Development of some Transcription Factor Regulatory Network Architectures for Differentially Expressed Genes

Thesis Submitted By

Aurpan Majumder

Doctor of Philosophy (Engineering)

Dept. of Instrumentation and Electronics Engineering
Faculty Council of Engineering and Technology
Jadavpur University
Kolkata- 700032, India

2023

JADAVPUR UNIVERSITY

KOLKATA - 700032, INDIA

Index no: 213/17/E

Title of the thesis: Development of some Transcription Factor Regulatory Network Architectures for Differentially Expressed Genes

Name & Dept. of the Supervisor

Dr. Prolay Sharma

Associate Professor
Department of Instrumentation and ElectronicsEngineering
Jadavpur University, Salt Lake Campus,Kolkata-700106, India

List of Publications

International Journals

[**1**] A. Majumder and M. Sarkar, "Exploring Different Stages of Alzheimer's Disease through Topological Analysis of Differentially Expressed Genetic Networks", International Journal of Computer Theory and Engineering, volume 6, issue 5, pages 386-391, October 2014. DOI: 10.7763/IJCTE.2014.V6.895 (SCOPUS)

[**2**] A. Majumder and M. Sarkar, "Paired Transcriptional Regulatory System for Differentially Expressed Genes", Lecture Notes on Information Theory, volume 2, issue 3, pages 266-272, September 2014. DOI: 10.12720/lnit.2.3.266-272

[**3**] M. Sarkar and A. Majumder, "Quantitative Trait Specific Differential Expression (qtDE)", Procedia Computer Science, volume 46, pages 706-718, April 2015. https://doi.org/10.1016/j.procs.2015.02.131 (SCOPUS)

[**4**] A. Majumder, M. Sarkar, H. Dash, and I. Akhilesh, "A Composite Entropy Model in a Multiobjective Framework for Gene Regulatory Networks", Current Bioinformatics, volume 13, issue 1, pages 85-94, 2018. http://dx.doi.org/10.2174/1574893611666161202104422 (SCI)

[**5**] A. Majumder, M. Sarkar, and P. Sharma, "A Composite Mode Differential Gene Regulatory Architecture based on Temporal Expression Profiles", IEEE/ACM Transactions on Computational Biology and Bioinformatics, volume 16, issue 6, pages 1785-1793, November-December 2019. https://doi.org/10.1109/tcbb.2018.2828418 (SCI)

[**6**] A. Majumder and P. Sharma, "Topologically Overlapped Fused LASSO Measure for Reconstructing Gene Regulation Networks", Published online: 13 Nov 2023, IETE Journal of Research, 2023. https://doi.org/10.1080/03772063.2023.2280620 (SCI)

International Conference Proceedings as Book Chapters

[**1**] M. Sarkar and A. Majumder, "Intelligent Topological Differential Gene Networks", In: Das, S., Pal, T., Kar, S., Satapathy, S., Mandal, J. (eds) Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015. Advances in Intelligent Systems and Computing, volume 404, Springer, New Delhi. https://doi.org/10.1007/978-81-322-2695-6_8 (SCOPUS)

[**2**] A. Majumder and M. Sarkar, “Dissimilar Regulatory Actions Between Neurodegenerative Disease Pairs Through Probablistic Differential Correlation”, In: P. Deiva Sundari et al. (eds.), Proceedings of 2nd International Conference on Intelligent Computing and Applications, Advances in Intelligent Systems and Computing, volume 467, pp. 59-74, October 2016, Springer, Singapore. http://dx.doi.org/10.1007/978-981-10-1645-5_6 (SCOPUS)

[**3**] M. Sarkar and A. Majumder, “Multiobjective Ranked Selection of Differentially Expressed Genes”, In: P. Deiva Sundari et al. (eds.), Proceedings of 2nd International Conference on Intelligent Computing and Applications, Advances in Intelligent Systems and Computing, volume 467, pp. 75-92, October 2016, Springer, Singapore. https://doi.org/10.1007/978-981-10-1645-5_7 (SCOPUS)

IEEE International Conference Proceeding

[**1**] A. Majumder and M. Sarkar, “Simple Transcriptional Networks for Differentially Expressed Genes”, In IEEE International Conference on Signal Propagation and Computer Technology, ICSPCT 2014, 12-13 July 2014, https://doi.org/10.1109/ICSPCT.2014.6885016

Statement of Originality

*I, **Aurpan Majumder**, registered on the 22nd of September, 2017, declare that this thesis entitled **"Development of some Transcription Factor Regulatory Network Architectures for Differentially Expressed Genes"** 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 following 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 per the **"Policy on Anti Plagiarism, Jadavpur University, 2019"**, and the level of similarity as checked by iThenticate software is* 4% .

Aurpan Majumder

Aurpan Majumder

Index No. 213/17/E

Date: August 23, 2023

Prolay Sharma

Dr. Prolay Sharma

Dr. Prolay Sharma
Associate Professor
Dept. of Instrumentation and Electronics Engg.
Jadavpur University, Salt Lake Campus,
Sector-III, Block-LB, Plot No.- 8, Kolkata-700 106

Associate Professor
Department of Instrumentation and Electronics Engineering
Jadavpur University, Salt Lake Campus
Kolkata-700106, India.

Certificate from the Supervisor

Date: August 23, 2023

This is to certify that the thesis titled "***Development of some Transcription Factor Regulatory Network Architectures for Differentially Expressed Genes***" submitted by **Mr. Aurpan Majumder**, who got his name registered on 22nd September 2017 for the award of Ph.D. (Engineering) degree of Jadavpur University, is based upon his work under the supervision of the undersigned; and neither that his thesis nor any part of it has been submitted for any degree/diploma or any other academic award anywhere before.

Prolay Sharma

Dr. Prolay Sharma
Associate Professor
Dept. of Instrumentation and Electronics Engg.
Jadavpur University, Salt Lake Campus,
Sector-III, Block-LB, Plot No.- 8, Kolkata-700 106

Dr. Prolay Sharma

Associate Professor

Department of Instrumentation and Electronics Engineering

Jadavpur University, Salt Lake Campus

Kolkata-700106, India.

Dedicated to my Family

Acknowledgements

At the beginning, I would like to express my sincere gratitude to my supervisor, Dr. Prolay Sharma, Associate Professor, Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata, for his indispensible advice, courteous support, and diligence during my PhD study. His avid practical skills and vivid experience did encourage me throughout my academic research maintaining a strong impression on every path that had been followed to ensure the true potential of the research work.

I would like to express my heartfelt thanks to Prof. Rajib Bandyopadhyay and Prof. Bipan Tudu, Professor, Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata, for their influential support in overcoming the initial hurdles of academic research. I am also grateful to Dr. Ratna Ghosh, Professor and Head, Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata, for the suggestions and encouragement during the research tenure. Additionally, I want to thank the entire Instrumentation & Electronics Engineering Department staff at Jadavpur University for their cooperation and support in completing this task.

At the other end, being an employee of NIT Durgapur, I am really grateful to all my fellow colleagues and the administrative body of the institute for granting the necessary permission required for carrying out this research work outside the institute.

Last, but not the least, the wholehearted appreciation goes out to my family members whose ardent support had always been a strong backbone during the research tenure.

Date: August 23, 2023

Place: Kolkata

Aurpan Majumder

(**Aurpan Majumder**)

Abstract

A fully functional gene regulatory network can be formed using gene-gene and/or gene-protein interactive patterns. To maintain a healthy cell cycle, it is necessary to have a proper control of the regulatory proteins in the network. Excess protein concentration may lead to beyond control division of the healthy cells causing cancer. Microarray gene expression profiles are frequently explored to understand the causal factors of regulation associated with some disease. Most of the significant research to interpret the regulations in a gene regulatory network is restricted to comparison of gene expression values across more than one condition or the discovery of genes having altered interaction levels with neighbours across conditions. Therefore, differential expression (DE), gene correlation and differential co-expression have been intensively studied using microarray gene expression profiles. In other words, the regulatory action of a gene in a complex network is guided by the differential functionalities of the gene acting under varied conditions. To gain better insight significant state of the art methodologies primarily explore the existence of differential co-expression patterns, where the co-expression level between genes alters across different states. Rigorous researches via such methodologies help in comparing the expression samples over normal and diseased states making us understand the pattern of various diseases.

Exploring the complex interactive mechanism in a Gene Regulatory Network (GRN) developed using transcriptome data obtained from standard microarray and/or RNA-seq experiments helps us to understand the triggering factors in diseased pathways. In this regard, the Transcription Factor (TF) genes generate protein complexes which affect the transcription of various target genes. These TF genes play a pivotal role in a Gene Regulatory Network (GRN) by differentially regulating genes across conditions. In some cases, it requires coordinated regulation of multiple TFs to control a Differentially Expressed (DE) gene. This form of regulation can be restricted to simple pairwise structures and may extend involving multiple TF genes regulating a target DE gene. This form of differential regulation can be expected to occur in

between two levels or even beyond following multiple hierarchical paths of regulation to the target DE gene. These regulatory situations helping in the reconstruction of differential TF regulatory networks (TRNs; can be considered as a subset of GRNs) emulating the biologically significant KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathways can be designed through various forms of transcriptome data utilizing static or time series expression profiles. Further extension of these approaches helps us in identifying the therapeutic targets, which happens to be open challenge in systems biology. In this regard, indirect gene regulatory hierarchical architectures may be promising enough considering varied topological structures and unknown gene regulation factors. Such causal regulations can be investigated, keeping in force all perturbation experiments of a dataset. Contemporary state of research primarily highlights direct interaction networks which mostly forego the inevitable presence of a third entity, if any, towards varied forms of causal regulations.

Table of Contents

Chapter 1 Introduction and Scope of the Thesis

Chapter 2 Nature of Differentially Expressed Gene and Transcription Factor Regulatory Networks

List of Figures

List of Tables

List of Abbreviations

A	**ACO**	**A**nt **C**olony **O**ptimization
	AD	**A**lzheimer's **D**isease
	ApDE	**A**periodic **D**ifferentially **E**xpressed
	ALS	**A**myotrophic **L**ateral **S**clerosis
	AN	**A**ctivator **N**ecessary
	ARACNE	**A**lgorithm for the **R**econstruction of **A**ccurate **C**ellular **NE**tworks
	AS	**A**ctivator **S**ufficient
	ATP	**A**denosine **T**ri**P**hosphate
	ApTF	**A**periodic **T**ranscription **F**actor
	AUROC	**A**rea **U**nder **R**egion **O**f **C**haracteristic
B	**BPNN**	**B**ack **P**ropagation **N**eural **N**etworks
C	**CDF**	**C**umulative **D**istribution **F**unction
	CNS	**C**entral **N**ervous **S**ystem
D	**DAVID**	**D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery
	DCE	**D**ifferential **C**o-**E**xpressed
	DCGLv2	**D**ifferential **C**o-expressed **G**ene **L**ink **version 2**
	DCL	**D**ifferential **C**oexpressed **L**ink
	DE	**D**ifferentially **E**xpressed
	DEGseq	**D**ifferentially **E**xpressed **G**enes from RNA-**seq** data
	DNA	**D**eoxyribo**N**ucleic **A**cid
	dCSA_r2t	**d**ifferential **C**orrelation **S**et **A**nalysis **r**egulator **to** **t**arget
	dCSA_t2t	**d**ifferential **C**orrelation **S**et **A**nalysis **t**arget **to** **t**arget
F	**FN**	**F**alse **N**egative
	FP	**F**alse **P**ositive
	FPR	**F**alse **P**ositive **R**ate
G	**GA**	**G**enetic **A**lgorithm
	GLMNET	Lasso and Elastic-**Net** Regularized **G**eneralized **L**inear **M**odels
	GO	**G**ene **O**ntology
	GRL	**G**ene **R**egulatory **L**ink
	GRN	**G**ene **R**egulatory **N**etwork
	GSAR	**G**ene **S**et **A**nalysis in **R**
	GSE	**G**ene **S**eries **E**xpression
	GSEA	**G**ene **S**et **E**nrichment **A**nalysis
	GTOM	**G**eneralized **T**opological **O**verlap **M**easure

H	**HCV**	**H**epatitis **C** **V**irus
	HD	**H**untington's **D**isease
	HeLa	cervical cancer cells of **He**nrietta **La**cks
I	**ILMN**	**IL**lu**M**i**N**a ID of a Gene
K	**KEGG**	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
L	**LASSO**	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
	LLR	**L**og **L**ikelihood **R**atio
M	**MS**	**M**ultiple **S**clerosis
	mTOM	**m**ultiple **T**opological **O**verlap **M**easure
N	**NMF**	**N**on-negative **M**atrix **F**actorization
	NSGA	**N**on-dominated **S**orting **G**enetic **A**lgorithm
O	**ORF**	**O**pen **R**eading **F**rame
P	**PCIT**	**P**artial **C**orrelation and **I**nformation **T**heory
	PD	**P**arkinson's **D**isease
	PDE	**P**eriodic **D**ifferentially **E**xpressed
	PESA	**P**areto **E**nvelope-based **S**election **A**lgorithm
	PRISM	**PR**otein **I**nformatics **S**ystem for **M**odeling
	PSO	**P**article **S**warm **O**ptimization
	PTF	**P**eriodic **T**ranscription **F**actor
Q	**qtDE**	**Q**uantitative **T**rait specific **D**ifferentially **E**xpressed
R	**RIF**	**R**egulatory **I**mpact **F**actor
	RIFT	**R**egulatory **I**mpact **F**actor with **T** score
	RMA	**R**obust **M**ulti-array **A**verage
	RN	**R**epressor **N**ecessary
	RNA	**R**ibo**N**ucleic **A**cid
	RS	**R**epressor **S**ufficient
S	**SCHIZ**	**Schiz**ophrenia
	SynTREN	**Syn**thetic **T**ranscriptional **RE**gulatory **N**etwork
T	**TED**	**T**omato **E**xpression **D**atabase
	TDD	**T**est **D**riven **D**evelopment
	TF	**T**ranscription **F**actor
	TF_DC	**T**ranscription **F**actor **D**ifferentially **C**oexpressed
	TF_DE	**T**ranscription **F**actor **D**ifferentially **E**xpressed
	TN	**T**rue **N**egative
	TO	**T**opological **O**verlap
	TOP	**T**opological **O**verlap with **P**-value significance

	TP	**T**rue **P**ositive
	TPR	**T**rue **P**ositive **R**ate
	TRIM	**T**ranscriptional **R**egulatory **I**nteraction **M**odel
	TRN	**T**ranscriptional **R**egulatory **N**etwork
	TRRUST	**T**ranscriptional **R**egulatory **R**elationships **U**nraveled by **S**entence-based **T**ext mining
W	**WGCNA**	**W**eighted **G**ene **C**o-expression **N**etwork **A**nalysis
Y	**YAGM**	**Y**east **A**ssociated **G**enes **M**iner
	YEASTRACT	**Yea**st **S**earch for **T**ranscriptional **R**egulators **A**nd **C**onsensus **T**racking

Chapter 1 Introduction and Scope of the Thesis

1.1 Introduction

Gene expression analysis making use of DNA microarray and RNA-seq experiments had always played an important role to explain complex biological interactions at the cellular level of different organisms [**1,2**]. In this regard, the substantial regulations involved can be defended with synergized role of genes and modular proteins in a particular genome experimented under different conditions or environments. This helps in instigating the effective causal reaction of genes in cancerous and normal states of a cell. A comprehensive picture of the dissimilar regulatory association of genes under varied metabolic conditions can be developed through Transcription Factor (TF), promoter DNA binding, protein-protein interaction (PPI) and protein-translational modification (PTM) [**3**]. Accordingly, differential regulation has emerged as a premier research module to interpret dynamic changes in Gene Regulation Networks (GRNs). In other words, the rewiring of GRNs in response to different environmental conditions can be made possible from the fundamental analysis of differential gene co-expression networks [**4-7**]. The differential analysis indicates either up or down regulation of genes compared to a control set of conditions. In this connection, the triggering factor or TF genes generate protein complexes that may have a direct [**8-12**] or indirect [**13-17**] causal effect on various target genes.

Understanding the nature of the target genes is crucial in the process of development of gene regulation networks. In this context, the healthy samples are compared with the diseased counterparts to interpret the working of the tissues having different gene expression levels across conditions. These types of genotypic variations are not only important in predicting the extent of disease but are inevitable to apprehend the heritable variations, an essential perspective in the evolutionary process. Evolution through natural selection motivates ecological changes across phenotypes [**18**]. Maintaining some specific patterns [**19**], the altered expression levels encountered by the differentially expressed genes (DE genes) have been studied across different development stages. Various salient methods proposed to infer the extent of differential expression are the single slide method [**20**], multiple slide method [**21**], Apo-AI and SR-BI [**20**], and DEGseq [**22**] to name a few.

A particularly challenging issue is to identify the regulatory network responsible for above kinds of target gene regulation in a given biological system [**23**]. In a generic sense any living cell exerts this control through interactions. This process allows limited combinations of ubiquitous, signal-specific transcription factors (TFs) binding to promoter DNA, to execute an exponentially larger number of regulatory decisions. The involvement of different kinds of TF genes in gene regulation makes possible the integration of several signalling pathways in the nucleus [**24**]. It is important to note that living cells function following the rules of cell growth and reproduction which highlight the orchestrated regulatory actions of genes and proteins [**25**]. In this context, some external factors such as UV rays, harmful chemical substances present in various food and beauty products, infiltration of viruses etc. may instigate the generation of signalling proteins causing the nucleus to stimulate cell division. These proteins do create a signal transduction cascade including a membrane receptor for the signal molecule, intermediary proteins that carry the signal through the cytoplasm, and transcription factors (TFs) in the nucleus thus activating the genes for cell division [**26**]. In each step of the transduction cascade, one TF gene may generate a protein complex influencing another TF gene or may have the capability of inducing the generation of multiple types of protein complexes and thus affecting the progress of regulation.

Extensive research guiding the process of regulatory control mentioned above not only requires the presence or importance of differentially expressed (DE) genes but also the involvement differential connectivity based on gene correlation or co-expression [**27**]. The amalgamation of these ideas in gene regulatory control can be apprehended in the true sense provided the differential co-expression of the participating DE genes or TF genes or TF and DE genes gives a better insight compared to standard DE gene expression analysis [**28**]. As given in [**9**], changes occurring in the coding region of some genes and posttranslational modifications (like phosphorylation, acylation, methylation, etc.) may lead to modification of protein activity without any significant change in the gene expression level but may encounter altered interaction pattern with other genes. In this regard, the effect of TF genes can be considered where a number of down- stream targets can be regulated by a master gene. Unfortunately, in diseased cells, the regulatory mechanism is dysfunctional because of high chance of random or unordered coalition of the genes present. Gene modules showing this kind of modified association can be detected primarily by the differential co-expression (DC) analysis as compared to standard differential expression (DE) analysis [**28**]. At the backdrop of differential expression (DE) and differential co-expression (DC), Transcription factors

(TFs) are proteins that bind to specific DNA sequences and play a crucial role in regulating gene expression. They are able to activate or repress the transcription of target genes, thereby influencing the production of specific proteins within a cell.

Differential gene regulation networks refer to the complex interactions between TFs and their target genes, which determine the specific gene expression patterns observed in different cell types, developmental stages, or physiological conditions. These networks are responsible for the dynamic and context-specific regulation of gene expression. In a differential gene regulation network, the activity of TFs is controlled by various factors, including environmental cues, signalling pathways, and interactions with other proteins. The binding of a TF to its target gene's regulatory region can either enhance or suppress gene expression. The specific DNA sequence recognized by a TF is known as its binding site. TFs can form regulatory cascades, where the expression of one TF is controlled by another TF. This hierarchical organization allows for precise control of gene expression patterns. Additionally, TFs can also form complex regulatory networks with feedback loops, feed-forward loops, and cross-regulatory interactions, further adding to the complexity of gene regulation. Advancements in high-throughput techniques, such as next-generation sequencing and chromatin immune-precipitation, have enabled the systematic identification of TF binding sites and the construction of genome-wide transcriptional regulatory networks. These networks provide valuable insights into the combinatorial interactions and regulatory logic underlying gene expression.

Studying transcription factor-controlled differential gene regulation networks is essential for understanding cellular processes, including development, disease progression, and response to environmental stimuli. By deciphering these networks, researchers can gain insights into the molecular mechanisms governing gene expression and identify potential therapeutic targets for various diseases.

1.2 Understanding DNA Microarray and RNA-seq Experiments

DNA microarray and RNA-seq are two commonly used experimental techniques for studying gene expression profiles on a genome-wide scale. They provide valuable insights into the transcriptional activity of genes and allow for the identification of differentially expressed genes under various conditions or in different cell types.

DNA Microarray:

A DNA microarray, also known as a gene chip, is a solid surface (such as a glass slide or silicon chip) on which thousands of DNA sequences or probes are immobilized in an ordered manner. The microarray contains specific probes that are designed to hybridize with complementary target sequences, typically cDNA or labelled RNA samples.

The workflow of a DNA microarray experiment typically involves the following steps, as given below.

Sample preparation: RNA is extracted from the cells or tissues of interest, and the RNA is reverse transcribed into complementary DNA (cDNA). If comparing two different conditions, two different sets of cDNA samples are prepared and labelled with different fluorescent dyes (e.g., Cy3 and Cy5).

Hybridization: The labelled cDNA samples are mixed and applied to the microarray slide. The cDNA binds to the complementary DNA probes on the microarray, forming a hybridization complex.

Scanning and data acquisition: The microarray slide is scanned using a fluorescence scanner to measure the intensity of the bound fluorescent dyes. The fluorescence intensity represents the relative abundance of the corresponding RNA transcripts in the original sample.

Data analysis: The fluorescence intensity values are processed and normalized to correct for technical variations. Statistical methods, such as t-tests or analysis of variance (ANOVA), are applied to identify genes that are differentially expressed between the compared conditions.

RNA-seq:

RNA-seq is a high-throughput sequencing-based technique that allows for the quantification of RNA transcripts in a sample. Unlike microarrays, RNA-seq provides a more comprehensive and unbiased view of the transcriptome, as it can detect both known and novel transcripts.

The general workflow of an RNA-seq experiment involves the following steps, as given below.

RNA extraction and purification: Total RNA is isolated from the cells or tissues of interest. The RNA may undergo additional purification steps to remove genomic DNA and other contaminants.

Library preparation: The isolated RNA is converted into complementary DNA (cDNA) through reverse transcription. The cDNA is then fragmented and sequenced library is generated. Different library preparation protocols exist, such as poly(A) enrichment for capturing mRNA or total RNA sequencing for capturing all RNA species.

Sequencing: The RNA-seq library is subjected to high-throughput sequencing using platforms like Illumina or Ion Torrent. The sequencing generates millions of short reads that represent fragments of the original RNA molecules.

Read mapping and quantification: The sequenced reads are aligned to a reference genome or transcriptome using bioinformatics tools. The aligned reads are then counted to determine the abundance of each transcript.

Data analysis: The read counts are normalized to correct for differences in library size and gene length. Statistical analysis, such as the calculation of fold changes or the application of specialized algorithms, is performed to identify genes that are differentially expressed between conditions.

It is notable that RNA-seq also allows for additional analyses beyond gene expression, such as alternative splicing, detection of novel transcripts, and identification of non-coding RNAs.

Both DNA microarray and RNA-seq provide valuable information about gene expression patterns. However, RNA-seq is generally considered more versatile and sensitive due to its ability to detect novel transcripts and quantify expression levels over a wider dynamic range. It has become the preferred choice for many gene expression studies in recent years.

1.3 Understanding Differentially Expressed Genes

Differentially expressed (DE) genes refer to genes that exhibit significant changes in their expression levels between different conditions or experimental groups. These conditions could be, for example, comparing gene expression in diseased versus healthy tissues, treated versus untreated samples, or different developmental stages.

Understanding DE genes is crucial in various areas of biological research, particularly in genomics, molecular biology, and biomedical sciences. Identifying DE genes helps researchers gain insights into the underlying molecular mechanisms associated with specific conditions or treatments. It can also provide valuable information about disease progression, biomarker discovery, and potential therapeutic targets.

Some key steps involved in understanding differentially expressed genes can be summarized as given below.

Experimental Design: A well-controlled experiment need to be designed with appropriate sample sizes and statistical power to ensure reliable results. There certain design factors such as treatment groups, replicates, and controls are required to be incorporated.

Data Generation: Gene expression data is required to be generated using high-throughput techniques such as microarrays or next-generation sequencing (RNA-seq). These methods quantify the expression levels of thousands of genes simultaneously.

Pre-processing and Normalization: The raw gene expression data requires pre-processing to remove technical artefacts and normalize the data across samples. This step ensures that the expression values are comparable between samples.

Statistical Analysis: At this stage, statistical analysis is an essential step to identify the DE genes. Various methods such as t-tests, analysis of variance (ANOVA), or more advanced techniques like edgeR, DEGseq, or DESeq2 can be used. These methods assess the statistical significance of gene expression differences between groups, considering factors such as fold change and adjusted p-values.

Multiple Testing Correction: After exploring the essential statistical analysis there may be a requirement of multiple hypotheses testing by adjusting p-values to control the false discovery rate (FDR). This correction helps reduce the chances of false positive results.

Gene Set Enrichment Analysis (GSEA): Applying this enrichment analysis it is possible to identify biological pathways, gene ontology terms, or other functional annotations that are overrepresented among the DE genes. This analysis helps to understand the biological relevance and potential functions of the identified genes.

Validation: Towards some fine tuning, the expression changes of selected DE genes may need to be validated using independent techniques such as quantitative PCR (qPCR) or immunohistochemistry. Validation experiments provide additional evidence of the differential expression observed in the initial analysis.

Functional Interpretation: The crux part of any research incorporating the DE genes includes interpretation of the biological significance of the DE genes in the context of the specific research question or condition under investigation. Available databases, literature, and bioinformatics tools are used to explore potential molecular interactions, signalling pathways, and biological processes associated with the identified DEGs.

By following these steps, researchers can gain a deeper understanding of the genes that play important roles in various biological processes and disease conditions. The identification and functional characterization of DE genes contribute to advancing our knowledge of gene regulation, disease mechanisms, and the development of new therapeutic interventions.

1.4 Understanding Differentially Co-expressed Genes

Differential co-expression analysis is a technique used in genomics to identify genes that show consistent changes in their expression levels across different conditions or samples. It aims to identify genes that have similar expression patterns, meaning they are co-expressed, but their levels of co-expression differ between conditions. These genes are referred to as differentially co-expressed genes.

The step-by-step overview of the process of understanding differentially co-expressed genes is given below.

Gene Expression Data: The gene expression data is obtained from different conditions or samples. This data is typically generated using technologies such as microarrays or RNA sequencing (RNA-seq). The data will consist of expression values for each gene across the conditions or samples.

Normalization: The gene expression data is to be normalized to correct for systematic biases and variations introduced during the experimental procedures. Common normalization methods include quantile normalization or variance stabilizing normalization (VSN).

Co-expression Analysis: The correlation or similarity between gene expression profiles across the conditions or samples is being computed. This can be done using various methods such as Pearson correlation, Spearman correlation, mutual information, polynomial regression, and spline regression. The result is a co-expression matrix that represents the pairwise relationships between genes.

Differential Co-expression Analysis: At this stage it is required to identify genes that exhibit differential co-expression patterns between conditions. This is typically done by comparing the co-expression values or correlations for each gene pair across conditions and identifying significant differences. Statistical tests such as t-tests, linear models, or non-parametric tests can be used to assess the significance.

Multiple Testing Correction: Due to the large number of gene pairs being tested, it is important to correct for multiple testing to reduce false positives. Methods like the Bonferroni correction, Benjamini-Hochberg procedure, or false discovery rate (FDR) control can be applied to adjust the p-values obtained from the statistical tests.

Functional Interpretation: Once differentially co-expressed genes are identified, it is crucial to assess their functional significance. Gene ontology (GO) analysis, pathway enrichment analysis, or network analysis can be performed to determine if these genes are involved in specific biological processes or pathways.

Validation: As per requirement, experimental validation of the identified differentially co-expressed genes may be necessary to confirm their biological relevance. Techniques such as qRT-PCR (quantitative real-time polymerase chain reaction) or independent gene expression profiling experiments can be performed to validate the results.

By identifying differentially co-expressed genes, researchers can gain insights into regulatory mechanisms, identify potential biomarkers, or discover key genes involved in specific biological processes or diseases.

1.5 Understanding the role of Transcription Factor Genes in a Gene Regulatory Network

Transcription factors (TFs) are proteins that play a crucial role in regulating gene expression by binding to specific DNA sequences in the promoter region of target genes. They act as molecular switches, turning genes on or off by facilitating or inhibiting the recruitment of RNA polymerase, which is necessary for gene transcription.

Gene regulatory networks (GRNs) are intricate systems of interacting genes and TFs that coordinate the expression of multiple genes, often in a specific spatial or temporal pattern. These networks are essential for various biological processes, such as development, cell differentiation, and response to environmental cues. The role of TF genes in a GRN is to control the activation or repression of target genes. They function by binding to DNA sequences called transcription factor binding sites (TFBS) in the promoter region of target genes. TF genes themselves can be regulated by other TFs or by external signals, forming complex regulatory cascades. Dysregulation of TFs can disrupt normal gene expression patterns and lead to abnormal cellular processes or developmental defects. For example, mutations in TF genes can be linked to cancer, developmental disorders, and metabolic diseases.

The interactions between TFs and their target genes can be categorized into activation and repression. Activator TFs enhance gene expression by recruiting co-activators and the RNA polymerase complex, while repressors TFs inhibit gene expression by blocking the binding of activators or recruiting co-repressors. TFs can act as master regulators that control the expression of multiple downstream target genes. A single TF gene can regulate the expression of numerous genes, while individual target genes can be regulated by multiple transcription factors. The combinatorial interactions between different transcription factors and their target genes form a complex network of gene regulation. The behaviour of such GRN depends on the specific TFs present, their binding affinities, and the combinatorial interactions between multiple TFs. During development, TF genes are often expressed in specific spatial and temporal patterns.

They play a vital role in specifying cell fate and guiding the differentiation of different cell types. By activating or repressing the expression of specific genes, TFs contribute to the establishment of cell identity and the formation of complex tissues and organs.

Understanding the role of TF genes in a GRN requires experimental techniques such as chromatin immune-precipitation (ChIP) to identify TF binding sites, as well as gene expression profiling to determine the target genes regulated by specific TFs. Additionally, computational modelling and bioinformatics analyses are employed to predict and study the dynamics of GRNs. By unravelling the intricate interactions between TF genes and target genes in a GRN, researchers can gain insights into the underlying mechanisms of complex biological processes. This knowledge is crucial for understanding normal development and disease states, and it has implications for fields such as synthetic biology and therapeutic interventions.

1.6 Motivation for the Work

Reconstruction of Gene Regulation Networks (GRNs) is done through analyzing static and time series gene expression data obtained from DNA microarray and/or RNA-seq experimental techniques used for studying gene expression profiles on a genome-wide scale. Analyzing the above forms of data through computational biology and varied types of algorithms in bioinformatics, the goal is to understand the relationships and interactions between genes and identify the regulatory mechanisms that control gene expression. In a broad perspective, incorporating both types of gene expression data, the following techniques are widely applied towards understanding the reconstruction of GRNs.

Data pre-processing: The first step is to pre-process the gene expression data. This includes normalizing the data to correct for technical variations, such as batch effects, background noise, and platform-specific biases. Common normalization methods that may be considered are like log transformation, quantile normalization, or robust z-score normalization.

Differential expression analysis: Differential expression analysis identifies genes that show significant changes in expression levels between different experimental conditions or sample groups. This step helps identify genes that are potentially regulated by specific factors or biological processes of interest.

Co-expression analysis: Co-expression analysis aims to identify groups of genes that show similar expression patterns across different samples. This approach assumes that co-regulated genes tend to have similar expression profiles. Various methods can be used for co-expression analysis, such as correlation-based methods (e.g., Pearson correlation coefficient) or more advanced approaches like weighted gene co-expression network analysis (WGCNA).

Network inference: Once co-expressed gene modules or clusters are identified with or without the differentially expressed (DE) genes, the next step is to infer regulatory relationships between genes to reconstruct a GRN. Several computational methods can be used for this purpose. Some of these are enlisted below.

Correlation-based methods- These methods infer regulatory relationships based on the correlation between gene expression profiles. For example, if the expression of gene A is highly correlated with gene B across samples, it suggests that gene A may regulate gene B or vice versa.

Bayesian networks- Bayesian network algorithms use probabilistic graphical models to infer regulatory relationships. These models consider dependencies between genes and estimate the probability of a gene being regulated by another gene based on the expression data.

Information theory-based methods- These methods measure the mutual information or entropy between genes to infer regulatory relationships. They quantify the amount of information shared between gene pairs and identify pairs with high information content, indicating potential regulatory interactions.

Machine learning approaches- Various machine learning algorithms, such as support vector machines (SVM), random forests, or neural networks, can be trained on gene expression data to predict regulatory relationships. These methods typically require a labelled training dataset with known regulatory interactions.

Network validation and refinement: Once the initial GRN is inferred, it is crucial to validate and refine the network. Experimental validation methods, such as ChIP-seq, DNase-seq, or reporter assays, can provide additional evidence for the predicted regulatory interactions. Integration of varied types of omics data, such as transcription factor binding data or protein-protein interaction data, can also help refine the GRN and improve its accuracy.

In the above forms of reconstruction of gene regulation networks from gene expression data, various forms of models prominent in literature primarily include topological design perspectives, Bayesian network designs, Boolean network approaches, and Hidden Markov and other stochastic models.

1.6.1 Topological Design: One of the topological methodologies that attracted the scientific community is on identification of the differences among the affected regions of a progressive neurogenerative disorder [**29**] involving variation in the transcriptome of many genes. In this perspective, the differential topology [**4,8**] of gene co-expression networks helped in apprehending the associations among the regions affected through progressive nature and the severity of the disease. The interactive components adopted in this view incorporate several node connectivity measures [**30-33**] that help in providing certain empirical evidence important in the prediction of the biological significance of a gene. Concepts related to weighted and un-weighted topological overlaps got justified [**10,34**] in search of biologically enriched differentially connective gene networks.

1.6.2 Bayesian Network Design: The Bayesian viewpoint [**35**] in the reconstruction process involves learning the structure and parameters of a Bayesian network from microarray data. The structure learning step aims to identify the dependencies and interactions between genes, while the parameter learning step estimates the conditional probabilities associated with these dependencies. A Bayesian network [**21,36**] is a graphical model that represents joint multivariate probability distributions and captures the conditional independence between variables. These models are useful for describing complex stochastic processes and provide a clear methodology for learning from noisy observations. The benefits of this design had made this a popular and powerful mode of reconstructing in silico signalling pathways. The Bayesian network structural learning [**37-39**] yields the provision to generate biologically constrained hierarchical gene regulatory pathways or signalling cascades.

1.6.3 Boolean Network Design: Probabilistic Boolean Networks (PBNs) are a modelling framework that combines the rule-based properties of Boolean networks with probabilistic considerations to handle uncertainty [**40,41**]. In PBNs, the state of each node (representing a gene or a variable) is determined by a Boolean function that takes into account the states of its parent nodes. However, unlike traditional Boolean networks where the transitions between states are deterministic, PBNs introduce probabilistic transitions. The dynamics of PBNs can be modelled using Markov chains [**42,43**] which represent a sequence of states where the probability of transitioning from one state to another depends only on the current state. By analyzing the transition probabilities, one can gain insights into the behaviour of PBNs. Standard Boolean networks can be seen as a special case of PBNs where the transition probabilities are either 0 or 1, representing deterministic behaviour.

1.6.4 Comparing Boolean and Bayesian Network Approaches: Bayesian network, mentioned earlier, on the other hand, are graphical models that explicitly represent probabilistic dependencies between variables. These dependencies are typically represented by directed edges between nodes, with each node representing a variable and each edge representing a probabilistic influence. In the context of PBNs, one can obtain the probabilistic dependencies between genes by considering the Boolean functions that determine their states.

Furthermore, within the framework of PBNs the influence of genes on other genes can be analyzed by the transition probabilities and conditional probabilities associated with the Boolean functions of the genes. By examining how changes in the state of one gene affect the probabilities of other genes transitioning to specific states, we can determine the dependencies between genes within the context of PBNs. Hence, PBNs provide a robust approach [**44-47**] to model gene regulatory networks by combining the rule-based properties of Boolean networks with probabilistic considerations. By studying PBNs as Markov chains and relating them to Bayesian networks [**48**], one can analyze the dynamics, probabilistic dependencies, and influence of genes on each other within the framework of PBNs.

1.6.5 Hidden Markov and other Stochastic Designs: In Transcriptional Regulatory Networks (TRNs; can be considered a subset of GRNs), the TF genes have the capability of regulating the differentially expressed target genes, either in individual or collaborative mode [**11,12**]. In this regard, a set of constraints can be defined to relate gene expression patterns to regulatory interaction models making use of Hidden Markov modelling [**11,12,49-52**] via two or more states. Applying such technique large scale

TRNs can be effectively developed for complex organisms introducing novel ideas in transcriptional dynamics and bio-activation. Deciphering the mechanisms of target gene regulation using multiple TF genes in a collaborative mode [**12**] has been proven to be a challenging task in systems biology. In this perspective, an earlier statistical model [**53**] mainly makes use of three key concepts for understanding the gene expression pattern. These are using of principal component analysis to suggest gene patterns, nested models to group genes in a hierarchical organization exhibiting similar expression patterns at different phases of the cell cycle, and compass plot which combines biological information (such as previously characterized cell cycle regulated genes) with statistical analysis to determine the phase [**54**] of the cell cycle for the genes of interest.

The static gene expression data obtained from any living cell mainly consists of independent and identically distributed profiles of gene expression. To figure out the true dynamicity of gene regulatory or transcriptional regulatory networks, time series gene expression profiles with high level of correlation existent between the profiles maintain optimum performance. The real dynamicity gets reflected through the translation time (protein formation time) of the source gene followed by translocation time of the end product (formed protein) binding with target genes at the promoter region, thus regulating the transcription of the concerned targets. In this segment of gene expression analysis, various forms of research that have accrued importance are like differential equation model architectures, dynamic Bayesian architectures, and time delayed gene regulatory architectures highlighting first and higher order target gene regulations.

1.6.6 Differential Equation Model based Architectures: One of the initial works reconstructs regulatory network based on an iterative reverse engineering approach using a minimal linear model [**55**]. This analysis reveals the steady state change in the gene expression obtained from the systematic perturbation of some nodes or genes in a regulatory model. This work highlights on the effect of multiple genetic perturbations increasing the average change in the gene expression levels without inducing secondary compensatory changes and other non-linear effects. Again, a study is there to estimate the parameters of a regulatory network, specifically focusing on the Hes1 system [**56**]. Here, the researchers employ Markov chain Monte Carlo (MCMC) methods [**57**], which are statistical techniques used for sampling from probability distributions, to analyze experimental data. The Hes1 system is modelled using stochastic differential equations

(SDEs), which take into account the inherent randomness and variability observed in biological systems. The researchers use rigorous likelihood-based inference methods to analyze the data and estimate the parameters of the model. This study specially addresses the challenge of sparse data, where the time intervals between observations are large; a common issue in biological experiments where data collection may be limited or infrequent. Further research in this domain enlightens the usage of ordinary differential equations in optimized reconstruction of dynamic gene networks utilizing synthetic time series gene expression data with noise and time delay [**58**]. Hence, this work presents an optimization-based approach for inferring gene regulatory networks, showcasing its effectiveness and applicability through simulations with synthetic data and an experimental case study involving gene expression data from the budding yeast cell cycle. In this differential equation model based approach, one of the significant research works addresses the decomposition of an N-dimensional biological system into N one-dimensional problems [**59**]. This decomposition allows for a more practical and scalable approach to determine candidate network interactions. By applying this algorithm to in silico networks based on known biological GRNs, the researchers were able to successfully predict candidate network topologies that reproduced the dynamics of the original networks. Here, using this algorithm, the computational complexity of the network identification process was shown to increase quadratically (N^2) with the size of the system. However, a parallel implementation of the algorithm achieved nearly linear speedup by utilizing multiple processing cores. This parallelization significantly reduced the computational demand required for reverse engineered GRNs. Reconstruction of time delayed GRNs are crucial because genes do not respond instantaneously to perturbations, and there can be various regulatory mechanisms with different delays involved. In this perspective, an algebraic equation [**60**] based on the observed expression changes had been developed to identify the sub-network structure around the perturbed gene measuring the overall expression changes resulting from the perturbation. This type of experimental setup could provide valuable information about the genes that are directly affected by the perturbation, the subsequent downstream effects on the gene network and to understand the time delay effects inherent in gene regulation. Some recent differential equation based architectures [**61,62**] incorporating the biochemical reaction diffusion aspects of miRNA, mRNA, and proteins have been used to define the stability and/or oscillatory properties of TRNs. These properties are primarily judged via the significant effects observed on the transcriptional and translational delay factors inherent in the dynamics of gene regulation.

1.6.7 Dynamic Bayesian Network Architectures: Bayesian network (BN) methods have high computational complexity and struggle to handle large-scale networks. Information theory-based methods have difficulties in identifying the directions of regulatory interactions and are prone to false positive/negative problems. To overcome these limitations, the approaches [**63-65**] followed primarily utilize network decomposition strategies and false-positive edge elimination schemes. These networks are a type of probabilistic graphical model used to model and analyze dynamic systems that evolve over time. These architectures help to capture the temporal dependencies and causal relationships among genes, allowing researchers to infer regulatory mechanisms and predict gene expression patterns. Among these approaches, the initial one [**63**] was into improving the accuracy and the computational efficiency of Dynamic Bayesian Network (DBN) methods for predicting gene regulatory networks. In this approach, by focusing on genes that show up- or down-regulation before or at the same time as their targets, the search space for potential regulators is reduced, allowing for more efficient analysis. However, in the latter half, one of the approaches [**64**] is focussed on developing new scoring functions based on the Bayesian Information Criterion (BIC) score, which aim to improve the accuracy of inferring GRNs by reducing the number of false positive edges. In this method, a combination of BNs and DBNs in the presence of the novel scoring functions helps to identify the optimal graph structure required to maximize the proposed scores. These scoring functions, when compared to the traditional BIC score, result in networks with fewer spurious edges and higher precision. The other method [**65**] in the latter half is about a framework for learning DBNs utilizing multiple groups of samples and comparing the GRNs between them. This is particularly valuable, as it can provide insights into the differences and similarities in gene regulation across different conditions or cell types instrumental in understanding the underlying biological mechanisms and identifying key regulators. Here, the selection of the optimal model based on cross-validated predictive accuracy is a sound approach, as it helps in preventing overfitting and ensures that the learned models generalize well to unseen data. In this context, the various DBN architectures which have been compared to understand to the structured learning process are G1DBN [**66**], dbnlearn [**67**], dbnR [**68**], ebdbNet [**69**], and bnstruct [**70**] respectively.

1.6.8 Time Delayed Gene Regulatory Architectures: Inferring time-delayed causal gene networks from time-series expression data is a challenging task in bioinformatics and systems biology. Several computational approaches [**14,15,58,60,71-75**] have been developed to address this problem. A general overview of the process is given below for having a better understanding of this specific area of research.

Data Collection: Time-series gene expression data is obtained, which typically involves measuring the expression levels of genes at multiple time points under different conditions or treatments. The data should include information about the expression levels of genes across time.

Pre-processing: The gene expression data need to be cleaned to remove noise, pre-processed to normalize the expression levels, and handle missing values or outliers. This step ensures that the data is suitable for subsequent analysis.

Time-Delay Embedding: The time-series expression data is transformed into a suitable representation for inferring causal relationships. One common approach is to perform time-delay embedding, which constructs a higher-dimensional space by incorporating previous time points as additional features. This captures the temporal dependencies between genes and allows for the detection of time-delayed causal relationships.

Causal Network Inference: Causal network inference algorithms are applied to the time-delay embedded data to identify causal relationships among genes. These algorithms can be broadly categorized into correlation-based methods, information theory-based methods, and machine learning-based methods. Some popular algorithms include Dynamic Bayesian Networks (DBN) [**63-70**], Granger causality [**1,76**], and Causal Entropy [**77**].

Model Selection and Validation: The most appropriate causal network model is selected based on statistical measures, such as goodness-of-fit metrics, likelihood ratios, or cross-validation techniques. Validation of the inferred network is done using experimental evidence or existing biological knowledge to ensure its biological relevance.

Network Analysis: There is a stepping need to analyze and interpret the inferred gene network to gain insights into the underlying biological mechanisms. This may involve identifying hub genes, functional modules, or key regulators within the network.

Several software tools and packages are available for inferring gene networks from time-series expression data, such as GeneNet [**78**], WGCNA [**79**], TIGRESS [**80**], and GENIE3 [**81**]. These tools provide implementations of various network inference algorithms and can assist in the analysis of gene regulatory networks.

Though inferring causal relationships solely based on gene expression data is done with lots of statistical and biological validations, still there remains ample scope to develop the gene regulatory or transcriptional regulatory networks (as the case may be). Integrating other types of data, such as protein-protein interactions, transcription factor binding data, or gene knockout experiments, can enhance the accuracy and reliability of the inferred gene network.

1.7 Open Research Issues

Gene Regulatory Networks explore the idea of biological collaborative mechanisms, which happens to be one of the most challenging issues of computational systems biology because of varied types of information sources (gene expression data, gene-protein interaction, protein-protein interaction, etc.). Predicting the nature and directivity of such collaborative regulations still remains an open source problem because of ever changing environmental conditions affecting the growth and development of living cells. To date various approaches (mentioned in Motivation for the Work) which have been developed to predict the nature of the reconstructed gene regulatory networks (GRNs) or transcriptional regulatory networks (TRNs) look toward the betterment of gene or transcriptional regulatory statistics following the various biological databases and try to find the key factors or therapeutic targets responsible in spreading of a particular disease. At this point, some research issues which need to be addressed further are enlisted below.

1.7.1 Research issue #1: It is known that transcription factor (TF) genes can be held responsible for altering the gene expression pattern of target genes. In this perspective, the target genes which effectively get altered in their gene expression levels to a significant extent under varied experimental conditions with or without the influence of the TF genes can be termed as the differentially expressed (DE) genes. To date, every significant research output focuses on obtaining the DE genes using some statistical measure that solely works on the expression level of the concerned gene. However, their lies scope to understand the differential pattern of a gene considering the physical trait of any organism. As an example, we can consider humans and chimpanzees; two different creatures having lots of dissimilarity in their physical features or traits but possessing the same set of genes at the genome level. However, with the renowned statistical dissimilarity measures, in most of the cases, we can expect to end up in the same set of target DE genes. But incorporating the physical feature information along with the gene expression level we can expect to have different sets of target DE genes, thus contributing to the different types of physical attributes corresponding to the growth and development of the various organs in the respective creatures.

1.7.2 Research issue #2: At the current point in time, a significant proportion of the human race is found to be suffering from various forms of neurodegenerative disorders. In this regard, varied types of therapies are getting proposed depending on the nature of development of the specific disorder or combination of similar types of disorder. From the clinical perspective, it becomes crucial to know the involvement of protein complexes (considering the source being the TF genes) along with the target DE genes responsible in the interactive structure and hence contributing to the beyond control development of any type of neurodegenerative disorder. Apart from the involvement of common TF or DE genes, it is equally important to gather knowledge about the mutually exclusive sets of DE genes that may contribute to different kinds of biomarkers like physiologic, radiographic, or histologic types.

1.7.3 Research issue #3: Understanding the normal and diseased states of a living cell is primarily based on the differential connectivity properties of any GRN or TRN, as the case may be. In this regard, inter and intra module differential co-expression pattern of the concerned TF and DE genes play a significant role. To explore this thought, while maintaining the level of computational burden and time complexity within reasonable limits, some generalized topological overlap measures have widely being used in different research articles. However, in most of the works, there is dearth in selection of a smart threshold that would help in locating the mutually exclusive gene pairs across normal and diseased states of a living cell. The management of this threshold selection happens to be crucial because the outcome in the form differentially connective gene regulatory pairs should be both statistically and biologically significant.

1.7.4 Research issue #4: Computational validation of various signalling pathways associated with the normal functioning or diseased state of a living cell gives us the opportunity to explore the topological importance of various gene-gene or protein-gene or protein-protein interaction structures in any regulatory network. In this regard, the ranking of any gene in the context of differential regulation under normal and diseased states, carry significant importance. Different algorithms proposed so far either takes up a global or a local network biased approach to convey the ranking or the importance of any gene in a particular regulatory cascade. Hence, the adopted ranking measures are unable to give us a clear and uniform ranking of a regulatory gene owing to varied forms of objective functions/measures which may not follow concurrent optimality. In this context, some multi-objective approach may deem suitable in order to understand the true role of any participating gene in a regulatory cascade. This may also make clear the individual or collaborative action of TF genes on target DE genes at every level of the reconstructed cascaded or hierarchical network.

1.7.5 Research issue #5: The individual or collaborative regulatory actions of TF genes on the target DE genes mainly stress on the activator or repressor role of any source TF gene following the activation sequence of any TF or DE gene obtained from the time series gene expression data. In this perspective, there is complete absence of significant research that contributes to time varied transcriptional regulation networks. Though the factor of time delay inherent in any regulation has widely been explored, contemporary research in this regard focuses on TF to DE gene regulation as an activator or repressor following time invariant attitude in a certain state or throughout all states of a living cell. Time dependent regulatory perspective may be existent at a certain state of the cell (provided for instance we look into the cell cycle of any organism guided through different stages of development in a time course experiment) indicating activator and/or repressor action existent in that state. Further the differential role of activator and repressor actions can also be confirmed across the same set of consecutive time points in different stages of the cell cycle. Hence, the matter of regulation can guide us more into the temporal dynamics of the transcriptional regulatory network which can be of prime importance in the design of suitable drugs based on the ever-changing environmental conditions, required for therapeutic targets.

1.7.6 Research issue #6: Following the time varied or temporal transcriptional regulatory network design significant research can be conducted to understand the major time delay components inherent in any regulation. This has the possibility to explore or give us some idea about the time involved in attacking a therapeutic target from a certain protein translated from any source TF gene following a specific statistically and biologically significant regulatory pathway.

1.7.7 Research issue #7: Estimation of hidden causal factors in the context of regulation has been put in force just a few years back. This indicates the existence of any hidden factor (not any physically existent gene entity or module) that may be there along with other genes toward the activation or repression in the transcription process of any target gene. Following a time variant and time delayed design of transcriptional regulatory network different forms of optimization algorithms can be used to predict the expression level of any target gene. In this process, there lies a chance of predicting the hidden factors responsible for prediction of the changes, if any, of the target genes.

1.7.8 Research issue #8: Mostly we refer to the importance of a direct regulatory action in a gene regulatory or transcriptional regulatory network. To stress on the fact, topologically we are more concerned about the significance of any direct regulation, even if there are various computational evidences of indirect regulation via one or more physical gene entities. At this background, the physical importance of a direct or semi-direct regulatory link can be checked in the presence of one or more intermediary regulatory genes and thus helping in determining the true specificity and sensitivity of any transcriptional network design. In this regard, it is an open challenge to carry the statistical estimation of some unknown gene combinations along with the primary regulator gene responsible in the regulation of therapeutic targets.

1.8 Research Objectives

1.8.1 Research Objective #1: Finding the quantitative trait specific differentially expressed (DE) genes. In other words, computation of DE genes considering the differential relation between the gene expression and physical trait factors of any organism. The statistical and biological significance of these DE genes can be checked making use Gene Ontology tools and KEGG pathway analysis.

1.8.2 Research Objective #2: Designing simple transcriptional regulatory networks involving a pair of transcription factor (TF) genes and a target DE gene. Through this approach, an initial framework depicting the power of collaborative TF regulations corresponding to a target DE gene can be checked making use linear and non-linear interactive measures.

1.8.3 Research Objective #3: The differential regulation of any gene or transcriptional regulatory network under varied conditions have primarily been checked across normal and diseased tissues using a widely applicable generalized topological overlap measure. In this regard, disjoint regulations highlight on the specific nature of the disease or the normal state of any living cell. To find these disjoint sets of regulations the selection a proper threshold demands utmost importance. Hence, a smart threshold selection for the network is required and accordingly the performance of the gene participation can be verified making use gene ontology and KEGG pathway analysis.

1.8.4 Research Objective #4: Understanding the effective role of differential regulation in complex networks under varied experimental conditions, can be thoroughly verified exploring through some studies on the nature of growth of different kinds of neurodegenerative disorders such as Alzheimer's disease (AD), Amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Multiple sclerosis (MS), Schizophrenia (SCZ), and Parkinson's disease (PD). In this regard, it is even more important to discover the extreme differential regulations between dissimilar types of disease pairs, where a particular disease pair is determined based on the parity of gene expression levels.

1.8.5 Research Objective #5: Multiobjective ranked selection of differentially expressed (DE) or transcription factor (TF) genes across different optimal fronts applying the concept of ranking the DE or the TF genes based on certain conflicting objectives. The optimal decisions in this regard are taken in the presence of trade-offs between two or more conflicting objectives. This step is vital to understand the true power of a TF or DE gene in the development of TF to DE gene regulatory networks in a signalling cascade.

1.8.6 Research Objective #6: Time delayed transcriptional or gene regulatory networks are inherent issues present in the process of formation of any biological network. Hence, in this regard, analyzing the functional effect (activation or repression) of regulation can be strengthened understanding the logical view (sufficient or necessary) of the same. Appreciating both the above ends, time varied transcriptional regulations (individual or collaborative) at any condition can be developed obtaining differential transcriptional regulations across conditions for the same set of time points in the different conditions of interest. The entire time regulation can be made more meaningful observing the periodicity of the transcription factor (TF) and differentially expressed (DE) gene expression levels.

***1.8.7 Research Objective** #7*: The existence of unknown transcriptional regulators inevitable in complex regulatory networks can be ascertained by comparing the statistical significance of known direct transcriptional regulatory links of differential type with the same in the indirect presence other transcription factor genes through a fused least absolute shrinkage and selection operator with a topological overlap measure as the interaction structure. This approach can throw new light in gene regulation statistics, thus helping us to develop further time critical collaborative or individual transcription factor to differentially expressed gene regulation.

