al, "Nf-kappab regulates androgen receptor expression and prostate cancer growth," *Am J Pathol*, vol. 175, no. 2, pp. 489–499, 2009.
- [64] L. Zhang and et. al, "Translationally regulated *c/ebp $\beta$*  isoform expression up-regulates metastatic genes in hormone-independent prostate cancer cells," *Prostate*, vol. 68, no. 12, pp. 1362–1371, 2008.
- [65] D. S. Wishart, Y. D. Feunang, A. C. Guo, and et al, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, p. D1074–D1082, 2018.
- [66] D. J. Sharpe, K. S. Orr, M. Moran, S. J. White, S. McQuaid, T. R. Lappin, A. Thompson, and J. A. James, "POU2F1 activity regulates HOXD10 and HOXD11 promoting a proliferative and invasive phenotype in Head and Neck cancer," *Oncotarget*, vol. 5, no. 18, pp. 8803–8815, 2014.
- [67] R. J. Holt, Y. Zhang, A. Bini, A. L. Dixon, C. Vandiedonck, W. O. Cookson, J. C. Knight, and M. F. Moffatt, "Allele-specific transcription of the asthma-associated phd finger protein 11 gene (*phf11*) modulated by octamer-binding transcription factor 1 (*oct-1*)," *Journal of Allergy Clinical Immunology*, vol. 127, no. 4, p. 1054–1062, 2011.
- [68] B. Andersen and M. G. Rosenfeld, "POU Domain Factors in the Neuroendocrine System: Lessons from Developmental Biology Provide Insights into Human Disease," *Endocrine Reviews*, vol. 22, no. 1, pp. 2–35, 2001.
- [69] H. Han, J. W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H. N. Jeon, H. Jung, S. Nam, M. Chung, J. H. Kim, and I. Lee, "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Research*, vol. 41, no. D1, pp. D380–D386, 2018.

- [70] S. Sen, A. Dey, and U. Maulik, "Identifying Potential Hubs for Kidney Renal Clear Cell Carcinoma from TF-miRNA-Gene Regulatory Networks," *In Proceedings of IEEE Applied Signal Processing Conference*, pp. 240–244, 2018.
- [71] B. Meszaros, G. Erdos, and Z. Dosztanyi, "IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding," *Nucleic Acids Research*, vol. 46, no. W1, pp. W329–W337, 2018.
- [72] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky, "PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids," *Biochimica et Biophysica Acta*, vol. 1804, no. 4, pp. 996–1010, 2010.
- [73] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, no. 40, 2008.
- [74] J. Yang, R. Yan, D. X. A Roy, J. Poisson, and Y. Zhang, "The I-TASSER Suite: Protein structure and function prediction," *Nature Methods*, vol. 12, pp. 7–8, 2015.
- [75] D. Croft, G. O'Kelly, and et.al., "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, p. D691–D697, 2011.
- [76] C. Chen, S. Zabad, H. Liu, W. Wang, and C. Jeffery, "MoonProt 2.0: an expansion and update of the moonlighting proteins database," *Biochimica et Biophysica Acta*, vol. 46, no. D1, pp. D640–D644, 2018.
- [77] X. W. Zhang and Y. L. Yap, "The 3D structure analysis of SARS-CoV S1 protein reveals a link to influenza virus neuraminidase and implications for drug and antibody discovery," *Computational and Theoretical Chemistry*, vol. 681, no. 1, p. 137–141, 2004.
- [78] W. Kamitani, C. Huang, K. Narayanan, K. G. Lokugamage, and S. Makino, "A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein," *Nature Structural & Molecular Biology*, vol. 16, p. 1134–1140, 2009.
- [79] K. Narayanan, S. I. Ramirez, K. G. Lokugamage, and S. Makino, "Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression," *Virus Research*, vol. 202, p. 89–100, 2015.
- [80] K. G. Lokugamage, K. Narayanan, C. Huang, and S. Makino, "Severe Acute Respiratory Syndrome Coronavirus Protein nsp1 Is a Novel Eukaryotic Translation Inhibitor That Represses Multiple Steps of Translation Initiation," *Journal of Virology*, vol. 86, pp. 13 598–13 608, 2012.

## BIBLIOGRAPHY

---

- [81] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 40, no. 9, pp. 1–8, 2008.
- [82] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite: Protein structure and function prediction," *Nature Methods*, vol. 12, pp. 7–8, 2015.
- [83] V. N. Uversky, "Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics," *Frontiers of Physics*, vol. 7, no. 10, pp. 1–18, 2019.
- [84] V. P. Kutysenko and et al, "Solution structure and dynamics of the chimeric SH3 domains, SHH- and SHA-"Bergeracs"," *Biochimica et Biophysica Acta*, vol. 1794, no. 12, pp. 1813–1822, 2009.
- [85] L. Li, V. N. Uversky, A. K. Dunker, and S. O. Meroueh, "A computational investigation of allostery in the catabolite activator protein," *Journal of the American Chemical Society*, vol. 129, no. 50, p. Journal of the American Chemical Society, 2007.
- [86] B. Xue, L. Li, S. O. Meroueh, V. N. Uversky, and A. K. Dunker, "Analysis of structured and intrinsically disordered regions of transmembrane proteins," *Molecular BioSystems*, vol. 5, no. 12, p. 1688–1702, 2009.
- [87] T. N. Melnik, T. V. Povarnitsyna†, A. S. Glukhov, V. N. Uversky, and B. S. Melnik, "Sequential Melting of Two Hydrophobic Clusters within the Green Fluorescent Protein GFP-cycle3," *Biochemistry*, vol. 50, no. 36, p. 7735–7744, 2011.
- [88] B. S. Melnik, T. V. Povarnitsyna, A. S. Glukhov, T. N. Melnik, V. N. Uversky, and R. H. Sarma, "SS-Stabilizing Proteins Rationally: Intrinsic Disorder-Based Design of Stabilizing Disulphide Bridges in GFP," *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 4, pp. 815–824, 2012.
- [89] B. S. Melnik, N. V. Molochkov, D. A. Prokhorova, V. N. Uversky, and V. P. Kutysenko, "Molecular mechanisms of the anomalous thermal aggregation of green fluorescent protein," *Biochim Biophys Acta*, vol. 1814, no. 12, pp. 1930–1939, 2011.
- [90] D. Schoeman and B. C. Fielding, "Coronavirus envelope protein: current knowledge," *Virology Journal*, vol. 16, no. 69, pp. 1–22, 2019.
- [91] H. J. Dyson, "Roles of Intrinsic Disorder in Protein-Nucleic Acid Interactions," *Molecular Omics*, vol. 8, no. 1, p. 97–104, 2012.