1.9 Research Scope

This work embodies the development of transcription factor gene regulatory networks for differentially expressed genes via certain statistically significant methodologies which include the following as per the organization of the thesis chapters, mentioned below.

Chapter 2: In this chapter, at the initial level, the significance of finding differentially expressed (DE) genes considering the physical traits of an organism helps us to define the evolutionary process. This also contributes to have a better understanding of the DE genes present across different stages of growth of any diseased cell. This is followed by development of paired transcription factor (TF) regulatory networks for DE genes obtained via normal statistical techniques and the technique employing the physical characteristics using linear correlative and non-linear interactive measures such as mutual information, polynomial regression, and spline regression. In the above context, finding individual differential regulatory links, explicitly indicating the presence or absence of a gene link under different conditions is unveiled using a smart threshold selection of statistically enriched topological overlap score metric.

Chapter 3: In this chapter, the application of a topological overlap score on the differentially expressed (DE) genes is initially checked in the context of Alzheimer’s disease that progresses through incipient, moderate, and severe stages. As the disease moves from mild or moderate to severe stage, the developed topological overlap metric show some form (weighted and un-weighted counterparts) of disjoint regulation with respect to each DE gene and thus segregating the mild from the severe stage. The significant participation of the DE genes having dissimilar regulation is also checked via KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathway analysis revealing the importance of the DE genes under consideration. This work led to thought of further understanding different types of neurogenerative disorders, their similarity level and extent of dissimilarity that may be present. In this context, a probabilistic differential regulatory score metric is developed that helps to find the extreme level of topologically disjoint regulations prevalent in between similar neurodegenerative disease pairs.

Chapter 4: The differentially expressed (DE) genes participating in various signalling cascades can be classified based on their nature of differential regulation scores. However, maintaining the individual regulation score as an objective function, multiple such conflicting objectives are considered for development of network paths consisting of DE genes placed in different optimal fronts. In each such front, the DE genes present are non-dominant to each other based on the considered objectives. This idea helps us to extend the concept to build up composite entropy minimized transcription factor (TF) regulatory networks following a hierarchical strategy in the process of development. In the latter case, the network is built up using TF genes of varied types (pure TFs, TF genes that are differentially expressed, i.e. TF-DE, and TF genes that are differentially co-expressed, i.e. TF-DC) across different optimal fronts with the presence of the DE genes at the last front, indicating the final target gene in the regulatory cascade.

Chapter 5: Using temporal gene expression profiles (time-series information), composite mode differential transcriptional regulatory networks for differentially expressed (DE) genes are developed incorporating the inherent presence of time varied regulatory effect (functional and logical roles of regulation are considered) in a certain condition and the periodicity property of any transcription factor (TF) or differentially expressed (DE) gene expression profile. The performance has been checked over individual as well as collaborative TF to DE gene regulatory networks showing significant statistical improvement in recognition of biological regulations along with adding a factor related to the type of composite mode differential regulation that may be present in a temporal perspective.

Chapter 6: Transcriptional factor (TF) regulatory networks may be of individual or collaborative interaction nature. Here, it has been verified through topologically overlapped fused least absolute shrinkage and selection operator approach that, there is a high chance of the inevitable presence of unknown TF regulatory modules working along with the primary TF gene in regulating a target differentially expressed (DE) gene. In other words, the direct interaction structure depicted in biological databases between a TF and target DE gene may involve additional unknown TF gene regulators working in tandem with the primary TF gene mentioned earlier.

***Chapter** 7*: This is the concluding portion of this thesis highlighting the crucial thoughts and benefits from the experimental outcomes discussed in the various chapters, starting from chapter no. 2 to chapter no. 6. In other words, this portion states the way the research has been conducted to understand the process of development of transcription factor (TF) network architectures for differentially expressed (DE) genes. In addition to the above, this chapter also points out some significant areas in the TF network architectural domain where further contribution can benefit therapeutic drug modelling on diseased tissues.

1.10 References

[**1**] A. Fujita, P. Severino, K. Kojima, J.R. Sato, A.G. Patriota, and S. Miyano, “Functional clustering of time series gene expression data by Granger causality”, BMC Systems Biology, 6, Article No.137, October 2012, https://doi.org/10.1186/1752-0509-6-137

[**2**] J. Qin, Y. Hu, F. Xu, H.K. Yalamanchili, and J. Wang, “Inferring gene regulatory networks by integrating ChIP-Seq/chip and transcriptome data via LASSO type regularization methods”, Methods, 67(3), 294-303, June 2014, https://doi.org/10.1016/j.ymeth.2014.03.006

[**3**] D. Guan, J. Shao, Z. Zhao, *et al.*, “PTHGRN: unraveling posttranslational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data”, Nucleic Acids Research, 42 (W1), W130-W136, July 2014, https://doi.org/10.1093/nar/gku471

[**4**] B.M. Tesson, R. Breitling, and R.C. Jansen, “DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules”, BMC Bioinformatics, 11, Article No.497, October 2010, https://doi.org/10.1186/1471-2105-11-497

[**5**] M. Watson, “CoXpress: differential co-expression in gene expression data”, BMC Bioinformatics, 7, Article No.509, November 2006, https://doi.org/10.1186/1471-2105-7-509

[**6**] J.K. Choi, U. Yu, O.J. Yoo, and S. Kim, “Differential coexpression analysis using microarray data and its application to human cancer”, Bioinformatics, 21(24), 4348-4355, December 2005, https://doi.org/10.1093/bioinformatics/bti722

[**7**] S.B. Cho, J. Kim, and J.H. Kim, "Identifying set-wise differential co-expression in gene expression microarray data", BMC Bioinformatics, 10, Article No.109, April 2009, https://doi.org/10.1186/1471-2105-10-109

[**8**] B. Zhang, H. Li, R.B. Riggins, *et al.*, "Differential dependency network analysis to identify condition-specific topological changes in biological networks", Bioinformatics, 25(4), 526-532, February 2009, https://doi.org/10.1093/bioinformatics/btn660

[**9**] A. de la Fuente, "From 'differential expression' to 'differential networking'- identification of dysfunctional regulatory networks in diseases", Trends in Genetics, 26(7), 326-333, July 2010, https://doi.org/10.1016/j.tig.2010.05.001

[**10**] G. Altay *et al.*, "Differential C3NET reveals disease networks of direct physical interactions", BMC Bioinformatics, 12, Article No.296, July 2011, https://doi.org/10.1186/1471-2105-12-296

[**11**] S. Awad, N. Panchy, S-K. Ng, and J. Chen, "Inferring the regulatory interaction types of transcription factors in transcriptional regulatory networks", Journal of Bioinformatics and Computational Biology, 10(5), Article No.1250012, October 2012, https://doi.org/10.1142/s0219720012500126

[**12**] S. Awad and J. Chen, "Inferring transcription factor collaborations in gene regulatory networks", BMC Systems Biology, 8, Article No.S1, January 2014, https://doi.org/10.1186/1752-0509-8-S1-S1

[**13**] A. Li and S. Horvarth, "Network neighborhood analysis with the multi-node topological overlap measure", Bioinformatics, 23(2), 222-231, January 2007, https://doi.org/10.1093/bioinformatics/btl581

[**14**] L-Y. Lo, M-L. Wong, K-H. Lee, and K-S. Leung, "Time Delayed Causal Gene Regulatory Network Inference with Hidden Common Causes", 10(9):e0138596, September 2015, https://doi.org/10.1371/journal.pone.0138596

[**15**] M.M. Kordmahalleh, M.G. Sefidmazgi, S.H. Harrison, *et al.*, "Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network", BioData Mining, 10, Article No.29, August 2017, https://doi.org/10.1186/s13040-017-0146-4

[**16**] L. Jia, W. Zhang, and B. Yang, "Identification of Nonlinear System Based on Complex-Valued Flexible Neural Network", In H. Yin et al. (Eds.): IDEAL 2017, LNCS 10585, 154–162, October 2017, http://dx.doi.org/10.1007/978-3-319-68935-7_18

[**17**] B. Yang, Y. Chen, W. Zhang, J. Lv, W. Bao, and D-S. Huang, "HSCVFNT: Inference of Time-Delayed Gene Regulatory Network Based on Complex-Valued Flexible Neural Tree Model", International Journal of Molecular Sciences, 19(10), Article No.3178, October 2018, https://doi.org/10.3390/ijms19103178

[**18**] M.T.J. Johnson, M. Vellend, and J.R. Stinchcombe, "Evolution in plant populations as a driver of ecological changes in arthropod communities", Philos Trans R Soc Lond B Biol Sci., 364(1523), 1593-1605, June 2009, https://doi.org/10.1098/rstb.2008.0334

[**19**] E. Kuhn, "From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in

plant”, Annals of Botany, 87(2), 139-155, Feb 2001, https://doi.org/10.1006/anbo.2000.1314

[**20**] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed, “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”, Statistica Sinica, 12,111-39, 2002

[**21**] R. Gottardo, A.E. Raftery, K.Y. Yeung, and R.E. Bumgarner, “Bayesian robust inference for differential gene expression in microarrays with multiple samples”, Biometrics, 62(1), 2006, https://doi.org/10.1111/j.1541-0420.2005.00397.x

[**22**] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, “DEGseq: an R package for identifying differentially expressed genes from RNA- seq data”, Bioinformatics, 26(1),136-138, January 2010, https://doi.org/10.1093/bioinformatics/btp612

[**23**] C. Cheng, Y. Fu, L. Shen, and M. Garstein, "Identification of yeast cell cycle regulated genes based on genomic features", BMC Systems Biology, 7, Article number: 70, July 2013, https://doi.org/10.1186/1752-0509-7-70

[**24**] N. Banerjee and M.Q. Zhang, "Identifying cooperativity among transcription factors controlling the cell cycle in yeast ", Nucleic Acids Research, 3l(23), 7024-7031, December 2003, https://doi.org/10.1093/nar/gkg894

[**25**] A.T. Kwon, H.H. Hoos, and R. Ng, “Inference of transcriptional regulation relationships from gene expression data”, Bioinformatics, 19(8), 905-912, May 2003, https://doi.org/10.1093/bioinformatics/btg106

[**26**] J.M.P. Desterro, M.S. Rodriguez, and R.T. Hay, “Regulation of transcription factors by protein degradation,” Cellular and Molecular Life Sciences, 57(8-9), 1207-1219, August 2000, https://doi.org/10.1007/PL00000760

[**27**] Y. Lai, B. Wu, L. Chen, and H. Zhao, “A statistical method for identifying differential gene-gene co-expression patterns”, Bioinformatics, 20(17), 3146–3155, November 2004, https://doi.org/10.1093/bioinformatics/bth379

[**28**] M. Bockmayr, F. Klauschen, B. Györffy, C. Denkert, and J. Budczies, “New network topology approaches reveal differential correlation patterns in breast cancer”, BMC Systems Biology,7, Article No.78, August 2013, https://doi.org/10.1186/1752-0509-7-78

[**29**] M. Ray and W. Zhang, “Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks”, BMC Systems Biology, 4, Article No.136, October 2010, https://doi.org/10.1186/1752-0509-4-136

[**30**] B. Zhang and S. Horvath, “A general framework for weighted gene co-expression network analysis”, Statistical Applications in Genetics and Molecular Biology, 4, Article No.17, August 2005, https://doi.org/10.2202/1544-6115.1128

[**31**] H. Yu, B-H. Liu, Z-Q. Ye, C. Li, Y-X. Li, and Y-Y. Li, “Link-based quantitative methods to identify differentially coexpressed genes and gene Pairs”, BMC Bioinformatics, 12, Article No.315, August 2011, https://doi.org/10.1186/1471-2105-12-315

[**32**] A. Reverter and E.K.F. Chan, "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks", Bioinformatics, 24(21), 2491-2497, November 2008, https://doi.org/10.1093/bioinformatics/btn482

[**33**] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera, and A. Califano, "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context", BMC Bioinformatics, 7, Article No.S7, March 2006, https://doi.org/10.1186/1471-2105-7-S1-S7

[**34**] M. Sarkar and A. Majumder, "TOP: An Algorithm in Search of Biologically Enriched Differentially Connective Gene Networks", In Proceedings of 5th Annual International Conference on Advances in Biotechnology (BIOTECH 2015), 124-133, March 2015, https://doi.org/10.5176/2251-2489_BIOTECH15.39

[**35**] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data", Journal of Computational Biology, 7(3-4), 601-620, July 2004, https://doi.org/10.1089/106652700750050961

[**36**] D. Heckerman, "A tutorial on learning with Bayesian networks", original version published in Learning in Graphical Models, M. Jordan, ed., MIT Press, Cambridge, MA, 1999, current revised version available at https://doi.org/10.48550/arXiv.2002.00269

[**37**] D. Zhu and H. Li, "Improved Bayesian Network inference using relaxed gene ordering", International Journal of Data Mining and Bioinformatics, 4(1), 44-59, January 2010, https://doi.org/10.1504/IJDMB.2010.030966

[**38**] F. Liu, S-W. Zhang, W-F. Guo, Z-G. Wei, and L. Chen, "Inference of Gene Regulatory Network Based on Local Bayesian Networks", PloS Computational Biology, 12(8):e1005024, August 2016, https://doi.org/10.1371/journal.pcbi.1005024

[**39**] L. Qu, Z. Wang, Y. Huo, Y. Zhou, J. Xin, and W. Qian, "Distributed Local Bayesian Network for Gene Regulatory Network Reconstruction", In IEEE Proceedings from 2020 6th International Conference on Big Data Computing and Communications (BIGCOM), 131-139, August 2020, https://doi.org/10.1109/BigCom51056.2020.00026

[**40**] I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks", Bioinformatics, 18(2), 261-274, February 2002, https://doi.org/10.1093/bioinformatics/18.2.261

[**41**] G. Karlebach and R. Shamir, "Constructing Logical Models of Gene Regulatory Networks by Integrating Transcription Factor–DNA Interactions with Expression Data: An Entropy-Based Approach", Journal of Computational Biology, 19(1), 30-41, January 2012, https://doi.org/10.1089/cmb.2011.0100

[**42**] M. Lluberes, J. Seguel, and J. Ramirez-Vick, "Markov Model Checking of Probabilistic Boolean Networks Representations of Genes", In Proceedings of International Conference on Bioinformatics & Computational Biology, Las Vegas, NV, USA, July 2011, Available online at: http://www.ece.uprm.edu/~domingo/teaching/ciic8996/BIC3596.pdf

[**43**] P. Trairatphisan, A. Mizera, J. Pang *et al.*, “Recent development and biomedical applications of probabilistic Boolean networks”, Cell Communication and Signaling, 11, Article No.46, July 2013, https://doi.org/10.1186/1478-811X-11-46

[**44**] J. Liang and J. Han, “Stochastic Boolean networks: An efficient approach to modeling gene regulatory networks”, BMC Systems Biology, 6, Article No.113, August 2012, https://doi.org/10.1186/1752-0509-6-113

[**45**] L. Chen, D. Kulasiri, and S. Samarasinghe, “A Novel Data-Driven Boolean Model for Genetic Regulatory Networks”, Frontiers in Physiology, Section- Systems Biology Archive, 9-2018, September 2018, https://doi.org/10.3389/fphys.2018.01328

[**46**] Z. Pusnik, M. Mraz, N. Zimic, and M. Moskon, “Review and assessment of Boolean approaches for inference of gene regulatory networks”, Heliyon, 8(8):e10222, August 2022, https://doi.org/10.1016/j.heliyon.2022.e10222

[**47**] J. Pan, J-e Feng, and M. Meng, “Steady-state analysis of probabilistic Boolean networks”, Journal of the Franklin Institute, 356(5), 2994-3009, March 2019, https://doi.org/10.1016/j.jfranklin.2019.01.039

[**48**] Y. Xiao, “A Tutorial on Analysis and Simulation of Boolean Gene Regulatory Network Models”, Current Genomics, 10(7), 511-525, November 2009, http://dx.doi.org/10.2174/138920209789208237

[**49**] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchinson, “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids”, Cambridge, United Kingdom: Cambridge Univ. Press, vol. 1, 1998.

[**50**] R-R Ji, D. Liu, and W. Zhang, “The application of hidden markov model in building genetic regulatory network”, Journal of Biomedical Science and Engineering, 3(6), 633-637, January 2010, http://dx.doi.org/10.4236/jbise.2010.36086

[**51**] B.K. Chu, M.J. Tse, R.R. Sato, and E.L. Read, “Markov State Models of gene regulatory networks”, BMC Systems Biology, 11, Article No.14, February 2017, https://doi.org/10.1186/s12918-017-0394-4

[**52**] S. Zhu and Y. Wang, “Hidden Markov induced Dynamic Bayesian Network for recovering time evolving gene regulatory networks”, Scientific Reports, 5, Article No.17841, December 2015, https://doi.org/10.1038/srep17841

[**53**] K-C Li, M. Yan, and S. Yuan, “A simple statistical model for depicting the cdc15-synchronized yeast cell-cycle regulated gene expression data”, Statistica Sinica, 12(1), 141-158, a special issue on Bioinformatics January 2002, https://www.jstor.org/stable/24307039

[**54**] G. Goelman, R. Dan, *et al.*, “Frequency-phase analysis of resting-state functional MRI”, Scientific Reports, 7, Article No.43743, March 2017, https://doi.org/10.1038/srep43743

[**55**] J. Tegner, M.K. Stephen Yeung, J. Hasty, and J.J. Collins, “Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling”, Proc. Nat. Acad. Sci. U S A, 100(10), 5944-5949, May 2003, https://doi.org/10.1073/pnas.0933416100

[**56**] E.A. Heron, B. Finkenstadt, and D.A. Rand, "Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study", Bioinformatics, 23(19), 2596–2603, October 2007, https://doi.org/10.1093/bioinformatics/btm367

[**57**] B. Eraker, "MCMC Analysis of Diffusion Models with Application to Finance", Journal of Business & Economic Statistics, 19(2), 177-191, April 2001, https://www.jstor.org/stable/1392162

[**58**] S. Kim, J. Kim, and K-H Cho, "Inferring gene regulatory networks from temporal expression profiles under time-delay and noise," Computational Biology and Chemistry, 31(4), 239-245, August 2007, https://doi.org/10.1016/j.compbiolchem.2007.03.013

[**59**] J.N. Bazil, F. Qi, and D.A. Beard, "A parallel algorithm for reverse engineering of biological networks," Integrative Biology, 3(12), 1215–1223, December 2011, https://doi.org/10.1039/c1ib00117e

[**60**] J-R Kim, S-M Choo, H-S Choi, and K-H Cho, "Identification of Gene Networks with Time Delayed Regulation Based on Temporal Expression Profiles", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(5), 1161-1168, September-October 2015, https://doi.org/10.1109/TCBB.2015.2394312

[**61**] Y. Zhang, H. Liu, F. Yan, and J. Zhou, "Oscillatory behaviors in genetic regulatory networks mediated by MicroRNA with time delays and reaction-diffusion terms", IEEE Transactions on Nanobioscience, 16(3), 166-176, April 2017, https://doi.org/10.1109/TNB.2017.2675446

[**62**] T. Dong and Q. Zhang, "Stability and oscillation analysis of a gene regulatory network with multiple time delays and diffusion Rate", IEEE Transactions on Nanobioscience, 19(2), 285-298, April 2020, https://doi.org/10.1109/TNB.2020.2964900

[**63**] M. Zou and S.D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data", Bioinformatics, 21(1), 71-79, January 2005, https://doi.org/10.1093/bioinformatics/bth463

[**64**] H.B. Ajmal and M.G. Madden "Dynamic Bayesian Network learning to infer sparse models from time series gene expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(5), 2794-2805, September-October 2022, https://doi.ieeecomputersociety.org/10.1109/TCBB.2021.3092879

[**65**] P. Suter, J. Kuipers, and N. Beerenwinkel, "Discovering gene regulatory networks of multiple phenotypic groups using dynamic Bayesian networks", Briefings in Bioinformatics, 23(4):bbac219, July 2022, https://doi.org/10.1093/bib/bbac219

[**66**] S. Lèbre, "Inferring Dynamic Genetic Networks with Low Order Independencies", Statistical Applications in Genetics and Molecular Biology, 8(1), 1-38, February 2009, https://doi.org/10.2202/1544-6115.1294

[**67**] R. Fernandes, "dbnlearn: Dynamic Bayesian Network Structure Learning, Parameter Learning and Forecasting", https://cran.r-project.org/web/packages/dbnlearn/index.html, 2020

[**68**] D. Quesada, “dbnR: Dynamic Bayesian Network Learning and Inference”, https://cran.r-project.org/web/packages/dbnR/index.html, version 0.7.5, 2022

[**69**] A. Rau, “ebdbNet: Empirical Bayes Estimation of Dynamic Bayesian Networks”, https://cran.r-project.org/web/packages/ebdbNet/index.html, version 1.2.6, 2022

[**70**] A.F.F. Sambo, “bnstruct: Bayesian Network Structure Learning from Data with Missing Values”, https://cran.r-project.org/web/packages/bnstruct/index.html, 2022

[**71**] L-Y Lo, K-S Leung, and K-H Lee, “Inferring time-delayed causal gene network using time-series expression data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(5), 1169-1182, September-October 2015, https://doi.org/10.1109/tcbb.2015.2394442

[**72**] H. Chen, P.A. Mundra, L.N. Zhao, F. Lin, and J. Zheng, “Highly sensitive inference of time-delayed gene regulation by network deconvolution”, BMC Systems Biology, 8, Article No.S6, https://doi.org/10.1186/1752-0509-8-S4-S6

[**73**] X. Zhang, L. Wu, and S. Cui, “An improved integral inequality to stability analysis of genetic regulatory networks with interval time-varying delays”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(2), 398-409, March-April 2015, https://doi.org/10.1109/TCBB.2014.2351815

[**74**] X. Fan, X. Zhang, L. Wu, and M. Shi, “Finite-Time stability analysis of reaction-diffusion genetic regulatory networks with time-varying delays”, 14(4), 868-879, July-August 2017, https://doi.org/10.1109/tcbb.2016.2552519

[**75**] T. Yu, J. Liu, Y. Zeng, X. Zhang, Q. Zheng, and L. Wu, “Stability analysis of genetic regulatory networks with switching parameters and time delays”, IEEE Transactions on Neural Networks and Learning Systems, 29(7), 3047-3058, July 2018, https://doi.org/10.1109/TNNLS.2016.2636185

[**76**] J.D. Finkle, J.J. Wu, and N. Bagheri, “Windowed Granger causal inference strategy improves discovery of gene regulatory networks”, PNAS, 115(9), 2252-2257, February 2018, https://doi.org/10.1073/pnas.1710936115

[**77**] H. Feng, R. Zheng, J. Wang, F-X Wu, and M. Li, “NIMCE: A Gene Regulatory Network Inference Approach Based on Multi Time Delays Causal Entropy”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(2), 1042-1049, March-April 2022, https://doi.org/10.1109/tcbb.2020.3029846

[**78**] K. Strimmer, “GeneNet: Modeling and Inferring Gene Networks”, https://cran.r-project.org/web/packages/GeneNet/index.html, 2021

[**79**] P. Langfelder, “WGCNA: Weighted Correlation Network Analysis”, https://cran.r-project.org/web/packages/WGCNA/index.html, 2023

[**80**] A-C Haury, F. Mordelet, P. Vera-Lincona, and J-P Vert, “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection”, BMC Systems Biology, 6, Article No.145, November 2012, https://doi.org/10.1186/1752-0509-6-145

[**81**] V.A. Huynh-Thu, “GENIE3: GEne Network Inference with Ensemble of trees”, Bioconductor version: Release (3.17), https://doi.org/doi:10.18129/B9.bioc.GENIE3

Chapter 2 Nature of Differentially Expressed Gene and Transcription Factor Regulatory Networks

Related Publications

[**1**] M. Sarkar and A. Majumder, “Quantitative Trait Specific Differential Expression (qtDE)”, Procedia Computer Science, volume 46, pages 706-718, April 2015.
https://doi.org/10.1016/j.procs.2015.02.131
[**2**] A. Majumder and M. Sarkar, “Simple transcriptional networks for differentially expressed genes”, In IEEE International Conference on Signal Propagation and Computer Technology, ICSPCT 2014, 12-13 July 2014, INSPEC Accession Number:14544879.
https://doi.org/10.1109/ICSPCT.2014.6885016
[**3**] A. Majumder and M. Sarkar, “Paired Transcriptional Regulatory System for Differentially Expressed Genes”, Lecture Notes on Information Theory, volume 2, issue 3, pages 266-272, September 2014.
doi: 10.12720/lnit.2.3.266-272
[**4**] M. Sarkar and A. Majumder, “Intelligent Topological Differential Gene Networks”, In: Das, S., Pal, T., Kar, S., Satapathy, S., Mandal, J. (eds) Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015. Advances in Intelligent Systems and Computing, volume 404. Springer, New Delhi.
https://doi.org/10.1007/978-81-322-2695-6_8

2.1 Introduction

The initial focus of this chapter is to introduce the concept of quantitative trait specific differentially expressed (DE) genes. DE genes can be obtained from the gene expression level (static or time series data) of any organism considering varied conditions of experimentation or the genotypic variations present across the phenotypes of the concerned subject. In this perspective, different forms of algorithms which have gained importance in understanding DE genes are [**1-4**]. However, in this chapter, the concern is about finding DE genes checking the differentially co-expressed [**5**] dependence between the gene expression level and some physical trait of an organism.

At the next stage, the matter of investigation lie on the reconstruction of paired transcription factor (TF) regulatory networks for the DE genes. A particularly challenging demand in this regard is to identify the gene regulatory network in a given biological system [**6**]. Here, through the involvement of different kinds of TF genes it is possible to integrate various forms of signalling pathways in the nucleus of a cell [**7**]. In a eukaryotic cell, there is the presence of multiple types of TF genes acting as regulators for different kinds of target genes. Here, the regulation of the target genes participating in the process of cell cycle division demands proper attention of the regulating TF genes because a number of such target genes show a possibility of transcription even before it is required [**8**]. Again the DE genes are physical entities that differ in the transcription level across conditions/ phases of the cell cycle mentioned above. Hence, it is important to study the basic nature of TF regulatory networks regulating target DE genes in such a way that maintains significant control over the level of differential expression of the target DE genes. The basic nature of the TF regulatory networks can be explored through reconstruction of paired TF networks for target DE genes making use of linear interaction or regulation measure like Pearson correlation [**9**], and non-linear interaction measures like mutual information [**9-15**], polynomial regression [**16**], and spline regression [**17**].

Though the involvement of significant regulators can be understood from the basic network architecture mentioned above, the crucial differences in regulatory architecture across conditions or states of the cell cycle, as the case may be, are dependent on the selection of a proper threshold level in a topological overlap measure. This demands attention because the differential functioning of a cell in varied conditions can be based on disjoint sets of inherent regulations. In this regard, it is worth mentioning that the study of differential co-expression yields a better understanding compared to differential expression analysis [**18**]. In diseased cells, having abnormal regulations the expression of the gene modules will be unordered or random. Gene modules showing this kind of pattern can be detected primarily by the differential co-expression approach. Hence, amalgamating the differential co-expression or connectivity along with differential expression can be used to develop the topological overlap measure [**19**] required for defining a smart threshold.

2.2 Finding Differentially Expressed Genes using Quantitative Traits

Here, the concept of differential dependencies between quantitative trait and gene expression profiles is explored to stress on natural selection based evolutionary growth driving ecological changes across phenotypes [**20**]. Existence of algorithms [**21**] prove the presence of cluster based segregation of genes via physical or quantitative traits through linear correlation strategies.

2.2.1 The basic findings: In this study, the focus is on identifying the differentially expressed (DE) genes based on traits using two types of significance measures: linear correlation and non-linear mutual information and polynomial regression. The purpose is to find DE genes specific to traits, referred to as qtDE (quantitative trait-specific differential expression).

To determine DE genes in the entire dataset, a statistical significance test is performed, specifically the Student T-test. Additionally, a well-known DE gene analysis tool called DEGseq [**4**] is utilized to compare the results obtained by qtDE with those from DEGseq. The results of the study demonstrates that the qtDE method to be promising compared to DEGseq, not only in terms of the number of DE genes identified (qtDE identified more genes than DEGseq) but also in the biological enrichment of the additional DE genes found by qtDE, as determined by KEGG pathway analysis. Furthermore, the number of DE genes involved in biologically enriched pathways, as well as the number of significant pathways themselves, is significantly higher when using the mutual information and polynomial regression-based trait-specific measures compared to the linear correlation-based measure. Overall, the study suggests that the qtDE method using non-linear mutual information and polynomial regression is more effective in identifying DE genes and biologically enriched pathways related to specific traits when compared to DEGseq and linear correlation-based approaches.

2.2.2 Methodology: An algorithm is devised to compute the gene significance values across two phenotypes using three different measures: linear correlation, non-linear mutual information, and non-linear polynomial regression. The algorithm calculates these gene significance values for each gene in order to determine differentially expressed (DE) genes between the two phenotypes.

In the algorithm, the gene expression matrices for phenotype 1 and phenotype 2 are represented by ExV1 and ExV2, respectively. The quantitative trait vectors for phenotype 1 and phenotype 2 are represented by T1 and T2, respectively. The algorithm also uses a soft threshold parameter denoted as β. The steps of the algorithm can be explained as follows:

Step 1: Computing gene significance values

The *LinCor* function is used to calculate the linear gene significance by computing the correlation between the expression profile of each gene and the quantitative trait. The *NLinMI* function is used to calculate the non-linear gene significance by computing the mutual information based uncertainty between the expression profile of each gene and the quantitative trait. The *NLinPR* function is used to calculate the non-linear gene significance by fitting a polynomial regression model between the expression profile of each gene and the quantitative trait. In each case, the obtained gene significance values are stored for phenotype 1 in GS1 matrix and for phenotype 2 in GS2 matrix.

Hence, at this stage, the algorithm calculates these gene significance values separately for both phenotypes.

Step 2: Difference Calculation

In this step, the algorithm computes the difference between the gene significance values obtained in Step 1 across the different phenotypes. This step quantifies the variation in significance between the phenotypes for each gene. The differences are stored in a variable called GS.

Step 3: T-Statistics Probability Distribution

In this step, the algorithm performs a T-statistics probability distribution of the gene significance difference values obtained in Step 2. This distribution allows for the calculation of cumulative distribution function (CDF) values across the entire set of genes. Each gene will have a corresponding CDF value based on its significance difference.

Step 4: Threshold Computation

Here, the algorithm computes the mean of the T-statistics CDF values obtained in Step 3. This mean value is used to determine a threshold level. The threshold level serves as a criterion for determining which genes are considered differentially expressed. It is important to note that the specific method for computing the threshold may vary depending on the algorithm or statistical approach used.

Step 5: Differential Expression Identification

In this final step, the T-statistics CDF (TcdV) values obtained in Step 3 are compared with the threshold value computed in Step 4. If the TcdV value of a particular gene is greater than the threshold, it suggests that there is a significant difference in the gene's significance values between the phenotypes. This indicates that the gene is differentially expressed.

The Algorithm described above is outlined below.

ALGORITHM: Quantitative Trait Specific Differential Gene Significance

```
Input: ExV1, ExV2, T1, T2, GN, β >1
Output: GS1, GS2, GS, TcdV, DE

Step1. s ← choose mode of computation
for i in 1 : N do
    r1 ← rbind (ExV1[i,],T1)
    nr1 ← transpose (r1)
    r2 ← rbind (ExV2[i,],T2)
    nr2 ← transpose (r2)
```

```
  if s = = 1 then
  GS1[i]←LinCor (nr1, β)
  GS2[i]←LinCor (nr2, β)
  else if s = = 2 then
  GS1[i]←NLinMI (nr1, β)
  GS2[i]←NLinMI (nr2, β)
  else
  GS1[i]←NLinPR (nr1, β)
  GS2[i]←NLinPR (nr2, β)
  end if
end for
Step2. GS← GS1 - GS2
Step3. TcdV← qt ((GS), set degree of freedom)
Step4. ThV← mean (TcdV)
Step5. for i in 1 : N do
            if TcdV[i] > ThV then
            s ←s+1
            DE[s]←GN[i]
            end if
        end for
%%%%%%%% End of Main Program %%%%%%%%
LinCor← function (nr, β) ….. Subroutine 1 corresponding to function call LinCor
V← cor (nr, set correlation method) ^ β
NLinMI← function (nr, β) ….. Subroutine 2 corresponding to function call NLinMI
V← mutualInfoAdjacency (nr, discretize columns, set entropy estimation method, set
the number of discretization bins) ^ β
NLinPR← function (nr, β) ….. Subroutine 3 corresponding to function call NLinPR
V← adjacency.polyreg (nr, set the degree of polynomial, specify the method to
symmetrise the pairwise model fitting index matrices) ^ β
```

2.2.3 Results: The dataset used [**22,23**] in this analysis consists of gene expression values for male and female mice across four different types of tissues: brain, muscle, liver, and adipose. This dataset contains a total of 3600 genes in each tissue.

At first, the mice weights have been segregated based on the male and female phenotypes, indicating the existence of separate weight values for male mice and female mice. Next, the mice weights are redistributed among the four tissues for both male and female mice, based on the mice ID and strain. This step involves assigning the weight values to each tissue sample according to the corresponding mouse ID and strain information. Once the data is properly organized, the analysis identifies genes that are differentially expressed across all four tissue types in male and female phenotypes. This is achieved using the algorithm mentioned above setting the value of β to 1.

2.2.3.1 Linear Correlation method: In this analysis, *LinCor* function, given in the algorithm, is followed. This user defined function as mentioned in the algorithm invokes another function *cor* associated with the R package named WGCNA [**24**] to compute the linear gene significance by correlative measure. The function *cor* corresponds to the Pearson correlation operation being performed between each gene of a particular tissue and the redistributed mice weight of that tissue for both phenotypes. In this segment, linear gene significance (GS1 for male and GS2 for female) is hence getting considered. Thereafter, going through the remaining steps of the algorithm the prediction of DE genes is done. In this context, 837,856, 1132 and 579 qtDE genes have been found in liver, adipose, muscle and brain respectively along with 213 common qtDE genes amongst these tissues. So, it can be assumed that with respect to weight, these genes are responsible for the evolution of two different sexes (male and female).

2.2.3.2 Non Linear Mutual Information method: In this segment, a user-defined function called *NLinMI* is used to compute the non-linear gene significance. The *NLinMI* function utilizes the *mutualInfoAdjacency* function from the R package WGCNA [**24**] to compute a symmetric uncertainty-based mutual information adjacency measure. The *mutualInfoAdjacency* function estimates entropy using maximum likelihood estimators with Miller-Madow bias correction. This operation is performed on each gene using the redistributed mice weight pair of the tissue being considered for both phenotypes. The result of this operation provides the non-linear gene significance (GS1 for males and GS2 for females). Continuing with the remaining sequence of operations, there are 1,479 qtDE genes in the liver, 1,236 in adipose tissue, 2,503 in muscle, and 1,499 in the brain. Additionally, there is a common set of 705 qtDE genes among these tissues.

2.2.3.3 Non Linear Polynomial Regression method: Here, the *NLinPR* function, which computes non-linear gene significance through polynomial regression, utilizes the *adjacency.polyReg* operation from the R package WGCNA [**24**]. This calculates a network adjacency matrix by fitting polynomial regression models to pairs of variables. In this case, the operation is applied between each gene of a particular tissue and the redistributed mice weight of the same for both phenotypes. This process results in two non-linear gene significance measures: GS1 for males and GS2 for females. Based on the analysis carried out in this segment, a total of 1395, 938, 1163, and 675 qtDE genes have been identified in the liver, adipose, muscle, and brain tissues respectively. Additionally, there are 364 common qtDE genes among these tissues.

On the over whole, KEGG Pathway analysis [**25**] has shown that non-linear methodologies outperform linear methods in terms of both p-value [**26**] and the number of participating genes. This suggests that non-linear approaches may be more effective in identifying biologically enriched pathways. In this analysis, a significant threshold of a p-value of at least 1E-03 and a minimum of 2 genes for a pathway has been set to be considered significant. This means that pathways meeting these criteria are considered to have a meaningful enrichment of genes and are worth further investigation.

Table 2.1, given below, highlights the significant pathways by the common qtDE genes among brain, muscle, liver, and adipose tissues for the linear (213 common qtDE genes) and non-linear (705 common qtDE genes by mutual information and 364 by polynomial regression) processes.

Table 2.2, given below, shows the significant biological pathway enrichment enlisting those genes which are not only common among all the three methods but also among the four tissues. Through this analysis, 9 common qtDE genes have been found and the significance of a crucial pathway has also been highlighted. From the table, it can be suggested that Ether Lipid Metabolism is the only notable pathway observed among the above mentioned common qtDE genes, but it only involves two such genes, namely pla2g7 and pld2. Additionally, it has also been observed that other KEGG pathways lacking these two genes show better biological enrichment compared to Ether Lipid Metabolism.

Table 2.3, shown below, depicts that the exclusion of these 9 common qtDE genes leads to improved biological enrichment of pathways through non-linear interactions. This suggests that by considering mutual information based differential dependency, it is possible to identify crucial KEGG pathways formed by qtDE genes more effectively compared to polynomial regression and correlative measures. Thus by excluding these 9 common qtDE genes and leveraging mutual information-based differential dependency, the research has demonstrated improved biological enrichment of pathways across different tissues. Hence, this finding highlights the importance of considering non-linear interactions and the potential limitations of linear regression and correlative measures in capturing complex biological processes.

Table 2.4, shown below, enlists the significant KEGG pathways formed by the DE genes excluding the 59 common ones found across the four tissues via DEGseq, the statistical technique used for comparing the effectiveness of the proposed method. It is notable to mention, that through DEGseq 732, 373, 424, 301 DE genes have been observed in adipose, brain, liver and muscle tissues respectively. DEGseq is an R package specifically designed for identifying DE genes using gene expression profiles from RNA-seq data. It does not consider any sample traits or any other additional information. DEGseq uses a statistical approach to compare gene expression levels between different samples or time points. It calculates p-values or q-values (adjusted p-values) to assess the statistical significance of differential expression. By setting a specific threshold for p-value or q-value, DEGseq determines which genes are differentially expressed.

Table 2.5, shown below, depicts the comparative results of KEGG pathway analysis considering the common and mutually exclusive sets of DE genes found from qtDE and DEGseq approaches. In adipose tissue, out of 856 qtDE genes identified by the correlative measure, 584 genes are common with the 732 DE genes found by DEGseq. Similarly, out of 938 qtDE genes identified by the polynomial regression measure, 498 genes are common with the 732 DE genes found by DEGseq. However, all these 732 DE genes are encompassed within the 1236 qtDE genes identified by the mutual information-based measure. In brain tissue, all 373 DE genes discovered by DEGseq are common in the qtDE gene sets identified by the correlative measure (579 genes), mutual information-based measure (1409 genes), and polynomial regression measure (675 genes). In liver tissue, out of 424 DE genes, 379 genes are common with the 837 qtDE genes identified by the correlative measure. Similarly, out of 424 DE genes, 392 genes are common with the 1395 qtDE genes identified by the polynomial regression measure.

However, all 424 DE genes are included in the 1479 qtDE genes identified by the mutual information-based approach. In muscle tissue, out of 1132 qtDE genes identified by the correlative measure, 283 genes are common with the 301 DE genes found by DEGseq. Similarly, out of 301 DE genes, 275 genes are common with the 1163 qtDE genes identified by the polynomial regression measure. In this case as well, all 301 DE genes are within the set of 2503 qtDE genes discovered by the mutual information-based measure. Comparing the results, as depicted in table 5, in some cases, no significant improvement is observed in the p-value after adding the disjoint set of DE genes (mutually exclusive to qtDE and DEGseq) to the common DE gene set found between qtDE and DEGseq. In this regard, there is null refinement in Drug metabolism cytochrome P450 in case of mutual information based differential interaction in brain, Proteasome in case of polynomial regression based differential interaction in muscle and Metabolism of xenobiotic by cytochrome P450 in case of mutual information based differential interaction in brain. Again, in Autoimmune thyroid disease (earned from correlative based differential interaction in adipose) the results rather deteriorate after adding an extra gene from DEGseq. Thus a promising role of qtDE over DEGseq can be claimed in the context of phenotypic variations.

Table 2.1: Significant pathways through linear correlative, nonlinear mutual informative and polynomial regression measure by the common qtDE genes across all tissues

Nonlinear Method						Linear Method		
Mutual Information			Polynomial Regression			Correlation		
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Olfactory transduction	6.6E-08	7	Metabolic pathways	1.8E-04	26	Cell cycle	1.1E-03	6
Leukocyte Transendothelial migration	1.9E-05	14	Nucleotide excision repair	2.3E-04	4	Chronic myeloid leukemia	2.8E-03	3
Complement and coagulation cascades	2.1E-03	8	Glutathione metabolism	4.6E-04	4	Ether Lipid metabolism	6.2E-03	2
Ether Lipid metabolism	3.4E-03	5	Ether Lipid metabolism	8.2E-04	3	--	--	--
Glycerolipid metabolism	4.4E-03	6	DNA replication	9.7E-04	3	--	--	--

Mutual Information			Polynomial Regression			Correlation		
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Glycero phospholipid metabolism	9.6E-03	7	Long Term Depression	1.1E-03	4	--	--	--
--	--	--	Mismatch repair	3.1E-03	2	--	--	--
--	--	--	Vascular smooth muscle contraction	6.2E-03	4	--	--	--

Table 2.2: Significant pathways by the 9 common qtDE genes found between linear and nonlinear methods across all tissues

Pathway	p-value	Gene names
Ether Lipid metabolism	1.6E-03	2(***pla2g7,pld2***)

Table 2.3: Significant pathways across different tissues through correlation, mutual information and polynomial regression based approach excluding the common DE genes in qtDE approach

Correlation			Mutual Information			Polynomial Regression		
ADIPOSE								
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Olfactory transduction	1.31E-06	12	Olfactory transduction	2.22E-11	9	Olfactory transduction	2.27E-17	12
Cell cycle	7.2E-05	24	Leukocyte transendothelial migration	2.82E-03	13	Amoebiasis	9.15E-04	8
Cytokine-cytokine receptor interaction	1.5E-03	2	Leishmaniasis	9.17E-03	8	Cytokine-cytokine receptor interaction	9.16E-04	9
Fc gamma R-mediated phagocytosis	1.8E-03	18	Glutathione metabolism	9.3E-03	7	Steroid biosynthesis	9.16E-04	7
Chagas disease	6.4E-03	17	Cell adhesion molecules	9.7E-03	7	Arginine and proline metabolism	8.8E-03	3
						Nitrogen metabolism	8.8E-03	7
BRAIN								
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Olfactory transduction	2.29E-09	6	Olfactory transduction	2.41E-15	12	Olfactory transduction	1.2E-09	9
Pyruvate metabolism	7.9E-03	4	Amoebiasis	3.36E-05	28	Metabolic pathways	6.3E-05	34
Metabolic pathways	9.6E-03	3	Leishmaniasis	2.73E-05	14	Maturity onset diabetes of the young	3.6E-04	8
Complement and coagulation cascades	9.97E-03	4	Focal adhesion	2.89E-04	49	Amino sugar and nucleotide sugar metabolism	1.2E-03	10
--	--	--	Cytokine-cytokine receptor interaction	3.23E-04	43	Fc gamma R-mediated phagocytosis	4.7E-03	13
--	--	--	--	--	--	Insulin signaling pathway	5.5E-03	8
--	--	--	--	--	--	Focal adhesion	7.9E-03	23

LIVER								
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Olfactory transduction	7.23E-07	8	Leishmaniasis	2.51E-04	12	Cytokine-cytokine receptor interaction	8.1E-05	31
Cytokine-cytokine receptor interaction	2.84E-05	32	Focal adhesion	4.05E-04	25	Chagas disease	2.4E-04	14
Complement and coagulation cascades	9.1E-03	13	Amoebiasis	1.6E-03	16	Focal adhesion	2.47E-04	27
Hematopoietic cell lineage	9.11E-03	14	ECM- receptor interaction	4.1E-03	12	Hematopoietic cell lineage	1.34E-03	13
--	--	--	Malaria	4.1E-03	10	Glutathione metabolism	4.2E-03	10
--	--	--	--	--	--	Arginine and proline metabolism	4.8E-03	18
MUSCLE								
Pathways	p-value	genes	Pathways	p-value	genes	Pathways	p-value	genes
Olfactory transduction	8.7E-11	4	Olfactory transduction	3.61E-21	19	Olfactory transduction	8.58E-12	8
Focal adhesion	3.7E-07	37	Cytokine-cytokine receptor interaction	7.4E-07	50	Focal adhesion	1.4E-05	12
Metabolic pathways	1.3E-03	48	Focal adhesion	8.9E-06	48	Amoebiasis	1.01E-04	23
Nitrogen metabolism	7.3E-03	8	Amoebiasis	7.6E-04	24	Arginine and proline metabolism	1.6E-03	13
Type-II diabetes mellitus	7.34E-03	12	Hematopoietic cell lineage	6.3E-03	21	Cell cycle	1.64E-03	23
--	--	--	--	--	--	Chagas disease	1.64E-03	20

Table 2.4: Significant pathways through DEGseq excluding the 59 common DE genes obtained among the four tissues

Region	Pathways	p-value	Genes
ADIPOSE	Leishmaniasis	1.05E-05	9
	Amoebiasis	4.7E-04	11
	TGF-beta signalling pathway	1.07E-03	9
	Olfactory transduction	5.46E-03	16
	Jak-Stat signalling pathway	5.8E-03	11
	Fc gamma R-mediated phagocytosis	8.8E-03	8
BRAIN	Tight junction	7.43E-05	9
	p53 signalling pathway	1.5E-05	4
	Fc gamma R mediated phagocytosis	1.6E-03	4
	Glycerolipid metabolism	1.6E-03	4
LIVER	Maturity onset diabetes of the young	6.03E-06	3
	Galactose metabolism	4.033E-05	5
	Olfactory transduction	7.43E-04	6

	Cytokine-cytokine receptor interaction	1.01E-03	9
	Focal adhesion	3.03E-03	7
	TGF-beta signaling pathway	4.25E-03	4
MUSCLE	Galactose metabolism	2.48E-04	3
	Focal adhesion	3.92E-04	7
	Olfactory transduction	8.4E-04	3
	Glycosaminoglycan biosynthesis –keratin sulfate	9.1E-04	3
	Fc gamma R-mediated phagocytosis	1.81E-03	4
	Chagas disease	1.64E-03	7

Table 2.5: Significant diseases formed by the common DE genes between our method (qtDE) and DEGseq along with the mutually exclusive sets of DE genes with respect to qtDE (Case1) and DEGseq (Case2) (the mutually exclusive sets of DE genes are given in bold)

Pathways	Method & Organ	Case1 (Common DE + mutually exclusive qtDE)		Case2 (Common DE + mutually exclusiveDEGseq)	
		p-value	Genes	p-value	Genes
Autoimmune thyroid disease	**Correlation** ADIPOSE	2.01E-03	5(*Cd86, H2-DMa, H2- T10, H2-Ab1,* ***H2- DMb1***)	1.3E-02	6(***H2-Aa***, ***H2-Q8***, *Cd86,H2-DMa, H2-T10, H2- Ab1*)
	Correlation MUSCLE	3.43E-02	4 (*H2-Ab1, Tnf, H2-Aa,H2-Eb1*)	6.3E-03	8(*H2-Eb1,* ***Cd86***, *H2-Aa,* ***H2-Q8***, ***H2-DMa***, ***H2-DMb1***, *H2-Ab1, Tnf*)
	Polynomial Regression ADIPOSE	2.3E-03	6(H2-*DMa,Tnf, H2-DMb1,Cd86, H2-Aa, H2-Eb1*)	5.8E-03	8(*Ifng,* ***H2-T10***, *H2-DMa, Tnf, H2-DMb1, Cd86, H2-Aa,* ***H2-Eb1***)
Cardiac muscle contraction	**Polynomial Regression** MUSCLE	1.06E-03	10(*Tpm3,Cacnb1, Myl3,Slc8a1, Cox6a2,Actc1, Cacna2d1,Tpm1, Cox7a1, Cox7a2*)	7.9E-04	12(***Cox7b***, ***Myh7***, *Tpm3, Cacnb1, Myl3, Slc8a1,Cox6a2,Actc1, Cacna2d1,Tpm1,Cox7a1, Cox7a2*)
Dilated cardiomyopathy	**Polynomial Regression** MUSCLE	1.7E-02	8(*Tpm3, Cacnb1, Myl3,Slc8a1,Actc1, Cacna2d1,Tpm1, Actb*)	4.08e-03	11(***Itga8***, ***Myh7***, ***Tnf***, *Tpm3,Cacnb1,Myl3,Slc8a1,Actc1,Cacna2d1, Tpm1, Actb*)
Graft-versus-host disease	**Correlation** ADIPOSE	2.01E-03	5(*Cd86, H2-DMa, H2- T10, H2-Ab1,* ***H2- DMb1***)	3.3E-03	7(***H2-Aa***, ***H2-Q8***, ***Il1b***, *Cd86, H2-DMa, H2- T10, H2-Ab1*)
	Correlation MUSCLE	1.4E-02	6(*H2-Aa, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Tnf*)	6.9E-03	8(***H2-Eb1***, ***Cd86***, *H2-Aa, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Tnf*)
Drug metabolism cytochrome P450	**Correlation** LIVER	4.14E-03	6(*Cyp2c40,Gstm2, Mgst2,Cyp2d10, Ugt1a9, Mgst3*)	1.7E-03	9(***Fmo3***,***Cyp2d22***, ***Gsta2***, *Cyp2c40, Gstm2, Mgst2,Cyp2d10, Ugt1a9, Mgst3*)

	Mutual Information BRAIN	1.3E-02	3(*Cyp2b9,Cyp2c55, Gstm1*)	1.3E-02	3(*Cyp2b9,Cyp2c55, Gstm1*)
	Polynomial Regression LIVER	6.8E-03	9(Gsta2,*Cyp2d22, Cyp2c40,Gstm2, Mgst2,Cyp2d10,Mgst3,Cyp2c54, Fmo3*)	2.1E-02	10(***Ugt1a9***, *Gsta2, Cyp2d22, Cyp2c40, Gstm2,Mgst2,Cyp2d10, Mgst3, Cyp2c54,Fmo3*)
Proteasome	**Polynomial Regression** MUSCLE	6.3E-03	6(*Psma4, Ifng, Psmb3,Psmc6, Psmb9, Psma2*)	2.3E-02	6(*Psma4, Ifng, Psmb3, Psmc6, Psmb9, Psma2*)
Viral myocarditis	**Correlation** ADIPOSE	1.4E-02	5(*Cd86, H2-DMa, H2- T10, H2-Ab1,* ***H2- DMb1***)	3.3E-03	9(***Rac2***, ***H2-Aa***, ***H2-Q8***,***Itgal***, ***Casp3***, *Cd86, H2-DMa, H2-T10, H2-Ab1*)
	Correlation MUSCLE	1.3E-02	8(*H2-Aa, H2-Q8, Myh7,H2-DMa,H2-DMb1,H2-Ab1, Fyn, Itgal*)	5.1E-03	11(***Rac2***, ***H2-Eb1***, ***Cd86***, *H2-Aa, H2-Q8, Myh7, H2-DMa, H2- DMb1, H2-Ab1,Fyn,Itgal*)
	Polynomial Regression ADIPOSE	2.1E-02	6(H2-*DMa, H2-DMb1,Cd86, H2-Aa, H2-Eb1, Myh2*)	9.03E-03	9(***Casp3***, ***H2-T10***, ***Itgal***, *H2-DMa, H2-DMb1, Cd86, H2-Aa, H2-Eb1,Myh2*)
Metabolism of xenobiotic by cytochrome P450	**Correlation** LIVER	1.06E-02	5(*Cyp2c40,Gstm2, Mgst2, Ugt1a9, Mgst3*)	3.4E-02	6(***Gsta2***,*Cyp2c40,Gstm2,Mgst2, Ugt1a9, Mgst3*)
	Mutual Information BRAIN	1.4E-02	3(*Cyp2b9,Cyp2c55, Gstm1*)	1.4E-02	3(*Cyp2b9,Cyp2c55, Gstm1*)

2.2.4 Discussion: Ether lipid metabolism is the only notable pathway identified from the common differentially expressed genes (qtDE genes) among four tissues and between linear and non-linear interactions based on phenotypic traits. This observation is supported by both Tables 2.1 and 2.2. Additionally, the importance of this pathway in mice and related primates has been discussed in [**27**], which highlights its significant role in various biological processes such as tumour cell invasiveness, energy storage, signalling molecules, and cardiovascular disease.

Another significant pathway, i.e. Olfactory transduction has been found from Tables 2.1, 2.3, and 2.4. In this regard, there are observations confirming the significant contribution disjoint set of differentially expressed (DE) genes in four tissues. [**28**] discusses the role of this pathway in relation to obesity development in both adipose and muscle tissues of mice. It provides further insight into the association between this pathway and the development of obesity in these specific tissues. Additionally, [**28**] mentions the functioning of the rodent olfactory epithelium in connection with the liver. On the other hand, [**29**] explores the role of this pathway in the functioning of olfactory sensory neurons (OSN) in the septal tissue.

The use of mutual information-based qtDE allows for the identification of various important pathways (like Leukocyte transendothelial migration, Amoebiasis, Focal adhesion, and Complement and coagulation cascades) and their potential contributions to different diseases and biological processes [**29-36**]. In this regard, some other notable pathways are Leishmaniasis (present in all tissues under study), and Cytokine-cytokine receptor interaction (present in muscle and brain). Significance of these pathways across different tissues are discussed thoroughly in [**37,38**]. From Table 2.4 we can have an idea of equivalent significance of these pathways using DEGseq model.