- [92] C. Wang, V. N. Uversky, and L. Kurgan, "Disordered nucleome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea," *Proteomics*, vol. 16, no. 10, pp. 1486–1498, 2016.
- [93] Z. Wu, G. Hu, J. Yang, Z. Peng, V. N. Uversky, and L. Kurgan, "In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces," *FEBS Letters*, vol. 589, no. 19, pp. 2561–2569, 2010.
- [94] P. M. Mishra, N. C. Verma, C. Rao, V. N. Uversky, and C. K. Nand, "Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis," *Progress in Molecular Biology and Translational Science*, vol. 174, pp. 1–78, 2020.
- [95] G. K. M. Goh, A. K. Dunker, J. A. Foster, and V. N. Uversky, "Shell disorder analysis predicts greater resilience of the SARS-CoV-2 (COVID-19) outside the body and in body fluids," *Microbial Pathogenesis*, vol. 174, no. 104177, pp. 1–6, 2020.
- [96] T. Tanaka, W. Kamitani, M. L. DeDiego, L. Enjuanes, and Y. Matsuura, "Severe Acute Respiratory Syndrome Coronavirus nsp1 Facilitates Efficient Propagation in Cells through a Specific Translational Shutoff of Host mRNA," *Journal of Virology*, vol. 86, no. 20, p. 11128–11137, 2012.
- [97] F. Thuaud, N. Ribeiro, C. G. Nebigil, and L. Desaubry, "Prohibitin Ligands in Cell Death and Survival: Mode of Action and Therapeutic Potential," *Chemistry & Biology*, vol. 20, no. 3, pp. 316–331, 2013.
- [98] R. Giri, T. Bhardwaj, M. Shegane, B. R. Gehi, P. Kumar, K. Gadhawe, C. J. Oldfield, and V. N. Uversky, "Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses," *Cellular and Molecular Life Sciences*, pp. 1–34, 2020.
- [99] H. Zhang, J. M. Penninger, Y. Li, N. Zhong, and A. S. Slutsky, "Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target," *Intensive Care Medicine* volume, vol. 46, p. 586–590, 2020.
- [100] P. Zhou, X. L. Yang, and Z. L. Shi, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, p. 270–273, 2020.

## BIBLIOGRAPHY

---

- [101] A. Vandelli, M. Monti, E. Milanetti, R. D. Ponti, Profile, and G. G. Tartaglia, "Structural analysis of SARS-CoV-2 and predictions of the human interactome," *bioRxiv*, pp. 1–30, 2020.
- [102] J. P. Davies, K. M. Almasy, E. F. McDonald, and L. Plate, "Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies," *bioRxiv*, pp. 1–44, 2020.
- [103] J. M. Lucas and et al, "The Androgen-Regulated Protease TMPRSS2 Activates a Proteolytic Cascade Involving Components of the Tumor Microenvironment and Promotes Prostate Cancer Metastasis," *Cancer Discovery*, vol. 4, no. 11, p. 1310–1325, 2014.
- [104] Y. Ming and L. Qiang, "Involvement of Spike Protein, Furin, and ACE2 in SARS-CoV-2-Related Cardiovascular Complications," *SN Comprehensive Clinical Medicine*, vol. 114, pp. 1–6, 2020.
- [105] S. W. Li, C. Y. Wang, Y. J. Jou, T. C. Yang, S. H. Huang, L. Wan, Y. J. Lin, and C. W. Lin, "SARS coronavirus papain-like protease induces Egr-1-dependent up-regulation of TGF- $\beta$ 1 via ROS/p38 MAPK/STAT3 pathway," *Scientific Reports*, vol. 6, pp. 1–13, 2016.
- [106] C. Y. Wang and et al, "SARS coronavirus papain-like protease up-regulates the collagen expression through non-Samd TGF- $\beta$ 1 signaling," *Virus Research*, vol. 235, pp. 58–66, 2017.
- [107] X. Zhao, J. M. Nicholls, and Y. G. Chen, "Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-beta signaling," *The Journal of Biological Chemistry*, vol. 283, no. 6, pp. 3272–3280, 2007.
- [108] M. Montopoli and et al, "Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532)," *Annals of Oncology*, vol. 31, no. 8, p. 1040–1045, 2020.
- [109] G. Magro, "SARS-CoV-2 and COVID-19: Is interleukin-6 (IL-6) the 'culprit lesion' of ARDS onset? What is there besides Tocilizumab? SGP130Fc," *Cytokine: X*, vol. 2, no. 2, p. 100029, 2020.
- [110] W. Liang and et al., "Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China," *The Lancet Oncology*, vol. 21, no. 3, pp. 335–337, 2020.

- [111] R. K. Gurram, W. Kujur, S. K. Maurya, and J. N. Agrewala, "Caerulomycin A Enhances Transforming Growth Factor- $\beta$  (TGF- $\beta$ )-Smad3 Protein Signaling by Suppressing Interferon- $\gamma$  (IFN- $\gamma$ )-Signal Transducer and Activator of Transcription 1 (STAT1) Protein Signaling to Expand Regulatory T Cells (Tregs)\*," *Journal of Biological Chemistry*, vol. 289, no. 25, pp. 17 515–17 528, 2014.
- [112] V. Mollica, A. Rizzo, and F. Massari, "The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer," *Future Oncology*, vol. 21, no. 3, pp. 1–5, 2020.
- [113] M. Deleidi and O. Isacson, "Viral and Inflammatory Triggers of Neurodegenerative Diseases," *Science Translational Medicine*, vol. 4, no. 121, pp. 1–9, 2012.
- [114] C. Porro, A. Cianciulli, and M. A. Panaro, "The Regulatory Role of IL-10 in Neurodegenerative Diseases," *Biomolecules*, vol. 10, no. 7, pp. 1–15, 2020.
- [115] X. W. Zhang and Y. L. Yap, "The 3D structure analysis of SARS-CoV S1 protein reveals a link to influenza virus neuraminidase and implications for drug and antibody discovery," *Computational and Theoretical Chemistry*, vol. 681, no. 1, p. 137–141, 2004.
- [116] C. M. S. Singal, P. Jaiswal, and P. Seth, "SARS-CoV-2, More than a Respiratory Virus: Its Potential Role in Neuropathogenesis," *ACS Chemical Neuroscience*, vol. 11, no. 13, p. 1887–1899, 2020.
- [117] J. Huang, M. Zheng, X. Tang, Y. Chen, A. Tong, and L. Zhou, "Potential of SARS-CoV-2 to Cause CNS Infection: Biologic Fundamental and Clinical Experience," *Frontiers in Neurology*, vol. 11, no. 659, pp. 1–9, 2020.
- [118] X. Sun and et al, "Cytokine storm intervention in the early stages of COVID-19 pneumonia," *Cytokine Growth Factor Reviews*, vol. 53, p. 38–42, 2020.
- [119] J. P. Rogers, E. Chesney, D. Oliver, T. A. Pollak, P. McGuire, P. F. Poli, M. S. Zandi, G. Lewis, and A. S. David, "Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic," *Lancet Psychiatry*, vol. 7, no. 7, p. 611–627, 2020.
- [120] M. Brandao, T. Simon, G. Critchley, and G. Giamas, "Astrocytes, the rising stars of the glioblastoma microenvironment," *Glia*, vol. 67, no. 5, pp. 779–790., 2019.

## BIBLIOGRAPHY

---

- [121] C. P. Couturier, S. Ayyadhury, P. U. Le, and et al., "Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy," *Nat Commun*, vol. 11, no. 3406, pp. 1–19, 2020.
- [122] Q. Cheng, J. Li, F. Fan, H. Cao, Z.-Y. Dai, Z.-Y. Wang, and S.-S. Feng, "Identification and analysis of glioblastoma biomarkers based on single cell sequencing," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.
- [123] J. Cha and I. Lee, "Single-cell network biology for resolving cellular heterogeneity in human diseases," *Experimental & Molecular Medicine*, vol. 52, p. 1798–1808, 2020.
- [124] H. Dai, L. Li, T. Zeng, and L. Chen, "Cell-specific network constructed by single-cell RNA sequencing data," *Nucleic Acids Research*, vol. 47, no. 11, pp. 1–14, 2019.
- [125] X. Wang, D. Choi, and K. Roeder, "Constructing local cell-specific networks from single-cell data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 51, pp. 1–8, 2021.
- [126] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech*, pp. 1–12, 2008.
- [127] N. Meghanathan, "Maximal Clique Size Versus Centrality: A Correlation Analysis for Complex Real-World Network Graphs," In *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, vol. 44, p. 95–101, 2016.
- [128] H. Han, H. Shim, and D. e. a. D. Shin, "TRRUST: a reference database of human transcriptional regulatory interactions," *Sci Rep*, vol. 5, no. 11432, pp. D561–D568, 2015.
- [129] X. Zhang and et al, "CellMarker: a manually curated resource of cell markers in human and mouse," *Nucleic Acids Research*, vol. 47, no. D1, p. D721–D728, 2019.
- [130] O. Franzen, L. Gan, and J. L. M. Bjorkegren, "PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data," *Database*, pp. 1–9, 2019.
- [131] A. Dey, S. Sen, and U. Maulik, "Study of transcription factor druggability for prostate cancer using structure information, gene regulatory networks and protein moonlighting," *Briefings in Bioinformatics*, vol. 23, 2022.