First column of Table 2.3 gives us some enriched KEGG pathways formed by the qtDE exclusive DE genes through linear correlative method. These are Cell cycle (present in the adipose tissue), Pyruvate metabolism (present in the brain tissue), Metabolic pathways (present in brain tissue), and Nitrogen metabolism (present in muscle tissue). Significance of these pathways is discussed in [**39-42**].

Third column of Table 2.3 exclusively shows some eloquent KEGG pathways by qtDE polynomial regression based method. The notable ones are Chagas disease (present in liver, also seen in table 4 but related to muscle), Arginine and proline metabolism (present in adipose, liver and muscle) and Cytokine-cytokine receptor interaction (present in liver and adipose). Importance of these pathways has been discussed in [**43,44**]. In this regard, another notable pathway present in Table 2.3 as well as in table 2.4 is Fc gamma R-mediated phagocytosis, the application of which has been discussed in [**45**].

Finally, the biological significance and differential roles of different disease related pathways enlisted in Table 2.5 are discussed in [**46-51**].

2.3 Reconstructing Paired Transcription Factor Regulatory Networks

Here, the matter of investigation is related to the formation of gene regulatory network (GRN) that may use gene-gene or gene-protein interaction patterns [**52**]. In this context, transcriptional regulators, which are proteins, play a crucial role in modulating gene expression levels during different stages of development. As a contributory step in this regard, the developed procedure helps to identify the pair or pairs of transcription factors (TFs) that have the potential to regulate a target gene in a linear or non-linear manner.

2.3.1 The basic findings: The target genes considered are the differentially expressed (DE) genes which are controlled by a pair or pairs of TF genes following a certain algorithm based on interaction measures defined by Pearson correlation, mutual information, and spline regression. In this regard, the Pearson correlation interactive approach is of linear nature, while mutual information and spline regression depict nonlinear relationships capable of capturing complex interactions between the TFs and the target DE gene. In this segment, corresponding to each target DE gene, the best pair/pairs of regulating TF genes is/are obtained through each of the three approaches mentioned above.

2.3.2 Methodology: The algorithm that has been followed to find the best possible combination of TF gene pairs for a target DE gene is outlined below.

ALGORITHM: Best Possible Combination of TF gene pairs

Input: gEmtx1, gEmtx2, TF1, TF2, TFg1, TFg2, β →1
Output: NLM1, NLM2, Mv1, Mv2, Mv, bTFp1, bTFp2, TFp1, TFp2

```
Step1. DE ←DEGexp (gEmtx1, set expression columns for conditition1, gEmtx2, set expression columns for condition2)
Step2. C← choose mode of operation
           for i in 1 : X do
             for j in 1:Y do
               r1← rbind(DE1[i,],TF1[j,])
               nr1← transpose(r1)
               r2←rbind(DE2[i,],TF2[j,])
               nr2 ← transpose(r2)
               if c = =1
               NLM1 [i,j]← NonLinMI(nr1, β)
               NLM2 [i,j]← NonLinMI(nr2, β)
               else if c= =2
               NLM1 [i,j]← NonLinSP (DE1[i,],TF1[j,])
               NLM2 [i,j]← NonLinSP (DE2[i,],TF2[j,])
               else
               NLM1 [i,j]←LinCor(nr1, β)
               NLM2 [i,j]←LinCor(nr2, β)
               end if
             end for
           end for
Step 3.    for i in 1:X do
             m←0
```

```
                    for j in 1:Y-1 do
                        for k in j+1:Y do
                            m ← m+1
                            Mv1[i,m] ← NLM1[i,k]*NLM1[i,j]
                            Mv2[i,m] ← NLM2[i,k]*NLM2[i,j]
                        end for
                    end for
                end for
Step 4.         Mv ← Mv1-Mv2
                for i in 1:X do
                M[i] ← min(Mv[i,])
                n ← 0
                  for j in 1:Y-1 do
                      for k in j+1:Y do
                          n ← n+1
                          if (M[i,1]==Mv[i,n])
                          Indx1 ← j
                          Indx2 ← k
                          bTFp1[i] ← TFg1[Indx1]
                          bTFp2[i] ← TFg2[Indx2]
                          end if
                      end for
                  end for
                end for
Step 5.         M ← $Y_{C_2}$
                for i in 1 : X do
                   for l in 1: M do
                       TFp1[i,l] ← 0
                       TFp2[i,l] ← 0
                   end for
                   V[i] ← M[i]+ Δv
                    n ← 0, s ← 0
                   for j in 1:Y-1 do
                       for k in j+1:Y do
                           n ← n+1
                           if MV[i,n] < V[i,1]
                           s ← s+1
                           TFp1[i,s] ← TFg1[k,1]
                           TFp2[i,s] ← TFg2[j,1]
                           end if
                   end for
                end for
            end for
%%%% End of main routine %%%%%%
```

```
NonLinMI ← function (nr, β) .. Subroutine 1 for function call NonLinMI
mutualInfoAdjacency (nr, discretize columns, set entropy estimation method,
the number of discretization beans) ^ β
NonLinSP← function (DE,TF) .. Subroutine 2 for function call NonLinSP
sm.spline (DE, TF, set the order of spline function)
LinCor← function (nr, β) .. Subroutine 3 for function call LinCor
cor (nr, set correlation method) ^ β
```

In the above outline, the stepwise details are given below.

Step 1: Identifying DE genes

Here, the DE genes are found initially making use an R package DEGseq. The analysis is carried out based on the expression values of genes under different time points or conditions. A specific threshold is set that can be a "p-value," a "z-score," or a "q-value." Comparing the expression values of genes using this this threshold, the genes which show significant changes in expression are considered the DE genes.

Step 2: Finding the individual level of linear and non-linear association

At this stage, the goal is to calculate the linear and nonlinear association between DE genes and TF genes across two conditions. The methods involved in the process are Pearson correlation, mutual information and spline regression. Initially, there are X DE genes identified in Step 1, and Y TFs. The gene expression matrix for the DE genes under condition 1 is denoted as DE1, and DE2 corresponds to condition 2. Similarly, TF1 represents the gene expression matrix of TFs under condition 1, and TF2 represents condition 2. The X DE genes are considered as the target genes for analysis. In addition to *LinCor* (invokes the correlation function in R package WGCNA), the algorithm uses two other functions; one is called *NonLinMI* to calculate the mutual information (MI) adjacency measure based on symmetric uncertainty between each TF and the target gene in each condition and the other is *NonLinSP* used to determine the nonlinear association between DE genes and TF genes through spline regression. In this particular work, a cubic spline with an order of 3 has been chosen. The result of these computations is the creation of two matrices: NLM1 and NLM2.

Step 3: Computing the pairwise values of linear and non-linear association

In this step, all possible pairs of row elements of each matrix separately are multiplied, indicating the multiplication between every pair of linear and nonlinear association values obtained between TF genes and the target DE gene. The multiplication results are stored in the set of matrices Mv1 and Mv2, respectively. Hence, multiplying each possible pair, for a given number of columns (in this case Y because Y TF genes have been found), there will be Yc_2 number of combinations. Therefore, the matrices Mv1 and Mv2 will have X number of rows and Yc_2 number of columns. In this regard, a high multiplication result suggests that the dependency of the target DE gene on two TF genes is high, indicating a strong regulatory action. On the other hand, a low multiplication result suggests a low dependency and weak regulation. In other words, the higher the value, the stronger the regulatory influence, and the lower the value, the weaker the regulatory influence between the TF genes and the target DE gene.

Step 4: Finding the best possible pair of TF genes for a certain target DE gene

At this level, subtraction of the two matrices, i.e. Mv1 and Mv2, is done. This subtraction is done to compare the regulation of a target DE gene by these TF genes in each condition. If the subtraction result is small, it suggests that the regulation of the target DE gene by both TF genes in each condition is approximately the same. On the other hand, if the subtraction value is relatively large, it indicates that the regulation of the target DE gene by both TF genes between conditions is not equal, highlighting a differential effect. As per the algorithm, to filter and select TF gene pairs with small subtraction values, the subtraction results between the two conditions are compared. The best TF pair is determined by finding the pair with the minimum subtraction value. These selected TF gene pairs are stored in the variables bTFp1 and bTFp2, representing the best TF gene pairs for a specific target DE gene.

Step 5: Finding TF gene pairs having almost similar level of association compared to the best TF gene pair

In this final step, the results from Step 4 are investigated a bit further. A limit is defined and a range of values are considered within that limit. The reference value is the subtracted result of the best TF pair found in Step 4. Let's say for a target DE gene, TF pair x and y have the minimum subtraction value, denoted as M [1]. Setting the limit as

Δv, it is easy to select TF pairs whose subtracted outcomes fall within the range of M [1] + Δv. In this regard, comparison of the differences between each TF pair's subtracted value and M [1] is conducted. If the subtracted value of a TF pair falls within the range M [1] + Δv, it is considered a candidate for biological interpretation.

2.3.3 Results: At the very initial stage, the gene expression matrix considered, may have zero and/or missing values at one or more profile measurements. In other words, a sparse expression matrix may be present. To resolve this issue, the zero value (wherever observed) is replaced by a very small number (in this case 10^30) and the missing value (wherever observed) is replaced using a value obtained through KNNimpute [**53**] method. The work here has involved two datasets, one being yeast information, i.e. budding yeast Saccharomyces Cerevisiae cell cycle data [**54**] containing expression matrix of 6178 genes across four conditions and the other being Affymetrix expression data of colon cancer tissues corresponding to human subjects containing 22,278 genes in two conditions, where 49 tissues are non-cancerous and 48 are of cancerous nature [**55**].

Among the four conditions depicted above for the Yeast data, the DE genes obtained via DEGseq gives us at the maximum 285 significant DE genes across the first and fourth condition, considering all possible pairs of conditions. On the other hand, for the colon cancer data from human subjects, 56 significant DE genes are observed via the same DEGseq approach. Coming to the identification of TF genes in Yeast, the same has been done making use of TF binding site, mutant, ChIP-chip, and the basic cell cycle expression data with a stringent p-value threshold (≤0.001) to determine TF promoter binding [**56,57**]. Accordingly, 17 TF genes are discovered in the context of Yeast cell cycle data. However, for the human colon cancer data the TF genes are found using [**58**]. By matching the IPI ids provided by [**58**] with the IPI ids of the referred colon cancer data, 1065 TF genes get effectively discovered.

Following the proposed algorithm, the best TF gene pairs significant for a target DE gene regulation is found in both the cases of yeast and human colon cancer data. The relevant tables in this regard are given below. Tables 2.6 and 2.7 depict the above significant findings with respect to mutual information and spline regression on one hand and linear correlation on the other for Yeast cell cycle data. Validation of the obtained interactions have been checked using a web based regulatory tool called YEASTRACT [**59**].

Table 2.6: Best Combination of TF gene pairs corresponding to target DE genes through nonlinear mutual information and spline regression methods for Yeast cell cycle data

Mutual Information		Spline Regression	
Target (DE)	TF pair	Target (DE)	TF pair
PRM5, BUL2	ACE2, FKH2	GRX7, PRB1, CNB1	ACE2, FKH1
YDR124W, YER010C	ACE2, MCM1	FUS3, SPI1, STF2, MOD5	ACE2, SWI5
PGM2	ACE2, CST6	UGA2, RTC3, SRP40, KTI12, YKL044W, AFR1, PRM10, COS9	MCM1, SWI4
STE2, TIF1, RSA1, YMR111C	ACE2, ASH1	LEU2	SWI5, ASH1
YPC1, BSC1, MCT1, YML119W	MCM1, ASH1	RCR1, YNL146W, YJR154W	CIN5, CST6
PRB1	STE12, SWI6	FIG2, NMA1, GSF2, HCH1, AFI1	TEC1, CST6
DSE12, RIM21, GPD1, BAP2	RLM1, STP1	YPR142C, YBR144C, CCT4, GUD1, ECM18, GGA1, YGL117W, YGR149W, DSE2, ERG24, HXT10, PIR3, BUR2, PGM2, IMA2, FMP21, KCC4, YDR249C, YML131W, OM14, SDH4, YIL108W, YJL068C	STE12, ASH1
YJL160C, MOG1, LTV1, YKL044W, MFG1, GYP7, SRB7, PCM1, IME4, LSB1, DIA4, YSC84, HXT4, AYR1, YJR026W,PSO2, MRPS18, IMA2, YOR029W, YOR053W, MKK1, YPL039W, YPL062W	TEC1, STB1	YBR144C, KCC4, FMP21, HXT10, GSY1, MST27, HXT4, AYR1, GUD1, YJR026W, YKL151C, CIK1, ERG24, ATO2, RPL25, YOR053W	STE12, STB1
GLK1, TDP1, ERG24, ESC8	SWI4, TEC1	PAM1, DIA4, SMD2, PAM16	STE12, RLM1
ATO2, DIP2, PIR3	CST6, SWI4	STP4, YET3, GYP7, YJL052W, SLT2, HXT6, HXT9, MYO3, YLR253W, PSO2, YNL043C, YMR317W, FDH2	TEC1, STB1
YBR225W, KCC4, DOT5, YJL068C, CYC2, YPK2, YIL108W	TEC1, RLM1	IME4, APE1, DIP2, YOR121C, MKK1, VPS38, FSH1, CAP2, GIP3, MF(alpha)1	ASH1, TEC1
OM14, DIA3, YET3, PAM1, ATP17, GSY1, YGL052W, YHR097C, PRM10, SMD2, VPS38, YML131W, SIP5, CIK1, BOR1, BSC6, GDH1, NTO1	ASH1, TEC1	TDP1, YBR225W, FUS1, YCR007C, ATP17, BCY1, BAP2, POR2AGA2, YSC84, TFA2, MCM5, NIT3, SIP5, VTI1, BOR1, ESC8, NTO1	ASH1, STP1
FMP30	ACE2, STE12		
FUS1, ECM4, YBR138C	CIN5, CST6		
MFA1, SSU1, FDH2	TEC1, CST6		

Table 2.7: Best Combination of TF gene pairs corresponding to target DE genes through linear correlative method for Yeast cell cycle data

Linear Correlation	
Target (DE)	TF pair
OLE1,YBR144C,HXT4	ACE2, SWI5
YCL076W,YKL044W,VPS38	SWI5, MCM1
PRM2,SLF1,YIL080W,ASK1,GSF2,AGP2	ACE2, CST6
GLK1,YJL068C	NDD1,CST6
AYR1,PSO2	SWI4, ASH1
TOR1,FIG2,ATG34	STE12, ACE2
YDL038C,CLK1	FKH1,STE12

ESC8,YHR138C	TEC1, MBP1
BAP2,GGA1	CST6, MBP1
RTC3,HCH1	SWI4,RLM1
SOV1,YCR007C,CCT4	SWI5,ASH1
NMA1,PRB1,YGR066C,ADI1	SWI6,CIN5
GDH1,GDP1,MFG1	SWI6,ASH1
FRA1,YGL052W,FSH1,GIP3,NTO1	CST6,ASH1
GUD1,GSY1,YJR026W	ASH1,STP1
PMC1,PRM8,SPC29,ECM18,HXT10,MST27,IME4,YKL151C,YPK2	RLM1,STE12
YSC84,SMD2,GPD2,YOR029W,YOR053W,RIM15,YPL062W	CIN5,STE12
SSB2,APE4,TAX4	CST6,STE12
YHR097C,FUS3,SST4,SDH4,BCY3,ECM4,YLR253W,CPA1	ASH1,STE12
RVS161,YNL146W,	MCM1,ASH1
RCR1,YET3,COQ6,LTV1	CIN5,STP1
YCL023C,OM14,PCL1,BUL2,ATG2,MKK1,YGR149W,YCL042W,YML131W,YMR111C	TEC1,CST6
BSC1,YPC1,AFI1,MOD5,SDS24,YMR317W,ZSP1,CAP2,RUP1	ASH1,TEC1

Some of the validated outcomes obtained through YEASTRACT are given in the form of TF regulatory networks shown in Figures 2.1 (mutual information based), 2.2 (spline regression based) and 2.3 (linear correlation based) respectively.

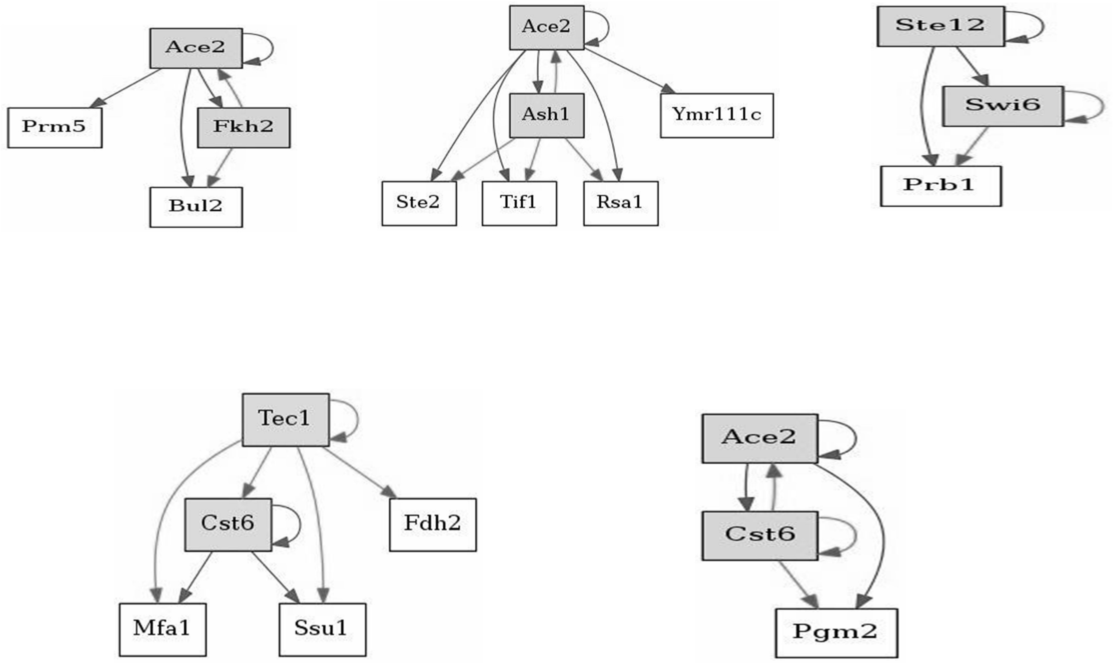


Figure 2.1: Corresponding to Yeast cell cycle data, biological validation of some TF pairs for DE genes obtained through Mutual Information (5 cases are shown)

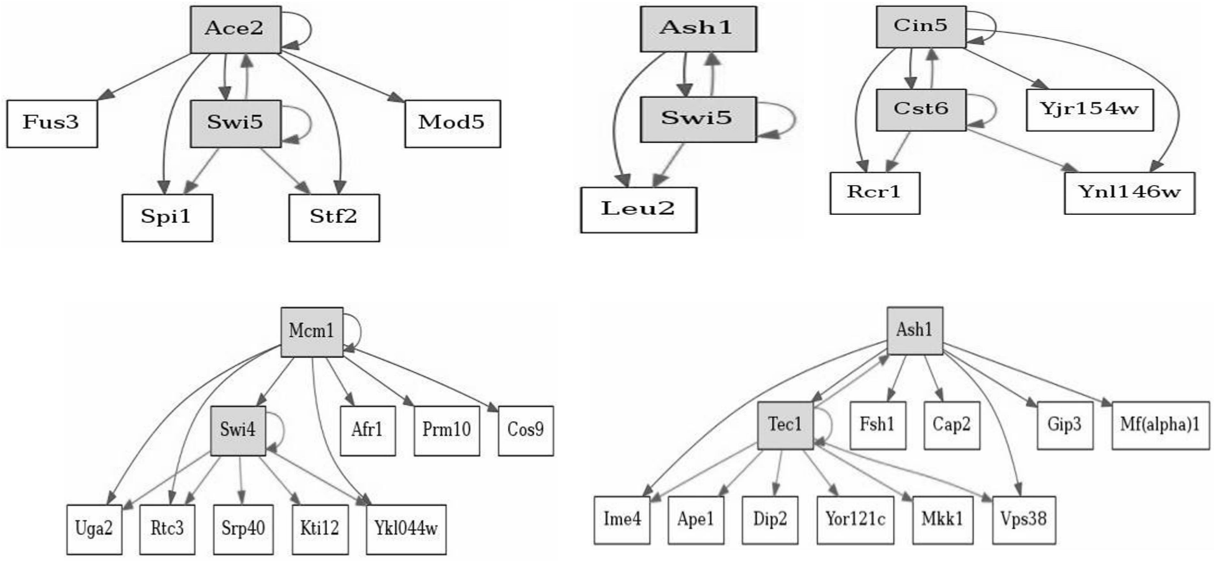


Figure 2.2: Corresponding to Yeast cell cycle data, biological validation of some TF pairs for DE genes obtained through Spline Regression (5 cases are shown)

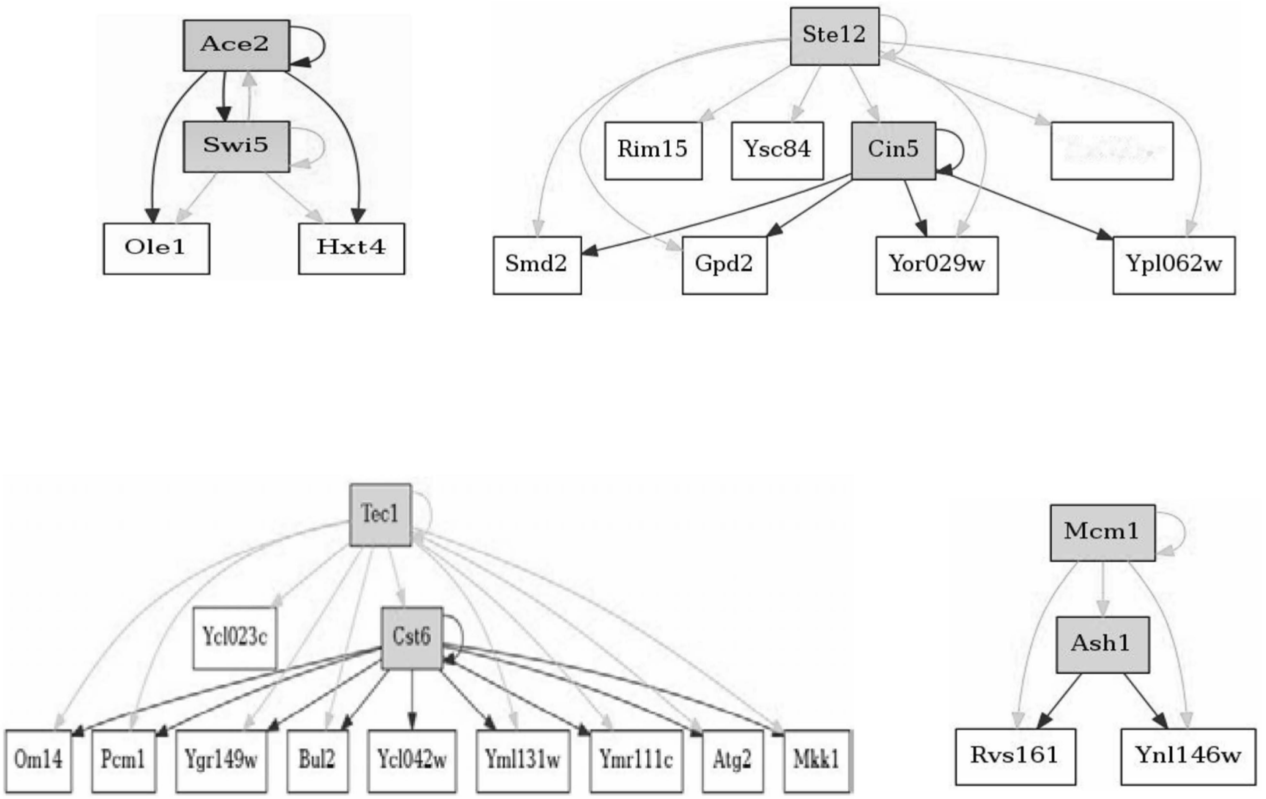


Figure 2.3: Corresponding to Yeast cell cycle data, biological validation of some TF pairs for DE genes obtained through Linear Correlation (4 cases are shown)

Similarly, in the case of human colon cancer data, Table 2.8 denotes the mutual information and spline regression based outcomes following the proposed approach. The validation of the corresponding interactions has been checked using two web based tools namely TFactS (https://www.tfacts.org/) and PRISM [**60**]. In TFactS, The p-value, e-value, q-value and FDR (Benjamini-Hochberg) thresholds are set as 0.01. They are given to control the rate of false positives for multiple testing conditions [**61**]. Remaining parameters are left at the default levels. In PRISM, a validated interaction shows additionally the ontology, biological context, e-value, p-value, fold enrichment, genes hit and binding sites as output. In this regard, Tables 2.9 and 2.10 show the validated examples from PRISM tool that tally with the outputs obtained from the proposed algorithm via mutual information and spline regression approaches respectively.

Table 2.8: Best Combination of TF gene pairs corresponding to target DE genes through nonlinear mutual information and spline regression methods for Human colon cancer data

Mutual Information		Spline Regression	
Target (DE)	TF pairs	Target (DE)	TF pairs
CA1	CDX2, PAX2	CA1	CRX, PAX2,PAX5, SP140 SEMA4A
GCG	CREB1,FOXA1,HOXC8, ST18 AHCTF1,STAT1	GCG	PAX2,POU6F2,ZNF236,NKX6-1, ZNF638,ZC3H10,FOXA1,CREB1
INHBA	ATF1,CREB1,NFYA,MEF2B,HIVEP3	INHBA	ATF1,CREB1,DHX57
CHGA	ATF1,CREB1,EGR1,ETS2,JUN,HOXC4, TFAP2A, STAT1	CHGA	NR2E3,MYF6,ZNF236,NEUROD6, ZSCAN12,ZNF155,LASS6,ZNF638, CBX2,EGR1,TFAP2A,ATF1
SPP1	TP53, FOXJ3 DEPDC6,EST1,GLI1,JUN, HOXC8	SPP1	HOXA9,MYB,SMAD1,KLF10,ETV4, ZNF750,ZSCAN16,ID4,POU5F1,TP5, DLX5,CTNNB1,ETS1,GLI1,HOXC8,
IL8	TP53,HSF2,LHX3,SOX21,IRF9, GATAD1,ZFR2,NFKB2,JUN,RELA, ZNF33B	IL8	TP53, NFE2L3,SMAD5,MSX2, ZNF444,STAT2,VENTX, BACH2,TCF20,MET, RORA,CDX1,TEAD4,ZNF257, TOX3, MET
ADH1C	TCF3,NFYA,ELF5,NFIC,TBP,DBP	ADH1C	CEBPB,NFYA,ELF5,SMARCA1,TBP,DBP
CHI3L1	SP4,MAX, PARP12	CHI3L1	ELF4,BACH2,SPI1,MLLT3,YEATS2, GTF3A,JRKL,USF1
ADH1B	ATF4,DBP,FOXC1,MTA1,CEBPB	SLC26A2	TFAP2C,CTNNB1,RBPJL,SP1,EMX1, ZNF257,NEUROG2,PLEKHA4,
MUC4	RCOR1,STAT5A,ZNF43,ZNF764,SMAD4	ADH1B	CEBPA,CEBPB,BACH2,HHEX, HOXB2,ZNF155,WNT8B
PDE9A	LHX6,ZNF665,ATXN7,DSP,GLI1	MUC4	ZNF236,ST18,HOXC10,ZNF750, SMAD7, TFAP2B
ANPEP	HOXD1,NFYA,NR2F1,ANKZF1,ETS2	PDE9A	ELF4,GLI1,TFAP2A,FOXJ3,BMP2
UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9	TBX21,NEUROD1,H1F0,HNF1A,RARA,SP1, RESTKLF2, NEUROD1, IRF7, ZMAT4, RBM22,SLC22A4,PPARG	CEACM7	EZH2,PLAGL2,SRY,RERE,HMGB3, MBNL2,GLI2,NFAT5,HOXB6, BCL11B,PBX1,ZNF177

UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9,IL8	IRF7,ZMAT4	ANPEP	ELF4,PHOX2B,IRF8,ESRRA,HLF, TSC22D2,CUL3,EST1,EST2
CLCA1	NR1H3,LHX6,RELA,GLI1	UGT1A1,UGT1A2, UGT1A3,UGT1A4, UGT1A5,UGT1A6, UGT1A7,UGT1A8, UGT1A9	RARA,CDX1,GATA6,HOXD12, CHD7,HNF1A,PPARG,HHEX, NPAS2, MNX1,ZBTB3,TOX3,RAPGEF
SST	ZNF287,FOXJ3,RNF113A,PAX6,CEBPE, ATF1,ATF2,ATF4,CREM	CLCA1	HOXA9,ATF2,HOXC13,GLI1,BMP2, ZC3H7B,
		SST	GATA1,DLX2,NR4A3,SRY, NEUROD4,C11orf9,PLEKHA4, CREB1,CEBPA,CEBPG
		HSD17B2	RBPJL,PGR,HOXC6,EN1,FBN1, CTNNB1,

Table 2.9: Validating some Mutual Information based TF genes corresponding to target DE genes using PRISM with certain significant scores

Target DE gene	TF gene	E-value	P-value	Fold enrichment
MUC4	STAT5A (Similar Protein to STAT1)	0.000	1.07E-16	2.16
IL8	JUN	0.116	2.71E-11	2.12
SPP1	JUN (Similar Protein to JPD2)	0.000	1.25E-42	2.02
GCG	STAT1	0.116	1.43E-22	2.57
SST	PAX6	0.349	6.21E-09	2.12

Table 2.10: Validating some Spline Regression based TF genes corresponding to target DE genes using PRISM with certain significant scores

Target DE gene	TF gene	E-value	P-value	Fold enrichment
MUC4	TFAP2B	0.697	7.16E-06	2.62
CA1	CRX, PAX2, and PAX5	0.697	2.78E-13	3.25
IL8	BACH2	0.116	4.29E-25	2.06
SST	DLX2 (Similar Protein to BARHL2)	0.349	5.99E-08	2.14

2.3.4 Discussion: In the context of Yeast cell cycle data, the obtained results can be biologically justified from the various gene interaction pathways involving TF and DE genes in [**62-69**]. For the human colon cancer information the two web tools that have been used to biologically validate the interaction outcomes obtained from the non-linear approaches similar protein generating TF gene names get suggested. For example, observing the TF gene entry in the first row of table 2.9, STAT5A is the actual TF gene that takes part in pairwise interactive regulation for MUC4 (target DE gene) through mutual information based adjacency measure. However, PRISM depicts STAT1 to be the interactive TF gene for the same target. The tool also clarifies that STAT1 and STAT5A are able to produce the similar kind of protein complex (hence can be considered alias of one another) required for the regulated transcription of the DE gene, MUC4.

The interactive regulatory associations involving datasets incorporating huge number of gene expression profiles are more likely to be nonlinear type rather than linear [**70**]. To explore this idea further, a part of this research is focused on investigating how TF genes nonlinearly regulate their target DE genes. The employed measures such as mutual information and spline regression are used to analyze these nonlinear regulatory interactions. The choice of the human colon cancer dataset was motivated by its high dimensionality, meaning it contained a large number of samples. This high dimensionality allowed the research to accurately capture the impact of nonlinear regulations on a large-scale dataset. To ensure that true regulatory relationships between TF and target DE genes are not missed, a pair of TF genes has been selected with the intention of minimizing false negatives [**56**].

2.4 Smart Threshold selection for reconstructing Gene Regulations

Proper identification of genes responsible in spreading complex diseases is always a matter of stringent investigation. In this context, in addition to the gene expression pattern the network connectivity across different states or conditions is also of prime importance. Some of the crucial network designs [**71,72**] that have marked a significant step in the understanding and development of the dynamicity involved also guides to explore the selection of suitable drug targets relevant to a certain pattern of a disease. Extending further, the significance of differential connectivity across different states or conditions of a living cell can be judged through the existence of the same on the basis of some defined threshold. The process of this threshold selection in network link determination can be done considering global or local perspective. Understanding the development of a regulatory network that may comprise of TF to target DE gene regulations at different stages require an optimistic approach toward selection of this threshold because network regulations can have the undoubted presence of both direct and indirect controls. Hence, investigating the differential connectivity having the inevitable presence of both such controls can be done through the selection of some smart threshold in a global (overall network) or local (certain well researched portion of a network to study the effects of any environmental or some external perturbation).

2.4.1 The basic findings: The concept of generalized topological overlap measure (GTOM) [**19,73**] applying linear correlative dependency among genes have been utilized to find the gene to gene interaction in every possible state of the gene regulatory network in the course of development of any disease. This measure has the capability of incorporating the importance of direct regulatory interactions in the presence of one or more indirect gene associations. To understand the significant difference in the topological connectivity of a certain gene (here referring to a DE gene), un-weighted TO (Topological Overlap) measure is resorted, which demands a smart threshold selection. This in turn helps in discovering the DE genes having minimal overlap between the regulatory networks across varied states or conditions. Accordingly, such DE genes can be marked as the biomarker genes or the significant hub genes responsible for the exclusive development of a particular diseased state or condition.

2.4.2 Methodology: The work has been executed in two different forms of gene regulatory network. In the first case, the alteration in interaction structure or the differential interaction has been explored between the control or normal state and the initial stage or the severe stage of influenza, as the case may be. In the latter one, the differential interactions have been studied between the initial and severe stages of influenza. For the earlier study, the DE genes between the control and initial or severe stages are taken into consideration. However for the latter, the effect of the common DE genes between those obtained above, are studied.

In the process, the intelligent or smart threshold selection algorithm that has been applied in the two different cases, mentioned above, is given next.

ALGORITHM

First Case: The Intelligent Threshold Selection for First/Second Network

Step1. Find DE genes between the conditions *control* and *day0/day6* of influenza.
Step2. Evaluation of GTOM measure corresponding to DE genes in each condition using equation

$$t_{ij} = \begin{cases} \dfrac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min\{|N_m(i) \cap N_m(j)|\} + 1 - a_{ij}} \ldots\ldots. i \neq j \\ 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. i = j \end{cases}$$

```
Step3. Go for the Fisher transformation of GTOM matrix and find the maximum value.
Step4. Choose the condition specific threshold value $\Delta th$ and set maximum threshold $th_{\max}$= |Fisher transformed GTOM values|$_{\max}$
Step5. n=1
Step6. while(n× $\Delta th \leq th_{\max}$)
Begin
    ➢ Find the specific interactions for each and every threshold $\Delta th$ in each condition.
    ➢ Evaluation of TO by equation $TO_i = (X_i \cap Y_i)/\max(X_i, Y_i)$
    ➢ Go for significance testing and evaluation of TOP by $TOPavg_i = (TO_i + pvalue_i)/2$
    ➢ Calculation of cumulative TOP or cTOP scores for each threshold.
    ➢ n = n+1.
End
Step7. Selection of best threshold pairs through comparison of cTOP values for all thresholds.
```

Second Case: The Intelligent Threshold Selection for Common Network

```
Step1.  Find the common DE genes between network 1 and network 2.
Step2.  Rest follows step 2 through 7 under First Case.
```

As per the algorithm, in the first case, two regulatory configurations involving DE genes are considered each in Network 1 and Network 2. Under Network 1, the two gene association networks considered are reconstructed in the control and initial or day 0 stage of influenza. Similarly in Network 2, the two gene association networks considered involve control and severe or day 6 stage of influenza. However, in the second case of Common Network, the two regulatory configurations involve initial (day 0) and severe (day 6) stages of influenza. The stepwise discussion of the above algorithm is as follows.

Step 1: Initially, in the first case, the DE genes are computed using the DEGseq package between control and day 0 and control and day 6 respectively. This gives two sets of DE genes, namely DE1 and DE2. Thereafter, in the second case, the common DE genes are found between DE1 and DE2. This yields the gene set cDE.

Step 2: The GTOM technique is utilized to check the gene to gene associations between each pair of DE genes (in both the cases). In other words, computation of the generalized topological overlap measure is conducted for each DE gene, be it an element of set DE 1 or DE2 or cDE for all the 6 regulatory configurations using the Pearson's correlation component value of the concerned gene with others. In this regard, a_{ij} is the correlation dependence between genes 'i' and 'j', $N_m(i)$ represents the neighbours of node 'i' excluding 'i' itself indicating no self-regulation, and $N_m(i) \cap N_m(j)$ means the set of common neighbours shared by genes 'i' and 'j', for an m order GTOM adjacency technique (suggesting the common node is reachable from node 'i' and node 'j' within m steps). In any case, as there are two conditions, calculation of GTOM is done between every pair of DE genes in both the conditions. In the implementation of this step, the order m is considered equal to 1. To compute the neighbours cited above with m=1, the concept of PCIT (Partial Correlation and Data Processing Inequality) is executed for each every DE gene. Through the application of PCIT, the importance of direct connectivity is highlighted in the presence of other necessary factors or physical entities or DE genes considered here, excluding the source and target DE genes.

Step 3: From this stage, the interest is in initiating the reconstruction of gene regulatory associations in the two conditions of interest corresponding to each of the two cases, outlined in the algorithm. To capture the idea, Fisher's transformation of the GTOM interaction matrix obtained in each of two conditions is done using $z_{ij} = 0.5 * \frac{1 + GTOM_{ij}^{q}}{1 - GTOM_{ij}^{q}}$ [**18**], where q=1 and the maximum z_{ij} value (z_{max}) in each condition is noted.

Step 4: At this point, 100 different thresholds are generated, and the best threshold is chosen amongst them. In order to obtain a specific threshold, it is chosen from the set of thresholds T (series of 100 equidistant values between 0 and z_{max}, defined in step3 above). Thus the minimum threshold (i.e. the difference between two thresholds) can be written as $\Delta th = \frac{z_{\max}}{100}$. So, denoting the series of thresholds as T = {T1, T2, T3,............, T100}, the individual thresholds will be T1=Δth , T2= T1+Δth, ,......................,T100=T1+99 Δth.

Steps 5 and 6: Here, the actual reconstruction of the DE gene regulatory associations are conducted in each condition of interest (control/day 0/day 6, as the case may be) based on a certain threshold selected from set T defined in step 4 above. If the z_{ij} value of a DE-DE GTOM based interaction is higher than the selected threshold, the adjacency value, A_{ij}, between the concerned pair is considered equal to 1. Otherwise, there is no interaction in between these two DE genes. Followed by this operation, the topological overlap (TO) of a certain interacting DE gene in the two conditions is found using the intersecting equation given in the algorithm. Here, X_i and Y_i indicate number of significant interactions (based on threshold level) respectively in the two conditions. In this regard, lower the TO value for a particular interacting DE gene, better is the disjoint condition specific regulatory property highlighted. The significance of the disjoint interacting property of any DE gene, i.e. the TO value obtained, is tested using a permutation test. In this context, random shuffling of the expression profiles is conducted between the two conditions followed by the computation of the GTOM value of the referred DE gene with every other DE gene and the corresponding TO attainment of the referred DE gene. This shuffling is conducted 1000 times. In the shuffled versions, the TO value obtained each time is compared with the actual TO value of the referred DE gene. The more number of times the shuffled TO values happen to be higher compared to the actual TO value, better is the statistical significance of the TO attainment of the referred DE gene. In short this is the p-value of the TO attainment. Combining the two scores, TO and its p-value, is one of the ultimate objectives in this segment, i.e. TOP as defined in the algorithm. Hence, corresponding to a selected threshold level, defined earlier, with N interacting DE genes, N TOP values are obtained in a particular case (i.e. states control and day0 resembling Network 1 conditions or control and day 6 resembling Network 2 conditions or day 0 and day 6 resembling Common Network conditions). In any case, First case or Second case, as highlighted in the algorithm as well clarified in the earlier statement, the N TOP values are added to find the cumulative TOP score (cTOP). Accordingly, for 100 defined threshold pairs (as per step 4; threshold defined is condition specific), 100 cTOP values are found.

Step 7: Finally, the best threshold pair is selected for a network structure (Network 1/Network 2/Common Network) corresponding to the minimum cTOP score. Thus the algorithm is able to clarify the selection of an intelligent threshold helping in discovering and hence understanding the disjoint or differential connectivity of the DE genes across conditions.

2.4.3 Results: Before implementing the algorithm as per discussion provided above, it is important to carry out the pre-processing, if any, is required for the publicly available gene expression data. Here, the experimentation is conducted on a peripheral blood data set [**74**] which shows the temporal analysis of patients having severe pdm (H1N1) influenza. The data consists of the expression profiles of 7 control (healthy) and 14 patient samples. The information shows patient samples are equally distributed across two kinds of conditions; condition1 pertains to the first day (i.e. day0, initial stage of influenza) and condition2 pertains to the seventh day (i.e. day6, severe stage of influenza) of peripheral blood expression profiles. Hence, there is the presence of 3×7= 21 time instants or gene expression profiles with 33,297 genes. On conducting some necessary pre-processing the data is reduced to have the 10,000 most variant genes. The pre-processing is done to understand the effect of the most variant genes (in this context the DE genes) in the identification and spreading of a certain stage of the influenza which in turn also help in the reduction of space and time complexity issues inherent in this approach.

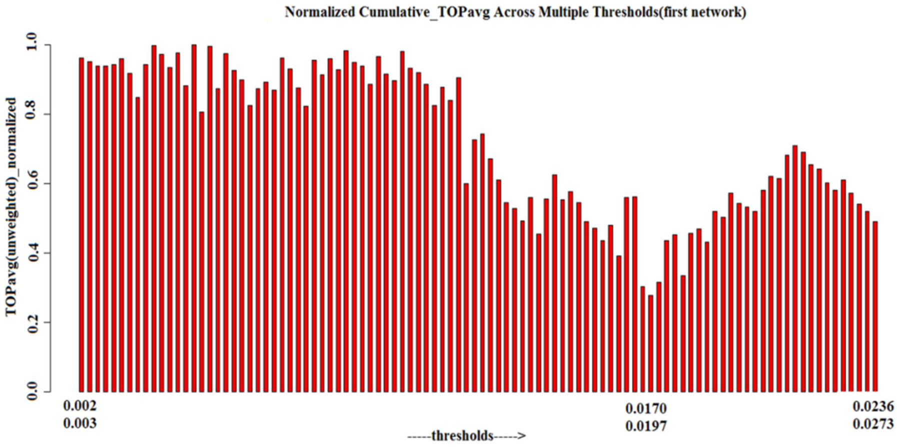
The necessary pre-processing involves conducting log normal distribution of the given gene expression data or matrix followed by finding the standard deviation of the gene expression profiles pertaining to a gene. The average of the set of standard deviation levels can be decided to be a threshold for selection of the most variant genes. Those genes are selected which maintain standard deviation values higher than the threshold.

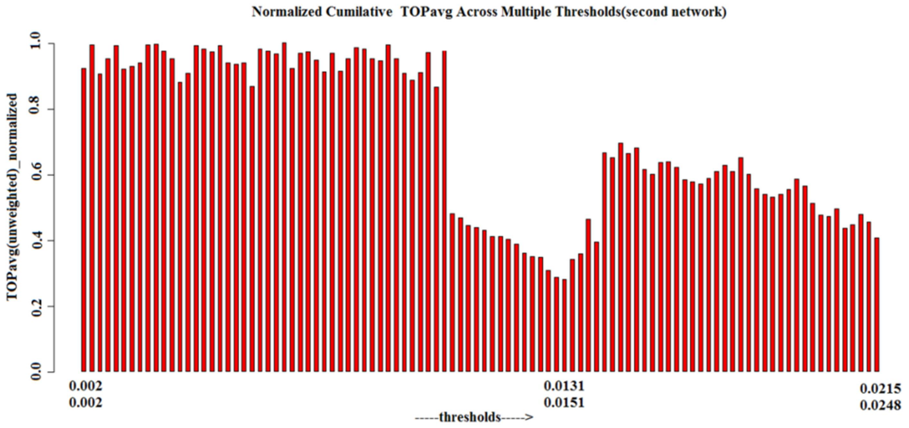
Post pre-processing the actual implementation is done. On the application of DEGseq to find the DE genes in Network 1 (comprising control and day 0 samples), Network 2 (comprising control and day 6 samples), and Common Network yields 73 (DE1 genes), 82 (DE2 genes), and 42 (cDE genes) respectively. The rest of the algorithm works on these sets of DE genes only depending on the type of network (Network 1/ Network 2/ Common Network) in consideration.

In Figure 2.4, the distribution of the cTOP values over the 100 equidistant threshold levels (based on z-score of GTOM mapping) are given for all the three kinds of networks, mentioned above. From this figure it can be seen that the best threshold values neither reside on the extreme left (minimum) nor toward the extreme right (maximum) side of the distribution; rather tends to be an intermediate value that is slightly skewed toward the right hand side of the distributions. For the first network or Network 1, the threshold values are distributed within 0.002 to 0.0236 and within 0.003 to 0.0273 in the two conditions respectively. The graphical bar plot depicts 0.0170 and 0.0197 to be the best intelligent threshold pair for the first network or Network 1 comprising samples from control and day o stage of influenza. Similarly, for the second network or Network 2, the threshold values are distributed from 0.002 to 0.0215 and from 0.002 to 0.0248 in control and day 6 stage of influenza respectively. The proposed algorithm in this case yields 0.0131 and 0.0151 to be the best intelligent threshold pair for this network. Following this thought, in the Common Network, the algorithm finds 0.0145 to be the best for both the conditions, where the thresholds are distributed from 0.002 to 0.0204 in both day 0 and day 6 stages of the gene expression profiles.

In Figure 2.5, corresponding to the best threshold pair (found above), the gene specific TOP score is depicted for each DE gene associated in developing the differential interactive structure across conditions in every network type. From here, it is also evident that the number of DE genes possessing statistically significant differential connectivity is much higher in the Common Network compared to the first network or Network 1. The same interpretation is applicable for the first network or Network 1 compared to the second network or Network 2.

To validate the graphical results obtained above, physical network wise validation has been done making use of BioLayout tool [**75**]. The same has been shown in Figures 2.6, 2.7, and 2.8 respectively for first network or Network 1, second network or Network 2, and Common Network. From these figures, it is clear that the DE genes with altered connectivity across conditions exhibit low TO values, while genes with consistent connectivity possess high TO values. In figure 2.6, the different connectivity properties are quite clear considering the DE1 genes namely ELL2, IGLV6-57, IGKC, HIST1H3J, DTL, and Chr2:89629867-89630178 from the first network or Network 1 possessing TOP values of 0, 0, 0, 0, 0, and 0.05, respectively. Similarly, in figure 2.7, the variation of connectivity is pretty evident considering the DE2 genes namely CEACAM6, SUCNR1, MPO, GGH, MS4A3, TDRD9, CLEC5A, LCORL, and ANLN possessing TOP values of 0.051, 0.293, 0.056, 0.316, 0.267, 0.131, 0.043, 0, and 0.17, respectively from the second network or Network 2. In the second network the DE2 genes HBZ, RAB13, and ANKRD22 have weak connectivity in both conditions and possess comparatively higher TOP values of 0.423, 0.56, and 0.47, respectively. In the context of the Common Network, as shown in Figure 2.8, the cDE genes C19orf59, HIST1H2AB, BUB1, HIST1H1B, and CD177 exhibit stringent differential connectivity across conditions possessing TOP value equal to 0.