- [132] T. Hide, Y. Komohara, Y. Miyasato, and et al., "Oligodendrocyte Progenitor Cells and Macrophages/Microglia Produce Glioma Stem Cell Niches at the Tumor Border," *EBioMedicine*, vol. 30, p. 94–104, 2018.
- [133] M. Gjorgjevski, R. Hannen, B. Carl, and et al., "Molecular profiling of the tumor microenvironment in glioblastoma patients: correlation of microglia/macrophage polarization state with metalloprotease expression profiles and survival," *Bioscience Reports*, vol. 39, no. 6, pp. 1–12, 2019.
- [134] J. D. Lathia, S. C. Mack, E. E. Mulkearns-Hubert, C. L. Valentim, and J. N. Rich, "Cancer stem cells in glioblastoma," *Genes Dev*, vol. 29, no. 12, p. 1203–1217, 2015.
- [135] Y. C. Ooi, P. Tran, and N. Ung, "The role of regulatory T-cells in glioma immunology," *Clin Neurol Neurosurg*, vol. 119, pp. 125–132, 2014.
- [136] Y. Soda, T. Marumoto, D. F. Morvinski, and et al., "Transdifferentiation of glioblastoma cells into vascular endothelial cells," *PNAS*, vol. 108, no. 11, pp. 4274–4280, 2011.
- [137] C. L. Chang, A. Rashidi, J. Miska, and et al., "Myeloid-Derived Suppressive Cells Promote B cell-Mediated Immunosuppression via Transfer of PD-L1 in Glioblastoma," *PNAS*, vol. 7, no. 12, pp. 1928–1943, 2019.
- [138] F. Yang, Y. Xie, J. Tang, B. Liu, Y. Luo, Q. He, L. Zhang, L. Xin, J. Wang, S. Wang, S. Zhang, Q. Cao, L. Wang, L. He, and L. Zhang, "Uncovering a distinct gene signature in endothelial cells associated with contrast enhancement in glioblastoma," *Frontiers in Oncology*, vol. 11, 2021.
- [139] R. Rispoli, C. Conti, P. Celli, E. Caroli, and S. Carletti, "Neural Stem Cells and Glioblastoma," *Neuroradiol J.*, vol. 27, p. 169–174, 2014.
- [140] T. Johung and M. Monje, "Neuronal Activity in the Glioma Microenvironment," *Curr Opin Neurobiol*, vol. 47, p. 156–161, 2017.
- [141] X. N. Zhang, K. D. Yang, and et al., "Pericytes augment glioblastoma cell resistance to temozolomide through CCL5-CCR5 paracrine signaling," *Cell Research*, vol. 31, pp. 1–16, 2021.
- [142] K. E. Cahill, R. A. Morshed, and B. Yamini, "Nuclear factor-KB in glioblastoma: insights into regulators and targeted therapy," *Neuro-Oncology*, vol. 18, no. 3, p. 329–339, 2015.
- [143] N. D. L. Iglesia, S. V. Puram, and A. Bonni, "STAT3 Regulation of Glioblastoma Pathogenesis," *Curr Mol Med*, vol. 9, no. 5, p. 580–590, 2009.