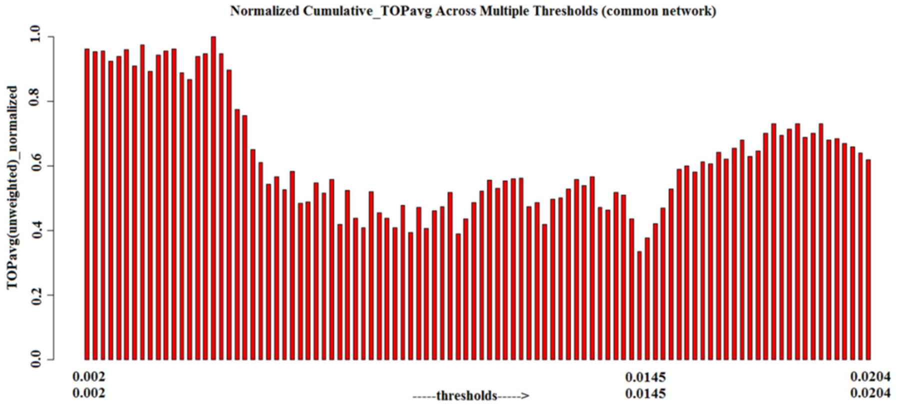


Figure 2.4: Plot of normalized cTOP value corresponding to each threshold pair in first network, second network, and common network

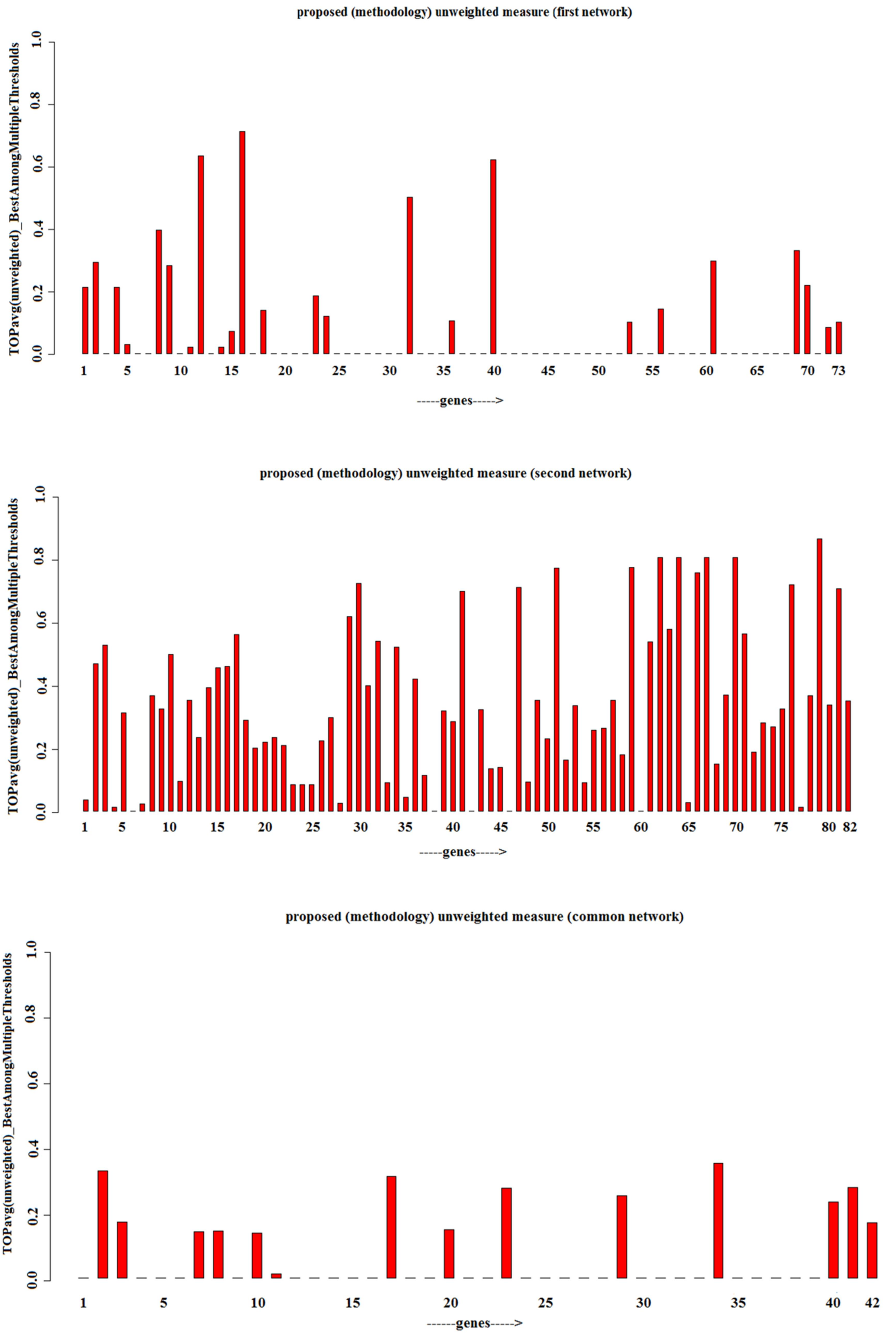


Figure 2.5: Plot of TOP score for each DE gene using the best threshold pairs across first network, second network, and common network

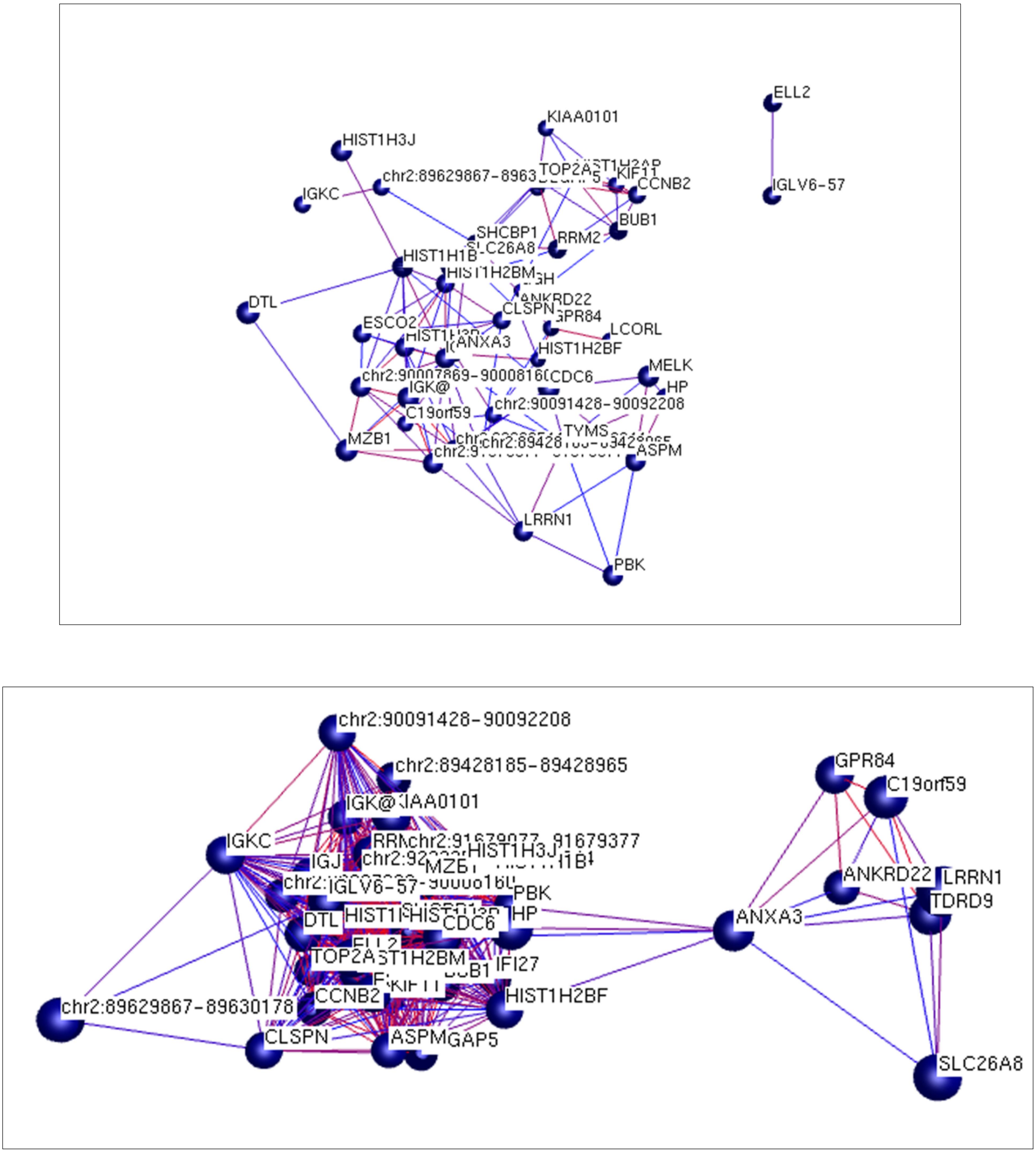


Figure 2.6: Physical realization of first network in control (top panel) and diseased (bottom panel) conditions

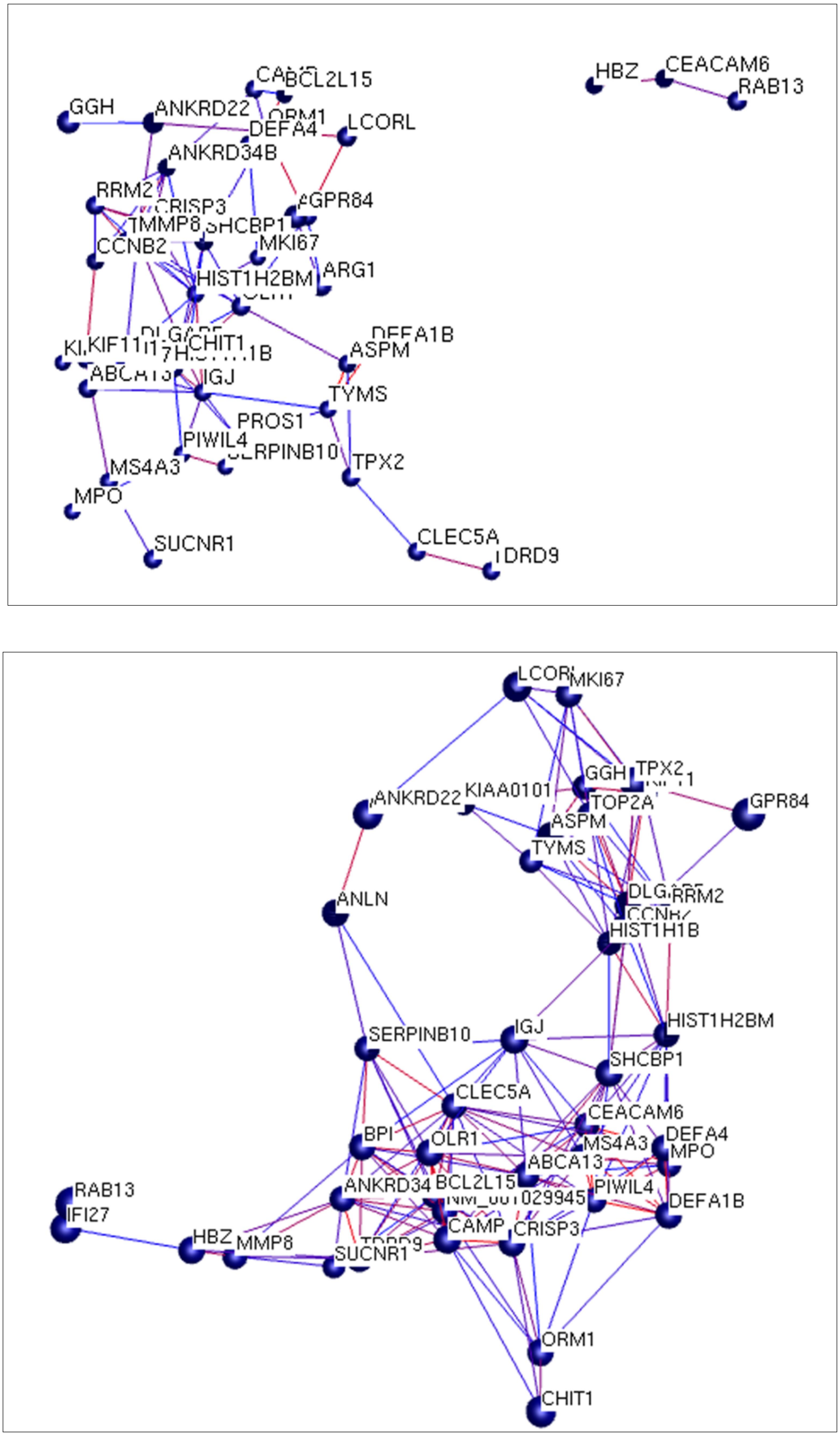


Figure 2.7: Physical realization of second network in control (top panel) and diseased (bottom panel) conditions

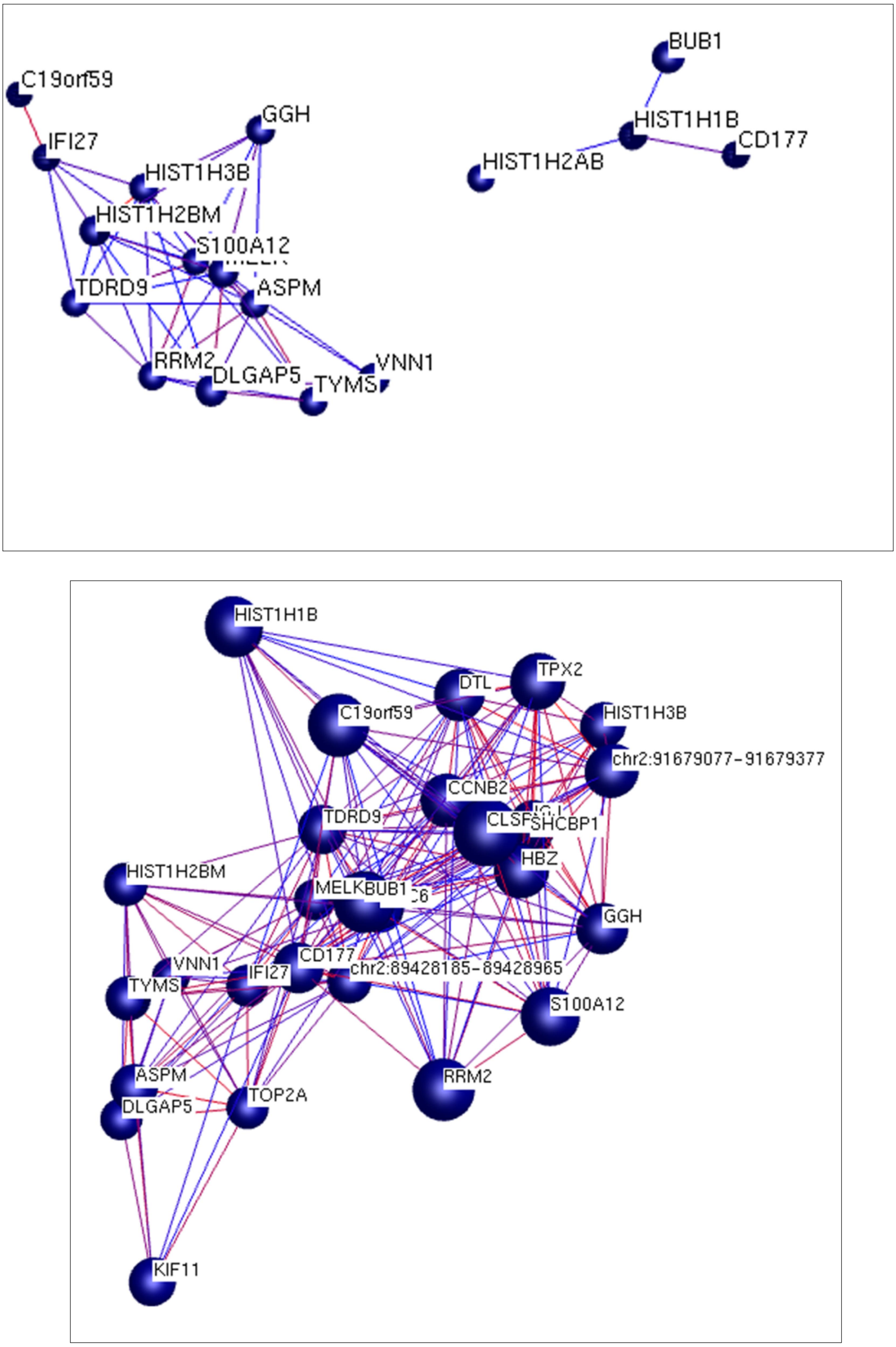


Figure 2.8: Physical realization of common network in day 0 (top panel) and day 6 (bottom panel) of influenza

2.4.4 Discussion: The search for DE genes possessing significantly strong differential connectivity across conditions guided by the topological overlap approach was primarily based on direct gene to gene interaction pattern, a concept which got extended in this work making use of an additional metric, GTOM. Through this metric, not only the direct interactions are being highlighted but the interactions involving shared neighbour genes are also taken into consideration. Hence, the differential connectivity pattern can be studied on a wider perspective. In this work, a two phase filtering strategy has been utilized. The first phase of filtering is executed framing the Fisher transformed GTOM interaction outcomes considering a series of equidistant thresholds. Here, among the 100 equidistant thresholds which have been considered, the best threshold value (as shown in Figure 2.4) is of the order of 10^{-4}. It is crucial to mention that further significant improvement in the best threshold value by considering greater than 100 thresholds is less likely to occur and inevitably will increase the time and space complexity of the proposed algorithm. Accordingly, a maximal restriction of 100 equidistant thresholds is considered. Computation of the TO measure corresponding to each such threshold defines the second phase of filtering. Based upon this TO result a significance testing is conducted, and finally by taking an average of the TO and p-value, TOP is calculated for each gene.

The change in network topology figured out by this approach can be attributed to the change in co-expression across different conditions. Thus it can be claimed that Differential Co-Expression (DCE) based analysis, applied on network topology based problems can help to identify biologically important genes. In this regard, it is beneficial to take help of some statistical properties like specificity and reproducibility [**18**] which could be very handful while comparing various algorithms depicting the point of differential connectivity in gene regulations.

2.5 Conclusion

The basis of this chapter lies on finding the DE genes from the most variant gene expression profiles of any given microarray/RNA-seq data. These DE genes can be found making use of statistically popular techniques like DEGseq or via some technique (here qtDE) which additionally incorporates the physical traits of an organism for attaining this objective. Next issue of importance is the regulation of these DE genes by the TF genes which are the responsible factors for protein complex generation binding to the promoter region of the target DE genes. However, this regulation demands some specific architecture, i.e. either individual or collaborative regulation, addressing the specific need of any living cell, as per the environmental constraint or any other external perturbation. Hence, the matter stands dynamic and involvement of linear as well non-linear regulatory controls may be of importance to understand the significance of the number of regulatory TF genes and nature of such regulatory aspects (direct or indirect regulation). The apprehension of the true dynamicity in TF to DE gene regulatory control depends on the extent of differential regulation across multiple states of any living cell. In this regard, the significance or importance of any source TF or target DE gene can be explored further along with development of different forms of regulatory measures (via linear / non-linear techniques), effectively helping in the development of TF to DE gene regulatory networks.

2.6 References

[**1**] E. Kuhn, “From library screening to microarray technology: Strategies to determine gene expression profiles and to identify differentially regulated genes in plant”, Annals of Botany, 87(2), 139-155, Feb 2001, https://doi.org/10.1006/anbo.2000.1314

[**2**] S. Dudoit, Y.H. Yang, M.J. Callow, T.P. Speed, “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”, Statistica Sinica, 12,111-39, 2002

[**3**] R. Gottardo, A.E. Raftery, K.Y. Yeung, R.E. Bumgarner, “Bayesian robust inference for differential gene expression in microarrays with multiple samples”, Biometrics,62(1), 2006, https://doi.org/10.1111/j.1541-0420.2005.00397.x

[**4**] L. Wang, Z. Feng, X. Wang, X. Wang, X. Zhang, “DEGseq: an R package for identifying differentially expressed genes from RNA- seq data”, Bioinformatics, 26(1),136-138, January 2010, https://doi.org/10.1093/bioinformatics/btp612

[**5**] S. Horvath, J. Dong, “Geometric interpretation of gene co-expression network analysis”, PLoS Computational Biology, August 2008, https://doi.org/10.1371/journal.pcbi.1000117

[**6**] C. Cheng, Y. Fu, L. Shen, and M. Garstein, "Identification of yeast cell cycle regulated genes based on genomic features", BMC Systems Biology, 7, Article number: 70, July 2013, https://doi.org/10.1186/1752-0509-7-70

[**7**] N. Banerjee and M.Q. Zhang, "Identifying cooperativity among transcription factors controlling the cell cycle in yeast ", Nucleic Acids Research, 3l(23), 7024-7031, December 2003, https://doi.org/10.1093/nar/gkg894

[**8**] C.J. Zopf, K. Quinn, J. Zeidman, and N. Mahesri, “Cell-cycle Dependence of Transcription Dominates Noise in Gene Expression", PLoS Computational Biology, 9(7): e1003161 , July 2013, https://doi.org/10.1371/journal.pcbi.1003161

[**9**] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: Mutual information, correlation, and model based indices," BMC Bioinformatics, 13, Article no. 328, December 2012, https://doi.org/10.1186/1471-2105-13-328

[**10**] L. Paninski, "Estimation of entropy and mutual information," Neural Computation, 15(6), 1191-1253, June 2003, https://doi.org/10.1162/089976603321780272

[**11**] J. Hausser and K. Strimmer, "Entropy inference and the James-Stein estimator, with application to Nonlinear Gene Association Network", The Journal of Machine Learning Research, 10, 1469-1484, December 2009, https://dl.acm.org/doi/10.5555/1577069.1755833

[**12**] G.A. Milier. "Note on the basis of information estimates", In. H. Quastler, editor, Information Theory in Psychology II-B, pp. 95-100.Free press, Glencoe, IL, 1955

[**13**] A. Agresti and D.B. Hitchcock, "Bayesian inference for categorical data analysis", Statistical Methods and Applications, 14, 297-330, December 2005, https://doi.org/10.1007/s10260-005-0121-y

[**14**] I. Nemenman, F. Shaffe, and W. Bialek. "Entropy and Inference revisited", In T.G. Dietterich, S. Becker, and Z. Ghahramani editors, Advances in Neural Information Processing Systems, 14, 471-478, Cambridge, MA, MIT press, 2002, https://dl.acm.org/doi/10.5555/2980539.2980601

[**15**] A. Chao and T-J Shen, "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample", Environmental and Ecological Statistics, 10, 429-443, December 2003, https://doi.org/10.1023/A:1026096204727

[**16**] Y.W. Chang, C.J. Hsieh, K-W Chang, M. Ringgaard, C-J Lin, "Training and testing low-degree polynomial data mappings via linear SVM", The Journal of Machine Learning Research, 11, 1471-1490, August 2010, https://dl.acm.org/doi/10.5555/1756006.1859899

[**17**] W.K. Chen. Feedback, Nonlinear and Distributed Circuits. CRC Press, 3rd Edition, pp. 9-20, 2009

[**18**] M. Bockmayr, F. Klauschen, B. Györffy, C. Denkert, and J. Budczies, "New network topology approaches reveal differential correlation patterns in breast cancer", BMC Systems Biology,7, Article No.78, August 2013, https://doi.org/10.1186/1752-0509-7-78

[**19**] M. Ray, W.X. Zhang, "Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks", BMC Systems Biology, 4, Article No.136, October 2010, https://doi.org/10.1186/1752-0509-4-136

[**20**] M.T.J. Johnson, M. Vellend, J.R. Stinchcombe, "Evolution in plant populations as a driver of ecological changes in arthropod communities", Philos Trans R Soc Lond B Biol Sci.,364(1523),1593-1605, June 2009, https://doi.org/10.1098/rstb.2008.0334

[**21**] J.H. Seo, Q. Li, A. Fatima, A. Eklund, Z. Szallasi, K. Polyak, A.L. Richardson, M.L. Freedman, "Deconvoluting complex tissues for expression quantitative trait locus-based analyses", Philos Trans R Soc Lond B Biol Sci., 368(1620), June 2013, https://doi.org/10.1098%2Frstb.2012.0363

[**22**] T.F. Fuller, A. Ghazalpour, J.E. Aten, T.A. Drake, A.J. Lusis, S. Horvath, "Weighted gene co-expression network analysis strategies applied to mouse weight", Mamm Genome, 18(6-7), 463-472, July 2007, https://doi.org/10.1007/s00335-007-9043-3

[**23**]Available at http://www.genetics.ucla.edu/labs/ horvath/CoexpressionNetwork/MouseWeight/ (Last accessed on February, 2014)

[**24**] P. Langfelder, S. Horvath, "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics, 9, Article No.559, December 2008, https://doi.org/10.1186/1471-2105-9-559

[**25**] M. Kanehisa, S. Goto, "KEGG: Kyoto encyclopaedia of genes and genomics", Nucleic Acids Research, 28(1), 27-30, January 2000, https://doi.org/10.1093/nar/28.1.27

[**26**] A. Gelman, "P values and statistical practice", Epidemiology, 24(1), 69-72, January 2013, https://doi.org/10.1097/ede.0b013e31827886f7

[**27**] Y. Zhang, X. Zou, Y. Ding, H. Wang, X. Wu, B. Liang, "Comparative genomics and functional study of lipid metabolic genes in caenorhabditis elegans". BMC Genomics, 14, Article No.164, March 2013, https://doi.org/10.1186/1471-2164-14-164

[**28**] Y. Choi, C.G. Hur, T. Park, "Induction of olfaction and cancer-related genes in mice fed a high-fat diet as assessed through the mode of action by network identification analysis", PLoS One, 8(3):e56610, March 2013, https://doi.org/10.1371/journal.pone.0056610

[**29**] A. Oshimoto, Y. Wakabayashi, A. Garske, R. Lopez, S. Rolen, M. Flowers, N. Arevalo, D. Restrepo, "Potential role of transient receptor potential channel M5 in sensing putative pheromones in mouse olfactory sensory neurons", PLoS One, 8(4):e61990, April 2013, https://doi.org/10.1371/journal.pone.0061990

[**30**] A. Molinas, G. Sicard, I. Jakob, "Functional evidence of multidrug resistance transporters (MDR) in rodent olfactory epithelium", PLoS One, 7(5): e36167, May 2012, https://doi.org/10.1371%2Fjournal.pone.0036167

[**31**] H.A. Valencia, S. Berdnikovs, J.M. Cook-Mills, "Mechanisms for vascular cell adhesion molecule-1 activation of ERK1/2 during leukocyte transendothelial migration", PLoS One, 6(10): e26706, October 2011, https://doi.org/10.1371%2Fjournal.pone.0026706

[**32**] N. Sawangjaroen, K. Sawangjaroen, P. Poonpanang, "Effects of piper longum fruit, piper sarmentosum root and quercus infectoria nut gall on caecal amoebiasis in mice", Journal of Ethnopharmacology, 91(2-3), 357-360, April 2004, https://doi.org/10.1016/j.jep.2004.01.014

[**33**] A.K. Pandey, S. Somvanshi, V.P. Singh, "Focal adhesion kinase: An old protein with new roles", Online Journal of Biological Sciences, 12(1), 11-14, March 2012, https://doi.org/10.3844/ojbsci.2012.11.14

[**34**] A.M. Carter, "Complement activation: An emerging player in the pathogenesis of cardiovascular disease", Scientifica (Cairo), 2012: 402783, December 2012, https://doi.org/10.6064/2012/402783

[**35**] D. Diamanti, E. Mori, D. Incarnato, F. Malusa, C. Fondelli, L. Magnoni, G. Pollio, "Whole gene expression profile in blood reveals multiple pathways deregulation in R6/2 mouse model", Biomarker Research, 1, Article No. 28, October 2013, https://doi.org/10.1186/2050-7771-1-28

[**36**] L.A. Thoren, G.A. Norgaard, J. Weischenfeldt, J. Waage, J.S. Jakobsen, I. Damgaard, F.C. Bergstrom, A.M. Blom, R. Borup, H.C. Bisgaard, Bo T. Porse, "UPF2 is a critical regulator of liver development, function and regeneration", PLoS One, July 2010, https://doi.org/10.1371/journal.pone.0011650

[**37**] I. Cruz, J. Nieto, J. Moreno, C. Canavate, P. Desjeux, J. Alvar, " Leishmania/HIV co-infections in the second decade", The Indian Journal of Medical Research, 123(3), 357- 388, March 2006, https://pubmed.ncbi.nlm.nih.gov/16778317/

[**38**] A. Patil, Y. Kumagai, K.C. Liang, Y. Suzuki, K. Nakai, " Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response", PLoS Computational Biology, 9(11):e1003323, November 2013, https://doi.org/10.1371/journal.pcbi.1003323

[**39**] E. Blanchet, J.S. Annicotte, L. Fajas, "Cell cycle regulators in the control of metabolism", Cell Cycle, 8(24), 4029-4031, December 2009, https://doi.org/10.4161/cc.8.24.10110

[**40**] Z.P. Xu, E.F. Wawrousek, J. Piatigorsky, "Transketolase haploinsufficiency reduces adipose tissue and female fertility in mice", Molecular and Cellular Biology, 22(17), 6142-6147, September 2002, https://doi.org/10.1128/mcb.22.17.6142-6147.2002

[**41**] M.K. Jha, S. Jeon, K. Suk, "Pyruvate dehydrogenase kinases in the nervous system: Their principal functions in neuronal-glial metabolic interaction and neuro-metabolic disorders", Current Neuropharmacology, 10(4), 393-403, December 2012, https://doi.org/10.2174/157015912804143586

[**42**] S. Krappmann, G.H. Braus, "Nitrogen metabolism of aspergillus and its role in pathogenicity", Medical Mycology, 43 Suppl 1:S31-40, May 2005, https://doi.org/10.1080/13693780400024271

[**43**] C. Friere-de-Lima, L.M. Pecanha, G.A. Dos Reis, "Chronic experimental chagas disease: functional syngeneic T-B-cell cooperation in vitro in the absence of an exogenous stimulus", Infection and Immunity, 64(7), 2861–2866, July 1996, https://doi.org/10.1128/iai.64.7.2861-2866.1996

[**44**] K. Racke, M. Warnken, "L-arginine metabolic pathways", The Open Nitric Oxide Journal, 2, 9-19, 2010

[**45**] J. Menche, A. Sharma, M.H. Cho, R.J. Mayer, S.I. Rennard, B. Celli, B.E. Miller, N. Locantore, R. Tal-Singer, S. Ghosh, C. Larminie, G. Bradley, J.H. Riley, A. Agusti, E.K. Silverman, A-L. Barabasi, "A diVIsive Shuffling Approach (VIStA) for gene expression analysis to identify subtypes in chronic obstructive pulmonary disease".

BMC Systems Biology, 8 (Suppl 2):S8, March 2014, https://doi.org/10.1186/1752-0509-8-s2-s8

[**46**] S.A. Ahmed, W.J. Penhale, N. Talal, “Sex hormones, immune responses, and autoimmune diseases: Mechanisms of sex hormone action”, The American Journal of Pathology, 121(3), 531–551, December 1985

[**47**] L.A. McKee, H. Chen, J.A. Regan, S.M. Behunin, J.W. Walker, J.S. Walker, J.P. Konhilas, “Sexually dimorphic myofilament function and cardiac troponin I phosphospecies distribution in hypertrophic cardiomyopathy mice”, Archives of Biochemistry and Biophysics, 535(1), 39-48, https://doi.org/10.1016/j.abb.2012.12.023

[**48**] S. Cvitic, M.S. Longtine, H. Hackl, K. Wagner, M.D. Nelson, G. Desoye, U. Hiden, “The human placental sexome differs between trophoblast epithelium and villous vessel endothelium”, PLOS One, October 2013, https://doi.org/10.1371/journal.pone.0079233

[**49**] H. Jeong, “Altered drug metabolism during pregnancy: Hormonal regulation of drug-metabolizing enzymes”, Expert Opinion on Drug Metabolism & Toxicology, 6(6), 689–699, June 2010, https://doi.org/10.1517/17425251003677755

[**50**] K. Li, W. Xu, Q. Guo, Z. Jiang, P. Wang, Y. Yue, S. Xiong, “Differential macrophage polarization in male and female BALB/c mice infected with coxsackievirus B3 defines susceptibility to viral myocarditis”, Circulation Research, 105(4), 353-364, August 2009, https://doi.org/10.1161/circresaha.109.195230

[**51**] A. Moskalev, M. Shaposhnikov, A. Snezhkina, V. Kogan, E. Plyusnina, D. Peregudova, N. Melnikova, L. Uroshlev, S. Mylnikov, A. Dmitriev, S. Plusnin, P. Fedichev, A. Kudryavtseva, “Mining gene expression data for pollutants (Dioxin, Toluene, Formaldehyde) and low dose of gamma irradiation”, PLOS One, January 2014, https://doi.org/10.1371/journal.pone.0086051

[**52**] A.T. Kwon, H.H. Hoos, R. Ng, “Inference of transcriptional regulation relationships from gene expression data”, Bioinformatics, 19(8), 905-912, May 2003, https://doi.org/10.1093/bioinformatics/btg106

[**53**] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hestie, R. Tibshirani, D. Botstein, R.B. Altman, "Missing value estimation methods for DNA microarrays", Bioinformatics, 17(6), 520-525, June 2001, https://doi.org/10.1093/bioinformatics/17.6.520

[**54**] http://genome-www.stanford.edu/ & http://www.yeastgenome.org/ (first one last accessed in May 2019 and the second one in April 2023)

[**55**] B.M. Ryan, *et al.*, “Germline variation in NCF4, an innate immunity gene, is associated with an increased risk of colorectal cancer”, International Journal of Cancer, 134(6), 1399-1407, March 2014, https://doi.org/10.1002/ijc.28457

[**56**] W-S. Wu, and W-H.Li, "Systematic identification of yeast cell-cycle transcription factors using multiple data source", BMC Bioinformatics, 9, Article No.522, December 2008, https://doi.org/10.1186/1471-2105-9-522

[**57**] M.C. Teixeira *et al.*, "The YEASTRACT Database : A tool for the analysis of Transcriptional Regulatory Association in Saccharomyces cerevisiae" , Nucleic Acids Research, 34, D446-D451 , January 2006, https://doi.org/10.1093/nar/gkj013

[**58**] J.M. Vaquerizas, S.K. Kummerfeld, S.A. Teichmann, N.M. Luscombe, “A census of human transcription factors: Function, expression and evolution”, Nature Reviews Genetics, 10, 252-263, April 2009, https://doi.org/10.1038/nrg2538

[**59**] P.T. Monteiro, *et al.*, “YEASTRACT-DISCOVERER: New tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae”, Nucleic Acids Research, 36, D132-136, January 2008, https://doi.org/10.1093/nar/gkm976

[**60**] A.M. Wenger, *et al.*, “PRISM offers a comprehensive genomic approach to transcription factor function prediction”, Genome Research, 23(5), 889-904, May 2013, https://doi.org/10.1101/gr.139071.112

[**61**] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, J.V. Helden, J.B. Demoulin, “Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data”, Nucleic Acids Research, 38(11):e120, June 2010, https://doi.org/10.1093/nar/gkq149

[**62**] S.D. Talia, H. Wang, J.M. Skotheim, A.P. Rosebrock, B. Futcher, F.R. Cross, "Daughter-Specific Transcription Factors Regulate Cell Size Control in Budding Yeast", PLoS BIOLOGY, 7(10):e1000221, October 2009, https://doi.org/10.1371/journal.pbio.1000221

[**63**] W. Zheng, H. Zhao, E. Mancera, L.M. Stelnmetz, M. Snyder, "Genetic Analysis of Variation in Transcription Factor Binding in Yeast", Nature, 464(7292), 1187-1191, April 2010, https://doi.org/10.1038/nature08934

[**64**] A.P. Capaldi *et al.*,“Structure and Function of a Transcriptional Network Activated by the MAPK Hogl”, Nature Genetics, 40(11), 1300-1306, November 2008, https://doi.org/10.1038/ng.235

[**65**] H.D. Madhani, T. Galitski, E.S. Lander, G.R. Fink, “Effectors of a developmental mitogen-activated protein kinase cascade revealed by expression signatures of signaling mutants”, PNAS, 96(22), 12530-12535, October 1999, https://doi.org/10.1073/pnas.96.22.12530

[**66**] J. Reimand, J.M. Vaquerizas, A.E. Todd. J. Vilo, N.M.Luscombe. “Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets” , Nucleic Acids Research, 38(14), 4768-4777, August 2010, https://doi.org/10.1093/nar/gkq232

[**67**] G. Chua *et al.*, “Identifying transcription factor functions and targets by phenotypic activation”, PNAS, 103(32), 12045-12050, August 2006, https://doi.org/10.1073/pnas.0605140103

[**68**] AB. Sanz *et al.* , “Chromatin remodeling by the SWI/SNF complex is essential for transcription mediated by the yeast cell wall integrity MAPK pathway”, Molecular Biology of the Cell, 23(14), 2805-2817, July 2012, https://doi.org/10.1091/mbc.e12-04-0278

[**69**] T.I. Lee *et al.*, "Transcriptional regulatory networks in Saccharomyces cerevisiae", Science, 298(5594), 799-804, October 2002, https://doi.org/10.1126/science.1075090

[**70**] S. Chatterjee and A. S. Hadi, Regression Analysis by Example, 4th ed., Wiley, 2006, ch.2, pp. 21-45

[**71**] D.Nitsch *et al.* "PINTA: a web server for network-based gene prioritization from expression data", Nucleic Acids Research, 39:W334-W338, May 2011, https://doi.org/10.1093/nar/gkr289

[**72**] I.W. Taylor *et al.* "Dynamic modularity in protein interaction networks predicts breast cancer outcome", Nature Biotechnology, 27, 199-204, February 2009, https://doi.org/10.1038/nbt.1522

[**73**] A.M. Yip, and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure", BMC Bioinformatics, 8, Article No.22, January 2007, https://doi.org/10.1186/1471-2105-8-22

[**74**] J.E. Berdal *et al.* "Excessive innate immune response and mutant D222G/N in severe A (H1N1) pandemic influenza", The Journal of Infection, 63(4), 308-316, October 2011, https://doi.org/10.1016/j.jinf.2011.07.004

[**75**] A. Theocharidis, S.v. Dongen, A.J. Enright, T.C. Freeman, "Network visualization and analysis of gene expression data using BioLayout Express3D", Nature Protocols, 4(10), 1535-1550, October 2009, https://doi.org/10.1038/nprot.2009.177

Chapter 3 Differential Gene Regulation in Neurodegenerative Disorders

Related Publications

[**1**] A. Majumder and M. Sarkar, "Exploring Different Stages of Alzheimer's Disease through Topological Analysis of Differentially Expressed Genetic Networks", International Journal of Computer Theory and Engineering, volume 6, issue 5, pages 386-391, October 2014.
DOI: 10.7763/IJCTE.2014.V6.895
[**2**] A. Majumder and M. Sarkar, "Dissimilar Regulatory Actions Between Neurodegenerative Disease Pairs Through Probablistic Differential Correlation", P. Deiva Sundari et al. (eds.), Proceedings of 2nd International Conference on Intelligent Computing and Applications, Advances in Intelligent Systems and Computing, volume 467, pp. 59-74, October 2016, Springer, Singapore.
http://dx.doi.org/10.1007/978-981-10-1645-5_6

3.1 Introduction

Differential gene regulation under varied constraints can predict the causal factors in any diseased cell compared to the normal condition through identification of certain target genes showing dysfunctional regulatory control in the presence of protein generating complex genes or transcription factor (TF) genes. These targets mostly exhibit significant differential expression (DE) profiles possessing extreme dissimilar type of regulations. Hence, the control of such target DE genes is being guided through the concept of differential co-expression. In this perspective, understanding the extremeness of differential regulation as well as the crucial role of the concerned DE genes at the onset and subsequent spreading of any disease can be utilized to identify the therapeutic targets. In this regard, experimenting on various types of neurodegenerative disorders is helpful to predict the form of treatment applicable at different stages of growth of one or more such disorders.

Alzheimer's disease is one of the most common forms of neurodegenerative disorders. It is indeed a medically popular form of dementia characterized by the progressive degeneration of brain cells. While the exact cause of Alzheimer's disease is still under investigation, various research contents have identified certain pathological features, such as the accumulation of amyloid plaques and neurofibrillary tangles, which are believed to contribute to the development and progression of the disease. The development of amyloid plaques, composed of abnormal protein fragments called beta-amyloid, and neurofibrillary tangles, formed by twisted fibers of a protein called tau, disrupts the normal functioning of neurons and leads to their eventual death. These changes result in the loss of connections between nerve cells and the subsequent impairment of cognitive functions, including memory, thinking, and behaviour. While aging is considered a significant risk factor for Alzheimer's disease, it is important to note that the onset of the disease typically occurs in people over the age of 65 [**1**,**2**]. However, there is also a less common early-onset form of Alzheimer's disease that can manifest between the ages of 30 and 60. The exact causes of Alzheimer's disease are still being investigated, and researchers have proposed various factors that may contribute to its development. Some studies have suggested a potential involvement of calcium-dependent potassium (K+) channels in platelets [**3**] and ryanodine receptors (RyRs), which are intracellular calcium release channels [**4**], as possible underlying mechanisms. Additionally, dietary factors, such as proteins, and hormones have also been implicated in the disease process. It is important to note that while these factors have been proposed as potential contributors to Alzheimer's disease, the disease is complex, and its development is likely influenced by a combination of genetic [**5-8**], environmental, and lifestyle[**9,10**] factors. The research output in this chapter aims to better understand the risk factors in relation to the nature of differential regulatory links associated with the DE genes involved in the development of Alzheimer's disease considering a combination of genetic and environmental constraints.

There are in effect other forms of neurodegenerative disorders in addition Alzheimer's disease. Though these other forms are not generally observed on a wide scale, functionally these maintain both similar and dissimilar regulatory properties compared to the widely popular Alzheimer's disease. The other forms are like Huntington's disease, Amyotrophic lateral sclerosis, multiple sclerosis, Schizophrenia, and Parkinson's disease. In this regard, pairing up these disorders based on some level of similarity [**11-17**] does help in validating the distinct gene regulatory properties specific to a disorder or a pair of disorders. In other words, identification of significantly paired gene networks possessing extremely different regulatory actions across multiple types of neurodegenerative disorders may help in revealing the biologically pivotal role of DE genes holding dynamic association in these disorders.

3.2 Assessing differential regulation in Alzheimer's disease

Alzheimer's disease (AD) normally progresses in the brain through incipient (Braak stages III-IV), mild or moderate (Braak stages IV-V), and severe (Braak stages V-VI) stages [**18**]. Focussed research in this domain handles differential topology prediction in gene regulation networks across varied regions of the brain [**19**]. This does help in understanding the severity of the disease in an elderly individual because AD is most commonly observed in senior citizens. However, it is equally important to explore the mode of progress of this disease through the various stages depicted above and take or propose corrective measures as per need. Here, the progression issue is addressed apprehending the role of the DE genes and their connectivity through the different stages of development.

3.2.1 The basic findings: Unveiling the inter dependency of all concerned DE genes in a regulatory network through a weighted topological overlapped (TO) approach helps in understanding the role of the DE genes in the generation and spreading of AD. The TO score of each DE gene goes through a significance testing phase comprising a permutation/T-test (yields p-value) followed by ranking the genes based on lower TO score as well as p-value. The biological significance of the highest ranked genes when compared with a popular technique [**20**] involved in ranking differential hubs in gene regulations shows better performance in the identification of DE genes participating in the crucial pathways pertaining to the progress of AD. For example, this work is capable of discovering DE genes responsible for modulation of the membrane potential involved in the movement of neurological signals. Again, there are DE genes which have been found responsible for facilitating the movement of extracellular fluids to and from tissues within a multicellular organism. This research output validates the hypothesis that the expansion of the Cerebral Spinal Fluid (CSF) space reduces the turnover rate of CSF, which in turn compromises its ability to act as a sink for clearing harmful metabolites, such as amyloid, from the Central Nervous System (CNS). This compromised CSF dynamics can have a significant negative impact on the interstitial environment of neurons [**21**], particularly as individuals age.

3.2.2 Methodology: The explicit flowchart given in Figure 3.1 below indicates both weighted and un-weighted TO dynamics incorporated in this differential regulatory approach.

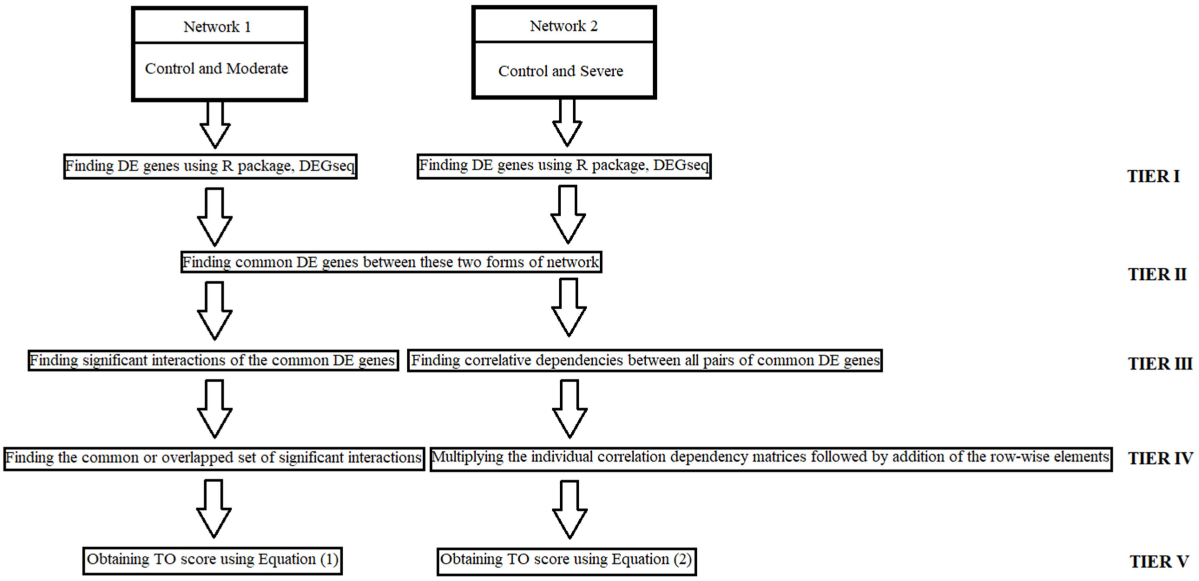

Network 1
Control and Moderate
Network 2
Control and Severe
Finding DE genes using R package, DEGseq
Finding DE genes using R package, DEGseq
TIER I
Finding common DE genes between these two forms of network
TIER II
Finding significant interactions of the common DE genes
Finding correlative dependencies between all pairs of common DE genes
TIER III
Finding the common or overlapped set of significant interactions
Multiplying the individual correlation dependency matrices followed by addition of the row-wise elements
TIER IV
Obtaining TO score using Equation (1)
Obtaining TO score using Equation (2)
TIER V

Figure 3.1: Flowchart of weighted and un-weighted TO analysis

Equation (1): $\mathrm{TO_i} = \frac{X_i \cap Y_i}{\max(X_i\ ,\ Y_i)}$ **Equation (2)**: $\mathrm{TO_i} = \frac{\min(A)}{\max(B_i\ ,C_i)}$

In the above figure, depicting the flowchart of the proposed process, the right wing starting from TIER I to TIER V is about the weighted TO analysis towards differential regulation. Similar comment is applicable to the left wing, but corresponding to un-weighted TO analysis.

In this implementation extended versions of the algorithms followed in [**19,22**] is gone through. The un-weighted approach is primarily dealing with the some significant DE-DE gene interactions, accordingly handling with a sparse differential dependency or regulatory matrix. However, the weighted counterpart is more into exploring all the DE-DE gene interactions, and thus bringing to limelight the importance of all differential connectivities present in a gene regulatory network. In this regard, the various TIERs of the flowchart can be explained as follows.

TIER I: From the Alzheimer's data present across three conditions, namely control, moderate, and severe, the DE genes between control and moderate in one hand and control and severe on the other can be computed making use of the suitable R package, DEGseq. These two sets of DE genes can be used to define two different regulatory networks depicting the variation in the state of growth of the disease.

TIER II: The weighted and un-weighted analysis both have been carried out on the common set of DE genes. Hence, it stands crucial to find the intersection of DE gene sets obtained above and have the common DE genes.

TIER III: At this stage, the differential co-expression of these common DE genes is computed in each network stated above. This means the difference of the linear correlated dependence between control and moderate states (thus forming the elements of the adjacency matrix for Network 1) OR control and severe states (thus forming the elements of the adjacency matrix for Network 2) corresponding to any pair of the common DE genes is found. Post this operation it is required to decide which wing of the flowchart is to be followed.

If the left wing is taken, then it is about un-weighted TO analysis. Hence, at this step, for moving forward with un-weighted approach, the PCIT technique [**23**] is followed to understand the significant gene-gene direct interactions in the presence of any third gene entity. The outcome of PCIT leads to the formation of a sparse adjacency matrix individually for Network 1 and Network 2 with an element value or the interaction weightage equalling 1 to convey significant interaction between the two concerned DE genes; otherwise the interaction weightage component equals 0.

On the other hand, following the right wing of the flowchart, the weighted analysis at this stage is all about the initial differential correlation or adjacency matrix formation for each of the networks, Network 1 and Network 2.

TIERs IV and V: In the un-weighted counterpart of the proposed approach, as per the left wing of the flowchart, the TO score is computed making use of equation (1). Through this it is possible to yield the common set of differentially interacting DE genes as well as the disjoint interacting capacity (considering Network 1 and Network 2) of the referred DE gene. Higher the disjoint interacting capacity or lower the TO score is, the better the DE gene contributes to differential connectivity required for the process of growth of the disease.

In the weighted segment of the proposed flowchart, as per the right wing, to design the TO score following equation (2), element wise multiplication of the adjacency matrices obtained for Networks 1 and 2 is conducted followed by addition of the row elements of the resultant matrix. Hence, for n common DE genes (obtained initially at TIER I), the addition stated above yields a matrix is of size [n×1]. From equation (2), as per the need of the design, it is clear to have the minimum value from the above [n×1] matrix, A. Again, the element wise addition following the i^{th} row of the initially obtained adjacency matrices yields the values B_i and C_i from the two networks. Hence, for a referred DE gene, the TO score, as proposed in the equation, is primarily dependent on the level of differential correlation or adjacency or control of the concerned common DE gene with respect to all the other (n-1) DE genes, thus signifying its topological weightage in each network.

3.2.3 Results: The data [**24**] taken for this research comprises of 54, 675 genes distributed over 30 time profiles. The information contains control, incipient, moderate, and severe stages of growth of Alzheimer's disease. However, the difference in expression level following the time profiles is not at all significant between control and incipient versions. Therefore, the gene expression matrices on which the DEGseq operation has been applied comprises of conditions control and moderate on one hand whereas control and severe on the other. Before finding the DE genes using the above procedure for the two networks (Network 1 and Network 2), necessary pre-processing of the initial data has been done yielding a resultant source gene expression matrix of 10,000 gene entries on which differential expression analysis is being conducted using the R package DEGseq. This matrix corrects the skewed representation of the original matrix following a log-normal distribution and removes the gene expression entries across all time profiles for gene vectors maintaining standard deviations below a certain threshold.

Following TIER II of the proposed flowchart given Figure 3.1, 66 common DE genes are obtained. After having the TO score for each DE gene, the statistical significance is analyzed using random permutation/'t'-test. Post significance treatment, 15 top ranked DE genes (close to 25% of 66) are considered for biological enrichment analysis (Gene Ontology or GO and Kyoto Encyclopaedia of Genes and Genomes or KEGG pathway analysis) [**25,26**]. These top 15 DE genes are not only having low TO score but are statistically significant (low p-value) as well. In the perspective of Alzheimer's disease, the betterment of the proposed approach with respect to a standard differential hub ranking scheme called Diffrank [**20**] can be understood from the outcomes of the biological enrichment analysis enlisted in Tables 3.1, 3.2, 3.3, and 3.4.

Table 3.1: Significant GO terms in proposed method Vs. Diffrank by Weighted TO measure

Proposed Algorithm			Diffrank		
GO-Terms	*p*-value	Genes	GO-Terms	*p*-value	Genes
GO:0046885	0.0346	2 ***trerf1, kynu***	GO:0006569	0.0348	3 ***mmp10, kynu, kcne1***
GO:0032350	0.0346	3 ***kynu, usp28, itgb3***	G0:0046218	0.0348	2 ***kynu, jak1***
GO:0006569	0.0349	2 ***jak1, efcab2***	GO:0002070	0.0352	2 ***cog3, ciao1***
GO:0046218	0.0349	2 ***cog3, efcab2***	GO:0005131	0.0359	1 ***efcab2***
GO:0004718	0.0352	1 ***fancd2***	GO:0002064	0.0359	1 ***arhgap24***

Table 3.2: Significant GO terms in proposed method Vs. Diffrank by Un-weighted TO measure

Proposed Algorithm			Diffrank		
GO-Terms	*p*-value	Genes	GO-Terms	*p*-value	Genes
GO:0015459	0.00095	2 ***kcne1, mmp10***	GO:0030303	0.019	2 ***ciao1, mmp10***
GO:0016247	0.00162	2 ***tp73, kcnmb2***	GO:0018676	0.019	2 ***cyp2c9, itgb3***

GO:0015457	0.00162	3 ***ciao1, kcne1, mmp10***	GO:0036767	0.019	2 ***kcnmb2, mmp10***
GO:0030303	0.0093	3 ***trim14, tp73, cdc20b***	GO:0019113	0.0271	2 ***kcnmb2, plxnd1***
GO:0008076	0.0093	2 ***mmp10, kcnmb2***	GO:0018675	0.0274	3 ***vsp53, tp73, itgb3***

Table 3.3: Significant KEGG pathways in proposed method Vs. Diffrank by Weighted TO measure

Proposed Algorithm			Diffrank		
KEGG Pathways	*p*-value	Genes	KEGG Pathways	*p*-value	Genes
Tryptophan metabolism	0.0220	3 ***kynu, cog3, efcab2***	Jak-STAT signalling pathway	0.0212	3 ***jak1, kcne1, trdv3***
Pancreatic cancer	0.0383	3 ***jak1, itgb3, usp28***	Tryptophan metabolism	0.0220	3 ***csh1, kynu, mmp10***
Leishmaniasis	0.0399	2 ***usp28, efcab2***	Pancreatic cancer	0.0383	2 ***ciao1, cog3***
Arrhythmogenic rightventricular cardiomyopathy	0.0416	2 ***efcab2, cog3***	Leishmaniasis	0.0399	1 ***efcab2***
ECM-receptor interaction	0.0459	2 ***trerf1, mgc3771***			
Hypertrophic cardiomyopathy (HCM)	0.0464	1 ***c20orf78***			
Hematopoietic cell lineage	0.0488	1 ***fancd2***			

Table 3.4: Significant KEGG pathways in proposed method Vs. Diffrank by Un-weighted TO measure

Proposed Algorithm			Diffrank		
KEGG Pathways	*p*-value	Genes	KEGG Pathways	*p*-value	Genes
p53 signaling pathway	0.0253	3 ***tp73, cdc20b, mmp10***	Linoleic acid metabolism	0.0212	2 ***cyp2c9, plxnd1***
Vascular smooth musclecontraction	0.0424	2 ***kcnmb2, efcab2***	Arachidonic acid Metabolism	0.0422	2 ***itgb3, ciao1***
Neurotrophin signalingpathway	0.0460	2 ***kcnmb2, trim14***	Retinol metabolism	0.0472	1 ***kcnmb2***

In each of the Tables (3.1, 3.2, 3.3, 3.4) given above, the number and ORF names of the participating DE genes in the various biologically enriched analyses are mentioned in the column titled ‘Genes’. It is important to mention that the enlisted enrichment of the top 15 genes from both the approaches (proposed vs. Diffrank) is far better than the remaining DE genes which are not a part of the top 15. From the above information, it is also to be noted that the proposed methodology works better compared to Diffrank from the point of KEGG pathway enrichment related to Alzheimer’s disease.

3.2.4 Discussion: The proposed method shows better result because the existing method, Diffrank [**20**], focuses on local (differential connectivity which is the local difference between two networks calculated by the number of genes associated with a particular gene) as well as on global concept (between centrality: which calculate the change in the expression levels of central genes). But as given in [**19**] AD does not affect all the brain regions simultaneously but there are differences in severity of AD across different regions of the brain. Hence, more of localized phenomena than the global one are observed. In general, most of the genetic diseases show this kind of pattern. It [**20**] also shows a problem of controlling the value of a trade-off parameter (λ), thus trying to maintain a balance between local and global connectivity. So the optimality of the result is solely dependent on this parameter, making the problem more parameter driven. Thus the proposed method which does not possess such constraint performs better in recognizing AD through the participation of significant DE genes.

In Tables 3.1 and 3.2 some significant GO terms have been enlisted. Some of the GO terms found by the proposed method are responsible for generation and spreading of AD in human and other primates. The biological verification is present in some literatures like : GO:0046885(regulation of hormone biosynthetic process) [**27**], GO:0032350 (regulation of hormone metabolic process) [**28**], GO:0006569 (tryptophan catabolic process) [**29**], GO:0046218 (indolalkylamine catabolic process) [**30**], GO:0004718 (protein tyrosine kinase activity) [**31**], GO:0015459 (potassium channel regulator activity) [**32**], GO:0016247 (channel regulator activity) [**33**], GO:0015457 (Transport) [**34**], GO:0008076 (voltage-gated potassium channel complex) [**35**].

Again in Tables 3.3 and 3.4 some of the significant KEGG pathways have been enlisted. The relation of these pathways responsible for generation and spreading of AD are available in literatures like: Tryptophan metabolism and Leishmaniasis [**36**], Arrhythmogenic right Ventricular cardiomyopathy [**37**], ECM-receptor interaction [**38**], Hypertrophic cardiomyopathy [**39**], Hematopoietic cell lineage [**40**], p53 signalling pathway [**41**], Neurotrophin signalling pathway [**42**], Retinol metabolism [**43**].

3.3 Extreme Differential regulations between Neurodegenerative Disease pairs

Similar forms of neurodegenerative diseases can be paired to understand the exclusive regulations associated in a certain pair that may be completely absent or presence may be quite insignificant in a different pair. In this regard, explicit importance of differential regulatory links in different forms of neurodegenerative diseases can be explored. The investigation gets even more interesting testing the impact of not only common (among multiple disease sets; a disease set here comprises of two pairs of neurodegenerative disorders) DE genes but the revealing role of mutually exclusive disease specific gene sets in the spreading of diseases. Thus the research work in this direction may turn fruitful scrutinizing the various disease specific problems and thus identifying the therapeutic targets involved.

3.3.1 The basic findings: Discovering the gene sets having extreme differential regulation between dissimilar disease pairs is the primary objective of this research. To unearth the differential co-expression characteristics between gene pairs an existing architecture based on development of a probabilistic score to detect differentially co-expressed gene modules has been implemented. Fundamentally a two tier analysis has been conducted where at first a probabilistic score [**44**] to select gene pairs possessing significant differential co-expression in at least one condition is executed. At the next tier, a non probabilistic differential regulatory structure is implemented for both common and uncommon (disease pair specific) DE genes. At the last part of the research output the pivotal role of the uncommon DE genes to illustrate the dynamic association of different neurodegenerative diseases has been revealed through a complete biological significance analysis. The various forms of neurodegenerative diseases [**45**] on which the proposed algorithm has worked are Alzheimer’s disease (AD), Amyotrophic lateral sclerosis (ALS), Huntington’s disease (HD), Multiple sclerosis (MS), Schizophrenia (SCZ), and Parkinson’s disease (PD). These have been paired up based on their proximities and then further pairing up of such different disease pairs form disease sets. Hence, the output of this research is able to figure out the onset and the growth process of any regulatory pathway specific to certain disease or disease pairs.

3.3.2 Methodology: The algorithm given below and its discussion primarily relies on the computation of probabilistic (LLR) and non-probabilistic (T) differential regulatory scores at different stages of implementation. This helps in revealing the participation of common and uncommon DE genes possessing extreme differential T scores across conditions or disease pairs.

ALGORITHM: Differentially Co-expressed Gene Set selection based on combination of Probabilistic and Non-Probabilistic frameworks

```
Step1: Finding DE genes between the conditions control and disease individually
for different diseases
Step2: Formation of multiple disease pairs based on proximity of different diseases
Step3: Further combination of disease pairs in order to form a disease set $d_i$,
yielding a total of D disease sets
Step4: While (each and every $d_i$ is considered; i=1, 2, 3,….., D)
        Begin
            ➢ Evaluation of T score for a particular pair by Equation (1)
            ➢ Disease pair specific $T_{OVA}$ calculation by comparison of T scores
              by Equation (2)
            ➢ Computation of LLR score, having both positive ($L_P$) and negative
              ($L_N$) values from $T_{OVA}$ using Equation (3)
            While (all possible $L_P$_$L_P$, $L_P$_$L_N$, $L_N$_$L_P$ combinations across disease
                  pairs in a disease set $d_i$ are taken into consideration)
            Begin
                    ➢ Extraction of significant gene pairs in terms of
                      possessing extreme LLR values across both disease
                      pairs
                    ➢ Filtering of common genes showing dysfunctional
                      regulation with that gene pair across disease pairs
                    ➢ Further extraction of disease pair specific exclusive
                      genes showing dissimilar regulations
            End
        End
```

Equation (1): $T^{u,v}_{C1,C2} = \frac{(R^{u,v}_{C2} - R^{u,v}_{C1}) - (\mu_2 - \mu_1)}{\sqrt[2]{{\sigma_2}^2 + {\sigma_1}^2}}$

Equation (2): $T_{OVA} = \operatorname{sign}\left(T^{u,v}_{C1,C2}\right) \min |T^{u,v}_{C1,C2}|$

Equation (3): $LLR_{1,0}(x)$

$$= \log \frac{pf(x/\mu_1, \sigma_1)}{(1-p)f(x/\mu_0, \sigma_0)} = \log \frac{p\sigma_0}{(1-p)\sigma_1} + \frac{(x-\mu_0)^2}{2{\sigma_0}^2} - \frac{(x-\mu_1)^2}{2{\sigma_1}^2}$$

Equation (4): p = $\Pr_{T^{real}_{OVA}}\left(x \geq \mu_{T^{random}_{OVA}} + k\sigma_{T^{random}_{OVA}}\right)$

The various steps of the algorithm are elucidated below.

Step1: The initiation comes up with finding the DE genes for each form neurodegenerative disorder mentioned earlier. For this the dataset in each disorder does contain two states, namely control and disease. The set of DE genes in each case is computed making use of the renowned R package DEGseq [**46**].

Step2: Following the earlier step, the next course of action is finding the similar pairs of neurodegenerative disorders. In this regard, a proximity survey is conducted between every pair of disorder.

In this regard, considering two disorders/diseases A and B containing n and m profiles under the control state ($n \leq m$), simple correlation is studied between every profile of one disease and each one from the other. In other words, profile 1 of A is correlated with each of the profiles, 1 to m, of B. Likewise the same is continued for all the other (n-1) profiles of A. Through this a [n×m] correlated matrix gets generated post which a threshold factor is formed equal to the average value of all the elements of this [n×m] matrix. This is followed by the selection of the top 'n' samples from pairs which exceed the threshold to yield a common control state. In this process, at least 'n' samples are required so that the disease pair (A,B) gets considered for further execution. If the above constraint is not satisfied by any disease pair, the same is not considered in the rest part of the proposed algorithm.

Step3: On completion of the previous step performed over the 6 experimental neurodegenerative diseases yields 5 possible disease pairs. After this, independent combination of these pairs is done producing 10 disease sets. The purpose of forming these individual sets is to study the strong differential regulation (up/down) of genes across different disease pairs.

Step4: At this step, the computation of T score followed by the LLR (Log-Likelihood Ratio; a probabilistic measure) metric is the major point of concern. Hence, obtaining the above measurement parameters and understanding the crucial roles of these in the process of execution is depicted below.

- ➢ The T score for a DE gene pair (u,v) is computed following Equation (1) under the presence of a common control state C_1 (obtained from Step 2 above) and diseased states C_{21} and C_{22}. Here, the T score is obtained for every pair of common DE genes between diseases A and B.

- The overall T score, T_{OVA}, is then found for the pair (u,v) following Equation (2). If the differential regulation of the pair (u,v) following Equation (1), show uniqueness (up/down regulation with respect to C_1) in the sign of regulation, i.e. the regulatory differences considering state pairs C_{21}-C_1 and C_{22}-C_1 both follow the same sign (up indicating positive sign, down indicating negative sign) with respect to C_1, then the DE gene pair (u,v) is taken forward for LLR significance analysis; otherwise it is not considered for the same.
- Considering the LLR (Log-Likelihood Ratio) framework given in [**44**], comparison of this unique T_{OVA} score for the DE gene pair (u,v) is done on real and random data sets. Assuming both the distributions (real and random) are normal (represented as $f(x/\mu_1,\sigma_1)$ and $f(x/\mu_0,\sigma_0)$), 'p' is the probability that the T_{OVA} score belongs to the first distribution. Given the mean and standard deviation of $f(x/\mu_0,\sigma_0)$ and $f(x/\mu_1,\sigma_1)$ are μ0, μ1 and σ0, σ1, the LLR score for the T score ($x=T_{OVA}$) is computed using Equations (3) and (4). Here, 100 shuffled versions of the parent gene expression matrix with respect C_1, C_{21}, and C_{22} are considered to generate the random matrices. From each random matrix, all relevant T_{OVA} scores are found along with the corresponding $\mu_{T_{OVA}^{random}}$ and $\sigma_{T_{OVA}^{random}}$ values. Considering 100 such shuffled distributions, the number of times the T_{OVA} score of the pair (u,v) is greater than or equal to $\mu_{T_{OVA}^{random}} + k\sigma_{T_{OVA}^{random}}$, where k equals 2 following [**44**], defines the probabilistic value 'p' mentioned above. Ultimately, a positive value of LLR for the pair (u,v) defines it as statistically significant, else not. In this regard, a statistically significant LLR is symbolized as L_P and an insignificant one or negative LLR pair is symbolized as L_N.
- The next challenge lies in finding DE gene pairs possessing extremely dissimilar LLR metric between two disease pairs of a disease set (obtained from Step 3 above) followed by the discovery of DE genes associated with the above gene pairs maintaining disjoint regulations between disease pairs in terms of T score. In this perspective, the types of LLR metric defined pairs taken in consideration are L_P_L_P, L_P_L_N, and L_N_L_P. If it is L_P_L_N or L_N_L_P kind of DE gene pair, then the pair is retaining highly positive LLR in one disease pair and minimum

negative LLR (close to zero) in the other disease pair. However, $L_N\_L_N$ is not considered because it indicates insignificant DE gene interaction in both the disease pairs.

- The discovery of DE genes associated with the DE gene pair (u,v) maintaining either $L_P\_L_P$ or $L_P\_L_N$ or $L_N\_L_P$ kind of significance between disease pairs in a disease set, is based on the fact that a DE gene is having positive T_{OVA} score in one disease pair and negative T_{OVA} score in the other with the gene 'u' and vice-versa for the gene 'v' of (u,v). Thus extreme differential regulations for both common and uncommon DE genes with respect to disease pairs can be discovered to an optimum extent.

3.3.3 Results: The research uses the gene expression data (Accession No.26927) [**46**] for Alzheimer's disease (AD), Amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), Multiple sclerosis (MS), Schizophrenia (SCHIZ), and Parkinson's disease (PD) sampled from an extensive cohort of well characterized post-mortem CNS tissues. In all the 6 forms of neurodegenerative diseases mentioned above, there are 20,859 genes, each containing 118 samples, distributed across normal or control and diseased states.

Initial requirement of DE genes yield 6236 for AD, 10831 for ALS, 9382 for HD, 9384 for MS, 9695 for SCHIZ, and 8892 for PD. Guided by the extensive clinical survey [**11-17**] and through the devised approach, the proximal disease pairs happen to be AD_HD, AD_SCHIZ, ALS_MS, MS_SCHIZ, and PD_SCHIZ, with 2794, 1937, 3749, 3403, and 2556 DE genes respectively. Followed by this output, the generated disease sets and the number of common DE genes associated with each disease set are:

AD_HD with AD_SCHIZ (AD_HD_AD_SCHIZ) having 1061 DE genes

AD_HD with ALS_MS (AD_HD_ALS_MS) having 856 DE genes

AD_HD with MS_SCHIZ (AD_HD_MS_SCHIZ) having 1023 DE genes

AD_HD with PD_SCHIZ (AD_HD_PD_SCHIZ) having 872 DE genes

AD_SCHIZ with ALS_MS (AD_SCHIZ_ALS_MS) having 1139 DE genes

AD_SCHIZ with MS_SCHIZ (AD_SCHIZ_MS_SCHIZ) having 785 DE genes

AD_SCHIZ with PD_SCHIZ (AD_SCHIZ_PD_SCHIZ) having 1147 DE genes

ALS_MS with MS_SCHIZ (ALS_MS_MS_SCHIZ) having 804 DE genes

ALS_MS with PD_SCHIZ (ALS_MS_PD_SCHIZ) having 1027 DE genes

MS_SCHIZ with PD_SCHIZ (MS_SCHIZ_PD_SCHIZ) having 1291 DE genes

This is followed by some significant differential regulation pair formations showing L_P_L_P / L_P_L_N / L_N_L_P under two conditions or disease pairs in any one of the disease sets given above. Considering one such regulation pair (u,v), the set of common DE genes which show extremely dissimilar regulatory pattern (for example less than 1059 DE genes for the disease set AD_HD_AD_SCHIZ discarding the remaining two genes because these together have formed the LLR pair (u,v)) can be divided into two subsets. One of the subsets shall comprise of DE genes having positive T_{OVA} score with gene 'u' in say AD_HD and negative T_{OVA} score with gene 'u' in AD_SCHIZ. The other subset shall have just the reverse order with respect to gene 'v' , i.e. negative T_{OVA} score with gene 'v' in say AD_HD and positive T_{OVA} score with gene 'v' in AD_SCHIZ. This kind of situation is depicted in second column of both Tables 3.5 and 3.6 as p1np-n2pn / n1pn-p2np / n1np-p2pn / p1pn-p2np / p1np-p2pn. Considering p1np-n2pn as a case, indicates L_P for gene pair (u,v) in first disease pair and L_N in the second disease pair. The other terms in the above nomenclature highlights the positive and negative T_{OVA} scores with respect to genes 'u' and 'v' at the corresponding disease pairs. On the other hand for the uncommon DE genes, i.e. subtracting disease set specific common DE genes from the disease pair specific DE genes, four disjoint subsets of DE genes are found interacting with the pair (u,v) following the proposed algorithm. In this case, two sets interact with gene 'u' and the other two sets interact with gene 'v' for a particular disease set.

The KEGG pathway [**47**] enrichment analysis given in Tables 3.5 and 3.6 depict the number and significance of DE genes participating in biological pathways relevant to the these neurological or neurodegenerative diseases used in this research. Explicitly handling the entry given in the first row of Table 3.5 shows the presence of 41 common and uncommon DE genes with respect to the first common DE gene pair index

Table 3.5: Some significant pathways in terms of enrichment score/number of participating genes, along with the specific disease combination set, LLR and T metric combination, common gene pair index (among multiple gene pairs found common between two LLR combinations), and common DE genes association

Disease set	LLR and T metric combination	Common gene pair index	Pathway	Enrichment _FDR score	Total number of participating genes	Participating number of common DE genes
AD_HD_AD_SCHIZ	n1np-p2pn	1	Parkinson’sdisease	8.42E−07	41	1 → 5, 2 → 3
			Huntington’sdisease	1.62E−04	46	1 → 4, 2 → 2
	p1np-p2pn	1	Oxidative phosphorylation	6.13E−04	36	1 → 2, 2 → 4
			Alzheimer’sdisease	1.33E−03	41	1 → 3, 2 → 3
	p1pn-p2np	1	Parkinson’sdisease	5.25E−03	36	1 → 2, 2 → 3
			Oxidative phosphorylation	7.7E−03	36	1 → 3, 2 → 2
AD_HD_ALS_MS	n1np-p2pn	3	Huntington’sdisease	3.5E−02	36	1 → 12, 2 → 6
	n1pn-p2np	1	Ribosome	3.24E−02	21	1 → 11, 2 → 3
	p1pn-p2np	2	Ribosome	5.3E−03	22	1 → 8
AD_HD_PD_SCHIZ	n1pn-p2np	1	Huntington’sdisease	3.9E−02	35	1 → 2, 2 → 4
			Oxidative phosphorylation	4.7E−02	28	1 → 2, 2 → 3
			Parkinson’sdisease	9.7E−02	27	1 → 3, 2 → 2

Table 3.5 (Continued)

Disease set	LLR and T metric combination	Common gene pair index	Pathway	Enrichment _FDR score	Total number of participating genes	Participating number of common DE genes
AD_SCHIZ_MS_SCHIZ	p1np-n2pn	2	Oxidative phosphorylation	1.67E−02	32	1 → 4, 2 → 9
			Alzheimer'sdisease	5.3E−02	36	1 → 3, 2 → 10
AD_SCHIZ_PD_SCHIZ	n1np-p2pn	1	Ribosome	4.43E−03	26	1 → 5, 2 → 3
		3	Ribosome	1.3E−03	21	1 → 4, 2 → 3
	n1pn-p2np	3	Ribosome	5.63E−05	28	1 → 4, 2 → 4
	p1np-n2pn	2	Ribosome	3.4E−03	24	1 → 4, 2 → 3
		7	Ribosome	8.45E−08	33	1 → 4, 2 → 4
	p1pn-n2np	4	Ribosome	7.31E−02	21	1 → 2, 2 → 5
		5	Ribosome	1.53E−02	24	1 → 3, 2 → 4
		6	Ribosome	5.37E−05	28	1 → 3, 2 → 4
		7	Ribosome	3.27E−02	23	1 → 4, 2 → 3
	p1pn-p2np	1	Ribosome	1.7E−04	26	1 → 5, 2 → 4
		2	Ribosome	7.5E−06	28	1 → 5, 2 → 3
		3	Ribosome	1.73E−04	27	1 → 3, 2 → 5
		4	Ribosome	2.4E−06	23	1 → 3, 2 → 5
		5	Ribosome	2.26E−04	27	1 → 6, 2 → 3
		6	Ribosome	9.91E−08	32	1 → 3, 2 → 7
ALS_MS_MS_SCHIZ	p1np-p2pn	8	Long-term potentiation	3.13E−03	22	1 → 1, 2 → 2

Table 3.6: Enrichment analysis of certain pathways without the significant common genes in first and second condition independently across gene pairs (here C1 stands for condition 1 and C2 stands for condition 2)

Pathways	Disease set with LLR, T metric combination for a gene pair index (mentioned in brackets)	No. of significant genes				Enrichment score excluding significant genes			
		1st gene specific		2nd gene specific		1st gene specific		2nd gene specific	
		C1	C2	C1	C2	C1	C2	C1	C2
Parkinson's disease	AD_HD_AD_SCHIZ n1np-p2pn (1)	Nil	3	1	Nil	*8.42E−07*	3.4E−05	3E−06	*8.42E−07*
Huntington's disease	AD_HD_ALS_MS n1np-p2pn (3)	3	4	2	2	2.8	5.12	0.77	0.77
Ribosome	AD_HD_ALS_MS n1pn-p2np (1)	7	4	2	Nil	6.4	1.16	0.22	*3.24E−02*
Oxidative phosphorylation	AD_SCHIZ_MS_SCHIZ p1np-n2pn (2)	1	1	4	3	4E−02	4E−02	0.61	0.46
Alzheimer's disease	AD_SCHIZ_MS_SCHIZ p1np-n2pn (2)	Nil	Nil	2	4	*5.3E−02*	*5.3E−02*	0.26	1.2
Ribosome	AD_SCHIZ_PD_SCHIZ n1pn-p2np (3)	1	Nil	Nil	Nil	2.3E−04	*5.63E−05*	*5.63E−05*	*5.63E−05*
Ribosome	AD_SCHIZ_PD_SCHIZ p1np-n2pn (2)	3	3	Nil	Nil	0.123	0.123	*3.4E−03*	*3.4E−03*
Ribosome	AD_SCHIZ_PD_SCHIZ p1np-n2pn (7)	1	Nil	3	Nil	4.1E−07	*8.45E−08*	8E−06	*8.45E−08*
Ribosome	AD_SCHIZ_PD_SCHIZ p1pn-p2np (6)	Nil	Nil	3	Nil	*9.91E−08*	*9.91E−08*	2.3E−06	*9.91E−08*

(any such DE gene pair can also be termed as the central controlling DE gene pair) in the context of AD_HD_AD_SCHIZ. Here, the LLR metric and T score combination, n1np-p2pn, satisfying the regulatory constraints of the 41 participating DE genes in the central presence of the first common DE gene pair index is observed in the pathway, Parkinson's disease, with a false discovery rate enrichment score of 8.42×10^{-7}. Again from here, it is observed that 5 common DE genes are associated with the first gene and 3 common DE genes are regulated by the second gene of the central controlling DE gene pair. Elucidating further, these 5 common DE genes maintain negative and positive T_{OVA} scores with the first gene of the central controlling DE gene pair in the first (AD_HD) and second (AD_SCHIZ) disease pairs respectively. Similar comments are applicable for the 3 common DE genes regulated by the second gene of the central controlling DE gene pair, but in the reverse order of T_{OVA} scores.

In order to gain a better insight on the functionalities of the common DE genes significance testing of the common genes (present in a pathway) individually across the disease pairs (genes having positive T_{OVA} to negative T_{OVA}, and negative T_{OVA} to positive T_{OVA} transitions from first to second pair) is conducted. Significance testing has been completed through random shuffling of the sample labels between the control and diseased states followed by computation and comparison of T_{OVA} scores. Table 3.6 shown above gives a detailed view of the significance testing of some selected pathways. In this table, the effect of removing significant genes over the enrichment of a pathway is of prime importance. In those cases where no significant gene is found, the original enrichment score (shown italicized) is retained. Continuing with the earlier elucidated example of Parkinson's disease (disease set in consideration being the same as cited above, AD_HD_AD_SCHIZ) out of the 5 common DE genes none has been found significantly associated with the first gene of the central controlling DE gene pair in first condition (C1), where as in second condition (C2), 3 genes are discovered significant. Removal of these 3 common DE genes worsens the FDR enrichment score to 3.4×10^{-5}. On the other hand, among 3 common DE genes, 1 gene is found significantly associated with the second gene of the central controlling DE gene pair in C1, but none has been found to be significant in C2. Here removal of the significant gene has given the enrichment score 3×10^{-6}. Hereafter, first and second condition suggests first and second disease pair of the disease set.

Rechecking the enrichment of the pathways (given in Table 3.6) only in the presence of the significant DE genes across conditions is one more major point of concern. This aspect has been worked on those pathways where this enrichment is far better compared to the context having all other DE genes excluding significant participants (given in Table 3.6). In Huntington's disease, corresponding to the first gene of the central controlling DE gene pair, in the second condition, 4 significant DE genes is obtained. These have Illumina IDs ILMN_4450, ILMN_10087, ILMN_13178, and ILMN_16327. Together, these have formed the same pathway with enrichment score of 3.03×10^{-2} against 5.12 as listed in Table 3.6. A similar situation is observed in Ribosome (obtained from AD_HD_ALS_MS), where corresponding to the first gene of the central controlling DE gene pair, in the first condition, there are 7 significant DE genes. These are ILMN_138835, ILMN_137528, ILMN_10289, ILMN_137046, ILMN_138635, ILMN_138392, and ILMN_13487. A combination of these genes did enrich the same pathway with a score of 7.99×10^{-6} compared to 6.4, given in Table 3.6. For the first gene of the central controlling DE gene pair with the same pathway in second condition, there are 4 significant DE genes, which are ILMN_137046, ILMN_138635, ILMN_139337 and ILMN_137876. Together these have given an FDR score of 4.84×10^{-4}, much better than 1.16, given in Table 3.6. Similar observations are present in the case of Oxidative Phosphorylation where enrichments corresponding to the second gene of the central controlling DE gene pair in first and second condition found for the DE genes namely ILMN_20286, ILMN_2295, ILMN_19166, ILMN_137342 and ILMN_20286, ILMN_19166, ILMN_16064 respectively, are far better compared to the enrichments excluding them. From Table 3.6, the FDR scores excluding these significant DE genes are 0.61 and 0.46, whereas simple combination of these significant DE genes yields FDR scores equal to 8.42×10^{-3} and 0.28 respectively. Significant genes obtained from Alzheimer's disease in connection to second condition of 2nd gene of the central controlling DE gene pair are ILMN_1351, ILMN_20286, ILMN_5679, and ILMN_19166. Collectively only these 4 DE genes have given an enrichment score of 3.1×10^{-2} compared to 1.2 excluding them. Finally, in the Ribosome pathway obtained from AD_SCHIZ_PD_SCHIZ with a LLR and T score combination of p1np-n2pn across first gene of the central controlling DE gene pair, 3 significant DE genes has been obtained in both conditions. These are ILMN_21554, ILMN_16945, and ILMN_16298 respectively. Combination of these genes only has given an enrichment score of 2.89×10^{-2} as compared to 0.123, except those.

However, there are some cases where the exclusive analysis of biological enrichment of pathways comprising of the significant DE genes are not fruitful compared to the situation excluding the significant DE genes. Hence, the biological contribution of the corresponding DE genes in the various pathways is further looked into. Like, from Table 3.6, in Parkinson's disease, specific significant DE genes associated with the first gene of the central controlling DE gene pair under second condition are ILMN_11281, ILMN_1167 and ILMN_17626. Here, first 2 genes are involved/activated in encoding of ubiquitin activated enzyme E1, and NADH dehydrogenase (ubiquinone) Fe-S protein 4, whereas the third gene is a cytochrome c oxidase subunit VIIc (COX7C) one. Involvement of them in activation/spreading of Parkinson's disease in different organisms are given in [**48,49**]. In the same disease, across second gene of the central controlling DE gene pair only 1 significant DE gene (ILMN_10929) under first condition experiences the same fate but is known to be involved in encoding of ubiquinol-cytochrome c reductase core protein II (UQCRC2). Now as given in [**50**] this protein actively participates in this diseased pathway. Again, from Table 3.6, for Huntington's disease, across first gene of the central controlling DE gene pair, in the first condition, 3 significant DE genes have been found which are respectively ILMN_22085, ILMN_10087, and ILMN_1167. Among these evidence of active participation in this disease is there for the latter two genes only. As given in [**51,52**] these DE genes participate in this diseased pathway via the encoding of cytochrome c, somatic (CYCS), nuclear gene encoding mitochondrial protein and Homo sapiens NADH dehydrogenase (ubiquinone) Fe-S protein 4. This is the same way through which ILMN_20348 (found across 2nd gene of the central controlling DE gene pair in both conditions) participates in this pathway [**52**]. Another gene found here is ILMN_138125, but surprisingly no existing literature describes its role in Huntington's disease. In Ribosome (by AD_HD_ALS_MS shown in Table 3.6), 2 significant DE genes across second gene of the central controlling DE gene pair in condition 1 are found, namely ILMN_137810 and ILMN_138613. These two are involved in encoding of different ribosomal proteins. In Oxidative Phosphorylation (OP), as per Table 3.6, ILMN_17626 is the only significant DE gene present across both conditions of the first gene of the central controlling DE gene pair. ILMN_17626 is a cytochrome c oxidase subunit VIIc (COX7C) gene, whose effect on OP is given in [**53**]. From Table 3.6 again only in Alzheimer's disease (AD) across first condition of second gene of the central controlling DE gene pair two significant genes are present, namely ILMN_1351, and

ILMN_20286. Individual assessment of these genes over AD via mitogen-activated protein kinase 1 (MAPK1) and NADH dehydrogenase (ubiquinone) 1 beta sub complex encoding is described in [**54-56**]. ILMN_21554 is the only significant DE gene found in Ribosome (obtained from AD_SCHIZ_PD_SCHIZ with a LLR and T score combination of n1pn_p2np and is third gene pair specific), and has got significant involvement in ribosomal protein formation whereas there are all total 3 significant DE genes obtained from another Ribosome pathway (this case is also from the same disease set and same LLR combination as the previous one but from seventh gene pair). These are ILMN_2271 (found with both the genes of the central controlling DE gene pair), ILMN_2500, and ILMN_1815 (last two are found exclusively with the second gene of the central controlling DE gene pair). Finally, the last combination is the group of 3 significant genes ILMN_11712, ILMN_15150, and ILMN_138613 (this time also the pathway is Ribosome, with the same disease set combination having the LLR and T score combination as p1pn_p2np, with sixth gene pair).

3.3.4 Discussion: Varieties of diseased pathways shown in Tables 3.5 and 3.6 are obtained feeding the required DE gene set into DAVID [**57,58**]. From the tables it can be seen that most of these pathways are mainly connected to neurodegenerative or neurological diseases. Only the three pathways, which do not bear any disease name, are Oxidative phosphorylation, Ribosome, and Long term potentiation. In order to check whether any/all of them have a role in such diseases an exhaustive clinical survey was done.

Oxidative phosphorylation basically regulates the neuronal actions via the help of Mitochondria (abbreviated as Mt). As given in [**59**] oxygen takes part in glucose break down in Mt through oxidative phosphorylation and generates ATP, which works as energy currency of the cell. Any form of mutation of Mt DNA (works as molecular machinery) enforces impaired ATP generation and perturbed oxidative phosphorylation cascade, further locking the neuronal function, which specifically leads to AD [**59,60**]. Again in Ribosome, absence of some binding partners (for an example GTPBP2) of the ribosome recycling protein may cause ribosome stalling and widespread neuro-degeneration [**61**]. Finally, some literatures suggest the role of Long term potentiation in different neurodegenerative or neurological disorders occurring due to synaptic dysfunctions [**62**].

It is also important to understand the bridging effect some KEGG pathways observed in Table 3.5 or 3.6 over different diseases of a corresponding disease set. For the pathway, Parkinson's disease (PD), under the disease set AD_HD_AD_SCHIZ, thorough study reveals TRANSGLUTAMINE (TG) kind of enzymes affecting (act as a common factor) different neurodegenerative disorders like PD and AD/HD [**63**]. As given in [**63**] for different kinds of TGs activated in AD and HD, CSF (Cerebrospinal Fluid) also contribute to the formation of proteinaceous deposits in PD. It is important to note that both PD and SCHIZ have a common originating link. As stated in [**64**] these are the results of the redox process (i.e. joint activity of Reactive Oxygen Species (ROS) and Oxidative Stress (OS)). This redox process also works as a common link between HD and AD, ALS, MS. In [**64**] a common redox association can be found between AD and MS. Apart from having a common chemically reactive baseline, all these different neurodegenerative diseases happen to be a subset of Neurodegenerative Misfolding Diseases (NMD) triggered by the misfolding of one or two proteins and their accumulation in the aggregated species toxic to neurons [**65**], especially due to the effect of protein disulphide isomers (PDI).

3.4 Conclusion

The contributions made in this chapter towards formation of topologically differential DE gene architectures instigate the research to get explored in the domain of Transcription Factor (TF) gene regulatory networks. The biological involvement of the DE genes in the process of reconstruction of differential gene regulation networks can be understood from the contribution of the DE genes in growth and spread of a disease across various stages as well as their involvement in different interlinked diseases. This contribution gets highlighted through the presence of extreme dissimilar regulatory controls executed in the progressive stages of any disease or in different kinds of biologically proximal interlinked disease pairs.

3.5 References

[**1**] E. Kensinger, "Early and late onset as subdivisions of Alzheimer's Disease", The Harvard Brain, pp. 26-29, 1996

[**2**] Alzheimer's Disease & Related Dementias, National Institute of Aging. Available online at: https://www.nia.nih.gov/health/alzheimers/basics

[**3**] H.A. de Silva, J.K. Aronson, D.G. Grahame-Smith, K.A. Jobst, A.D. Smith, "Abnormal function of potassium channels in platelets of patients with Alzheimer's disease", Lancet, 352(9140), 1590-1593, November 1998, https://doi.org/10.1016/s0140-6736(98)03200-0

[**4**] J.T. Lanner, "Ryanodine receptor physiology and its role in disease", Advances in Experimental Medicine and Biology, 740, 217-234, 2012, https://doi.org/10.1007/978-94-007-2888-2_9

[**5**] N. Bogdanovic *et al.* "On the turnover of brain cholesterol in patients with Alzheimer's disease. Abnormal induction of the cholesterol-catabolic enzyme CYP46 in glial cells", Neuroscience Letters, 314(1-2),45-48, November 2001, https://doi.org/10.1016/s0304-3940(01)02277-7

[**6**] N. Gustin, "Researchers discover link between insulin and Alzheimer's", News Release, March 2005. Available online at: https://www.eurekalert.org/news-releases/792767

[**7**] A. Morinaga, K. Ono, J. Takasaki, T. Ikeda, M. Hirohata, and M. Yamada, "Effects of sex hormones on Alzheimer's disease-associated β-amyloid oligomer formation in vitro", Experimental Neurology, 228(2), 298-302, April 2011, https://doi.org/10.1016/j.expneurol.2011.01.011

[**8**] V. Hosur, R.H. Loring, "α4β2 nicotinic receptors partially mediate anti-inflammatory effects through Janus kinase 2-signal transducer and activator of transcription 3 but not calcium or cAMP signaling," Molecular Pharmacology, 79(1), 167-174, January 2011, https://doi.org/10.1124/mol.110.066381

[**9**] S. Seneff, G. Wainwright, and L. Mascitelli, "Nutrition and Alzheimer's disease: The detrimental role of a high carbohydrate diet", European Journal of Internal Medicine, 22(2), 134-140, April 2011, https://doi.org/10.1016/j.ejim.2010.12.017

[**10**] D.J. Selkoe, "Alzheimer's disease: Genes, proteins, and therapy", Physiological Reviews, 81(2), April 2001, https://doi.org/10.1152/physrev.2001.81.2.741

[**11**] B.W. Palmer *et al.* "Assessment of capacity to consent to research among older persons with Schizophrenia, Alzheimer disease, or Diabetes Mellitus: Comparison of a 3-Item Questionnaire with a Comprehensive Standardized Capacity Instrument", Archives of General Psychiatry, 62(7), 726–733, July 2005, http://dx.doi.org/10.1001/archpsyc.62.7.726

[**12**] M.K. Sutherland, M.J. Somerville, L.K.K. Yoong, C. Bergeron, M.R. Haussler, , D.R.C. McLachlan, "Reduction of vitamin D hormone receptor mRNA levels in Alzheimer as compared to Huntington hippocampus: correlation with calbindin-28k mRNA levels", Molecular Brain Research, 13 (3), 239–250, April 1992, https://doi.org/10.1016/0169-328X(92)90032-7

[**13**] S.M. Rao, S.J. Huber, R.A. Bornstein, "Emotional changes with multiple sclerosis and Parkinson's disease", Journal of Consulting and Clinical Psychology, 60(3), 369–378, June 1992, https://psycnet.apa.org/doi/10.1037/0022-006X.60.3.369

[**14**] Healthline, Available online at https://www.healthline.com/

[**15**] G.B. Frisoni, M. Filippi, "Multiple sclerosis and Alzheimer disease through the looking glass of MR imaging", American Journal of Neuroradiology, 26(10), 2488–2491, November 2005

[**16**] O.A. Andreassen, *et al.* "Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci", Molecular Psychiatry, 20(2), 207-214, February 2015, https://doi.org/10.1038/mp.2013.195

[**17**] A. Ghanemi, "Schizophrenia and Parkinson's disease: selected therapeutic advances beyond the dopaminergic etiologies", Alexandria Journal of Medicine, 49(4), 287–291, December 2013, https://doi.org/10.1016/j.ajme.2013.03.005

[**18**] H. Braak and E. Braak, "Neuropathological stageing of Alzheimer related changes", Acta Neuropathologica, 82(4), 239-259, September 1991, https://doi.org/10.1007/BF00308809

[**19**] M. Ray and W.X. Zhang, "Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks", BMC Systems Biology, 4, Article No. 136, October 2010, https://doi.org/10.1186/1752-0509-4-136

[**20**] O. Odibat and C.K. Reddy, "Ranking differential hubs in gene co-expression networks", Journal of Bioinformatics and Computational Biology, 10(1):1240002, February 2012, https://doi.org/10.1142/S0219720012400021

[**21**] C.E. Johanson, J.A. Duncan III, P.M. Klinge, T. Brinker, E.G. Stopa, G.D. Silverberg, "Multiplicity of cerebrospinal fluid functions: New challenges in health and disease," Cerebrospinal Fluid Research, 5, Article No.10, May 2008, https://doi.org/10.1186/1743-8454-5-10

[**22**] J. Ruan, A.K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications," BMC Systems Biology, 4, Article No. 8, February 2010, https://doi.org/10.1186/1752-0509-4-8

[**23**] A. Reverter and E.K.F. Chan, "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks", Bioinformatics, 24(21), 2491-2497, November 2008, https://doi.org/10.1093/bioinformatics/btn482

[**24**] Gene Expression Omnibus, Available online at: https://www.ncbi.nlm.nih.gov/geo/

[**25**] GOstats: Tools for manipulating GO and microarrays, Available online at: https://bioconductor.org/packages/release/bioc/html/GOstats.html

[**26**] GeneTrail: Advanced high-throughput enrichment analysis, Available online at: https://genetrail.bioinf.uni-sb.de/

[**27**] J. Lin *et al.*, "Genetic ablation of luteinizing hormone receptor improves the amyloid pathology in a mouse model of Alzheimer disease", Journal of Neuropathology and Experimental Neurology, 69(3), 253-261, March 2010, https://doi.org/10.1097/NEN.0b013e3181d072cf

[**28**] M.M. Hasan, "Comparative systems-level analysis of the G1/S transition in yeast and higher eukaryotes: focusing on the Whi5/Rb network and initiation of DNA

replication", Ph.D. Dissertation, Department of Biotechnology and Biosciences, University of Milan-Bicocca, Italy, December 2011.

[**29**] A. Kowarsch *et al.*, "Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation", BMC Bioinformatics, 11, Article No.585, November 2010, https://doi.org/10.1186/1471-2105-11-585

[**30**] K. James, "Knowledge derivation and data mining strategies for probabilistic functional integrated networks", PhD Thesis, School of Computing Science, Newcastle University, U.K., July 2011

[**31**] X. Han, T. Shen, and H. Lou, "Dietary polyphenols and their biological significance", International Journal of Molecular Sciences, 8(9), 950-988, September 2007

[**32**] K-S. Lynn, C-H. Lu, H-Y. Yang, W-L. Hsu, W-H. Pan, "Construction of gene clusters resembling genetic causal mechanisms for common complex disease with an application to young-onset hypertension", BMC Genomics, 14, Article No.497, July 2013, https://doi.org/10.1186/1471-2164-14-497

[**33**] K. Bossers *et al.*, "Concerted changes in transcripts in the prefrontal cortex precede neuropathology in Alzheimer's disease", Brain, 133(12), 3699-3723, December 2010, https://doi.org/10.1093/brain/awq258

[**34**] A.B. Goodman and A.B. Pardee, "Evidence for defective retinoid transport and function in late onset Alzheimer's disease", PNAS, 100(5), 2901-2905, February 2003, https://doi.org/10.1073/pnas.0437937100

[**35**] Y. Pertzov *et al.*, "Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis", Brain, 136(8), 2474-2485, June 2013, https://doi.org/10.1093/brain/awt129

[**36**] K.E.J. Tripodi, S.M.M. Bravo, and J.A. Cricco, "Role of Heme and Heme-Proteins in trypanosomatid essential metabolic pathways", Enzyme Research, 2011, Article No. 873230, April 2011, https://doi.org/10.4061/2011/873230

[**37**] NCBI. Available online at: https://www.ncbi.nlm.nih.gov/

[**38**] J.C. Bis *et al.*, "Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation", Molecular Psychiatry, 25, 1859-1875 (2020), August 2018, https://doi.org/10.1038/s41380-018-0112-7

[**39**] T. Force, K. Kuida, M. Numchuk, K. Parang, J.M. Kyriakis, "Inhibitors of protein kinase signaling pathways", Circulation, 109, 1196-1205, March 2004, https://doi.org/10.1161/01.CIR.0000118538.21306.A9

[**40**] M. Squillario and A. Barla, "A computational procedure for functional characterization of potential marker genes from molecular data: Alzheimer's as a case study", BMC Medical Genomics, 4, Article No.55, July 2011, https://doi.org/10.1186/1755-8794-4-55

[**41**] A.V. Gudkov and E.A. Komarova, “Pathologies associated with the p53 Response”, Cold Spring Harbor Perspectives in Biology, 2(7): a001180, July 2010, https://doi.org/10.1101%2Fcshperspect.a001180

[**42**] A.L. Torre *et al.*, “A role for the tyrosine kinase ACK1 in neurotrophin signaling and neuronal extension and branching,” Cell Death and Disease, 4(4):e602, April 2013, https://doi.org/10.1038/cddis.2013.99

[**43**] M.A. Lane and S.J. Bailey, “Role of retinoid signalling in the adult brain”, Progress in Neurobiology, 75(4), 275–293, March 2005, https://doi.org/10.1016/j.pneurobio.2005.03.002

[**44**] D. Amar, H. Safer, R. Shamir, “Dissection of regulatory networks that are altered in disease via differential co-expression”, PLOS Computational Biology, March 2013, https://doi.org/10.1371/journal.pcbi.1002955

[**45**] P.F. Durrenberger *et al.*, “Selection of novel reference genes for use in the human central nervous system: a BrainNet Europe Study”, Acta Neuropathologica, 124(6), 893–903, December 2012, https://doi.org/10.1007/s00401-012-1027-z

[**46**] L. Wang, Z. Fenq, X. Wang, X. Wang, X. Zhang, “DEGseq: an R package for identifying differentially expressed genes from RNA-seq data”, Bioinformatics, 26(1), 136–138, January 2010, https://doi.org/10.1093/bioinformatics/btp612

[**47**] M. Kanehisa, S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes”, Nucleic. Acids Research, 28(1), 27–30, January 2000, https://doi.org/10.1093/nar/28.1.27

[**48**] O.M. Viquez, S.W. Caito, W.H. McDonald, D.B. Friedman, W.M. Valentine, “Electrophilic adduction of ubiquitin activating enzyme E1 by N,N-diethyldithiocarbamate inhibits ubiquitin activation and is accompanied by striatal injury in the rat”, Chemical Research in Toxicology, 25(11), 2310–2321, November 2012, https://doi.org/10.1021/tx300198h

[**49**] S. Arnold, “Cytochrome c oxidase and its role in neurodegeneration and neuroprotection”, Part of the Advances in Experimental Medicine and Biology book series, 748, 305–339, January 2012, https://doi.org/10.1007/978-1-4614-3573-0_13

[**50**] NCBI database, https://www.ncbi.nlm.nih.gov (Gene ID: 7385)

[**51**] X. Wang *et al.*, “Inhibitors of cytochrome c release with therapeutic potential for Huntington’s disease”, The Journal of Neuroscience, 28(38), 9473–9485, September 2008, https://doi.org/10.1523/JNEUROSCI.1867-08.2008

[**52**] HMDB. Available online at: https://hmdb.ca/proteins/HMDBP00180

[**53**] GeneCards. Available online at: https://www.genecards.org/cgi-bin/carddisp.pl?gene=COX7C

[**54**] X. Zhu, H.G. Lee, A.K. Raina, G. Perry, M.A. Smith, “The role of mitogen-activated protein kinase pathways in Alzheimer’s disease”, Neuro-Signals 11(5), 270–281, September-October 2002, https://doi.org/10.1159/000067426

[**55**] NCBI database, https://www.ncbi.nlm.nih.gov (Gene ID: 4729)

[**56**] S.H. Kim, R. Vlkolinsky, N. Crains, M. Fountoulakis, G. Lubec, "The reduction of NADH Ubiquinone oxidoreductase 24- and 75-kDa subunits in brains of patients with Down syndrome and Alzheimer's disease", Life Sciences, 68(24), 2741–2750, May 2001, https://doi.org/10.1016/s0024-3205(01)01074-8

[**57**] D.W. Huang, B.T. Sherman, R.M. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources", Nature Protocols, 4(1), 44–57, December 2008, https://doi.org/10.1038/nprot.2008.211

[**58**] D.W. Huang, B.T. Sherman, R.M. Lempicki, "Bioinformatics enrichment tools: paths towards the comprehensive functional analysis of large gene lists", Nucleic. Acids Research, 37(1), 1–13, January 2009, https://doi.org/10.1093/nar/gkn923

[**59**] B. Uttara, A.V. Singh, P. Zamboni, R.T. Mahajan, "Oxidative stress and neurodegenerative diseases: a review of upstream and downstream antioxidant therapeutic options", Current Neuropharmacology, 7(1), 65–74, March 2009, https://doi.org/10.2174/157015909787602823

[**60**] J. Hroudová, N. Singh, Z. Fišar, "Mitochondrial dysfunctions in neurodegenerative diseases: relevance to Alzheimer's disease", BioMed Research International, 2014, Article ID 175062, May 2014, https://doi.org/10.1155/2014/175062

[**61**] R. Ishimura *et al.*, "Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration", Science, 345(6195), 455–459, July 2014, https://doi.org/10.1126/science.1249749

[**62**] M. Marttinen, K.M. Kurkinen, H. Soininen, A. Haapasalo, M. Hiltunen, "Synaptic dysfunction and septin protein family members in neurodegenerative diseases", Molecular Neurodegeneration, 10, Article No.16, April 2015, https://doi.org/10.1186/s13024-015-0013-z

[**63**] A. Martin, G.D. Vivo, V. Gentile, "Possible role of the transglutaminases in the pathogenesis of Alzheimer's disease and other neurodegenerative diseases", International Journal of Alzheimer's Disease, 2011, Article ID 865432, February 2011, https://doi.org/10.4061/2011/865432

[**64**] P. Kovacic, R. Somanathan, "Redox processes in neurodegenerative disease involving reactive oxygen species", Current Neuropharmacology, 10(4), 289–302, December 2012, https://doi.org/10.2174/157015912804143487

[**65**] M.F. Mossuto, "Disulfide bonding in neurodegenerative misfolding diseases", International Journal of Cell Biology, 2013, Article ID 318319, August 2013, https://doi.org/10.1155/2013/318319

Chapter 4 Multiobjective Ranked Selection of Transcription Factor and Differentially Expressed Genes and its importance in Gene Regulatory Networks

Related Publications

[**1**] M. Sarkar and A. Majumder, “Multiobjective Ranked Selection of Differentially Expressed Genes”, In: Deiva Sundari, P., Dash, S., Das, S., Panigrahi, B. (eds) Proceedings of 2nd International Conference on Intelligent Computing and Applications, Advances in Intelligent Systems and Computing, volume 467, pp. 75-92, October 2016, Springer, Singapore.
https://doi.org/10.1007/978-981-10-1645-5_7
[**2**] A. Majumder, M. Sarkar, H. Dash, and I. Akhilesh, “A Composite Entropy Model in a Multiobjective Framework for Gene Regulatory Networks”, Current Bioinformatics, volume 13, issue 1, pages 85-94, 2018.
http://dx.doi.org/10.2174/1574893611666161202104422

4.1 Introduction

Formation of a gene regulatory network can be biologically validated by the existence of significant KEGG pathways. In such pathways, individual or multiple collaborative interactions made by one or more genes or proteins over some target genes can be computationally verified. However, considering the differential networking perspective under varied environmental situations or perturbation characteristics, the importance of such pathways or regulatory cascades might show altering statistical enrichments. In this regard, to statistically validate a regulatory cascade, requires strong computational background to understand the crucial role (single or collaborative) and the level of existence of any particular gene present in the regulatory pathway. To meet these requirements, in the differential networking paradigm, the significant importance of any gene can be developed based on its differential regulatory capabilities and the dynamic stability of the same.

This chapter addresses the above need through the development of significant regulatory cascades comprising of Differentially Expressed (DE) genes and/or Transcription Factor (TF) genes. The differential regulatory scores of these genes can maintain conflicting objectives or in other words the regulatory objectives may not be at par with one another. Hence, a multi-objective optimized approach may be useful in designing the non-dominated sets of participating DE genes or TF genes at various levels of any regulatory network. Thus a hierarchical architecture comprising TF and/or DE genes can be developed with the provision of dynamically handling the regulations in certain environmental or externally perturbed situations. This development can possibly validate any regulatory network or cascade designing composite entropy minimized regulatory architectures containing single or collaborative differential interactions between TF and DE genes or between TF and Differentially Coexpressed (DC) genes.

4.2 Ranking Differentially Expressed Genes in a Multiobjective Framework

The initial challenge in developing gene regulatory cascades or pathways is to find the significant role of TF and/or DE genes at various stages of this hierarchical structure. Developing such structures can be guided by the differential regulatory capabilities of the participating TF and/or DE genes. Many existent methodologies judge the differential regulatory actions of any gene in a complex regulatory network. However in this context, the gene ranking obtained via such different methodologies according to their significance, is quite dissimilar to one another making regulatory assessment of genes very difficult. Hence, the challenge comes in determining the contributory role of genes in a complex gene regulatory network when different methods yield separate sets of significant genes. Thus, it can be difficult to reconcile the different results and identify the specific genes that play a role in a network's differential functionality. One possible approach to address this challenge is to perform a comprehensive analysis that takes into account the results from multiple methodologies. By integrating the results and prioritizing genes known to have connections to the biological processes or pathways relevant to the network under study, it is possible to identify common genes or gene sets (obtained by evaluating multi-objective constraints or ranking measures defined by the separate methodologies) consistent and robust enough to indulge in specific biological investigations. As this work is centred on the differential regulatory controls of TF and/or DE genes, the prima facie of this research followed by the next in this chapter lies in finding the significant ranking of the TF (may be DE too) and/or DE genes through non-dominance analysis applying multi-objective approach on the different ranking measures.

Application of multi-objective evolutionary algorithms has got an extensive research in the domain of gene classification [**1**] by clustering [**2-5**]. In this context, most of the works try to optimize clustering indices in a multi-objective paradigm. The commonly used algorithms have been GA-II, NSGA-II, PESA-II [**6**], to name a few. Extensively it focuses on methods to improve clustering via number of clusters present in a chromosome, intra-cluster compactness, inter-cluster separation, and cluster size [**7**]. However, one of the major drawbacks of clustering approach is time complexity, irrespective of optimization techniques such as GA [**8**], PSO [**9**], BPNN [**10**], ACO [**11**] etc. This aspect occurs mainly due to clustering with generation and validation of new populations in different iterations [**4,5**]. This time challenging constraint does not arise in this research as the differential regulatory problem is throughout managed with the parent physically existent population of genes only. The intention is to find the best possible combination of DE genes from the existing ones on the basis of ranking and not to generate any new set of solutions with predicted gene expression levels.

4.2.1 The basic findings: In this research, a novel procedure for computing significant DE genes by utilizing multiple ranking strategies is presented. The concept is based on the creation of multiple non-dominated sets consisting of solutions from different Pareto optimal fronts. The main goal of this specific research is to identify a set of non-dominated DE genes in the primary Pareto optimal front, where each DE gene possesses an optimal combination of significance rankings across various ranking algorithms. The main goal of this research has been extended in the next research output (i.e. developing entropy minimized transcriptional regulatory networks) presented in this chapter.

The results of this research demonstrate that the majority of KEGG pathways [**12**] based on the defined set of differentially expressed (DE) genes contain a maximum of two DE genes from the non-dominated primary Pareto optimal set. This finding aids in understanding the independent regulatory function of a gene from this non-dominated set compared to the set of dominated genes. In other words, the presence of enriched control pathways with significantly ranked DE genes that are non-dominant to each other is nearly absent. This indicates that the identified non-dominated genes, with their optimal combination of rankings, play distinct roles in the regulatory processes and are not redundant or overlapping with each other.

4.2.2 Methodology: The objective of this research is to find non-dominated DE genes residing in the primary Pareto optimal front. To achieve this, the different ranking algorithms like rank sum statistics [**13**], gene significance based enrichment analysis, GSEA [**14**], activity score (AS) [**15**], and TOP based gene significance [**16**] have been utilized as the four objective functions. Considering a minimization problem with these four conflicting objectives (these are either maximization or minimization problems), a feasible DE gene pair (u,v) can have the gene ‘u’ dominating gene ‘v’, provided $z_i(u) \leq z_i(v)$ for i = 1,2,3,4 with at least one $z_i(u) < z_i(v)$ for the objective function ‘z_i’ [**6**]. In other words, the DE gene ‘u’ will possess a higher rank or lower score with the paired DE gene ‘v’ in one, two, three or four objectives with equal score in the other three, two, one or zero objectives respectively. If the gene pair (u,v) does not meet any of the above criteria, then the DE genes ‘u’ and ‘v’ can be termed non-dominant to one another.

The flowchart of the proposed algorithm is shown in Figure 4.1 and elucidated thereafter.

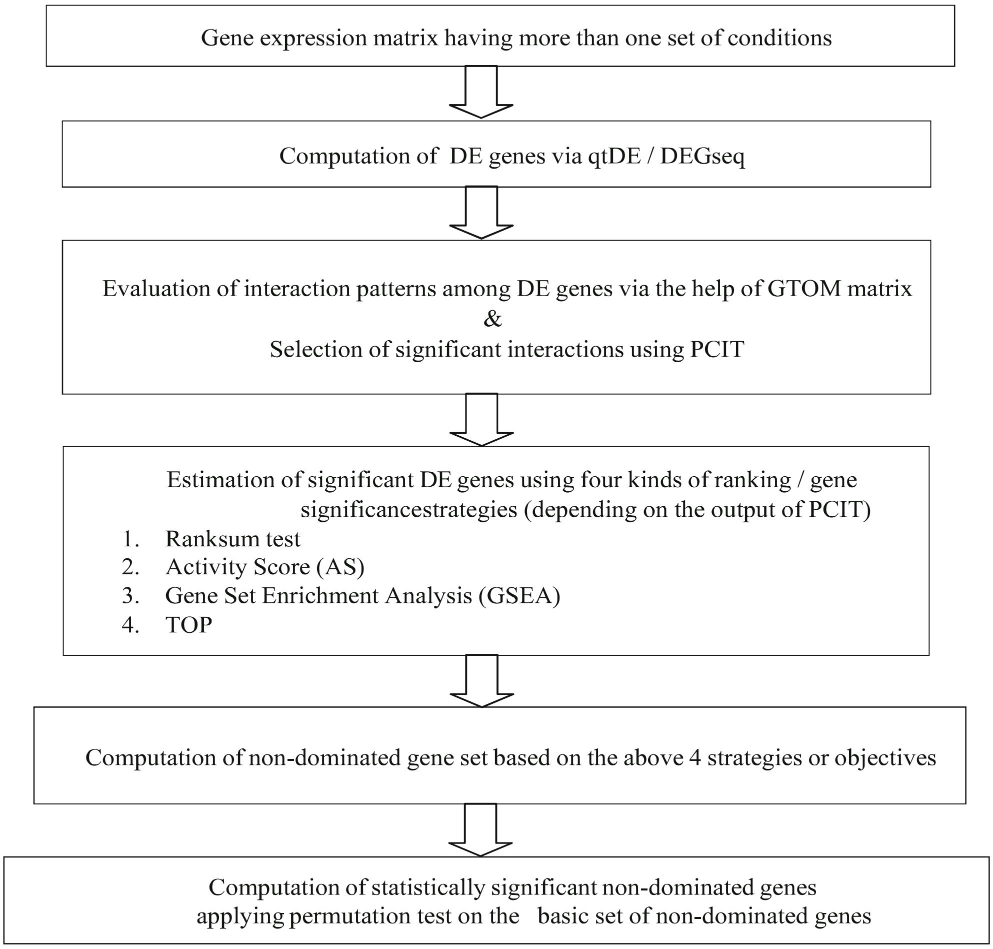

Gene expression matrix having more than one set of conditions
Computation of DE genes via qtDE / DEGseq
Evaluation of interaction patterns among DE genes via the help of GTOM matrix
&
Selection of significant interactions using PCIT
Estimation of significant DE genes using four kinds of ranking / gene significancestrategies (depending on the output of PCIT)
1. Ranksum test
2. Activity Score (AS)
3. Gene Set Enrichment Analysis (GSEA)
4. TOP
Computation of non-dominated gene set based on the above 4 strategies or objectives
Computation of statistically significant non-dominated genes applying permutation test on the basic set of non-dominated genes

Figure 4.1: Flowchart for evaluating significantly non-dominated DE genes

Elucidation of the above steps present in the flowchart is depicted below.

- Initially, as per the flowchart, any gene expression matrix is required containing more than one condition of interest.
- The interest is on finding any DE gene which possesses the property of differential connectivity when paired with other DE genes for the purpose of developing differentially connective regulatory pathways. For this DE genes are computed making use of the same statistical technique (used in earlier research works present in the previous chapters), DEGseq [**17**], and the quantitative trait specific DE approach, qtDE [**18**], discussed in detail in chapter 2.
- The interaction analysis under different conditions in both sets of DE genes (DE genes are computed using two measures) are done using the renowned GTOM technique [**19**] followed by finding the DE genes which maintain significant direct connectivity in the presence of other DE genes using the PCIT [**20**] approach.

- The next operation happens to be computing the differentially significant genes that can participate in a regulatory pathway based on the four multi-objective ranking functions or strategies given in the flowchart (Ranksum, AS, GSEA, and TOP). The brief discussion of each of these functions is given later.
- In the above multi-objective approach, the primary intention lies in finding the non-dominated set of DE genes that fall on the primary Pareto-optimal front. In order to do this, any of these corresponding genes should have better ranking (or lower score) in any one, two, or three of the four objective functions in comparison with all other DE genes. In other words, it indicates, a particular DE gene residing in the primary Pareto-optimal front can have either equal or higher score (worse ranking) compared to all other DE genes in three, two, or one of the four objectives respectively. Thus, the number of possible combinations to look through for discovering the non-dominated set of DE genes placed in the primary Pareto-optimal front equals $4_{C_1} + 4_{C_2} + 4_{C_3}$, i.e. 14 combinations.
- At the final step, significantly non-dominated DE genes (found above) are computed conducting permutation test [**21**] on the PCIT filtered GTOM matrices. Thereafter, through KEGG analysis, this helps in understanding the solo participation of these DE genes acting as primary controllers in one or more regulatory pathways, dominating the other DE genes (not a part of the primary Pareto-optimal front).