## BIBLIOGRAPHY

---

- [144] Y. Gao, B. Liu, and L. F. and et al., "Targeting JUN, CEBPB, and HDAC3: A Novel Strategy to Overcome Drug Resistance in Hypoxic Glioblastoma," *Frontiers in Oncology*, vol. 9, 2019.
- [145] H. Zhang, J. M. Penninger, Y. Li, N. Zhong, and A. S. Slutsky, "Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target," *Intensive Care Medicine*, vol. 46, p. 586–590, 2020.
- [146] S. Lukassen and et.al, "SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells," *The EMBO Journal*, vol. 88, no. 2, p. 913–924, 2020.
- [147] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.
- [148] Z. Yu and et.al, "Single-Cell Transcriptomic Map of the Human and Mouse Bladders," *Journal of the American Society of Nephrology*, vol. 30, no. 11, pp. 2159–2176, 2019.
- [149] Y. Wang, W. Song, J. Wang, T. Wang, X. Xiong, Z. Qi, W. Fu, X. Yang, and Y. G. Chen, "Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine," *Journal of Experimental Medicine*, vol. 217, no. 2, pp. 1–15, 2020.
- [150] J. Liao, Z. Yu, Y. Chen, M. Bao, C. Zou, H. Zhang, D. Liu, T. Li, Q. Zhang, J. Li, J. Cheng, and Z. Mo, "Single-cell RNA sequencing of human kidney," *Scientific Data*, vol. 7, no. 4, pp. 1–9, 2020.
- [151] S. A. MacParland, J. C. Liu, and I. D. McGilvray, "Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations," *Nature Communications*, vol. 9, no. 4383, pp. 1–21, 2015.
- [152] P. A. Reyfman and et.al, "Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis," *American Journal of Respiratory and Critical Care Medicine*, vol. 199, no. 12, pp. 1517–1536, 2016.
- [153] J. Z. Wang and et.al, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, 2007.
- [154] C. Pesquita and et.al, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, 2008.

- 
- [155] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.
- [156] D. Croft and et.al, "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, vol. 39, 2011.
- [157] M. Kanehisaa and S. Goto, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 28, 2000.
- [158] M. V. Kuleshov and et.al, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, pp. 90–97, 2016.
- [159] H. Zhang and et.al, "Digestive system is a potential route of covid-19: an analysis of single-cell coexpression pattern of key proteins in viral entry process," *Gut*, pp. 1–9, 2000.
- [160] A. Erol, "Pioglitazone treatment for the COVID-19-associated cytokine storm," *OSFPREPRINTS*, pp. 1–8, 2020.
- [161] S. Singla and J. R. Jacobson, "Statins as a Novel Therapeutic Strategy in Acute Lung Injury," *Pulmonary Circulation*, vol. 2, no. 4, pp. 397–406, 2012.
- [162] F. Taghizadeh-Hesary and H. Akbari, "The Powerful Immune System Against Powerful COVID-19: A Hypothesis," *Preprints 2020*, pp. 1–6, 2020.
- [163] A. M. South, D. I. Diz, and M. C. Chappell, "COVID-19, ACE2, and the cardiovascular consequences," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 318, p. 1084–1090, 2020.
- [164] A. M. South, L. Tomlinson, D. Edmonston, S. Hiremath, and M. A. Sparks, "Controversies of renin–angiotensin system inhibition during the covid-19 pandemic," *Nature Reviews Nephrology*, vol. 1, no. 3, pp. 111–119, 2020.
- [165] F. Huang and et.al, "Angiotensin II plasma levels are linked to disease severity and predict fatal outcomes in H7N9-infected patients," *Nature Communications*, vol. 5, no. 3595, pp. 1–7, 2014.
- [166] G. M. Kuster, O. Pfister, T. Burkard, Q. Zhou, R. Twerenbold, P. Haaf, A. F. Widmer, and S. Osswald, "Ras inhibition as a therapeutic chance associated with covid-19," *European Heart Journal*, 2020.
- [167] P. Mehta, D. F. McAuley, M. Brown, E. Sanchez, R. S. Tattersall, and J. J. Manson, "COVID-19: consider cytokine storm syndromes and immunosuppression," *The Lancet*, vol. 395, no. 10229, 2020.
-