The necessary details of the four objective functions utilized in the above computation are as follows:

Ranksum: In this context, Wilcoxon Ranksum test [**13**] is conducted on the PCIT filtered weighted [**16**] GTOM matrices obtained from two or more conditions of interest. If two conditions are considered, then it leads to the formation of two matrices, A and B.

This is a non-parametric test of the null hypothesis where the two populations are same against an alternative hypothesis, where a particular population puts up a skewed effect to the combined distribution. In general considering two populations C and D with independent random samples $c_1,c_2,\ldots.,c_m$ and $d_1,d_2,\ldots..,d_n$ are merged and the measurements are ranked from lowest to highest values. Here, the mean (μ) and standard deviation (σ) of the merged data can be found as $\mu = m(m+n+1)/2$ and $\sigma = \sqrt[2]{mn(m+n+1)/12}$.

Now based on the merged data two kinds of hypothesis are checked. First one being the null hypothesis H0: C = D, which means the distribution of X measurements in population C is same as that of D or in other words the ranked distribution pattern in the merged data happens to be an association of samples taken at random from the individual distributions. In this case, the differential pattern of the data vectors (here, the filtered GTOM vectors for each gene) cannot be predicted. On the other hand, the alternative hypothesis is of two kinds. First one being H1: C > D, which means in the ranked merged distribution, samples from C is right shifted compared to D, and the second one being H2: C < D, suggesting the samples from C is left shifted compared to D. In both of the cases, it is possible to ascertain the differential pattern of the filtered GTOM vectors highlighting the differential contribution of the concerned gene.

Here, the application of PCIT [**20**] yields filtered GTOM matrices. In this connection, two matrices (one for each condition) are obtained with entries 1 and 0. As a next step, common set of interactions are searched for every gene across both conditions indicating the search for those entries possessing 1 at the same location in both matrices for each individual gene. Finally, these 1's get replaced by the original filtered GTOM values in both conditions and each other entry (uncommon 1's as well as 0's) is made equal to 0. At this stage, the Wilcoxon-rank-sum test is applied over these modified versions of the filtered weighted GTOM matrices A and B.

Activity Score: Like the Ranksum test here also at first the common set of interactions between two conditions is considered followed by replacing the (common) 1's by the filtered GTOM values in both conditions and other fields by 0.

Next, as per [**15**], the activity score (AS) score is calculated. Here, corresponding to any row (i.e. a gene) of the filtered GTOM matrix, X_i means the genes having a non-zero entry. Z_j is the Z score of the differential expression of any such gene using Ranksum test. Following the equations $w_i = \sum_{j \in X_i} \mathrm{rank}(Z_j)$, $u_i = \frac{i(N+1)}{2}$, $\sigma_i = \sqrt[2]{\frac{i(N-i)(N+1)}{12}}$ from [**15**] calculations of w_i, u_i, and σ_i are done, where N is the total number of genes used in the analysis. Thus utilising the values produced by $w_i = \sum_{j \in X_i} \mathrm{rank}(Z_j)$, $u_i =$

$\frac{i(N+1)}{2}$, $\sigma_i = \sqrt[2]{\frac{i(N-i)(N+1)}{12}}$ the AS score is computed as per the equation $AS = (-1)^{\propto} \times \max_{i\in neighbour}(\frac{w_i-u_i}{\sigma_i} \times \frac{w_i-u_i}{w_{max}-u_i})$ of [**15**], provided $\alpha = 0$ if $\frac{w_i-u_i}{\sigma_i} > 0$; otherwise α=1. This is followed by the ranking of the DE genes using the AS scores obtained from the above technique.

GSEA: Starting with PCIT filtered weighted GTOM matrices obtained in the two conditions maintaining non-zero entries for the significant set of interactions, here comes finding the Z scores and ranking the genes accordingly.

At first, corresponding to every gene, separately in each condition, ranking of the gene interaction values from lowest to highest level is followed by assigning rank labels for every significant interaction. At the next step a random vector is generated for every gene comprising of values meant to index locations of the significant GTOM matrix. In each case, the rank of the indexed gene interaction is checked followed up by addition of the weights (filtered GTOM value) of the genes having less rank than the indexed gene interaction. This concept follows the method of enrichment score calculation [**14**] and is framed as per the pair of equations given below:

$$P_{hit}(S,i) = \sum_{E_j\in S, j\leq i} \frac{{r_j}^p}{N_R} \text{ where } N_R = \sum_{E_j\in S} {r_j}^p$$

and

$$P_{miss}(S,i) = \sum_{E_j\in S, j\leq i} \frac{1}{N - N_H}$$

Above, where ‘i’ is the indexed rank and ‘j’ are those genes possessing lower rank with ‘S’ being the neighbourhood of significant interactions for a gene. The term ‘N’ in the above pair of equations represents the total number of genes and N_H is the number of significant gene interactions corresponding to a gene. From this, the enrichment score for every indexed gene interaction is calculated as $ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i)$ and accordingly the ES matrices are formed in each condition.

The differential attitude of the generated ES vectors is calculated via Ranksum test yielding the rES vector. As a next step, the permuted ES and rES scores are found considering 200 cases of random shuffling of the parent GTOM matrices. For each gene the average (rES′) and the standard deviation (S′) of the permuted rES scores is obtained and accordingly the final Z score that helps in ranking the genes is computed using the equation $\mathbf{Z} = \frac{\mathbf{rES-rES'}}{\mathbf{S'}}$.

TOP: This analysis can be performed using two approaches: un-weighted and weighted [**22**] (developed and discussed in detail in the last two chapters). However as discussed in [**16**] results obtained by un-weighted measure being better than weighted counterpart, here the TOP score has been computed via the un-weighted measure only.

At the practical end, it is simply a combination of TO value [**22**] with its significance using T test [**21**]. Assuming A_1 and A_2 to be the PCIT filtered GTOM matrices with entries 1 and 0 corresponding to significant and insignificant gene to gene interactions in the two different conditions, it is possible to find the TO of gene 'i' using the equation $\mathrm{TO_i = (X_i \cap Y_i)/max(X_i, Y_i)}$ where for gene i in condition 1, X_i no. of interaction(s) is/are significant, and in condition 2 it is Y_i. Next as in [**16**] the average of TO measure and the p-value (using permutation/t test [**21**]) is used as the ranking measure. It is calculated for a particular gene 'i' as $\mathrm{TOP_i = (TO_i + pvalue_i)/2}$.

4.2.3 Results: The algorithm has been tested on mice data set containing gene expression levels of male and female phenotypes across four tissues, namely adipose, brain, liver, muscle. Details of the data set along with required pre-processing [**23,24**] proved useful to initiate the projected analysis.

Applying qtDE and DEGseq, the number of DE genes obtained at each phase of the proposed algorithm is given below in Tables 4.1 and 4.2 respectively. Here, the three different phases at which the DE genes have been computed are as follows:

1. Initial number of DE genes
2. Number of DE genes in the primary Pareto optimal front (i.e. the non-dominated set of DE genes extracted from the entire bunch)
3. Number of ***significant*** DE genes in the primary Pareto optimal front (i.e. the non-dominated DE genes which have been proved statistically significant)

Table 4.1: Number of DE genes obtained using qtDE based on the three different methods at each phase

Mice Tissue	Linear Correlation			Mutual Information			Polynomial Regression		
	Initial number of DE genes	Number of DE genes in the primary Pareto optimal front	Number of ***significant*** DE genes in the primary Pareto-optimal front	Initial number of DE genes	Number of DE genes in the primary Pareto optimal front	Number of ***significant*** DE genes in the primary Pareto-optimal front	Initial number of DE genes	Number of DE genes in the primary Pareto optimal front	Number of ***significant*** DE genes in the primary Pareto-optimal front
Adipose	856	77	12	1236	28	6	938	20	7
Brain	579	105	91	1499	21	5	675	112	67
Liver	837	8	6	1479	8	2	1395	33	9
Muscle	1132	114	99	2503	99	33	1163	160	124

Table 4.2: Number of DE genes obtained using DEGseq at each phase

Mice Tissue	DEGseq		
	Initial number of DE genes	Number of DE genes in the primary Pareto optimal front	Number of ***significant*** DE genes in the primary Pareto-optimal front
Adipose	732	9	3
Brain	373	21	9
Liver	424	7	3
Muscle	301	3	0

An important observation or exception comes from Table 4.2 above. Through DEGseq technique of computing DE genes no significant non-dominated DE gene is found in muscle.

In Tables 4.3, 4.4, 4.5 and 4.6, the detailed KEGG pathway analysis on the total DE genes detected via qtDE and DEGseq are presented. In these tables, genes from basic non-dominated set or in other words the DE genes participating from primary Pareto optimal front are highlighted in bold characters, and the significantly non-dominated DE genes are given in bold and italics.

Apart from the information portrayed in Tables 4.3 to 4.6, there are some more significant KEGG pathways which have not shown any contribution or participation of DE genes from the primary Pareto optimal front. These are depicted across the three tissues (adipose, brain, and liver) corresponding to the qtDE approach in Table 4.7.

Table 4.3: Significant KEGG pathways based on DE genes obtained from linear qtDE approach

Pathways	*p* values	Genes
Tissue: ADIPOSE		
Olfactory transduction	3.49E−07	**Olfr584**, Olfr599, Olfr957, Clca1
Leishmaniasis	5.51E−03	Jun, **H2-Aa**, Mapk1, **Prkcb**, Jak2, Tgfb3, Il1b,H2-DMa, H2-Ab1
ErbB signalling pathway	8.31E−02	Jun, Rps6kb2, Mapk1, Stat5a, Erbb3, **Prkcb**, Cdkn1b, Areg
Graft versus host disease	8.31E−02	**H2-Aa**, H2-Q8, Il1b, Cd86, H2-DMa, H2-T10,H2-Ab1
Malaria	8.31E−02	Itgal, Tgfb3, Il1b, ***Ccl2***, Hgf, Vcam1
Type I diabetes mellitus	8.31E−02	**H2-Aa**, H2-Q8, Il1b, Cd86, H2-DMa, H2-T10,H2-Ab1
Viral myocarditis	8.31E−02	Rac2, H2-Aa, H2-Q8, Itgal, Casp3, Cd86, **H2-DMa**, H2-T10, H2-Ab1
Cell adhesion molecules(CAMs)	8.31E−02	**H2-Aa**, H2-Q8, Itgal, Cdh2, Cntnap2, Cd86,H2-DMa, H2-T10, Jam2, H2-Ab1, Vcam1
Ribosome	0.10	Rpl15, Rps3a, Rps3, Rps8, Rpl3l, Rpl29, Rpl6, Rps14, **Rpl35**
Tissue: BRAIN		
Olfactory transduction	6.69E−04	Olfr1226, **Olfr206**, Olfr380, Olfr599, Olfr1234
Leukocyte transendothelial migration	4.03E−02	***Txk***, Jam2, Itgal, Vegfb
Galactose metabolism	8.79E−02	***Gck***, Galt, **Gaa**, Ugp2
Tissue: LIVER		
Olfactory transduction	4.27E−07	Clca1, Olfr535, Olfr380, **Olfr599**, Olfr1234, Clca2
Drug metabolism– cytochrome P450	3.59E−02	Fmo3, ***Cyp2d22***, Gsta2, Cyp2c40, Gstm2, Mgst2, Cyp2d10, Ugt1a9, Mgst3
Glutathione metabolism	3.59E−02	Ggt1, Gsta2, Gstm2, Mgst2, ***Pgd***, Gclm, Mgst3
Chagas disease	6.39E−02	***C1qb***, Cd3g, Car, Jun, Cd3d, Tgfb3, Il1b, Pik3r1,Tlr2, Tnf
Tissue: MUSCLE		
Olfactory transduction	6.96E−09	***Olfr571***, **Olfr957**, Olfr380, Clca2, Olfr584, Olfr1234, Olfr535, ***Clca1***
Asthma	4.3E−02	H2-Eb1, **H2-Aa**, H2-DMa, H2-DMb1, H2-Ab1, Tnf
Cell adhesion molecules(CAMs)	4.3E−02	Pvrl2, Jam2, H2-Eb1, Cd86, Cd22, Cntnap2, **H2-Aa**, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Cdh2, Itga8, Ptprc, Itgal
Complement and coagulation cascades	4.3E−02	Vwf, F11, Masp1, C1qa, Thbd, Cd59a, Serpine1, **C8b**, F5
Focal adhesion	4.3E−02	***Parva***, Vwf, Chad, Rac2, Rap1a, Pdgfrb, Col5a3, Vegfb, Pik3r1, Vtn, Hgf, Ccnd2, Fyn, Myl2, Itga8, Pak7, Kdr, Flnb
Glutathione metabolism	4.3E−02	Gstm1, Gclm, **Pgd**, Gstm2, Gpx7, Ggt1, Gpx4, Gsta2
Intestinal immune network for IgA production	4.3E−02	H2-Eb1, Cd86, **H2-Aa**, H2-DMa, Pigr, H2-DMb1, H2-Ab1, Tnfsf13b
Renal cell carcinoma	4.3E−02	Slc2a1, Rap1a, ***Tceb1***, Vegfb, Pik3r1, Hgf, Pak7, Ets1, ***Cul2***
Type I diabetes mellitus	4.3E−02	H2-Eb1, Cd86, **H2-Aa**, H2-Q8, H2-DMa, H2-DMb1, H2-Ab1, Tnf, Ins1, Hspd1
Viral myocarditis	5.8E−02	Rac2, H2-Eb1, Cd86, **H2-Aa**, H2-Q8, Myh7,H2-DMa, H2-DMb1, H2-Ab1, Fyn, Itgal
Allograft rejection	6.3E−02	H2-Eb1, Cd86, **H2-Aa**, H2-Q8, H2-DMa,H2-DMb1, H2-Ab1, Tnf
Graft-versus-host disease	6.3E−02	H2-Eb1, Cd86, **H2-Aa**, H2-Q8, H2-DMa,H2-DMb1, H2-Ab1, Tnf

Table 4.4: Significant KEGG pathways based on DE genes obtained from polynomial regression guided qtDE approach

Pathways	*p* values	Genes
Tissue: ADIPOSE		
Olfactory transduction	7.23E−11	Olfr584, Olfr380, Olfr1234, Clca1, **Olfr957**
Cytokine-cytokine receptor interaction	2.88E−03	***Ifng***, Bmp2, Tnfsf13b, Hgf, Ccl5, Ccl2, Ccr5, Ccl7,Il7r, Csf3r, Cxcl14, Cxcl5, Tnfrsf18, Tgfb3, Tnf, Kitl, Il10rb, Tnfsf8, Inhbb, Il22ra2, Tnfrsf12a, Csf2rb2, Ltbr, Kdr
TGF-beta signaling pathway	2.73E−02	***Ifng***, Bmp2, Rbl1, Bmp5, Dcn, Tgfb3, Rps6kb2,Tnf, Inhbb, Bmp8b, Ltbp1
Malaria	4.17E−02	***Ifng***, Hgf, Ccl2, Lrp1, Itgal, Tgfb3, Tnf, Tlr2
Leishmaniasis	7.55E−02	***Ifng***, Tgfb3, H2-DMa, Tnf, H2-DMb1, H2-Aa,H2-Eb1, Tlr2
Allograft rejection	7.9E−02	***Ifng***, H2-T10, H2-DMa, Tnf, H2-DMb1, Cd86,H2-Aa, H2-Eb1
Graft versus host disease	7.9E−02	***Ifng***, H2-T10, H2-DMa, Tnf, H2-DMb1, Cd86,H2-Aa, H2-Eb1
Amino sugar and nucleotide sugar metabolism	9.19E−02	Hexa, **Gnpnat1**, Gne, Gck, Galt, Hexb
Tissue: BRAIN		
Olfactory transduction	3.21E−07	Olfr1226, **Olfr1234**, Olfr894, Olfr380
Steroid biosynthesis	5.73E−02	***Sc4mol***, Dhcr7, Sqle, ***Nsdhl***
Glycerolipid metabolism	7.31E−02	Akr1b8, Ppap2b, **Gpam**, Lpl, Agpat2, Gyk
Tissue: LIVER		
Olfactory transduction	6.46E−15	Olfr1234, Olfr599, Olfr584, Clca1, **Olfr535**,Olfr380, Olfr957, Olfr894
Leishmaniasis	3.07E−03	Tlr2, Ncf2, H2-Ab1, Il1b, H2-Eb1, Jun, Ifng, H2-Aa, Tgfb3, ***Prkcb***, **Tnf**, Jak2, H2-DMa
Focal adhesion	3.06E−02	Vtn, Rac2, Jun, Tnc, Col1a2, Pdgfc, Chad, Fyn, Rap1a, Pik3r1, Kdr, Pdgfrb, Parva, Col5a3, Vwf, Itga8, Lamb3, Prkca, Mylpf, Flnb, Actn2, ***Prkcb***,Hgf
Amoebiasis	4.3E−02	Tlr2, Il1b, Serpinb6a, Ifng, Col1a2, Casp3, Pik3r1, C8b, Col5a3, Tgfb3, Lamb3, Prkca, Actn2, ***Prkcb***, **Tnf**
Hematopoietic cell lineage	4.3E−02	Il1b, H2-Eb1, Cd2, Cd22, Il7r, Csf3r, Cd59a, Mme, Kitl, Cd3d, Kit, **Tnf**
Type I diabetes mellitus	4.3E−02	Cd86, H2-Ab1, Ins1, Il1b, H2-Eb1, Ifng, H2-Q8, H2-Aa, **Tnf**, H2-DMa, H2-T10
Graft-versus-host disease	5.61E−02	Cd86, H2-Ab1, Il1b, H2-Eb1, Ifng, H2-Q8, H2-Aa, **Tnf**, H2-DMa, H2-T10
Malaria	5.61E−02	Tlr2, Il1b, Ifng, Itgal, Lrp1, Tgfb3, **Tnf**, Hgf, Ccl2
TGF-beta signaling pathway	5.61E−02	Bmp8b, Rps6kb2, Rbl1, Ifng, Inhbb, Bmp2, Bmp5, Dcn, Ltbp1, Tgfb3, **Tnf**, Gdf5
Galactose metabolism	9.14E−02	Pfkp, Gaa, Galt, **Akr1b8**, Gck
Tissue: MUSCLE		
Olfactory transduction	1.76E−04	Olfr957, Olfr894, Olfr535, Olfr571, ***Olfr599***,Clca1, Clca2, Olfr893, Olfr584
Cardiac muscle contraction	3.86E−02	**Cox7b**, Myh7, **Tpm3**, Cacnb1, Myl3, ***Slc8a1***, Cox6a2, Actc1, Cacna2d1, Tpm1, Cox7a1, Cox7a2
Cytokine-cytokine receptor interaction	3.86E−02	Kdr, Tnf, Cxcl14, Il10ra, Bmp2, ***Vegfb***, Csf1r, Il7r,Flt4, Inhbb, Tnfrsf12a, Tnfsf13b, Gdf5, Ifng, Cxcr3, Egfr, Tnfrsf21, Kitl, Tnfrsf18, Ccl6, Ccr5, Il1b, Ccl4, Cxcl1
Hematopoietic cell lineage	5.87E−02	***Cd22***, Tnf, H2-Eb1, Csf1r, Il7r, Cd14, Mme, Kitl,Cd2, Il1b, Cd34

Table 4.4 (continued)

Pathways	*p* values	Genes
Tissue: MUSCLE		
Olfactory transduction	1.76E−04	Olfr957, Olfr894, Olfr535, Olfr571, ***Olfr599***,Clca1, Clca2, Olfr893, Olfr584
Cardiac muscle contraction	3.86E−02	**Cox7b**, Myh7, **Tpm3**, Cacnb1, Myl3, ***Slc8a1***, Cox6a2, Actc1, Cacna2d1, Tpm1, Cox7a1, Cox7a2
Cytokine-cytokine receptor interaction	3.86E−02	Kdr, Tnf, Cxcl14, Il10ra, Bmp2, ***Vegfb***, Csf1r, Il7r,Flt4, Inhbb, Tnfrsf12a, Tnfsf13b, Gdf5, Ifng, Cxcr3, Egfr, Tnfrsf21, Kitl, Tnfrsf18, Ccl6, Ccr5, Il1b, Ccl4, Cxcl1
Hematopoietic cell lineage	5.87E−02	***Cd22***, Tnf, H2-Eb1, Csf1r, Il7r, Cd14, Mme, Kitl,Cd2, Il1b, Cd34
Hypertrophic cardiomyopathy (HCM)	5.87E−02	Itga8, Myh7, Tnf, **Tpm3**, Cacnb1, Myl3, ***Slc8a1***,Actc1, Cacna2d1, Tpm1, Actb
Dilated cardiomyopathy	9.92E−02	Itga8, Myh7, Tnf, **Tpm3**, Cacnb1, Myl3, ***Slc8a1***,Actc1, Cacna2d1, Tpm1, Actb

Table 4.5: Significant KEGG pathways based on DE genes obtained from mutual information guided qtDE approach

Pathways	*p* values	Genes
Tissue: ADIPOSE		
Selenoamino acid metabolism	3.76E−02	Ahcy, ***Mat1a***
Cysteine and methionine metabolism	4.46E−02	Ahcy, ***Mat1a***
Intestinal immune network for IgA production	4.46E−02	Pigr, ***H2-Aa***
Ribosome	8.14E−02	Rpl3 l, **Rps3a**
Tissue: LIVER		
Ubiquitin mediated proteolysis	2.37E−03	Tceb1, Ube2m, ***Aco2***
Tissue: MUSCLE		
Olfactory transductions	1.838E−06	Olfr837, Olfr918, ***Olfr1450***, ***Olfr307***

Table 4.6: Significant KEGG pathways based on DE genes obtained from DEGseq

Pathways	*p* values	Genes
Tissue: ADIPOSE		
Amoebiasis	2.75E−02	Serpinb6a, Prkcb, **Serpinb9e**, Tlr4, Casp3,Rab5a, Tfgb3, Adcy1, Il1b, Prkca
Leishmaniasis	3.59E−02	Mapk1, Tlr4, H2-DMa, Tgfb3, Il1b, **Prkcb**,Jak2
ErbB signaling pathway	3.68E−02	Mapk1, Nck1, Crk, **Prkcb**, Rps6kb1, Areg,Prkca, Pak7
Fc gamma R-mediated phagocytosis	5.34E−02	Mapk1, Arpc3, Myo10, Crk, **Prkcb**, Rps6kb1,Prkca, Pla2g6
Amino sugar and nucleotide sugar metabolism	7.2E−02	**Hk1**, Ugdh, Gpi1, Ugp2, Galt
Renal cell carcinoma	9.12E−02	Mapk1, **Rap1a**, Tgfb3, Crk, Tceb1, Pak7
Tissue: BRAIN		
Tight junction	3.57E−03	Myh7, ***Jam2***, Myh2, Mylpf, Jam3, Prkcd,Ash1 l, Csnk2a1
Tissue: LIVER		
Galactose metabolism	3.07E−03	Hk2, Gaa, Galt, **Pfkp**, Akr1b8

Table 4.7: Significant KEGG pathways having NIL contribution of non-dominated DE genes (observed in qtDE approach)

Mice Tissue	Linear Correlation		Mutual Information		Polynomial Regression	
	KEGG pathway	P value	KEGG pathway	P value	KEGG pathway	P value
Adipose	--	--	Axon guidance	4.46E−02	Complement and coagulation cascades	2.73E−02
	--	--	Cytosolic DNA-sensing pathway	4.46E−02	Asthma	4.17E−02
	--	--	--	--	Hematopoietic cell lineage	4.17E−02
	--	--	--	--	Intestinal immune network for IgA production	8.9E−02
Brain	Leishmaniasis	3.07E−02	--	--	Fc gamma R-mediated phagocytosis	9.11E−02
Liver	Leishmaniasis	2.83E−02	--	--	Cytokine-cytokine receptor interaction	3.07E−03
	Focal Adhesion	3.59E−02	--	--	Complement and coagulation cascades	3.06E−02
	Hematopoietic cell lineage	3.59E−02	--	--	Maturity onset diabetes of the young	4.3E−02
	Asthma	7.27E−02	--	--	Cell adhesion molecules (CAMs)	5.61E−02

4.2.4 Discussion: From Tables 4.3, 4.4, 4.5, and 4.6 it is observed that the significant KEGG pathways have one or at most two genes from the non-dominated DE gene set. In other words, the outputs portrayed in these tables signify a rare probability of grouping more than one significant non-dominated gene in a single pathway. Exceptions are Renal cell carcinoma (present in muscle using linear qtDE), Steroid Biosynthesis (present in brain using polynomial regression), and Olfactory Transduction (present in muscle using linear qtDE and mutual information guided qtDE).

It is to be noted that the permutation test [**21**] conducted on the non-dominated DE gene set present in the primary Pareto optimal front yields the significant non-dominated DE genes with p-value less than 0.2 (the threshold). In the case of the above exceptions, apart from a single DE gene the other DE gene present in the significant non-dominated set possess high p value very close to the cut-off level, i.e. 0.2. From Table 4.3 it can be seen that for Olfactory Transduction the significantly non-dominated genes are *Olfr571* and *Clca1*. These are having p-values 0.07, and 0.188; for Renal Cell Carcinoma the significant non-dominant DE genes are *Tceb1* and *Cul2* with p-values of 0.084 and 0.147. This shows us an appreciable difference between the p-values of *Olfr571* and *Clca1* in one case and between *Tceb1* and *Cul2* in the other. In both the cases, based the p-value threshold selection (i.e. 0.2), the second gene is on the verge of elimination from the significantly non-dominated set. A similar pattern is observed for Steroid Biosynthesis from Table 4.4 and in Olfactory Transduction from Table 4.5. The participating non-dominated significant DE genes from the primary Pareto optimal front are *Sc4mol* and *Nsdhl* in the previous one and *Olfr1450* and *Olfr307* in the latter. Significance analysis for non-dominated DE genes yields p-values 0.185 and 0.084 for *Sc4mol* and *Nsdhl* followed by 0.108 and 0.179 for *Olfr1450* and *Olfr307* respectively.

Most of the above pathways enlisted in the various tables are associated with the differential evolution of mice. Examples are like Olfactory Transduction [**25,26**], Leishmaniasis and Cell adhesion molecules (CAMs) [**27,28**], Graft versus Host disease and Allograft rejection [**29**]. Some other significant pathways (like Hematopoietic cell lineage, TGF-beta signalling pathway, Complement and coagulation cascade, Cytokine-cytokine receptor interaction, Ubiquitin mediated proteolysis, and Fc gamma R-mediated phagocytosis) also work on the differential development like gastrulation, axis symmetry of the body, organ morphogenesis, liver development and tissue homeostasis in adults of mice/other mammals across different sex [**30-35**].

4.3 Developing Entropy minimized Transcriptional Regulatory Networks

Transcriptional Regulatory Networks (TRNs) are essentially complex biological regulatory systems involving various kinds of genes and modular proteins where some of the genes act differently under varied types of environmental conditions or external perturbations. This dissimilar attitude of the genes (in other words the phenomenon of differentially expressed or differentially co-expressed properties in genes) is mostly contributed by the effect of molecular interactions like Transcription Factor (TF) promoter DNA binding, protein-protein interaction and protein translational modification [**36**].

In the above context, the regulatory design through association of TF genes mostly comes up in significant research outcomes; thus playing a pivotal role in differentially regulating genes across varied types of conditions. In this regard, the kinds of TF regulatory networks or TRNs for differentially expressed (DE) or differentially co-expressed (DC) genes that have been studied or developed are mostly of individual or collaborative forms of regulation. Though some important contribution in this perspective is present in this thesis work (the forms of network developed and validated in the previous chapters), the variety of significant research already done in this domain are mainly into understanding stepwise TF gene collaborations towards gene regulation [**37,38**].

To contribute in the step regulatory architecture present across various biological pathways incorporating TF, DC, and DE genes, active research is required to develop multiple serial TF regulatory paths to any target gene maintaining a stable architecture portraying the inherent differential regulations under varied states or conditions. The stability of the developed network can be adjusted putting in force information theoretic approach such as entropy minimization [**39**]. A significant research [**40**] in which a novel concept of generating stable gene regulatory architectures between direct interacting source and target genes using the concept of entropy is helpful in apprehending the non-redundant existence of regulator genes. This thought has been extended in this research work through realizing the undoubted role of multiple types of TF genes in disjoint serial regulatory pathways; thus recreating a hierarchical structure in line with complex biological regulations.

4.3.1 The basic findings: Various kinds of regulatory measures which have proven to be of significant importance to assess the differential regulation capacity of genes are like RIF I and II [**41**], TED and TDD [**42**], TFactS [**43**], dCSA_t2t, and dCSA_r2t [**44**]. Among these seven network-based regulatory algorithms TED, TDD, and TFactS deal with the number of DE/ DC genes targeted by a TF gene. At a later stage, the proposed algorithm demands random shuffling of gene expression levels in order to calculate the significance of the formed pathways. In this context, at any instance, with a different set of altered gene expression profiles there is need to recompute the new set of DC genes, DE genes, and Differentially Coexpressed Links (DCLs). As this operation increases the space and time complexity of the proposed algorithm, hence TED, TDD and TFactS have not been considered as probable objectives of gene score evaluation. On the contrary, the other four algorithms work on the expression levels of TF genes/targets; hence these regulatory measures are vulnerable to changes in expression profiles. As the alterations in the expression profiles directly helps in the computing the evaluation scores, this set comprising of RIF II, dCSA_t2t, and dCSA_r2t are considered as probable objectives. RIF-I does not work on self-regulated genes (TFs which are DE too). As the intention of this research work highlights interactions involving DE and DC TFs (can also be termed TF_DE and TF_DC genes), hence RIF-I does not suit the need. Thus, the objective functions applied here are Regulatory Impact Factor II (RIF-II), Differential Correlation Set Analysis between Regulatees (dCSA_t2t), and Differential Correlation Set Analysis between Regulator and Regulatees (dCSA_r2t). These three objective functions or evaluation scores are utilized to initiate with the multi-objective approach designed to develop the serial TF regulatory networks.

Apart from this above finding which happens to be the significant initiation of this research problem, the ultimate hierarchical regulatory structure based on placing the various kinds of TF genes across multiple Pareto optimal fronts is an open challenge in the context of biological regulations. In this regard, the research done here portrays a novel idea of formation of TRNs for a target DE gene maintaining entropy minimized stable regulatory structures compositely throughout the network, i.e. reconstructing stabilized network pathways starting from one or more source TF genes (should be TF genes that are neither TF_DE nor TF_DC type), mediating through TF_DC and/or TF_DE genes, and finally ending with a target DE gene. In the process of this reconstruction, the genuine significance (both statistically and biologically) of TF, TF_DC, and TF_DE genes taking part in the regulation at any stage of the hierarchical network is also being judged keeping in account various biological regulatory databases [**42,45**]. Here, TF_DC genes indicate those TF genes which are significantly differentially co-expressed with other genes. On the other hand, TF_DE genes indicate those TF genes which are significantly differentially expressed between various experimental conditions or external perturbations under consideration.

4.3.2 Methodology: The brief outline of the algorithm that has been used to build up composite entropy minimized TF regulatory paths for target DE genes is given below. The specific inputs required are the gene expression matrices under different conditions and the biological databases supporting TF to other gene interactions. Using these two informational inputs, single or collaborative non-redundant regulations can be found or ascertained maintaining minimal entropy as the network design progresses forward from one or more source TF genes.

ALGORITHM: Composite Entropy minimized TF regulatory network reconstruction

Inputs:
I. Genes from GEO database.
II. TF to target database.

Outputs:
I. Set of differentially expressed (DE) and co-expressed (DC) genes comprising of TF as well as non-TF genes.
II. The transcription factor (TF) genes across multi-objective fronts.
III. Statistically enriched composite entropy minimized TF regulatory paths to a target non-TF differentially expressed (DE) gene.

Step1: Computation of DE and DC genes using DEGseq and DCGLv2 R packages respectively.
Step2: Computation of network regulatory measures (e.g. RIF II, dCSA_r2t, dCSA_t2t) for each and every TF using the TF to target database.
Step3: Placing the TFs across various optimal fronts based on the multi-objective network regulatory measures defined in step 2.
Step4: Computation of PCIT paths to target DE genes comprising of TFs obtained from the various optimal fronts.
Step5: Starting with TFs from the non-dominated Pareto optimal front, formation of minimal composite entropy regulatory structures targeting a DE gene.
Step6: Verifying that the minimal entropy regulatory paths are either a subset or set of PCIT paths.

The elucidation of the above algorithm following the given steps is the matter of concern in the next half of this module, mentioned hereafter.

Estimation of DE and DC genes: Step1 of the algorithm is concerned with the estimation of DE and DC genes from the TF and target DE genes. The estimation of the DE genes has been done using the R package DEGseq [**17**]. On the other hand, the R package DCGL (version 2) has been utilized to identify the DC genes. The functions *DEGexp* from DEGseq and *DCp* from DCGLv2 are used for finding the above kinds of genes.

The basic formation of TF regulatory paths: Steps 2 and 3 of the algorithm are involved in this segment. Here, the regulatory scores are computed for each and every TF gene followed by placing the TF genes at the various Pareto optimal fronts, based on the multi-objective approach, depicted in the earlier research work present in this chapter. In this regard, the three regulatory scores acting as conflicting objectives in a multi-objective framework are RIF II, dCSA_r2t, and dCSA_t2t. The equations involved to compute these scores are:

$$RIF2(TF_i) = \left| \frac{1}{n_{de}} \sum_{j=1}^{n_{de}} \left[(e1_j \times r1_{ij})^2 - \left(e2_j \times r2_{ij} \right)^2 \right] \right|$$ depicting the regulatory score, RIF II

$$dCSA\_r2t(TF_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \left| r1_{ij} - r2_{ij} \right|$$ depicting the regulatory score, dCSA_r2t

$$dCSA\_t2t(TF_i) = \frac{2}{n_i (n_i - 1)} \sum_{k=1}^{n_i - 1} \sum_{j=k+1}^{n_i} \left| r1_{jk} - r2_{jk} \right|$$ depicting the regulatory score, dCSA_t2t

In the above equations, gene with index ‘i’ refers to the TF gene in consideration. Three different scores are hence obtained for a particular TF gene ‘i’. Again, genes with indices ‘j’ and ‘k’ are the direct regulatees of the TF gene index ‘i’ and the same is found from the TF2target database [**42**] present as an essential input to the devised algorithm. Apart from these notations, $e1_j$, $e2_j$, $r1_{ij}$, $r2_{ij}$, n_i, n_{de} denote the average log gene expression levels of DE gene ‘j’ in conditions 1 and 2, the interaction scores or dependency scores between TF gene ‘i’ and DE gene ‘j’ under the two conditions, the total number of targets regulated by TF gene ‘i’ as per the TF2target database or any other biological database (depends on the experimented organism), and the total number of DE genes obtained respectively. From the equations, it is clear that every regulatory score is about understanding the power of differential regulation of a TF gene considering all published targets of the concerned TF gene from the databases mentioned above. Hence, it can be claimed that TF genes giving high objective scores shall be having more controlling capability for a given set of targets.

Here comes the challenge of rearranging the TF genes according to their regulation strength by these three objectives yielding almost three dissimilar TF orderings. In other words, it suggests independent TF regulatory assessment based on the above three objectives and hence making difficult to understand the contributory role of TF genes in gene regulatory networks. To solve this issue, the multi-objective technique discussed in the earlier research is used to find out the non-dominated TF gene sets distributed across multiple Pareto optimal fronts. Now while framing the serial regulatory architectures, (involving single or collaborative regulations) maintaining dominance of TF genes present in the upper fronts over the lower ones, the following constraints are being applied.

- A pure TF (i.e. neither TF_DC nor TF_DE) gene can control TF_DE genes as well as TF_DC genes. [**41,42**]
- A TF_DC gene can control another TF_DC gene as clarified in [**42**]. Again from the TF2target database [**42**] a a TF_DE gene getting controlled by a TF_DC gene can be verified.
- A TF_DE gene can be executing control over another TF_DE gene [**41**].

Due to lack of proper biological evidences, any other form of controlling action is disregarded. In this perspective comes a pure TF gene being regulated by any TF_DC or TF_DE gene and a TF_DC gene getting controlled by any TF_DE gene.

Based on the above constraints, prior to the finalizing the basic TF regulatory paths incorporating a TF gene from every possible Pareto optimal front, pure TF genes that are placed at lower fronts with respect TF_DC or TF_DE genes are removed from further analysis. Similarly, removal of TF_DC gene is executed if found getting dominated by a TF_DE gene present in an upper front. Hence, after such removals, a significance analysis is carried out by shuffling the expression profiles of the existent TF genes (pure, TF_DC, and TF_DE types) and the three objective scores are re-evaluated in each shuffled context. Higher significances of the existence of the basic TF regulatory paths are ascertained provided less number of shuffled paths match with the actual ones obtained at this stage of the algorithm.

<u>Formation of DE gene specific paths using PCIT</u>: This is about Step 4 of the designed algorithm. The realization of the connectivity from the source TF gene to any target DE gene via intermediate TF genes (maintaining the constraints stated above) present at almost every non-primary Pareto optimal front, is executed taking help of the basic PCIT [**20**] concept. However, with DE and DC genes forming the underlying pillars for further analysis, instead of applying simple Pearson's correlation measure in PCIT the T score [**46**] measure designed using the differential correlation between TFs genes is utilized. In this regard, the equations followed are $r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz})^2(1-r_{yz})^2}}$ computing the correlation dependency between x and y in the presence of z,

$T^{x,y}_{C_1,C_2} = \frac{(r^{x,y}_{C_2} - r^{x,y}_{C_1}) - (\mu_2 - \mu_1)}{\sqrt{\sigma_2^2 + \sigma_1^2}}$ depicting the T score between x and y present across conditions C_1 and C_2, and $r_{xy,z} = \frac{T^{x,y}_{C_1,C_2} - T^{x,z}_{C_1,C_2} T^{y,z}_{C_1,C_2}}{\sqrt{\left(1 - T^{x,z}_{C_1,C_2}\right)^2 \left(1 - T^{y,z}_{C_1,C_2}\right)^2}}$ depicting the same dependency pattern between x and y in the presence of z, but replacing the correlation component by T score across conditions. Now via the application of PCIT, in any trio combination of genes (x,y,z), a direct connectivity between a pair of genes, for example (x,y), is not considered significant if with $\varepsilon = \frac{1}{3}\left(\frac{r_{xy,z}}{r_{xy}} + \frac{r_{xz,y}}{r_{xz}} + \frac{r_{yz,x}}{r_{yz}}\right)$, the dependence r_{xy} satisfies the relations $\left|r_{xy}\right| \le \left|\varepsilon \times r_{xz}\right|$ and $\left|r_{xy}\right| \le \left|\varepsilon \times r_{yz}\right|$. Thus to check the strength of direct regulation over indirect ones, this procedure is repeated via all other indirect genes for the pair (x,y). In this context, if direct association (x,y) satisfies the relations $\left|r_{xy}\right| \le \left|\varepsilon \times r_{xz}\right|$ and $\left|r_{xy}\right| \le \left|\varepsilon \times r_{yz}\right|$ for a higher proportion of indirect genes (like 'z'), the direct connectivity or dependency path between x and y is considered insignificant and hence neglected.

A synthetic example to explain the above phenomenon in relation to the formation of DE gene specific regulatory pathways is given in Figure 4.2. This figure depicts the application of PCIT to form the DE gene specific paths for the DE gene 'X' with 'A', 'B', 'C', and 'D' as the TF genes present at the different Pareto optimal fronts. Here the direct connectivities, A→X, A→D, and B→X turn out to be insignificant and hence are marked for deletion (shown as dashed lines in the figure) in the presence of the trio of genes that can be seen in each case from the figure. To be more specific, considering one of these probable deletions, A→X is insignificant and is marked for deletion considering the network trios A→B→X, A→C→X and A→D→X. After deleting the insignificant paths, for DE gene 'X', the different paths of interaction following the Pareto optimal fronts are A→B→C→X, A→B→D→X, A→B→C→D→X and A→C→X, and A→C→D→X. Among these, the path showing the maximum average weight is chosen for the target DE gene 'X'. In other words, the path possessing the maximum average weight among W1avg = 1/3*(w1+w6+w7), W2avg = 1/3*(w1+w4+w10), W3avg = 1/4*(w1+w7+w8+w10), W4avg = 1/2*(w2+w6) and W5avg = 1/3*(w2+w8+w10) is considered to be the regulatory path for DE gene 'X'.

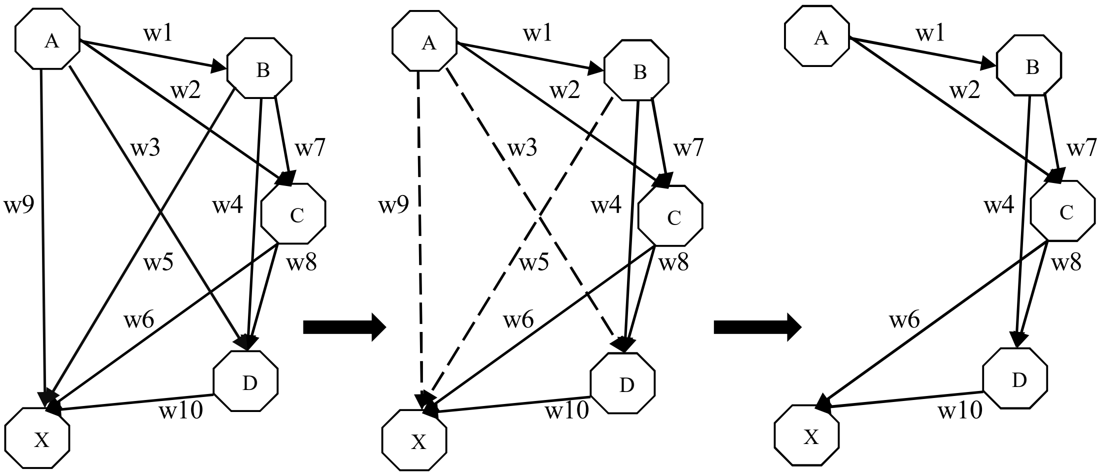


Figure 4.2: Stepwise analysis using PCIT for a target DE gene X with A, B, C, D as front wise TF genes

<u>Developing entropy minimized regulatory network</u>: In this segment, step 5 of the proposed algorithm is discussed in detail followed by simple verification of the condition stated in step 6; thus validating the entropy minimized TF regulatory network formation from a multi-objective perspective.

Initiating with the conditional entropy based network formation process given in [**40**], composite conditional entropy technique has been proposed and implemented in this research work involving TF genes from multiple Pareto optimal fronts. As per the algorithm, the one or more TF genes placed at an upper front is/are dominating one or more TF genes at the immediate lower front and likewise the dominating or regulation capability of all the TF genes involved in DE gene specific pathways can be optimally structured using the novel concept composite conditional entropy followed by application of steepest descent technique [**40**] to remove redundant TF gene regulators across the various Pareto optimal fronts.

Here, all real valued expression vectors $\bar{c}$ present in a certain gene expression profile can be converted to a defined Boolean vector $\bar{b}$ with a probability function, $p(\bar{b}/\bar{c}) = \prod_{i/b_i=1} \frac{1}{1+e^{-c_i}} \prod_{i/b_i=0} (1 - \frac{1}{1+e^{-c_i}})$, where $\frac{1}{1+e^{-c}}$ is the probability that one particular real value c corresponds to the Boolean level 1. In a dataset possessing 'n' independent and identical gene expression profiles, the probability of having the defined $\bar{b}$ can be expressed as $P(\bar{b}) = \frac{1}{n} \sum_{\bar{c}^j \in profiles} p(\bar{b} / \bar{c}^j)$. Considering all possible $\bar{b}$ vectors, the basic entropy considering the distribution of real valued entities can be described as $H = \sum_{all\ possible\ combinations\ of\ \bar{b}} P(\bar{b}) \log\,[1/P(\bar{b})]$. This basic entropy between a regulatee TF gene placed at a lower front with immediate upper front TF genes acting as regulators is computed in two different contexts. In one case, considering the regulatee gene ‘x’ with set of regulators ‘Y_x’, the entropy component H^C (Y_x, x) is calculated and in the other only considering the regulator genes ‘Y_x’, the entropy component H^C (Y_x) is found. From these two findings, the conditional entropy comprising the gene ‘x’ and its regulators ‘Y_x’ gets defined as $H^C(x/Y_x) = H^C(x,Y_x) - H^C(Y_x)$ [**40**]. This process is explained using the sample network given in Figure 4.3 with Tables 4.8 and 4.9.

Table 4.8: Initial Boolean TF to target regulatory vectors (column wise)

B	0	0	1	1
A2	0	1	0	1
A3	0	1	0	1

Table 4.9: Revised Boolean TF to target regulatory vectors (column wise)

B	0	0	1	1
A1	0	1	0	1
A2	0	1	0	1
A3	0	1	0	1

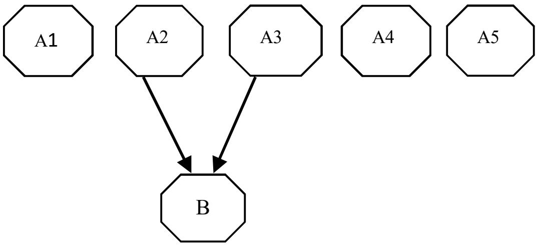


Figure 4.3: Initial portion of a composite TRN showing reported interactions from biological databases

From the reported set of interactions obtained from various biological databases, it is being assumed that Y_x comprises initially of TF genes A2 and A3, i.e. Y_x = {A2,A3} and the regulated TF gene at the immediate lower front is B. Applying the Boolean vectors (column wise) from each of the Tables 4.8 and 4.9 individually on the gene expression profiles, the initial and revised conditional entropies are computed. Here, the initial conditional entropy works on the combination (x=B, Y_x= {A2, A3}) and the revised one with the supposed inclusion (not present in the reported biological databases) of TF gene, A1, i.e. (x=B, Y_x= {A1, A2, A3}) defines the revised combination. As per the Figure 4.3, the other six combinations of TF genes that may work at the immediate upper front are {A2, A3, A4}, {A2, A3, A5}, {A1, A2, A3, A4}, {A1, A2, A3, A5}, {A2, A3, A4, A5}, and {A1, A2, A3, A4, A5}. Thus in total as per the figure, 1(initial) and 7 (revised) TF combinations are possible in the regulation process. It is evident that with more inclusions the conditional entropy for all these revised combinations will be better than the initial one, i.e. H^C (x/Y_x). Hence, to devise a concrete regulation only those revised combinations are chosen for which the improvement is greater than the average by three standard deviations; finally choosing the combination with minimal conditional entropy. If no such revised combination comes out then the initial 'Y_x' retains the regulation of 'x'.

The extension of this concept is going to add the regulations present at all the other fronts and thus the situation does not remain restricted to two levels only and helps in developing a composite regulatory network. Developing the composite regulatory network through composite conditional entropy can be discussed using Figure 4.4 with Tables 4.10 and 4.11. Here, the solid lined interactions shown in Figure 4.4 indicate the reported interactions present in the various biological databases and the dashed lined interactions have been added maintaining minimal conditional entropy in between the regulatee and the combination of regulators. From the figure, this has been shown for the combinations, (x=B1, Y_x= {A1, A2, A5}) and (x=C2, Y_x= {B1, B2, B3}). The extension of the initial conditional entropy with the inclusion of additional layers of TF genes has been made composite using, $H^C(x/Y_x/Y_{Y_x}) = H^C(x,Y_x/Y_{Y_x}) - H^C(Y_x/Y_{Y_x})$.

Table 4.10: Initial composite Boolean TF to target regulatory vectors (column wise)

C2	0	0	0	0	1	1	1	1
B2	1	1	1	1	1	1	1	1
B3	1	1	1	1	1	1	1	1
A2	0	0	1	1	0	0	1	1
A3	0	0	1	1	0	0	1	1
A2	0	1	0	1	0	1	0	1
A3	0	1	0	1	0	1	0	1
A4	0	1	0	1	0	1	0	1

Table 4.11: Revised composite Boolean TF to target regulatory vectors (column wise)

C2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
B1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
A2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
A5	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
A2	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
A3	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
A2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
A3	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
A4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

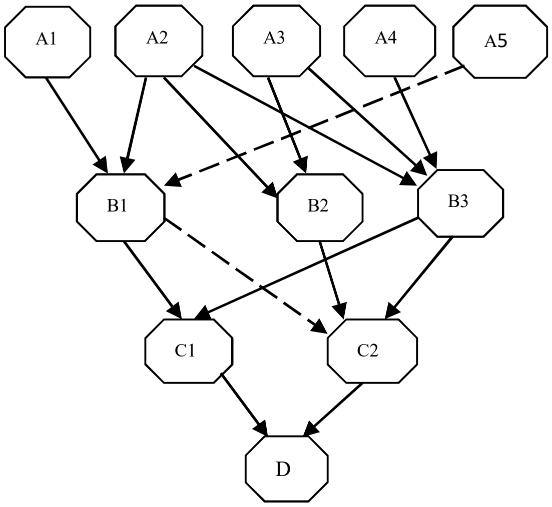


Figure 4.4: The entropy minimized composite TRN showing all sorts of gene interactions

In this equation, considering x=C2 would have Y_x = {B1, B2, B3} and the regulators of Y_x as {A1, A2}, {A2, A3}, {A2, A3, A4}. In other words, $H^C(Y_x/Y_{Y_x})$ computes the conditional entropy considering the involvement of TF genes from the layers A and B, and $H^C(x, Y_x/Y_{Y_x})$ considering the involvement of all the concerned genes from layers A, B, and C, i.e. (x=C2, Y_x= {B1, B2, B3}, Y_{Y_x} = {A1, A2}, {A2, A3}, {A2, A3, A4}) combination. From these two conditional entropy components, the composite conditional entropy is calculated as $H^C(x/Y_x/Y_{Y_x})$. While generating the Boolean vectors it has been considered that the target or the source TF gene can either be in on or off state (1/0), but in order to pass the required message to the destination the channel must always be in on condition. Investigating through the various Boolean matrices given in the relevant tables, 0/1 parallel regulatory control executed by the source Pareto-optimal nodes gets clarified. All the intermediary nodes apart from those in the last front are assigned 1 irrespective of individual or parallel control. In the last revised matrix, given in Table 4.11, there are 3 intermediate nodes. Thus for any particular condition of gene C2 (i.e., 0/1) there are 2^3 combinations of [{A1, A2, A5},{A2, A3},{A2, A3, A4}]. This is similarly followed by checking the improvement in entropy as discussed earlier to select the best regulator set for composite regulations. Now while checking the improvement in entropy, if the addition of gene B1 is worthy enough to be included as a part of the composite regulatory path for gene C2, then it is being considered to take part in further regulation. Otherwise, the current regulatory structure excluding B1 is retained (which has not happened in this structure). Another case as shown in Figure 4.4 is the inclusion of A5 with reported regulators A1 and A2 for B1. In a similar way it is possible to form the regulatory architectures for C1 and D shown in Figure 4.4.

The final challenge after the creation of the above composite entropy minimized regulatory network is to check redundancies in the interactions between TF genes at various Pareto optimal fronts. In other words, it is about understanding the significant role of every TF gene in an upper front regulating another TF gene present at the immediate lower front. For example, considering the combination (x=B3, Y_x= {A2, A3, A4}), the gene expression values of any one of the regulator genes is altered across all profiles/samples using the formula, $x^{new} = x^{old} - \Delta\lambda$, where x^{new} indicates the altered or new expression level and x^{old} being the old value. Here, Δ is the partial derivative of the conditional entropy, $H^C(x/Y_x)$, with respect to c_{ij} (expression value of ith gene in jth profile), i.e. $\Delta = \frac{\partial}{\partial c_{ij}} H^C(x/Y_x)$. The alteration of expression levels is conducted till some steady state entropy value gets attained. This operation is conducted for each TF gene acting as regulator, i.e. A2, A3, A4, one at a time in the above example. The number of iterations through which the steady state conditional entropy is attained is the deciding factor of existence of the TF gene acting as a regulator in a collaborative framework. In the above context, λ =1 to reflect the effect on the stability of the network due to the actual change in entropy contributed by the factor, Δ.

4.3.3 Results: The implementation of the proposed algorithm has been carried on one synthetic dataset followed by three real expression matrices. In very case, the information is contained across more than one condition. The purpose of taking such varied information is to check the robustness of the algorithm in forming stable TRNs.

In the synthetic context, a dataset is generated using the tool, SynTREN [**47**]. Here, an expression matrix is generated for 200 genes defined over 100 samples. NMF [**48**] analysis on the samples classifies 56 from condition 1 and the remaining 44 from condition 2. Total number of TF genes found from the tool is 20, out of which 3 are TF_DE and 7 are TF_DC. From the remaining 180 genes, there are 36 DE genes and 37 DC genes.

The first real data used is budding yeast, Saccharomyces cerevisiae, cell cycle data consisting of 6,178 genes across four conditions. From this set, 17 TF genes are found. The necessary details of this data is available at [**49**]. While comparing every pair of conditions it gets revealed that conditions 1 and 4 give maximum number of DE genes with 6 TF and 235 non-TF genes. Again, 9TF and 320 non-TF genes are found through DC analysis of the dataset.

The second real data used is mice embryonic testis development dataset comprising of 12,488 genes for both male and female phenotypes (GSE 1358) [**50**]. Here, 90 TF genes are found, out of which 37 are TF_DE and 23 are TF_DC genes. From the remaining non-TF genes, 3,320 are DE in nature and 734 of DC type.

The third and final real data used for this research is the RMA expression data of liver samples from human subjects with HCV cirrhosis with and without concomitant HCC (GSE 17967) [**42**]. In total there are 22,277 genes, among which 131 are TF genes, 260 are DE genes, and 598 are DC genes. Again, out of 131 TF genes, 8 are DE and 57 are DC in nature.

After discovering the initial inputs (the physically existent TF, TF_DE, TF_DC, DE genes) in each of the above cases, the differential regulatory power of the TF, TF_DC, and TF_DE genes are assessed through multi-objective analysis based on the three regulatory scores, namely RIF II, dCSA_r2t, and dCSA_t2t. In the context of yeast and synthetic data the TF to target gene information required for assessing the regulatory powers is fetched from the web based tool, YEASTRACT [**45**]. However, the same operation is conducted for mice and human data utilizing the information present in TF2target database [**42**] and TRRUST [**51**]. Post this, sequence of results following the proposed algorithm is given next for each of the above experimentations (here unless specified, TF genes mean pure TF or TF_DC or TF_DE genes).

Phase 1:

- 7 TF genes distributed across 5 Pareto optimal fronts for the synthetic dataset, containing 1, 1, 1, 1, 3 TF genes in the respective fronts starting from primary non-dominated optimal front.
- 12 TF genes distributed across 3 Pareto optimal fronts for yeast data, containing 5, 6, 1 TF genes in the respective fronts starting from primary non-dominated optimal front.
- 64 TF genes distributed across 4 Pareto optimal fronts for mice data, containing 57, 4, 1, 2 TF genes in the respective fronts starting from primary non-dominated optimal front.
- 18 TF genes distributed across 8 Pareto optimal fronts for human data, containing 3, 3, 6, 2, 1, 1, 1, 1 TF genes in the respective fronts starting from primary non-dominated optimal front.

Phase 2:

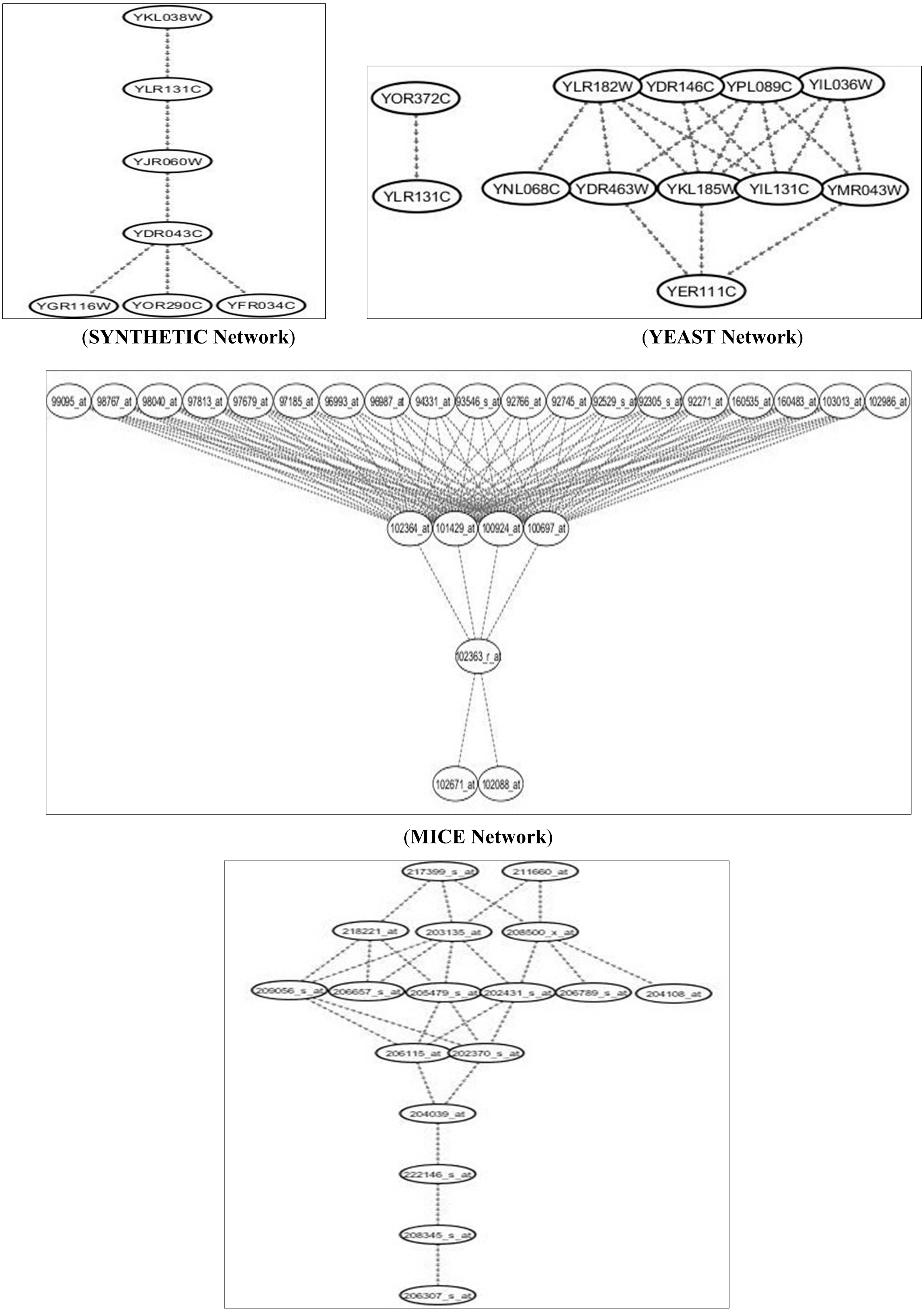
- The number of TF regulatory paths corresponding to synthetic, yeast, mice, and human data is 3, 30, 456, and 108 respectively after application of TF gene regulation constraints through PCIT analysis discussed earlier followed by significance assessment. In this regard, the number of DE gene specific pathways for the respective cases happen to be 3×36=108, 30×235=7,050, 456×3320 = 15,13,920, 108×260=28,080.
- For mice, because of having a humongous number of regulatory pathways, only the first 100 significant DE genes are considered for further analysis. This point will be clarified at a later phase for each of the cases.
- Apart from a single instance in yeast for all the other datasets there is no biological evidence of regulation skipping intermediate front TFs. In other words, the TF genes involved in different biological (KEGG) pathways strictly follow the sequence of regulation devised through this algorithm.

Phase 3:

- Composite entropy validation of these regulatory paths obtained for each case yields the stable entropy minimized networks given in Figure 4.5.
- In the entropy stabilized networks, the assessment of parallel redundant regulatory paths in the collaborative regulations between consecutive front pairs, based on the number of iterations following the steepest descent approach, is given Figures 4.6(A), 4.6(B), 4.6(C), 4.6(D) respectively. In this context, the grey bars correspond to those TF gene regulators possessing consistent stable regulatory action with the regulatee and black bars highlight regulators having far less stable regulatory control over the regulatee and hence neglected for further composite entropy analysis.
- The number of common and different TF genes for each regulatee between entropy validated networks and corresponding databases [**42,45,51**] is given Figure 4.7. Here, the grey bars indicate the database reported interactions within which the black portions signify the number of overlapped TF genes.

Phase 4:

- The statistical enrichment of these entropy minimized regulatory pathways is done which yields p-value = 0 for the synthetic data TF gene regulatory paths, with the p-value enrichments for yeast, mice and humans enlisted in Tables 4.12, 4.13, and 4.14 respectively.
- It is also noticed that final pruned TF regulatory networks obtained for all the datasets are mostly identical or subsets of the corresponding regulatory structures devised through PCIT in the proposed algorithm.
- As per the entropy minimized structure, only synthetic and mice data yields two or more TF genes at the last Pareto optimal front. For both yeast and human data, only one TF gene is found significantly existent at the last Pareto optimal front. Hence, understanding the importance of the TF regulators in controlling the DE genes that come next, is relevant for synthetic and mice data only.
- In case of the synthetic data, all the 3 TF genes lying at the last front are equally important in controlling each of the 30 target DE genes. However, in case of mice, considering the first 100 significant DE genes (reason stated in Phase 2 above), only 1 target DE gene has a single controller TF gene, 102671_at.



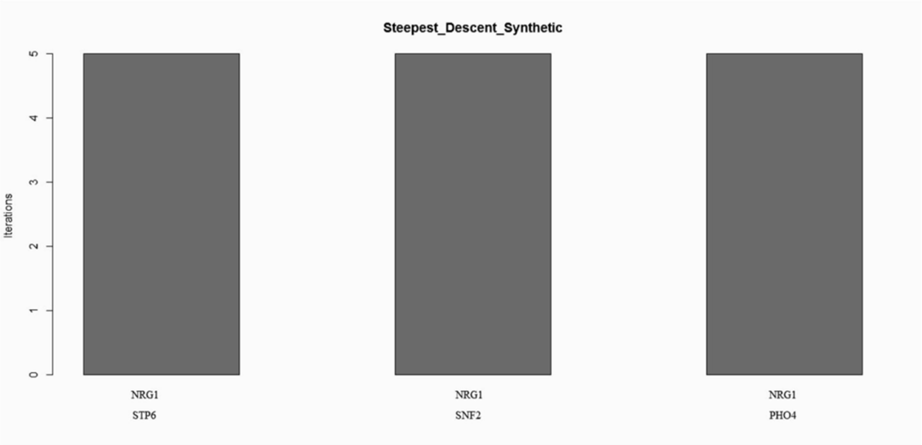
(**SYNTHETIC Network**)

(**YEAST Network**)

(**MICE Network**)

(**HUMAN Network**)

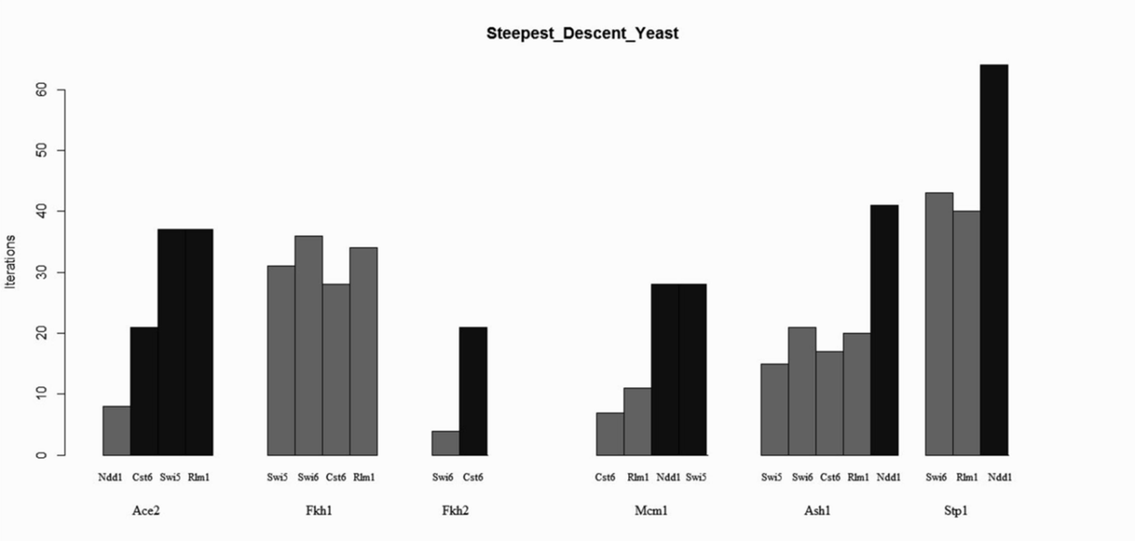
Figure 4.5: Entropy based TF Regulatory Architectures in the Multiobjective Framework



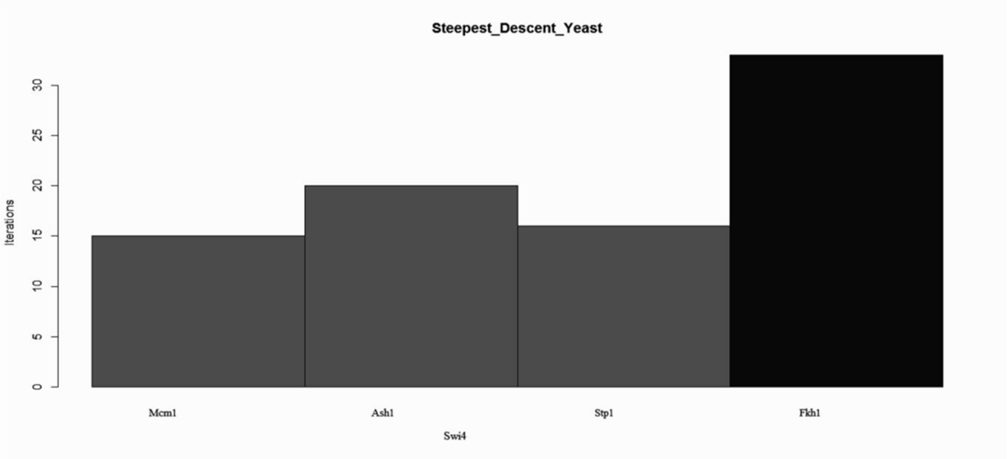
(In between 4th and 5th Pareto optimal fronts)

Figure 4.6 (A): Assessing the redundant regulatory paths between consecutive front pairs

SYNTHETIC Network



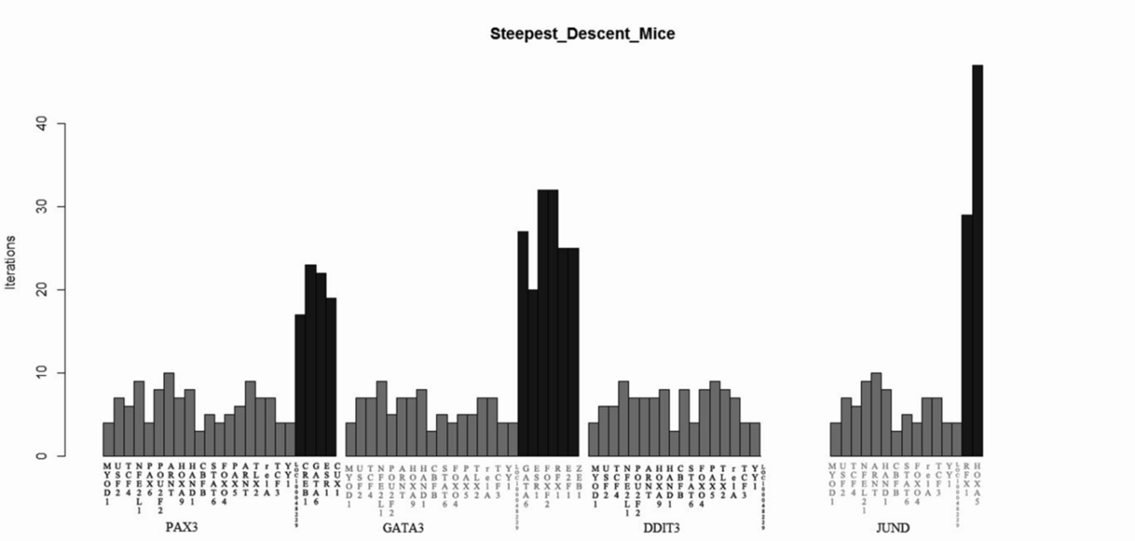
(In between 1st and 2nd Pareto optimal fronts)



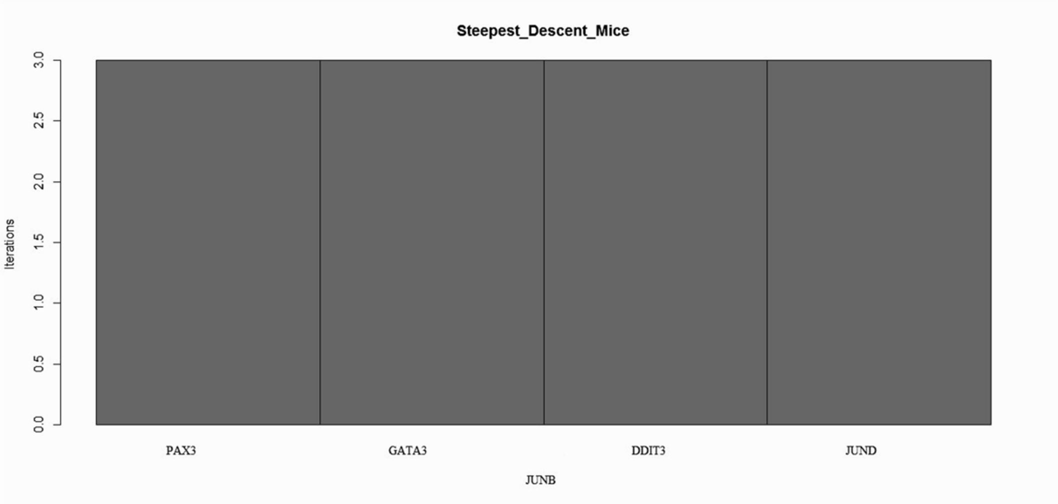
(In between 2nd and 3rd Pareto optimal fronts)

Figure 4.6 (B): Assessing the redundant regulatory paths between consecutive front pairs

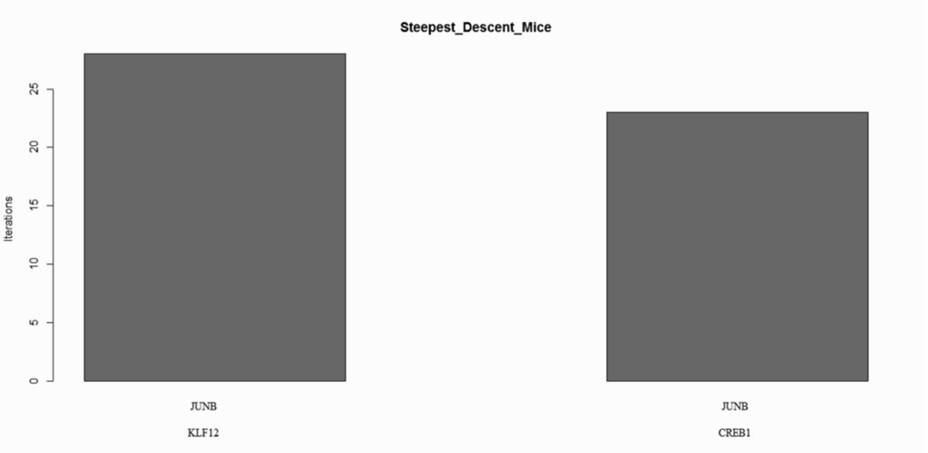
YEAST Network



(In between 1st and 2nd Pareto optimal fronts)



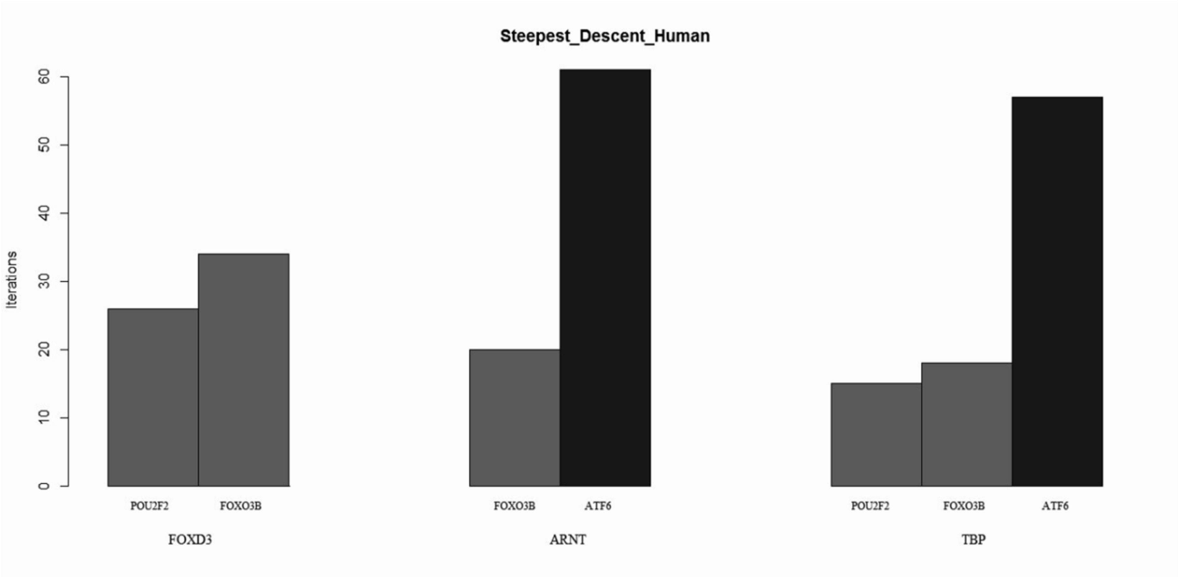
(In between 2nd and 3rd Pareto optimal fronts)



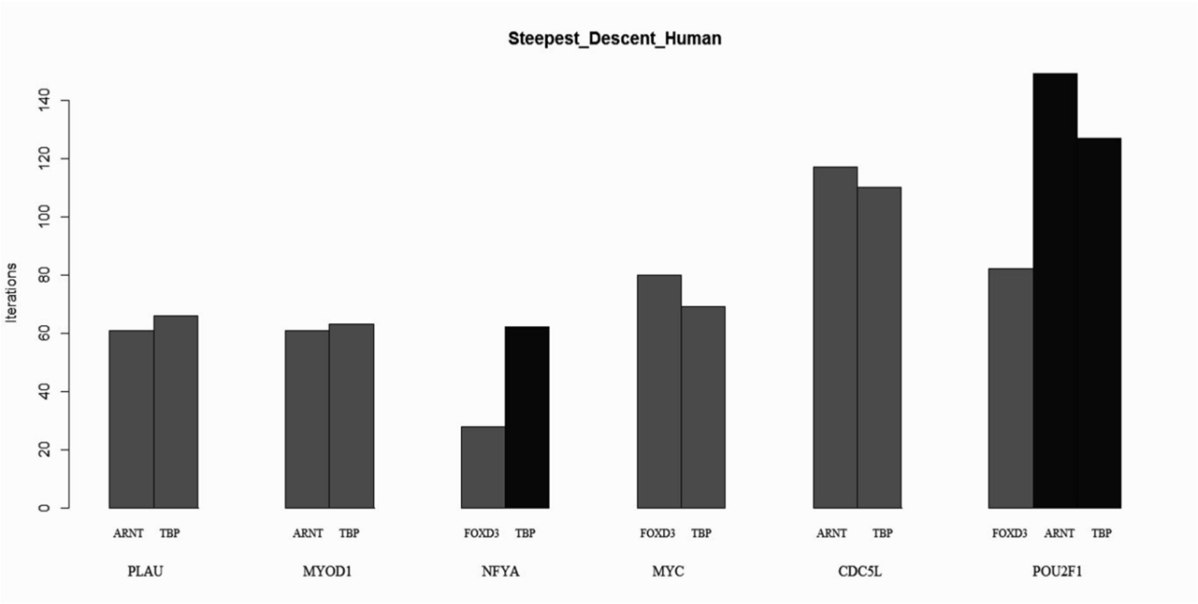
(In between 3rd and 4th Pareto optimal fronts)

Figure 4.6 (C): Assessing the redundant regulatory paths between consecutive front pairs

MICE Network

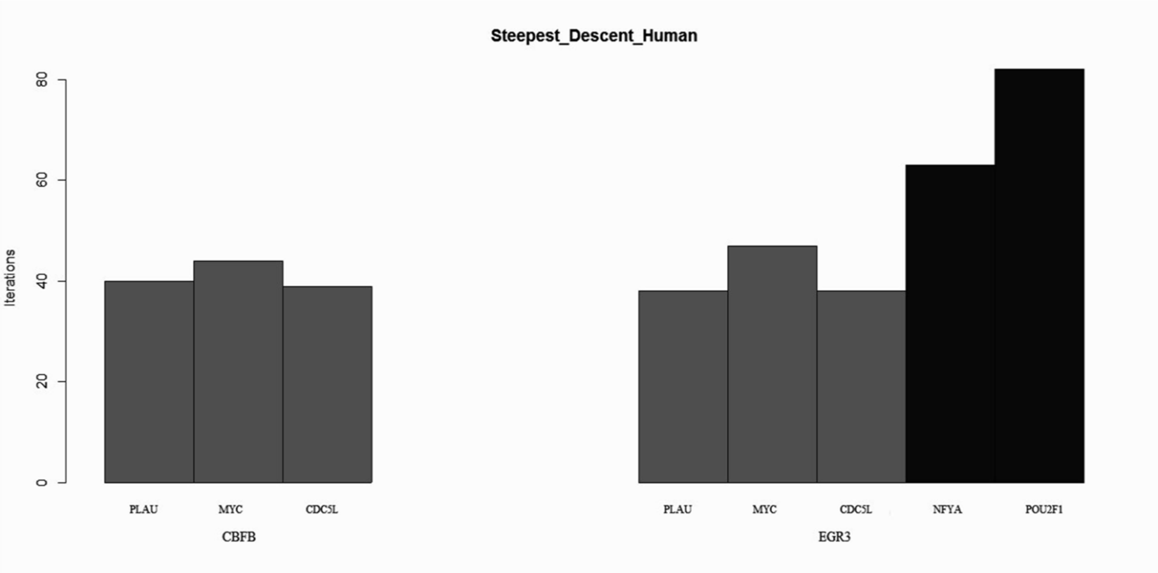


(In between 1st and 2nd Pareto optimal fronts)

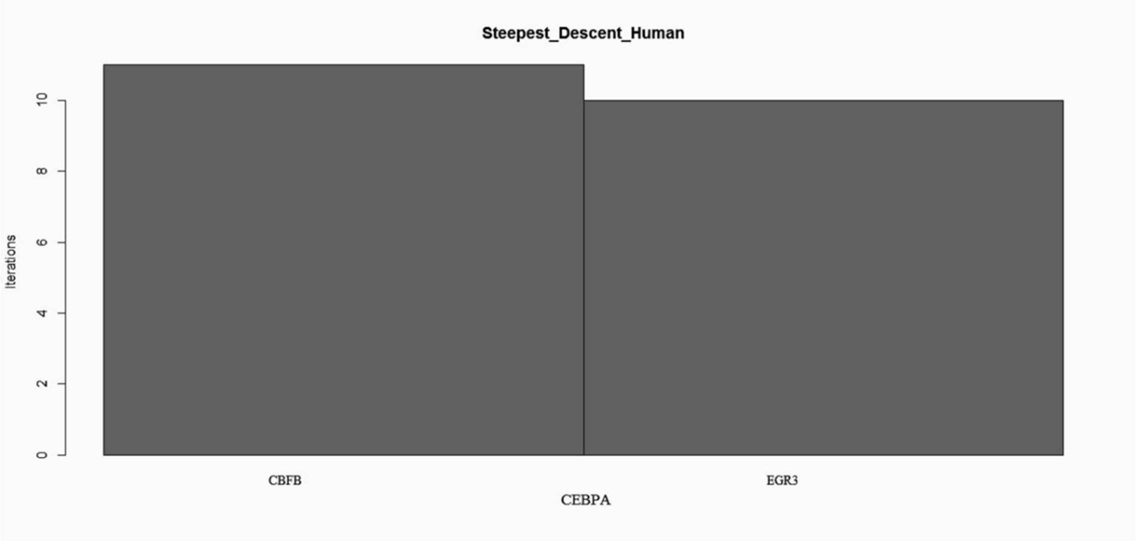


(In between 2nd and 3rd Pareto optimal fronts)

Figure 4.6 (D): Assessing the redundant regulatory paths between consecutive front pairs **HUMAN Network** (Continued)



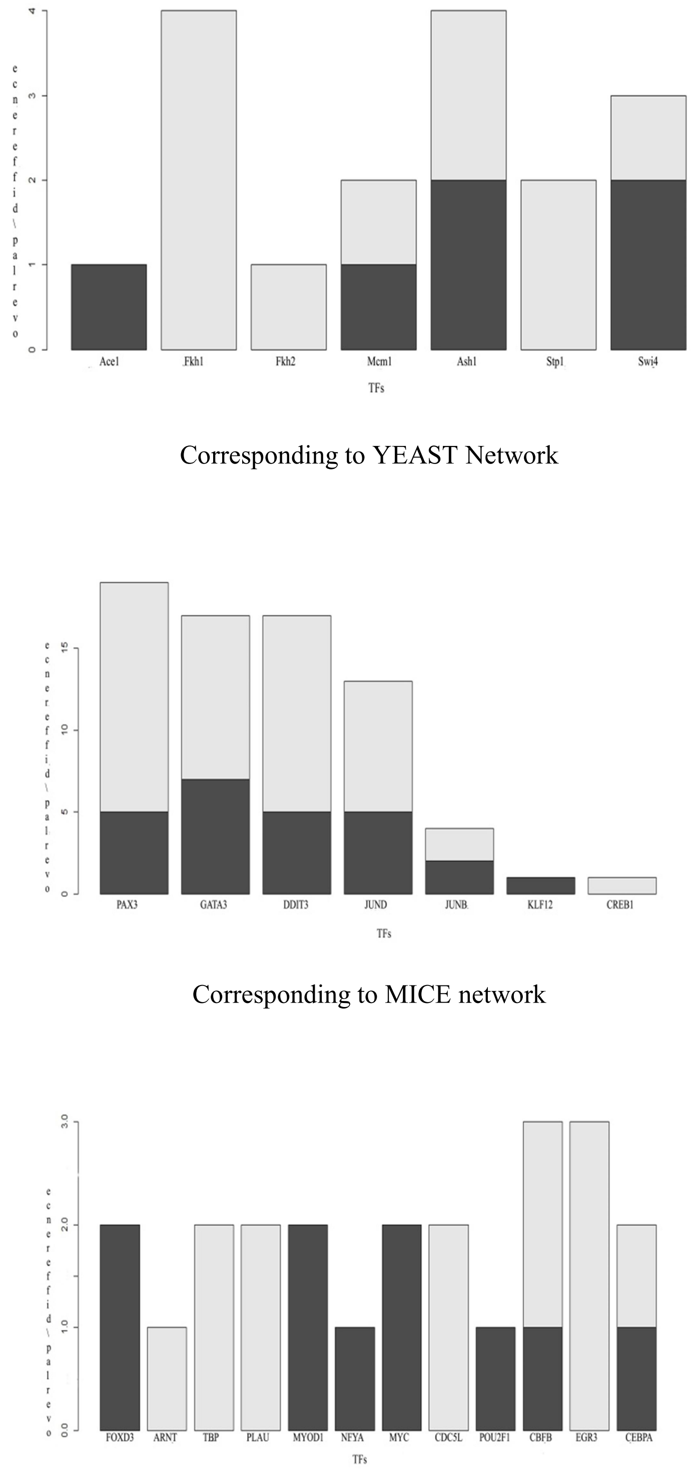
(In between 3rd and 4th Pareto optimal fronts)



(In between 4th and 5th Pareto optimal fronts)

Figure 4.6 (D): Assessing the redundant regulatory paths between consecutive front pairs

HUMAN Network



Corresponding to YEAST Network

Corresponding to MICE network

Corresponding to HUMAN network

Figure 4.7: The number of common and different TF genes for each regulatee between entropy validated networks and corresponding databases.

Table 4.12: Statistical enrichment scores of entropy minimized TF gene regulatory pathways (YEAST)

Yeast TF genes			
Front1	**Front2**	**Front3**	**p-value**
YOR372C	YLR131C	-	0.504
YDR146C	YIL131C	-	0.444
YDR146C	YKL185W	YER111C	0.01556009
YLR182W	YIL131C	-	0.468
YLR182W	YNL068C	-	0.288
YLR182W	YKL185W	YER111C	0.01825004
YLR182W	YDR463W	YER111C	0.00796159
YIL036W	YIL131C	-	0.656
YIL036W	YMR043W	YER111C	0.02232027
YIL036W	YKL185W	YER111C	0.02486027
YPL089C	YIL131C	-	0.596
YPL089C	YMR043W	YER111C	0.01988202
YPL089C	YKL185W	YER111C	0.02211508
YPL089C	YDR463W	YER111C	0.00879171

Table 4.13: Statistical enrichment scores of entropy minimized TF gene regulatory pathways (MICE)

Mice TF genes				
Front1	**Front2**	**Front3**	**Front4**	**p-value**
102986_at	100697_at	102363_r_at	102088_at	0.000038
102986_at	100697_at	102363_r_at	102671_at	0.000116
103013_at	100697_at	102363_r_at	102088_at	0.000038
103013_at	100697_at	102363_r_at	102671_at	0.000116
160483_at	100697_at	102363_r_at	102088_at	0.000038
160483_at	100697_at	102363_r_at	102671_at	0.000116
160535_at	100697_at	102363_r_at	102088_at	0.000038
160535_at	100697_at	102363_r_at	102671_at	0.000116
92271_at	100697_at	102363_r_at	102088_at	0.000038
92271_at	100697_at	102363_r_at	102671_at	0.000116
92305_s_at	100697_at	102363_r_at	102088_at	0.000038
92305_s_at	100697_at	102363_r_at	102671_at	0.000116
92529_s_at	100697_at	102363_r_at	102088_at	0.000038
92529_s_at	100697_at	102363_r_at	102671_at	0.000116
92745_at	100697_at	102363_r_at	102088_at	0.000038
92745_at	100697_at	102363_r_at	102671_at	0.000116
92766_at	100697_at	102363_r_at	102088_at	0.000038
92766_at	100697_at	102363_r_at	102671_at	0.000116
93546_s_at	100697_at	102363_r_at	102088_at	0.000038
93546_s_at	100697_at	102363_r_at	102671_at	0.000116
94331_at	100697_at	102363_r_at	102088_at	0.000038
94331_at	100697_at	102363_r_at	102671_at	0.000116
96987_at	100697_at	102363_r_at	102088_at	0.000038
96987_at	100697_at	102363_r_at	102671_at	0.000116
96993_at	100697_at	102363_r_at	102088_at	0.000038
96993_at	100697_at	102363_r_at	102671_at	0.000116
97185_at	100697_at	102363_r_at	102088_at	0.000038
97185_at	100697_at	102363_r_at	102671_at	0.000116
97679_at	100697_at	102363_r_at	102088_at	0.000038
97679_at	100697_at	102363_r_at	102671_at	0.000116
97813_at	100697_at	102363_r_at	102088_at	0.000038
97813_at	100697_at	102363_r_at	102671_at	0.000116
98040_at	100697_at	102363_r_at	102088_at	0.000038
98040_at	100697_at	102363_r_at	102671_at	0.000116
98767_at	100697_at	102363_r_at	102088_at	0.000038
98767_at	100697_at	102363_r_at	102671_at	0.000116
99095_at	100697_at	102363_r_at	102088_at	0.000038
99095_at	100697_at	102363_r_at	102671_at	0.000116
102986_at	100924_at	102363_r_at	102088_at	0.000046
102986_at	100924_at	102363_r_at	102671_at	0.000148
103013_at	100924_at	102363_r_at	102088_at	0.000046
103013_at	100924_at	102363_r_at	102671_at	0.000148
160483_at	100924_at	102363_r_at	102088_at	0.000046
160483_at	100924_at	102363_r_at	102671_at	0.000148
160535_at	100924_at	102363_r_at	102088_at	0.000046
160535_at	100924_at	102363_r_at	102671_at	0.000148

Mice TF genes				
Front1	Front2	Front3	Front4	p-value
92305_s_at	100924_at	102363_r_at	102088_at	0.000046
92305_s_at	100924_at	102363_r_at	102671_at	0.000148
92529_s_at	100924_at	102363_r_at	102088_at	0.000046
92529_s_at	100924_at	102363_r_at	102671_at	0.000148
92745_at	100924_at	102363_r_at	102088_at	0.000046
92745_at	100924_at	102363_r_at	102671_at	0.000148
92766_at	100924_at	102363_r_at	102088_at	0.000046
92766_at	100924_at	102363_r_at	102671_at	0.000148
93546_s_at	100924_at	102363_r_at	102088_at	0.000046
93546_s_at	100924_at	102363_r_at	102671_at	0.000148
94331_at	100924_at	102363_r_at	102088_at	0.000046
94331_at	100924_at	102363_r_at	102671_at	0.000148
96987_at	100924_at	102363_r_at	102088_at	0.000046
96987_at	100924_at	102363_r_at	102671_at	0.000148
96993_at	100924_at	102363_r_at	102088_at	0.000046
96993_at	100924_at	102363_r_at	102671_at	0.000148
97679_at	100924_at	102363_r_at	102088_at	0.000046
97679_at	100924_at	102363_r_at	102671_at	0.000148
97813_at	100924_at	102363_r_at	102088_at	0.000046
97813_at	100924_at	102363_r_at	102671_at	0.000148
98040_at	100924_at	102363_r_at	102088_at	0.000046
98040_at	100924_at	102363_r_at	102671_at	0.000148
98767_at	100924_at	102363_r_at	102088_at	0.000046
98767_at	100924_at	102363_r_at	102671_at	0.000148
99095_at	100924_at	102363_r_at	102088_at	0.000046
99095_at	100924_at	102363_r_at	102671_at	0.000148
102986_at	101429_at	102363_r_at	102088_at	0.000048
102986_at	101429_at	102363_r_at	102671_at	0.000163
103013_at	101429_at	102363_r_at	102088_at	0.000048
103013_at	101429_at	102363_r_at	102671_at	0.000163
160483_at	101429_at	102363_r_at	102088_at	0.000048
160483_at	101429_at	102363_r_at	102671_at	0.000163
160535_at	101429_at	102363_r_at	102088_at	0.000048
160535_at	101429_at	102363_r_at	102671_at	0.000163
92305_s_at	101429_at	102363_r_at	102088_at	0.000048
92305_s_at	101429_at	102363_r_at	102671_at	0.000163
92529_s_at	101429_at	102363_r_at	102088_at	0.000048
92529_s_at	101429_at	102363_r_at	102671_at	0.000163
92745_at	101429_at	102363_r_at	102088_at	0.000048
92745_at	101429_at	102363_r_at	102671_at	0.000163
92766_at	101429_at	102363_r_at	102088_at	0.000048
92766_at	101429_at	102363_r_at	102671_at	0.000163
93546_s_at	101429_at	102363_r_at	102088_at	0.000048
93546_s_at	101429_at	102363_r_at	102671_at	0.000163
94331_at	101429_at	102363_r_at	102088_at	0.000048
94331_at	101429_at	102363_r_at	102671_at	0.000163
96987_at	101429_at	102363_r_at	102088_at	0.000048
96987_at	101429_at	102363_r_at	102671_at	0.000163
96993_at	101429_at	102363_r_at	102088_at	0.000048
96993_at	101429_at	102363_r_at	102671_at	0.000163
97679_at	101429_at	102363_r_at	102088_at	0.000048
97679_at	101429_at	102363_r_at	102671_at	0.000163
97813_at	101429_at	102363_r_at	102088_at	0.000048
97813_at	101429_at	102363_r_at	102671_at	0.000163
98040_at	101429_at	102363_r_at	102088_at	0.000048
98040_at	101429_at	102363_r_at	102671_at	0.000163
98767_at	101429_at	102363_r_at	102088_at	0.000048
98767_at	101429_at	102363_r_at	102671_at	0.000163
99095_at	101429_at	102363_r_at	102088_at	0.000048
99095_at	101429_at	102363_r_at	102671_at	0.000163
102986_at	102364_at	102363_r_at	102088_at	0.000102
102986_at	102364_at	102363_r_at	102671_at	0.000374
103013_at	102364_at	102363_r_at	102088_at	0.000102
103013_at	102364_at	102363_r_at	102671_at	0.000374
160483_at	102364_at	102363_r_at	102088_at	0.000102
160483_at	102364_at	102363_r_at	102671_at	0.000374
160535_at	102364_at	102363_r_at	102088_at	0.000102
160535_at	102364_at	102363_r_at	102671_at	0.000374
92529_s_at	102364_at	102363_r_at	102088_at	0.000102
92529_s_at	102364_at	102363_r_at	102671_at	0.000374
92766_at	102364_at	102363_r_at	102088_at	0.000102
92766_at	102364_at	102363_r_at	102671_at	0.000374

Mice TF genes				
Front1	**Front2**	**Front3**	**Front4**	**p-value**
93546_s_at	102364_at	102363_r_at	102088_at	0.000102
93546_s_at	102364_at	102363_r_at	102671_at	0.000374
94331_at	102364_at	102363_r_at	102088_at	0.000102
94331_at	102364_at	102363_r_at	102671_at	0.000374
96987_at	102364_at	102363_r_at	102088_at	0.000102
96987_at	102364_at	102363_r_at	102671_at	0.000374
97813_at	102364_at	102363_r_at	102088_at	0.000102
97813_at	102364_at	102363_r_at	102671_at	0.000374
98040_at	102364_at	102363_r_at	102088_at	0.000102
98040_at	102364_at	102363_r_at	102671_at	0.000374
98767_at	102364_at	102363_r_at	102088_at	0.000102
98767_at	102364_at	102363_r_at	102671_at	0.000374
99095_at	102364_at	102363_r_at	102088_at	0.000102
99095_at	102364_at	102363_r_at	102671_at	0.000374

Table 4.14: Statistical enrichment scores of entropy minimized TF gene regulatory pathways (HUMANs)

Human TF genes								
Front1	**Front2**	**Front3**	**Front4**	**Front5**	**Front6**	**Front7**	**Front8**	**p-value**
217399_s_at	208500_x_at	202431_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	208500_x_at	202431_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	208500_x_at	204108_at	-	-	-	-	-	5.56E-05
217399_s_at	208500_x_at	206789_s_at	-	-	-	-	-	2.95E-04
217399_s_at	218221_at	205479_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	218221_at	205479_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	218221_at	206657_s_at	-	-	-	-	-	5.56E-05
217399_s_at	218221_at	209056_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	218221_at	209056_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	205479_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	205479_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	209056_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	209056_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	202431_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	202431_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
217399_s_at	203135_at	206657_s_at	-	-	-	-	-	1.11E-04
211660_at	208500_x_at	202431_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	208500_x_at	202431_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	208500_x_at	206789_s_at	-	-	-	-	-	0
211660_at	208500_x_at	204108_at	-	-	-	-	-	7.79E-05
211660_at	203135_at	209056_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	209056_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	202431_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	202431_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	205479_s_at	206115_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	205479_s_at	202370_s_at	204039_at	222146_s_at	208345_s_at	206307_s_at	0
211660_at	203135_at	206657_s_at	-	-	-	-	-	4.67E-04

4.3.4 Discussion: The outcomes of this research either fully match with the biological databases [**42,45,51**] or add some significant regulations in a biologically existent single or collaborative framework or delete some existing regulations present in the biological databases based on the weakness of regulations revealed through the steepest descent technique applied in the proposed algorithm. Thus this composite entropy minimized approach highlights new interactions which can be studied or analyzed further or predict deletions of some existent interactions reported in the databases. This will definitely help the biological community in the true understanding of the differential role of any

TF or DE gene. Some of the major cases where additions and deletions have been reported by this implementation are given below.

Significant Additions:

- The regulation of *YMR043W* in yeast can be taken as an example. According to YEASTRACT, it is regulated by only *YIL036W*, but the composite entropy minimized algorithm shows us another TF gene, *YPL089C* along with *YIL036W* regulating *YMR043W* in parallel.
- Another example is the regulation of *100697_at* in mice. Formal database shows *102986_at*, *160483_at*, *93546_s_at*, *92305_s_at*, and *98040_at* to be its potential regulators. Through the application of composite entropy minimized algorithm, TF genes such as *103013_at*, *160535_at*, *92745_at*, *92766_at*, *94331_at*, *96987_at*, *96993_at*, *97185_at*, *97679_at*, *97813_at*, *98767_at*, *92271_at*, *92529_a_at*, and *99095_at* are also added to be its potential collaborative regulators in the TRN.

Significant Deletion:

- Here an example can be the regulation of *YLR131C* in yeast. From the database *YOR372C*, *YDR146C* and *YIL036W* are its possible regulators. However with the application of composite entropy, *YIL036W* and *YDR146C* gets knocked out due to weak regulation, yielding *YOR372C* as the sole regulator of *YLR131C*.

Significant Deletions and Addition:

- In this context, the regulation of *204108_at* in humans can be considered as an example. Reported interactions from the database show *203135_at* and *218221_at* to be its potential regulators. However, the composite entropy approach predicts both as weak regulators, compared to *208500_x_at*, absent in the database corresponding to this regulation. Hence, the traditional regulators are better to be ignored and *208500_x_at* can be considered to be the sole regulator for further biological study.

Apart from these types of alterations, there are some interactions in mice and human composite entropy minimized networks which report bidirectional regulation in the biological database, TRRUST [**51**]. The important examples in this regard are given below.

- The interactive regulation between *206789_s_at* and *208500_x_at* in humans is one example. Here, both the TF genes are DC in nature. Hence, any one is capable of controlling the other. But as per the composite entropy minimized network, *208500_x_at* is the controller of *206789_s_at*, but not vice-versa.
- A similar interactive case is present between *208500_x_at* and *211660_at* in humans. Here, the first one is a TF_DC gene and the second is a pure TF gene. Thus, as per the composite entropy minimized approach, regulation of *211660_at* by *208500_x_at* has not been observed.
- A collaborative exception found in mice network can be treated as another example in this regard. Here, as per TRRUST, *92305_s_at*, *97185_at*, *96993_at*, *98767_at* act as interactive TF genes with *100924_at*. Among these TF genes *97185_at*, *96993_at*, and *100924_at* all are TF_DC in nature. Hence, theoretical bidirectional regulation is possible within these TF genes. On the contrary, *92305_s_at* and *98767_at* are identified as pure TF genes making the regulation unidirectional with respect to the regulatee TF gene, *100924_at*. However, as per the composite entropy minimized network in mice, all the four regulator TF genes collaboratively control the regulatee TF gene, *100924_at*.

4.4 Conclusion

In this chapter, the concept of single and collaborative gene regulations has been brought to limelight through minimization of multi-objective constraints which help to determine the differential regulatory power of a TF or DE gene. In this regard, obtaining various types of genes across multiple Pareto optimal fronts and their differential regulatory properties helps in understanding the formation of regulatory pathways in any complex biological network.

Here, the challenge in reconstruction of TRNs is dependent on the initiator or source regulator in any pathway. The pure TF genes (neither TF_DC nor TF_DE types) have been considered as the source of the various regulatory cascades present in the composite entropy minimized network. In this regard, a number of regulatory cascades or pathways could have started from a pure TF gene placed at a lower Pareto optimal front. But with the presence of TF_DC and/or TF_DE genes placed at higher Pareto optimal fronts according to the outcome of the minimization of conflicting differential regulatory scores in a multi-objective paradigm, a significant number of pure TF genes placed at the lower fronts got discarded from further analysis. Due to this an appreciable number of TF gene regulatory pathways could not be studied. This aspect has primarily affected the composite entropy minimized network design with respect to human liver expression data with HCV cirrhosis. Hence, to understand the regulatory statistics, formation of the network is expected to incorporate pure TF genes from the lower Pareto optimal fronts as well.

4.5 References

[**1**] J. Handl, J. Knowles, "On semi-supervised clustering via multi-objective optimization", In Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, GECCO'06, 1465–1472, ACM, New York, July 2006, https://doi.org/10.1145/1143997.1144238

[**2**] P. Mitra, C.A. Murthy, S.K. Pal, "Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3), 301–312, March 2002, https://doi.org/10.1109/34.990133

[**3**] U. Maulik, A. Mukhopadhyay, S. Bandyopadhyay, "Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes", BMC Bioinformatics, 10, Article No. 27, January 2009, https://doi.org/10.1186/1471-2105-10-27

[**4**] S. Bandyopadhyay, A. Mukhopadhyay, U. Maulik, "An improved algorithm for clustering gene expression data", Bioinformatics, 23(21), 2859–2865, November 2007, https://doi.org/10.1093/bioinformatics/btm418

[**5**] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, "A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions". PLoS ONE, 7(4):e32289, April 2012, https://doi.org/10.1371/journal.pone.0032289

[**6**] A. Konak, D.W. Coit, A.E. Smith, "Multi-objective optimization using genetic algorithms: a tutorial", Reliability Engineering and System Safety, 91(9), 992–1007, September 2006, https://doi.org/10.1016/j.ress.2005.11.018

[**7**] A.K. Alok, S. Saha, A. Ekbal, "Feature selection and semi-supervised clustering using multiobjective optimization", In 2014 International Conference on Soft Computing and Machine Intelligence, New Delhi, India, 126-129, September 2014, https://doi.org/10.1109/ISCMI.2014.19

[**8**] D.E. Goldberg, "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison-Wesley, Boston, 1989

[**9**] M.A. Khanesar, M. Teshnehlab, M.A. Shoorehdeli, "A novel binary particle swarm optimization", In Proceedings of the 15th Mediterranean Conference on Control and

Automation, Athens, Greece, 1-6, June 2007, https://doi.org/10.1109/MED.2007.4433821

[**10**] D.F. Specht, "A general regression neural network", IEEE Transactions on Neural Networks, 2(6), 568–576, November 1991, https://doi.org/10.1109/72.97934

[**11**] M. Dorigo, C. Blum, "Ant colony optimization theory: a survey", Theoretical Computer. Science, 344(2-3), 243–278, November 2005, https://doi.org/10.1016/j.tcs.2005.05.020

[**12**] M. Kanehisa, S. Goto, "KEGG: Kyoto encyclopaedia of genes and genomics", Nucleic Acids Research, 28(1), 27-30, January 2000, https://doi.org/10.1093/nar/28.1.27

[**13**] F. Wilcoxon, "Individual comparisons by ranking methods", Biometrics Bulletin, 1(6), 80–83, December 1945, https://doi.org/10.2307/3001968

[**14**] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", Proc. Natl. Acad. Sci. USA 102(43), 15545–15550, October 2005, https://doi.org/10.1073/pnas.0506580102

[**15**] C. Wu, J. Zhu, X. Zhang, "Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma", BMC Bioinformatics, 14, Article No.365, December 2013, https://doi.org/10.1186/1471-2105-14-365

[**16**] M. Sarkar, A. Majumder, "TOP: an algorithm in search of biologically enriched differentially connective gene networks", In 5th Annual International Conference on Advances in Biotechnology, BioTech 2015, Kanpur, India, 2015, https://doi.org/10.5176/2251-2489_BIOTECH15.39

[**17**] L. Wang, Z. Fenq, X. Wang, X. Wang, X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data", Bioinformatics, 26(1), 136–138, January 2010, https://doi.org/10.1093/bioinformatics/btp612

[**18**] M. Sarkar, A. Majumder, "Quantitative trait specific differential expression (qtDE)", Procedia Computer Science, 46, 706–718, 2015, https://doi.org/10.1016/j.procs.2015.02.131

[**19**] A.M. Yip, S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure", BMC Bioinformatics, 8, Article No.22, January 2007, https://doi.org/10.1186/1471-2105-8-22

[**20**] A. Reverter, E.K.F. Chan, "Combining partial correlation and an information theory approach to the reversed engineering of gene co expression networks", Bioinformatics, 24(21), 2491–2497, November 2008, https://doi.org/10.1093/bioinformatics/btn482

[**21**] Wolfram Mathworld. http://www.mathworld.wolfram.com

[**22**] A. Majumder, M. Sarkar, "Exploring different stages of Alzheimer's disease through topological analysis of differentially expressed genetic networks", International Journal of Computer Theory and Engineering, 6(5), 386–391, October 2014, DOI: 10.7763/IJCTE.2014.V6.895

[**23**] A. Ghazalpour *et al.*, "Integrating genetic and network analysis to characterize genes related to mouse weight", PLoS Genetics, 2(8):e130, August 2006, https://doi.org/10.1371/journal.pgen.0020130

[**24**] http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/MouseWeight/

[**25**] Y. Choi, C-G. Hur, T. Park, "Induction of olfaction and cancer-related genes in mice fed a high-fat diet as assessed through the mode of action by network identification analysis", PLoS ONE, 8(3):e56610, March 2013, https://doi.org/10.1371/journal.pone.0056610

[**26**] A. Oshimoto *et al.,* "Potential role of transient receptor potential channel M5 in sensing putative pheromones in mouse olfactory sensory neurons", PLoS ONE, 8(4): e61990, April 2013, https://doi.org/10.1371/journal.pone.0061990

[**27**] A. Cruz, J. Nieto, J. Moreno, C. Canavate, P. Desjeux, J. Alvar, "Leishmania/HIV co-infections in the second decade", The Indian Journal of Medical Research, 123(3), 357–388, March 2006

[**28**] W. Zhao, X. Ji, F. Zhang, L. Li, L. Ma, "Embryonic stem cell markers", Molecules, 17(6), 6196–6246, June 2012, https://doi.org/10.3390/molecules17066196

[**29**] D.G. Baker, "Natural pathogens of laboratory mice, rats, and rabbits and their effects on research", Clinical Microbiology Reviews, 11(2), 231–266, April 1998, https://doi.org/10.1128/cmr.11.2.231

[**30**] T. Jaffredo, L. Yvernogeau, "How the avian model has pioneered the field of hematopoietic development", Experimental Hematology, 42(8), 661–668, August 2014, https://doi.org/10.1016/j.exphem.2014.05.009

[**31**] A. Bandyopadhyay, K. Tsuji, K. Cox, B.D. Harfe, V. Rosen, C.J. Tabin, "Genetic analysis of the roles of BMP2, BMP4, and BMP7 in limb patterning and skeletogenesis", PLoS Genetics, 2(12):e216, December 2006, https://doi.org/10.1371/journal.pgen.0020216

[**32**] J.L. Wynn, H.R. Wong, "Pathophysiology and treatment of septic shock in neonates", Clinics in Perinatology, 37(2), 439–479, June 2010, https://doi.org/10.1016/j.clp.2010.04.002

[**33**] A. Patil, Y. Kumaga, K-C. Liang, Y. Suzuki, K. Nakai, "Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response", PLoS Computational Biology, 9(11):e1003323, November 2013, https://doi.org/10.1371/journal.pcbi.1003323

[**34**] J. Hamazaki, K. Sasaki, H. Kawahara, S-I. Hisanaga, K. Tanaka, S. Murata, "Rpn10-mediated degradation of ubiquitinated proteins is essential for mouse development", Molecular and Cellular Biology, 27(19), 6629–6638, October 2007, https://doi.org/10.1128/mcb.00509-07

[**35**] J. Menche *et al.*, "A diVIsive shuffling approach (VIStA) for gene expression analysis to identify subtypes in chronic obstructive pulmonary disease", BMC Systems Biology, 8, Article No.S8, March 2014, https://doi.org/10.1186/1752-0509-8-S2-S8

[**36**] D. Guan, J. Shao, Z. Zhao, *et al.*, "PTHGRN: unravelling posttranslational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data", Nucleic Acids Research, 42(W1), W130-W136, July 2014, https://doi.org/10.1093/nar/gku471

[**37**] I. Simon, J. Barnett, N. Hannett, *et al.*, "Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle", Cell, 106(6): 697-708, September 2001, https://doi.org/10.1016/S0092-8674(01)00494-9

[**38**] A. Blias, B.D. Dynlacht, "Constructing transcriptional regulatory networks", Genes & Development, 19: 1499-511, 2005, doi:10.1101/gad.1325605

[**39**] I. Nemenman, F. Shaffe, and W. Bialek. "Entropy and Inference revisited", In T.G. Dietterich, S. Becker, and Z. Ghahramani editors, Advances in Neural Information Processing Systems, 14, 471-478, Cambridge, MA, MIT press, 2002, https://dl.acm.org/doi/10.5555/2980539.2980601

[**40**] G. Karlebach, R. Shamir, "Constructing Logical Models of Gene Regulatory Networks by Integrating Transcription Factor–DNA Interactions with Expression Data: An Entropy-Based Approach", Journal of Computational Biology, 19(1), 30-41, January 2012, https://doi.org/10.1089/cmb.2011.0100

[**41**] A. Reverter, N.J. Hudson, S.H. Nagaraj, M. Pérez-Enciso, B.P. Dalrymple, "Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data", Bioinformatics, 26(7), 896-904, April 2010, https://doi.org/10.1093/bioinformatics/btq051

[**42**] J. Yang, H. Yu, B.H. Liu, *et al.*, "DCGLv2.0: An R package for unveiling differential regulation from differential co expression", PLoS ONE, 8(11): e79729, November 2013, https://doi.org/10.1371/journal.pone.0079729

[**43**] A. Essaghir, F. Toffalini, L. Knoops, A. Kallin, Jv. Helden, J.B. Demoulin, "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data", Nucleic Acids Research, 38(11): e120, June 2010, https://doi.org/10.1093/nar/gkq149

[**44**] H. Yu, R. Mitra, J. Yang, Y. Li, Z. Zhao, "Algorithms for network based identification of differential regulators from transcriptome data: a systematic evaluation", Science China Life Sciences, 57, 1090-1102, November 2014, https://doi.org/10.1007/s11427-014-4762-7

[**45**] P.T. Monteiro, N.D. Mendes, M.C. Teixeira, *et al.*, "YEASTRACT DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae", Nucleic Acids Research, 36 (Database Issue), D132-136, January 2008, https://doi.org/10.1093/nar/gkm976

[**46**] D. Amar, H. Safar, R. Shamir, "Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression", PLoS Computational Biology, 9(3): e1002955, March 2013, https://doi.org/10.1371/journal.pcbi.1002955

[**47**] T. Van den Bulcke, K. Van Leemput, B. Naudts, *et al.*, "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms",

BMC Bioinformatics, 7, Article No. 43, January 2006, https://doi.org/10.1186/1471-2105-7-43

[**48**] Q. Qi, Y. Zhao, M. Li, R. Simon, "Non-negative matrix factorization of gene expression profiles: a plug-in for BRB-Array Tools", Bioinformatics, 25(4), 545-547, February 2009, https://doi.org/10.1093/bioinformatics/btp009

[**49**] P.T. Spellman, G. Sherlock, M.Q. Zhang, *et al.*, "Comprehensive identifications of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", Molecular Biology of the Cell, 9(12), 3273-3297, December 1998, published ONLINE in October 2017, https://doi.org/10.1091/mbc.9.12.3273

[**50**] C.L. Small, J.E. Shima, M. Uzumcu, M.K. Skinner, M.D. Griswold, "Profiling gene expression during the differentiation and development of the murine embryonic gonad", Biology of Reproduction, 72(2), 492-501, February 2005, https://doi.org/10.1095/biolreprod.104.033696

[**51**] H. Han, H. Shim, D. Shin, et al., "TRRUST: a reference database of human transcriptional regulatory interactions", Scientific Reports, 5, Article No.11432, June 2015, https://doi.org/10.1038/srep11432

Chapter 5 Developing Time Variant Transcription Factor Regulatory Networks for Differentially Expressed Genes

Related Publication

[**1**] A. Majumder, M. Sarkar, and P. Sharma, “A Composite Mode Differential Gene Regulatory Architecture based on Temporal Expression Profiles”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, volume 16, issue 6, pages 1785-1793, November-December 2019.
https://doi.org/10.1109/tcbb.2018.2828418

5.1 Introduction

Transcriptional Regulatory Networks (TRNs) are essential complex biological interaction mechanisms used to monitor the differential gene activities across various cell lines and conditions [**1**]. In this perspective, the differential gene analyses in cell division cycles [**2**] carry optimum importance to monitor the reasons behind different types of malignant and neural diseases and to find suitable drug targets [**3**] in therapeutic investigations. One of the primary ways to study the differential gene activity is investigating the differential expression (DE) capability of genes followed by the topological differences of interconnectivity among these DE genes under some defined cell lines and conditions (extensively researched and documented in the previous chapters). Another way to look into the differential gene activity in complex interactive networks is about understanding the role of any specific type of regulator gene responsible for the alterations in gene network functionalities. This has also been looked into in the previous chapter, but disregarding any specific time properties involved in the regulatory process. The transcription factor (TF) genes that are mostly responsible for such regulations happens to create regulatory proteins that govern and control the cell cycle across various stages or conditions [**4**]. Moving into the scope of time dependent regulations, it is expected that TF genes can control any target gene activity within a certain timeline (a matter of extensive research) maintaining unique regulatory effects following a delay characteristic which mainly addresses translation time of one or more TF genes, corresponding protein folding time, translocation time, promoter binding time, and transcription time of the target gene.