## BIBLIOGRAPHY

---

- [168] S. Madsbad, "COVID-19 Infection in People with Diabetes," *Endocrinology*, 2020.
- [169] D. Sinha, A. Kumar, H. Kumar, S. Bandyopadhyay, and D. Sengupta, "drop-clust: efficient clustering of ultra-large scrna-seq data," *Nucleic Acids Research*, vol. 46, no. 6, p. e36, 2018.
- [170] J. Kellner, S. Li, P. A. Zweidler-McKay, E. J. Shpall, and I. McNiece, "Phenotypic and functional comparison of mobilized peripheral blood versus umbilical cord blood megakaryocyte populations," *Cytotherapy*, vol. 17, no. 4, pp. 418–427, 2015.
- [171] S. S. Mostafa, E. T. Papoutsakis, and W. M. Miller, "Oxygen tension modulates the expression of cytokine receptors, transcription factors, and lineage-specific markers in cultured human megakaryocytes," *Experimental Hematology*, vol. 29, no. 7, pp. 873–883, 2001.
- [172] M. Schelker, S. Feau, J. Du, N. Ranu, E. Klipp, G. MacBeath, B. Schoeberl, and A. Raue, "Estimation of immune cell content in tumour tissue using single-cell rna-seq," *Nature Communications*, vol. 8, no. 1, p. 2032, 2017.
- [173] C. Zheng and et. al., "Landscape of infiltrating t cells in liver cancer revealed by single-cell sequencing," *Cell*, vol. 169, no. 7, pp. 1342–1356, 2017.
- [174] E. Carrasco and et al., "Human cd6 down-modulation following t-cell activation compromises lymphocyte survival and proliferative responses," *Frontiers in Immunology*, vol. 8, no. 769, 2017.
- [175] S. S. Skanland, K. Moltu, T. Berge, E. M. Aandahl, and K. Tasken, "T-cell co-stimulation through the cd2 and cd28 co-receptors induces distinct signalling responses," *The Biochemical Journal*, vol. 460, no. 3, pp. 399–410, 2014.
- [176] A. C. Villani and et al., "Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors," *Science*, vol. 356, no. 6335, pp. 517–526, 2017.
- [177] N. Gil-Yarom and et al., "Cd74 is a novel transcription regulator," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 3, p. 562–567, 2017.
- [178] H. Rabinowich, L. Pricop, R. B. Herberman, and T. L. Whiteside, "Expression and function of cd7 molecule on human natural killer cells," *The Journal of Immunology*, vol. 152, no. 2, pp. 517–526, 1994.
- [179] F. Wolf, P. Angerer, and F. Theis, "SCANPY: large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, pp. 1–5, 2018.

- 
- [180] M. Mandal, J. Mondal, and A. Mukhopadhyay, "A PSO-based Approach for Pathway Marker Identification from Gene Expression Data," *IEEE transactions on nanobioscience*, vol. 14, pp. 591–597, 2015.
- [181] J. H. Levine, E. F. Simonds, S. C. Bendall, and et al, "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis," *Cell*, vol. 162, pp. 184–197, 2015.
- [182] G. Zheng, J. Terry, and P. B. et al, "Massively parallel digital transcriptional profiling of single cells," *Nat Commun*, vol. 8, no. 14049, pp. 1–12, 2017.
- [183] M. Litviňuková, C. Talavera-López, H. Maatz, D. Reichart, C. L. Worth, E. L. Lindberg, M. Kanda, K. Polanski, M. Heinig, M. Lee *et al.*, "Cells of the adult human heart," *Nature*, vol. 588, no. 7838, pp. 466–472, 2020.
- [184] N. X. Vinh, J. Epp, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.
- [185] N. N. Kachouie and M. Shutaywi, "Weighted Mutual Information for Aggregated Kernel Clustering," *Entropy*, vol. 22, no. 3, pp. 1–15, 2020.
- [186] T. Kan, S. Zhang, and S. Z. et al., "Single-cell RNA-seq recognized the initiator of epithelial ovarian cancer recurrence," *Oncogene*, vol. 41, p. 895–906, 2022.
- [187] Y. Guo, Y. Li, B. Cai, Q. He, G. Chen, M. Wang, K. Wang, X. Wan, and Q. Yan, "Phenotyping of immune and endometrial epithelial cells in endometrial carcinomas revealed by single-cell RNA sequencing," *Aging*, vol. 13, no. 5, p. 6565–6591, 2021.
- [188] W. Shan, C. Wang, Z. Zhang, and et al, "ATM may be a protective factor in endometrial carcinogenesis with the progesterone pathway," *Tumour Biology*, vol. 36, pp. 1529–1537, 2015.
- [189] L. McGlorthan, A. Paucarmayta, Y. Casablanca, and et al, "Progesterone induces apoptosis by activation of caspase-8 and calcitriol via activation of caspase-9 pathways in ovarian and endometrial cancer cells in vitro," *Apoptosis*, vol. 3, pp. 184–194, 2021.
- [190] T. Kan, S. Zhang, and S. Z. et al., "Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and transcriptional activities of ECs in CC," *Molecular Therapy: Nucleic Acid*, vol. 24, pp. 682–694, 2021.
- [191] D. DeTomaso, M. G. Jones, M. Subramaniam, and et al, "Functional interpretation of single cell similarity maps," *Nature Communications*, vol. 10, pp. 1–11, 2019.
-

## BIBLIOGRAPHY

---

- [192] B. Lake, S. Chen, B. Sos, and et al., "Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain," *Nature Biotechnology*, vol. 36, pp. 70–80, 2018.
- [193] S. Hanzelmann, R. Castelo, and J. Guinney, "GSVA: gene set variation analysis for microarray and RNA-Seq data," *BMC Bioinformatics*, vol. 14, pp. 1–15, 2017.
- [194] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, pp. 1–11, 2005.
- [195] E. Lee, H. Y. Chuang, J. W. Kim, and et al, "Inferring Pathway Activity toward Precise Disease Classification," *PLoS Computational Biology*, vol. 4, pp. 1–9, 2008.
- [196] S. Aibar, C. B. Gonzalez-Blas, T. Moerman, and et al., "SCENIC: single-cell regulatory network inference and clustering," *Nature Methods*, vol. 14, p. 1083–1086, 2017.
- [197] Y. Xia, S. Shen, and I. M. Verma, "NF- $\kappa$ B, an active player in human cancers," *Cancer immunology research*, vol. 2, p. 823–830, 2014.
- [198] B. S. Harrington and C. M. Annunziata, "NF- $\kappa$ B Signaling in Ovarian Cancer," *Cancers*, vol. 11, pp. 1–16, 2019.
- [199] T. Zhang, C. Ma, Z. Zhang, and et al, "NF- $\kappa$ B signaling in inflammation and cancer," *MedComm*, vol. 2, p. 618–653, 2021.
- [200] L. Zdrojkowski, T. Jasinski, G. F. Dias, and et al, "The Role of NF- $\kappa$ B in Endometrial Diseases in Humans and Animals: A Review," *International Journal of Molecular Sciences*, vol. 24, pp. 1–14, 2023.
- [201] S. Gupta, P. Kumar, and B. C. Das, "HPV: Molecular pathways and targets," *Current Problems in Cancer*, vol. 42, pp. 161–174, 2018.
- [202] S. Tilborghs, J. Corthouts, Y. Verhoeven, and et al, "The role of Nuclear Factor-kappa B signaling in human cervical cancer," *Critical Reviews in Oncology / Hematology*, vol. 120, pp. 141–150, 2017.