This form of time dependent regulation exhibited by the TF genes can be of two types, such as time invariant and time variant categories. The time invariant category (referenced later) considers the entire time course of gene expression activity to understand the dominating regulatory effect in a certain state of the cell cycle governed by environmental stimulus or different forms of external perturbation factors. Thus either activator or repressor actions (i.e. the dominating regulatory effect mentioned above) on the target genes get defined in a certain state with specific delay properties. On the other hand, the time variant category (the matter of research presented in this chapter) considers the activation or repression of the target genes by the TF genes at some specific activation points [**5,6**] of the entire time course information. Hence, in the latter case, the dominating regulatory effect considering time delay characteristics, if any, may alter within a certain state or condition of the cell cycle. Thus the differentially regulatory effect (between a TF and DE gene for example) under different stages or conditions, considering the importance of the time course data, can also portray significant alteration in the time variant architecture; an occurrence not possible to observe and study further through time invariant regulation at certain state or condition. Hence, the matter of research indulged in this chapter is related to design of time variant TF regulatory networks; thus contributing to the development dynamic TF regulatory architectures for the DE genes.

5.2 Time Variant Transcriptional Regulatory Networks

In any Transcriptional Regulatory Network (TRN), be it designed on a dataset consisting of independent and identical sample profiles or highly correlative time course profiles, the TF genes can act individually or may possess a collaborative action to control one or more target DE genes. Fundamental aspects of time regulation based on highly correlative time course profiles given in Table 5.1 are discussed in [**5**,**7**]. Here, the regulatory action of a TF gene on any target gene is classified based on the TF gene's functional role as an activator or repressor on one hand and its logical role as necessary or sufficient on the other. The inference on TF to target gene association in TRNs can be framed as per the prominent regulatory aspects shared in [**8-12**]. However most of these are associated with certain significant shortcomings. Among them the cost and time complexity to check for all possible combinations of higher order gene knockout [**5**] cannot be ignored. One possible solution to this problem is considering those TF genes having proper biological evidence to bring phenotypic changes and the other can be restricting TF gene inclusion to form a higher order group, provided their interaction pattern with target genes are found significant at an individual level or in groups of lower order [**5**]. Another significant limitation, highlighted earlier as well, is the usage of the entire time course information in a certain condition for studying the regulatory control of TFs on the target DE genes. This, as mentioned earlier, can be resolved checking the importance of the TF to DE gene interaction at some specific time points depicting a particular type of temporal expression pattern.

The onset of the specific time course analysis is about recording the significant changes in the expression pattern of all concerned genes (here the interest is concentrated on TF and DE genes corresponding to the development of any TRN) as the time course profile is surpassed at a certain state or condition. In this context, the significant up-expression or down-expression of any gene with respect to consecutive time points of interest helps in understanding the combinations of functional and logical roles of regulation within a specific time frame. The combination or mode of regulation that can be defined in a particular context at a certain time point is shared in Table 5.2.

Table 5.1: Different regulatory interactions of a time course TRN design

Role	Regulatory Interaction	Function
Functional	Activator	Expression change of target gene is homogenous with its TF gene
	Repressor	Expression change of target gene is heterogeneous with its TF gene
Logical or Directional	Necessary	Down expression of TF genes leads to the response opposite of the functional role
	Sufficient	Up expression of TF genes leads to the response analogous to the functional role
	Necessary and Sufficient	Up expression of TF genes leads to the response analogous to the functional role and down expression initiates to the response just opposite of the functional role

Table 5.2: Different modes based on expression change of TF and target gene at any time point

TF gene		Target gene		Mode
⬆	⬇	⬆	⬇	
Yes	--	Yes	--	AS
Yes	--	--	Yes	RS
--	Yes	--	Yes	AN
--	Yes	Yes	--	RN

From Table 5.2, the interaction between the up/ down expression level of TF gene and target gene yields a one to one association between the two logical and functional roles extracting four independent sets of dependence, namely Activator Sufficient (AS), Repressor Sufficient (RS), Activator Necessary (AN), and Repressor Necessary (RN).

The interaction study involving the above sets of dependence or modes of regulation did get its significance through the seminal research works [**5,7,13,14**]. In these seminal works the model TRIM followed mTRIM got realized. TRIM works on hidden Markov modelling which can at the most develop regulatory network involving a maximum of 2 TF genes. As the biological pathways are complex in nature, this happens to be a trivial approach. Hence, the extended framework mTRIM which incorporates the enhanced expectation maximization based Bayesian Inference approach was implemented to handle multiple TF genes (up to the order of 6) regulatory interactions.

However, none of these methods did investigate the inherent periodicity of gene expression data or quantitatively estimated the strength of regulation involved in the temporal or time regulatory interaction process. These algorithms primarily rely on the fraction of time instants or activation points where the various modes function in a TF to target interaction model. The possible solution of this limitation can be addressed using the algorithms developed in [**15**], which incorporate the change of correlation between TF and DE genes between conditions while searching for significant TF genes. But, the later phase of algorithms [**15**] consider the entire time sequence of TF genes and target genes in the course of regulation and hence time-dependent or time variant models cannot be realized. Inspired by these ideas, a novel concept of computing the TF to target gene regulation strength has been shown in this research chapter using linear correlation at the specific activation time instants depending on the mode of regulation. In this regard, it is noteworthy to mention that non-linear regulation architectures exist in collaborative networks [**16-19**] utilizing the entire time course information under different states or conditions. But the regulation being limited to certain activation time instants and not over the entire time course of the time series sequence, the assumption of linear regulatory effect in single as well as collaborative TF models is expected to highlight true biological regulations in complex interactive TRNs.

5.2.1 The basic findings: This experimentation has brought two new ideas in the formation of TRNs. One of these is about the quantification of the differential regulation in TF to target DE gene interaction within the same specific activation time sequence present in different conditions followed by checking the consistency of this temporal differential regulation with respect to some control or reference state. The other novel concept is focussed on understanding the modes (AS/RS/AN/RN) of this temporal differential regulation between periodic and aperiodic combinations of TF and target DE genes. In the latter perspective, the significance of the research is exemplified through single and collaborative TF to target DE gene direct regulations. In other words, as described in the methodology segment, single interactions modes of temporal differential regulations can be like AS-RS-AS and AS-RN-AS in periodic TF-aperiodic DE gene and aperiodic TF-periodic DE gene regulatory combinations respectively. Again in multi TF regulations or collaborative interactions, the modes of temporal differential regulation that may work together can be like RS-RN-RS and AS-AN-AS in 2 TF genes regulatory combination for a target DE gene in aperiodic TF-aperiodic DE pairs. Thus the stringency of this novel dynamic network reconstruction process is not only limited to checking the consistency of the temporal differential regulatory context across varied conditions with a reference condition in existence but also includes multiple TF genes in a collaborative context where specific time activation sequences gets shorter corresponding to particular temporal patterns of the target DE gene's variation (i.e. up or down expressed) present across different conditions (though this is a true novelty highlighting the actual regulatory pattern in a time series experiment). Hence, the number of regulatory TF genes that may be present in a collaborative architecture is lower than renowned architectures like mTRIM [**5**]. This particular observational difference signifies the importance of any TF to DE gene regulation (present in single or collaborative gene networks) in a temporal perspective on the application of the proposed research work. This indicates the apprehension followed by necessary removal of redundant regulator genes in further analysis, specifically in a collaborative biological interaction environment.

5.2.2 Methodology: The prime novelty introduced in this research chapter is about the underneath logic and framework of the proposed differential regulatory score, RIFT or the *Regulatory Impact Factor with T score*, based on the temporal differential pattern of TF and target DE genes at some specific consecutive time instants similar to one another across different time frames of the cell cycle process. The algorithm in this regard starts off with finding the TF and DE genes present in the cell cycle database along with the discovery of periodicity of these genes from the temporal differential expression patterns.

The RIFT score is judged by the contribution of Affinity score as a mere weightage factor of differential regulatory power exhibited on a target DE gene by one or more TF genes. Prior to this research, the concept of Affinity score (discussed later) was the crux part in the qualitative analysis (i.e. just understanding the probabilistic attitude of time course regulations with respect to any of combinational modes involving functional and logical attributes: AS, RS, AN and RN) of regulatory interactions between TF and target genes. In this research contribution, the presence of Affinity score as a weightage factor in the RIFT score design does turn important in understanding the significance of any regulation occurrence in a single or collaborative TF to DE gene regulatory framework. In this regard, the proposed algorithm for devising time variant TRNs is given next.

ALGORITHM: Forming time variant temporal differential regulatory networks

Input:

I. Time series HeLa Cancer data.

II. TF2target and TRRUST database.

Output:

Significant TF to target DE gene composite interactions in single and collaborative mode.

Step1: Time series data of TF genes is extracted from the HeLa cancer data.

Step2: Time series data of DE genes are extracted from HeLa cancer data using R package *maSigPro*.

Step3: R package *GeneCycle* is utilized to search for periodically expressed TF and DE genes which yield 4 sets of expression patterns (given in Figure 5.1) across the 3 different stages of the periodic cell cycle.

Step4: Differential values are computed between the consecutive time instant expression levels in every phase of the HeLa cancer cell cycle data (in total 3 phases or conditions).

Step5: The pattern of significant expression change per gene from one time instant to the next is depicted as ⬆ (up expression), ⬇(down expression) or ▬ (no change).

Step6: Based on step 5, various TF to target DE gene differential regulatory modes are computed (an example given in Figure 5.1 for illustration purpose).

Step7: A novel interactive score RIFT (given in Equations 2 and 3) is devised between TF and target DE gene utilising the differential valued gene vectors (obtained in Step 4) between phase 1 and 2 (yields RIFT^{21}) and phase 2 and 3 (yields RIFT^{23}), where phase 2 is considered to be the control state.

Step8: If RIFT^{21} and RIFT^{23} depict the same mode of differential regulation (similar sign) for a TF gene to target DE gene combination, such a pair is reserved for significance analysis (single TF gene regulatory significance).

Step9: All qualified combinations (obtained from Step 8) for a target DE gene are checked for significant collaborative architectures (multiple TF genes regulatory significance).

The various equations in relevance to the RIFT framework are given below:

Equation (1A): $AfnScore(I) = \dfrac{P(E_{tf1,}E_{tf2,......,}E_{tfm,}E_{g}) * P(E_{g})}{P(E_{tf1,}E_{tf2,......,}E_{tfm}\ \ )}$

Equation (1B): *AfnScore*(*RS*) = [(Number of time points where the TF gene is up-expressed and target is down-expressed /total number of time points) × (Number of time points where target is down-expressed /total number of time points)] / [(Number of time points where TF gene is up-expressed / total number of time points)]

Equation (1C): *AfnScore*(*RS_multi-TF*) = [(Number of time points where TF$_1$,TF$_2$,....,TF$_m$ genes are up-expressed and target is down-expressed / total number of time points) × (Number of time points where target is down-expressed / total number of time points)] / [(Number of time points where TF$_1$,TF$_2$,....,TF$_m$ genes are up-expressed / total number of time points)]

Equation (2): RIFT score formation

$$RIF_i = \left| \frac{1}{n_{DE}} \sum_{j=1}^{n_{DE}} [(e_{2j}.I_{2ij})^2 - (e_{1j}.I_{1ij})^2] \right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(A)}$$

$$T = \frac{(I_{2ij} - I_{1ij}) - (\mu_2 - \mu_1)}{\sqrt{(\sigma_2^2 + \sigma_1^2)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(B)}$$

where, $T_{overall} = sign(T)\min\left|T_{c_i,c_j}^{u,v}\right|$

$$T = \frac{(I_{2ij} - \mu_2) - (I_{1ij} - \mu_1)}{\sqrt{(\sigma_2^2 + \sigma_1^2)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(C)}$$

$$RIFT = \frac{e_{2j}^2(I_{2ij} - \mu_2)^2 - e_{1j}^2(I_{1ij} - \mu_1)^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(D)}$$

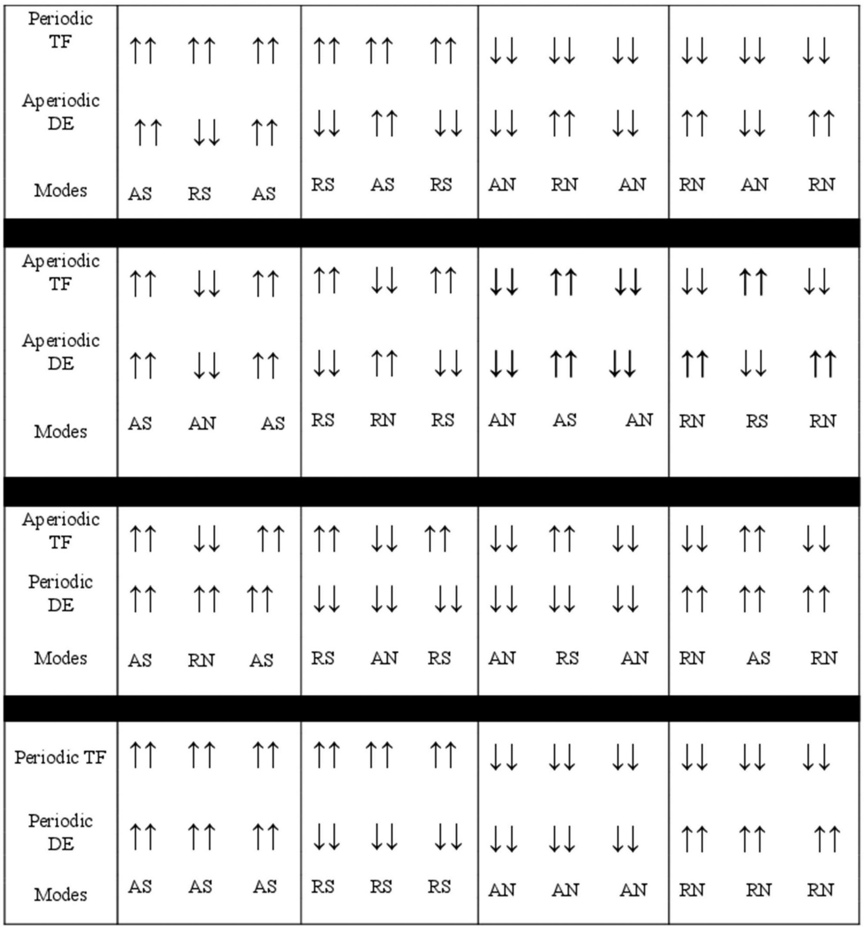
Equation (3): RIFT score utilizing the concept of Affinity score between conditions

$$RIFT_{ij}^{21} = \frac{e_{2j}^2(I_{2ij} \times Afs2 - \mu_2)^2 - e_{1j}^2(I_{1ij} \times Afs1 - \mu_1)^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(A)}$$

$$RIFT_{ij}^{23} = \frac{e_{2j}^2(I_{2ij} \times Afs2 - \mu_2)^2 - e_{3j}^2(I_{3ij} \times Afs3 - \mu_3)^2}{\sqrt{\sigma_3^2 + \sigma_2^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{part(B)}$$

Equation (4): $cor^{ij} = \frac{\Delta t_1}{t} cor_{AS}^{ij} + \frac{\Delta t_2}{t} cor_{RS}^{ij} + \frac{\Delta t_3}{t} cor_{AN}^{ij} + \frac{\Delta t_4}{t} cor_{RN}^{ij}$

The differential regulatory pattern for different combinations of TF and DE genes is given in Figure 5.1.



Periodic TF	↑↑	↑↑	↑↑	↑↑	↑↑	↑↑	↓↓	↓↓	↓↓	↓↓	↓↓	↓↓
Aperiodic DE	↑↑	↓↓	↑↑	↓↓	↑↑	↓↓	↓↓	↑↑	↓↓	↑↑	↓↓	↑↑
Modes	AS	RS	AS	RS	AS	RS	AN	RN	AN	RN	AN	RN
Aperiodic TF	↑↑	↓↓	↑↑	↑↑	↓↓	↑↑	↓↓	↑↑	↓↓	↓↓	↑↑	↓↓
Aperiodic DE	↑↑	↓↓	↑↑	↓↓	↑↑	↓↓	↓↓	↑↑	↓↓	↑↑	↓↓	↑↑
Modes	AS	AN	AS	RS	RN	RS	AN	AS	AN	RN	RS	RN
Aperiodic TF	↑↑	↓↓	↑↑	↑↑	↓↓	↑↑	↓↓	↑↑	↓↓	↓↓	↑↑	↓↓
Periodic DE	↑↑	↑↑	↑↑	↓↓	↓↓	↓↓	↓↓	↓↓	↓↓	↑↑	↑↑	↑↑
Modes	AS	RN	AS	RS	AN	RS	AN	RS	AN	RN	AS	RN
Periodic TF	↑↑	↑↑	↑↑	↑↑	↑↑	↑↑	↓↓	↓↓	↓↓	↓↓	↓↓	↓↓
Periodic DE	↑↑	↑↑	↑↑	↓↓	↓↓	↓↓	↓↓	↓↓	↓↓	↑↑	↑↑	↑↑
Modes	AS	AS	AS	RS	RS	RS	AN	AN	AN	RN	RN	RN

Figure 5.1: Different temporal combinations across conditions depicting differential regulatory patterns

Elucidation of each step followed in the design of the proposed algorithm is given next.

Step 1: The TF gene information of HeLa cancer data is obtained from the TF2target and TRRUST databases [**9,20**], post which the time series information of all the TF genes are extracted from the HeLa dataset.

Step 2: The R package *maSigPro* [**21**] is utilized to extract the time series profiles of the DE genes. Essentially the function utilized here is a regression based phenomenon used to find genes which show significant expression profile difference between experimental groups, especially in time course data. As the HeLa cell cycle information consists of three different phases, the DE genes are computed between conditions or phases 1 and 2 on one side and between phases 2 and 3 on the other side. This is followed by extracting the time profiles conducting intersection of the above results, yielding the expression profiles of common DE genes that will act as target gene expression profiles for further analysis.

Step 3: The R package *GeneCycle* [**22**] used to find periodically expressed genes from a time series experiment, has been utilized to detect the periodic TF and DE genes (from common DE genes obtained in step 2). In those cases, where the package fails to detect any hidden periodicities are classified as aperiodic TF and aperiodic DE genes. After obtaining the periodic information of TF and DE genes, four regulatory combinations can be realized, namely periodic TF-aperiodic DE, aperiodic TF-aperiodic DE, aperiodic TF-periodic DE, and periodic TF-periodic DE respectively.

Step 4: The differential values between consecutive pairs of time instants are calculated in every phase of the cell cycle information to obtain the differential gene vector at each phase for every TF and DE gene.

Step 5: The pattern of significant change of expression level between consecutive time instants gets depicted with the notations ⬆ for up-expression, ⬇ for down expression, and for no significant change in expression level it is ▬.

Step 6: In continuation with step 3, looking into the four types of TF to DE gene regulatory combinations shown in Figure 5.1, it is understandable that the temporal pattern of up or down expression in the same set of consecutive time points (in the figure two consecutive time points have been assumed) across conditions are expected to be similar for both periodic TF and DE genes respectively.

However, for the aperiodic context, there may be the existence of different types of temporal patterns of up and down expression in the same set of consecutive time points across conditions. In this regard, the situations depicting the expression sequence across conditions are (up-expression)-(down-expression)-(up-expression), (down-expression)-(up-expression)-(down-expression), (up-expression)-(down-expression)-(down-expression), (down-expression)-(up-expression)-(up-expression), (up-expression)-(up-expression)-(down-expression), and (down-expression)-(down-expression)-(up-expression) respectively. As per Figure 5.1, to maintain uniformity in the representation process across all the four regulatory combinations, the first two types of expression sequence has been considered for studying the aperiodic TF gene and aperiodic DE gene time course profiles. Considering the last four types of expression sequence, stated above, in periodic TF-aperiodic DE or aperiodic TF-periodic DE mode twelve temporal regulatory combinations having six each per periodic TF or periodic DE representation can be obtained. Again, in aperiodic TF-aperiodic DE mode thirty-six temporal regulatory combinations can be realized having six each per aperiodic TF and aperiodic DE representation. However, for periodic TF-periodic DE, only four temporal modes with two each per periodic TF representation can exist. Hence, the objective of two each per periodic or aperiodic TF representation leads to four temporal regulatory combinations as shown in each row wise segment of Figure 5.1. It is also important to note that a differential temporal pattern in periodic TF-aperiodic DE and aperiodic TF-periodic DE modes (from the point of view of functional role across the 3 stages: activator-repressor-activator or repressor-activator-repressor) and in aperiodic TF-aperiodic DE mode (from the point of view of logical role across the 3 stages: necessary-sufficient-necessary or sufficient-necessary-sufficient) are observed keeping in account the first two types of expression sequence only for aperiodic TF and aperiodic DE genes. This unique property is not observed in case any of the rest four temporal regulatory combinations is considered.

Last but not the least, as an alma mater of this step, it is crucial to mention that though periodic TF-periodic DE mode does not show a differential temporal pattern, it works through the differential regulation utilizing the concept of RIFT just like the other periodic and aperiodic combinations of TF and DE genes shown in Figure 5.1.

The relevant details pertaining to next step, i.e. Step 7, of the algorithm are divided in two parts, Step 7(A) and Step 7(B) respectively. The same is given below.

Step 7A: Here, two RIFT scores are computed. One is between first and second phase and the other between second and third phase of the cell cycle information. The formation of the RIFT score utilizes two differential regulatory concepts together. In this regard, there is the quantitative measurement of differential regulation through the concept of RIF (i.e. the Regulatory Impact Factor) as depicted in Equation (2)-part(A) and the quanti-qualitative measurement of differential regulation via T score depicted in Equation (2)-part(B)&(C).

In the RIF score, the differential regulatory power (a quantitative aspect) of any TF gene can be computed knowing the set of target DE genes under its control. This score does not have any role played by a specific section of time points (the background on which the RIFT scores are designed). From Equation (2)-part(A), the various factors of importance are the ‘e’ and ‘I’ components in addition to n_{DE}, the number of DE genes. In this equation, ‘e’ components indicate the mean expression level of the target DE genes and ‘I’ components indicate the regulatory level of the TF to DE gene interactions under two different conditions. However, the T score (both quantitative as well as qualitative) depicts the change in the actual differential interaction in between two genes with respect to mean differential interaction considering all pairs of genes in presence of the distributions of all differential interaction values under two conditions of interest. This is followed by checking the similarity in sign or consistency of the differential scores obtained across two pairs of conditions within which one condition is treated or assumed to be in the reference or control state. The first part of the T score analysis is quantitative and the latter part verifies the consistency of the quantitative score, i.e. a qualitative assessment. In this regard, the T scores across condition pairs showing similarity in sign (either positive or negative in both pairs) are considered for further evaluations. This perspective of T score can be seen from Equation (2)-part(B). A simple modification or rearrangement of the involved parameters present in the above equation leads to the formation of Equation (2)-part(C).

The basic RIFT score assessing the differential regulatory power between conditions is given in Equation (2)-part(D). This can be seen to be a justified combination of the two different types of regulatory scores, RIF and T depicted in Equation (2)-part(A) and Equation (2)-part(C) respectively.

Step 7B: In this segment the assessment of the differential regulation in between TF and DE genes is done using a modified version of the above RIFT score computed in the two pairs of conditions (in this case in between phase 1 and phase 2 & phase 2 and phase 3 of the cell cycle data). This modification incorporates the concept of Affinity Score defined in Equation (1A). The affinity perspective addresses the temporal patterns of regulation, i.e. AS or RS or AN or RN, present in between periodic and aperiodic combinations of TF and DE genes (discussed earlier) at different time zones, each comprising of a set of consecutive time points, in a certain phase or state of the cell cycle example considered for this research. The relevant description of the parameters involved in Equation (1A) is clearly given in the examples for the case of RS, i.e. Repressor Sufficient, interactions shown in Equation (1B) in between a TF and DE gene and in Equation (1C) in between more than one TF (i.e. multi-TF) and DE gene.

The importance of the affinity in the differential regulation perspective under the two sets of conditions or phases of the cell cycle can be understood from Equation (3)-

part(A) and Equation (3)-part(B) respectively. To be more specific, the Affinity Score in a certain phase acts as a factor of weightage for the co-expression level of interaction between a TF and possible target DE gene. For example, considering a temporal regulatory combination such as AS_RS_AS (found in periodic TF-aperiodic DE combination shown in Figure 5.1), the similar set of consecutive time points at the three different phases are checked for AS in phase 1, RS in phase 2, and AS in phase 3. The Affinity scores in each phase corresponding to AS, RS, and AS are found to put up as the respective weightage factors of regulatory interaction values involved and thus computing the necessary RIFT scores, RIFT^{21} and RIFT^{23}.

In this regard, it is extremely important to define the basis of computing the co-expression interactive values (more than one value because there may be the existence of more than one temporal regulatory combination) in each phase of the cell cycle. In other words, the level of linear correlations computed in these small consecutive sets of time instants are based on the values of the differential gene vectors obtained from step 4 above.

Step 8: The RIFT^{21} and RIFT^{23} scores corresponding to each pair of TF and DE gene, in the various periodic and aperiodic combinations obtained from step 3 above, are checked for similarity in sign, i.e. either both positive or both negative with respect to the control state or phase 2 of the HeLa cell cycle information considered in this research. Such TF-DE gene pairs exhibiting similarity in sign on the basis of any one or more of the temporal mode combinations (all relevant combinations of AS, RS, AN, RN considered in this research are shown in Figure 5.1) are processed for significance analysis. Here, shuffling of the expression profiles obtained from the time course information of the HeLa cell division cycle is done followed by computation of the RIFT scores for every such TF-DE gene pair. Any such TF-DE gene pair is declared significant corresponding to one or more temporal mode combination, provided there are very less number of shuffled cases showing similarity in sign with respect to RIFT^{21} and RIFT^{23}. In this regard, the insignificant pairs can be considered for further pairing and significance check, if applicable. The context of further pairing can also consider significant and insignificant pair combinations.

Step 9: This being the final step of the algorithm, it is possible to obtain more than two TF genes regulating a DE gene (following the various temporal modes of regulation existent in the periodic-aperiodic combinations). Thus this step contributes to the formation of time variant multi TF to DE gene collaborative networks. The earlier step, i.e. step 8, is restricted to the formation of time variant single TF to DE interactive networks.

As an example, considering the second row (aperiodic TF-aperiodic DE combination) of Figure 5.1, the various temporal modes can be marked as AS_AN_AS = 1, RS_RN_RS = 2, AN_AS_AN = 3, and RN_RS_RN = 4. In order to design 2 TF groups from 2 insignificant TF genes or 1 insignificant with 1 significant TF gene, the following maximum number of temporal regulatory subsets can be formed {{1 1}, {2 2}, {3 3}, {4 4}, {1 4}, {4 1}, {2 3}, {3 2}}. Here, the first element in any temporal subset shows the logical and functional role of first TF gene and the second element shows the same but for the second TF gene. In other words, it is possible to have for an aperiodic target DE gene Z, two insignificant aperiodic TF regulators X and Y grouped together with individual roles as AS_AN_AS and RN_RS_RN i.e. {1 4} above. This paired temporal subset and other similar ones shown above are solely dependent on the differential expression pattern sequence of the target DE gene, Z. On assuming t1, t2, and t3 as the three set of time points in the three successive stages of the cell cycle where the corresponding temporal composite modes are activated across conditions, the X-Y grouping for Z can be treated to be a heterogeneous regulatory group based on the pattern of change of the TF expression levels in the three successive stages. Next, considering another network trio from the first row (periodic TF-aperiodic DE combination) of Figure 5.1 where the target Z is aperiodic but the regulator duo (X,Y) is periodic, the similar temporal subset = {1 4} portrays the individual roles as AS_RS_AS and RN_AN_RN. In this case, X-Y grouping is homogeneous. In this way, for 3 TF regulations number of such possible temporal subsets can at the most be 16. Generalizing this possibility can yield the number of such temporal subsets for ‘m’ TF regulations $\leq 2^{m+1}$.

In all such temporal subsets corresponding to collaborative regulations, significance testing can be conducted in the presence of the shuffled cases. However, this significance operation is conducted provided all the TF-DE gene pairs in a collaborative group show consistency or similarity in sign of RIFT scores. In this regard, as an example, Equation (1C) portrays the level of affinity of the RS (Repressor Sufficient) mode at any phase of the cell cycle information in a multi TF interaction context.

Delving into the contribution of this kind of Affinity Score as a weightage factor may bring out an interesting result. Considering an earlier example of 2 TF groups with temporal subset {AS_RS_AS , RN_AN_RN}, it may happen that single TF affinity weightage applied in the RIFT score for any one element of the subset is not able to classify the corresponding element or in other words the concerned TF-DE gene pair significant. However, when grouped together, as mentioned above, the Affinity Score of AS and RN together in phase 1, RS and AN together in phase 2, and AS and RN together in phase 3, acting as the corresponding phase dependent weightage factors in the RIFT score analysis may turn fruitful in deciding the above temporal subset to be a significant collaborative regulation. Thus the contribution of the Affinity Score is bound to vary depending on the type of regulation (single or collaborative).

5.2.3 Results: The experiment has been conducted on the gene expression data obtained from the cell division cycle of HeLa cancer cell line [**1,3**]. This is time series profile information distributed across three different phases, where each phase comprises of 16 time instants of experimentation and there lies a 1 hour gap between each such experimentation conducted in a phase. Hence, in total there are 48 time instants or time course profiles distributed over three phases of the cell cycle.

The fundamental results related to the discovery of TF and DE genes are enlisted below in Table 5.3.

Table 5.3: Number of TF and DE genes

Category of the GENE	Periodicity Issue	Number of genes	Any overlap
TF gene	Periodic	37	4 genes are TF as well as DE
	Aperiodic	100	
DE gene	Periodic	496	
	Aperiodic	123	

Considering phase 2 of the HeLa cancer cell line to be the reference or control state, the number of DE genes obtained between phase 1 and phase 2 is 747 and the same obtained between phase 2 and phase 3 is 697. Between these 619 common DE genes are obtained. These common DE genes are again distributed between periodic and aperiodic profiles as shown in Table 5.3. The 496 periodic DE genes are further classified into 8 clusters following similar pattern of differential distributions in each cluster with 51, 68, 48, 60, 61, 62, 74, and 72 genes in the respective clusters. The same operation when conducted on aperiodic DE genes yields 2 clusters with 69 and 54 genes respectively. The clustering operation [**23**] is conducted to understand DE genes of almost similar distribution and their composite differential regulation pattern through temporal modes with one or more TF genes.

As the seminal work of [**5**] (called mTRIM) developed a concrete time dependent framework considering functional and logical roles of TF to DE gene regulation under single and collaborative situations, the research done and presented in this chapter is compared with mTRIM to understand the fruitfulness, if any, followed by new directions of exploration of further research. The crucial points absent in mTRIM are the periodicity of the gene expression profiles getting considered in the regulatory assessment process and the time varying differential regulation study considering various phases of any cell cycle data. But both of these are of prime importance in the proposed research.

In order to compare these two algorithms, the mTRIM simulation is conducted on the four pattern specific (periodic / aperiodic) regulatory combinations considered in the presented research. To do so, the mTRIM algorithm is applied on each target DE gene present in the respective clusters of aperiodic and periodic DE genes mentioned above. A novel exploration in this regard considers the HUB genes present in each aperiodic and periodic cluster of DE genes. The selection of HUB genes is justified by the fact that any such HUB gene retains maximal regulation in an interconnected network corresponding to a cluster. Accordingly, the HUB genes are searched in each periodic and aperiodic DE cluster using the R package GSAR [**24**] based on linear correlative measure. As both algorithms (mTRIM and the proposed RIFT architecture) under consideration deal with the specific time point regulation of the target DE genes by TF genes, instead of measuring linear correlative dependence over a flat time range the logical and functional attributes of every regulator-regulatee pair is considered while measuring correlation. As an illustration, a cluster with ‘n’ DE genes expressed across ‘t’ time points can be considered to find the HUB gene. Here, between gene ‘i’ and gene ‘j’ it can be considered that all the four possible temporal modes (AS, RS, AN, RN) in some specific time ranges exist, like Δt_1 for AS, Δt_2 for RS, Δt_3 for AN, and Δt_4 for

RN. This is followed by the measurement of temporal mode specific correlations using the differential expression level profiles as retrieved from the proposed algorithm. Thus considering the fraction of time points as a weight factor the overall correlation between gene ‘i’ and gene ‘j’ can be devised as given in Equation (4) under Methodology, i.e. section 5.2.2 of this chapter. This novel form of correlation given in Equation (4) is developed only to highlight and incorporate functional and logical regulatory roles in any basic correlative model. Hence, instead of simple addition of the correlation values at the respective temporal components (highlighting equal significance of functional and logical modes obtained from any time series regulatory sequence), the novel equation does include true weighted contribution or proportional significance of every mode involved.

Comparison of the two algorithms, as stated earlier, based on TF to target matching obtained from TF2target [**9**] and TRRUST [**20**] databases considering each and every target DE gene as well as the HUB gene in each cluster is one of the crucial outcomes of this research. This is followed by the AUROC (Area under Region of Characteristic) plots which prove the proposed approach to be better than mTRIM. However, the humongous amount of information pertaining to TF-DE gene pairs in single and collaborative networks corresponding to all possible differential composite mode regulatory combinations are not possible to be documented as a part of this thesis. Therefore to have an idea of all such possible combinations, one may refer to the appendix of the journal publication associated with this chapter.

The necessary comparison results related to matching with databases and AUROC plots can be represented as shown below.

In Tables 5.4 and 5.5, the percentage values under each case (mTRIM and the proposed RIFT model) indicates the ratio of successful matches considering TF2target database and TTRUST to the total number of statistically significant interactions obtained. In this context, two categories of investigation have been considered. In Category I, all the DE genes present in a cluster are considered. In Category II, only the HUB gene as decided via the novel temporal correlative model (discussed earlier) is taken into consideration. Extending further, under 2 TF genes, *50% / 100%* values report cases where there is evidence of 1 TF or 2 TFs matching present. Similarly for 3 TF genes, *33% / 66% / 100%* values report cases where there is evidence of 1 TF or 2 TFs or 3 TFs matching present. Finally for 4 TF genes, *25% / 50% / 75% / 100%* report cases where there is evidence of 1 TF or 2 TFs or 3 TFs or 4 TFs matching present.

Table 5.4: Category I_TF to target matching considering all DE genes in a cluster

2 TFs (50 % / 100%)			
Combinational Pattern of TF and DE genes	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	12.99 / 9.14%	41.17 / 0%
	2	12.84 / 0.828%	13.51 / 0%
Aptf_Pde	1	13.86 / 1.23 %	12.33 / 0%
	2	13.71 / 1.02 %	5.88 / 0%
	3	16.14 / 1.67 %	20 / 0 %
	4	19.22 / 1.84 %	11.76 / 0 %
	5	15.44 / 1.91 %	5.12 / 0%
	6	8.42 / 0.49 %	3.3 / 0%
	7	11.52 / 0.55 %	6 / 0%
	8	5.06 / 0 %	13.95 / 2.32 %
Ptf_Apde	1	15.77 / 2.56 %	5 / 0 %
	2	16.18 / 0.15 %	37.5 / 0 %
Ptf_Pde	1	14.08 / 0.15 %	15.38 / 0 %
	2	12.65 / 1 %	22.22 / 0 %
	3	15.18 / 1.66 %	0 / 0 %
	4	20.74 / 2.53 %	7.9 / 7.9 %
	5	16.11 / 1.4 %	9.09 / 0 %
	6	13.98 / 1 %	4.76 / 0 %
	7	16.68 / 1.56 %	20.68 / 0 %
	8	17.96 / 1.5 %	12.9 / 0 %
3 TFs (33 % / 66 % /100%)			
Combinational Pattern of TF and DE gene	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	18.27 / 1.63 / 0 %	50 / 20 / 0 %
	2	14.91 / 1.79 / 0.131 %	0 / 0 / 0 %

Aptf_Pde	1	14.08 / 1.53 / 0.05 %	11.15 / 0 / 0 %
	2	16.78 / 2.8 / 0.23 %	8.33 / 8.33 / 0 %
	3	12.63 / 3.08 / 0.37 %	47.36 / 0 / 0 %
	4	25.68 / 5.58 / 0.63 %	0 / 0 / 0 %
	5	18.76 / 4.44 / 0.52 %	14.28 / 0 / 0 %
	6	10.91 / 1.35 / 0.046 %	0 / 0 / 0 %
	7	16.55 / 2 / 0.2 %	5 / 0 / 0 %
	8	19.8 / 3.17 / 0.19 %	28.57 / 9.52 / 0 %
Ptf_Apde	1	25 / 5.55 / 0.25 %	0 / 0 / 0 %
	2	25.83 / 3.47 / 0.41 %	0 / 0 / 0 %
Ptf_Pde	1	20.58 / 0.25 / 0.11 %	14.28 / 14.28 / 0 %
	2	16.16 / 1.6 / 0.079 %	20 / 0 / 0 %
	3	21 / 6.12 / 0.67 %	0 / 0 / 0 %
	4	27.5 / 4.18 / 0.16 %	0 / 25 / 0 %
	5	16.8 / 4.24 / 0.48 %	12.5 / 0 / 0 %
	6	13.9 / 11.5 / 0 %	0 / 0 / 0 %
	7	12.78 / 1.92 / 0.0916 %	10 / 0 / 0 %
	8	19.9 / 3.69 / 0 %	0 / 0 / 0 %

4TFs
(25 % / 50 % / 75 % / 100%)

Combinational Pattern of TF and DE gene	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	17.42 / 1.51 / 0 / 0 %	0 / 100 / 0 / 0 %
	2	18.31 / 3.01 / 0.2 / 0.01 %	0 / 0 / 0 / 0 %
Aptf_Pde	1	21.53 / 7.69 / 0.51 / 0 %	0 / 0 / 0 / 0 %
	2	18.52 / 4.93 / 0.85 / 0.06 %	0 / 33.33 / 33.33 / 0 %
	3	15.5 / 0.18 / 0.27 / 0 %	77.78 / 0 / 0 / 0 %
	4	28.45 / 8.72 / 1.95 / 0.075 %	0 / 0 / 0 / 0 %
	5	20.7 / 5.79 / 1.04 / 0.17 %	0 / 0 / 0 / 0 %
	6	12.16 / 1.84 / 0.19 / 0 %	0 / 0 / 0 / 0 %
	7	15.81 / 5.32 / 0.24 / 0.08 %	0 / 0 / 0 / 0 %

	8	25.74 / 5.36 / 0.62 / 0 %	50 / 0 / 0 / 0 %
Ptf_Apde	1	INSIGNIFICANT	INSIGNIFICANT
	2	20.13 / 0.34 / 0 / 0 %	
Ptf_Pde	1	22.6 / 5.36 / 0.69 / 0 %	0 / 50 / 0 / 0 %
	2	20.25 / 3.24 / 0.42 / 0 %	50 / 0 / 0 / 0 %
	3	24.7 / 6.5 / 2.57 / 0.18 %	0 / 0 / 0 / 0 %
	4	31.93 / 6.8 / 0.86 / 0.058 %	INSIGNIFICANT
	5	19.28 / 6.71 / 1 / 0 %	0 / 0 / 0 / 0 %
	6	16.42 / 16.9 / 0 / 0 %	0 / 0 / 0 / 0 %
	7	14.4 / 2.2 / 0.16 / 0 %	0 / 0 / 0 / 0 %
	8	0 / 0 / 0 / 0 %	INSIGNIFICANT

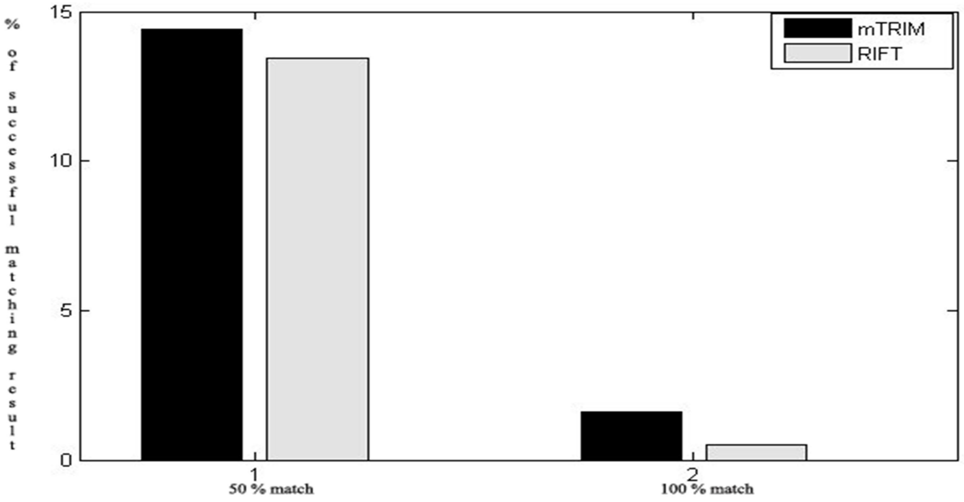
Table 5.5: Category II_TF to target gene matching considering the HUB gene in each cluster

2 TFs **(50 % / 100%)**			
Combinational Pattern of TF and DE gene	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	10.5 / 0.15%	INVALID
	2	9.91 / 0.22%	0 / 0%
Aptf_Pde	1	8.1 / 0.088 %	INVALID
	2	0 / 0 %	INVALID
	3	6.22 / 0 %	INVALID
	4	0 / 0 %	INVALID
	5	6.54 / 0 %	INVALID
	6	48.11 / 12.85 %	INVALID
	7	6.65 / 0.0924 %	0 / 0%
	8	5.06 / 0 %	INVALID
Ptf_Apde	1	0 / 0 %	INVALID
	2	27.18 / 9.17 %	50 / 0 %

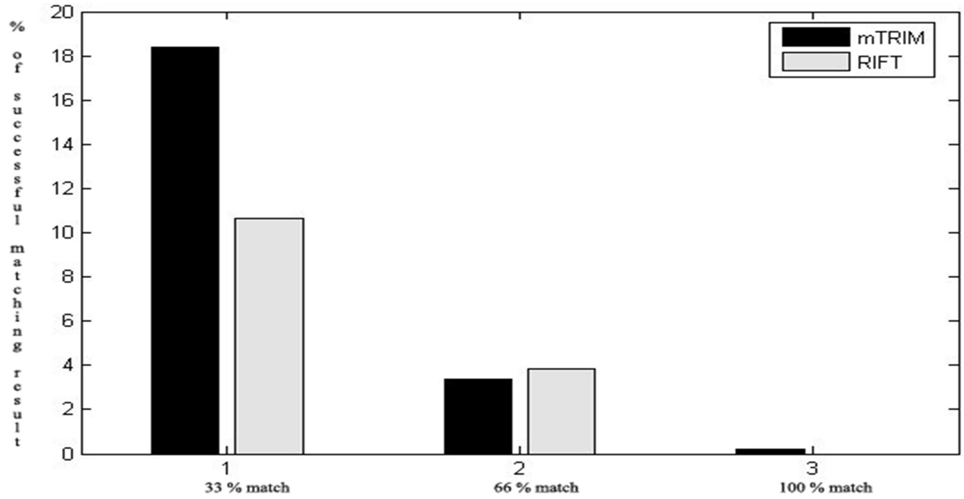
Ptf_Pde	1	6.49 / 0 %	INVALID
	2	0 / 0 %	INVALID
	3	11.07 / 0 %	INVALID
	4	0 / 0 %	INVALID
	5	21.77 / 0.81 %	INVALID
	6	33.96 / 5.6 %	INVALID
	7	27.67 / 11.1 %	50 / 0 %
	8	0 / 0 %	INVALID
		3 TFs **(33 % / 66 % /100%)**	
Combinational Pattern of TF and DE gene	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	4 / 0 / 0 %	INVALID
	2	11.85 / 0 / 0 %	0 / 0 / 0 %
Aptf_Pde	1	11.11 / 0 / 0 %	INVALID
	2	0 / 0 / 0 %	INVALID
	3	0 / 0 / 0 %	INVALID
	4	0 / 0 / 0 %	INVALID
	5	13.67 / 0.87 / 0 %	INVALID
	6	46.55 / 31.76 / 5.15 %	INVALID
	7	19.63 / 0 / 0 %	0 / 0 / 0 %
	8	0 / 0 / 0 %	INVALID
Ptf_Apde	1	0 / 0 / 0 %	INVALID
	2	5 / 0 / 0 %	0 / 0 / 0 %
Ptf_Pde	1	20 / 0 / 0 %	INVALID
	2	0 / 0 / 0 %	INVALID
	3	17.21 / 0 / 0 %	INVALID
	4	0 / 0 / 0 %	INVALID
	5	48.78 / 2.77 / 0 %	INVALID
	6	41.18 / 0 / 0 %	INVALID
	7	0 / 0 / 0 %	0 / 0 / 0 %
	8	0 / 0 / 0 %	INVALID

4TFs **(25 % / 50 % / 75 % / 100%)**			
Combinational Pattern of TF and DE gene	Cluster Index	mTRIM	RIFT
Aptf_Apde	1	0 / 0 / 0 / 0 %	INVALID
	2	18.53 / 0 / 0 / 0 %	0 / 0 / 0 / 0 %
Aptf_Pde	1	0 / 0 / 0 / 0 %	INVALID
	2	0 / 0 / 0 / 0 %	INVALID
	3	0 / 0 / 0 / 0 %	INVALID
	4	0 / 0 / 0 / 0 %	INVALID
	5	11.32 / 0.41 / 0 / 0 %	INVALID
	6	43.78 / 35.75 / 10.86 / 0.45 %	INVALID
	7	23.81 / 0 / 0 / 0 %	0 / 0 / 0 / 0 %
	8	0 / 0 / 0 / 0 %	INVALID
Ptf_Apde	1	INSIGNIFICANT	INSIGNIFICANT
	2		
Ptf_Pde	1	0 / 0 / 0 / 0 %	INVALID
	2	0 / 0 / 0 / 0 %	INVALID
	3	40.63 / 0 / 0 / 0 %	INVALID
	4	0 / 0 / 0 / 0 %	INSIGNIFICANT
	5	57.14 / 7.14 / 0 / 0 %	INVALID
	6	66.67 / 0 / 0 / 0 %	INVALID
	7	0 / 0 / 0 / 0 %	0 / 0 / 0 / 0 %
	8	0 / 0 / 0 / 0 %	INSIGNIFICANT

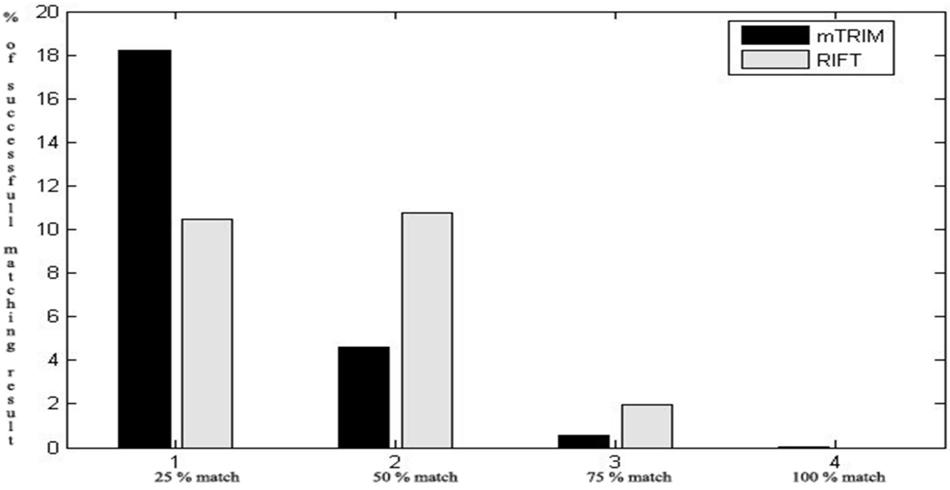
In the above tables, INVALID means that HUB gene is not present as a possible target for any set of regulators (TF) and INSIGNIFICANT means unavailability of significant TF combinations corresponding to a target DE gene. In a different view, the matching perspective on the basis of comparative bar charts between mTRIM and RIFT representing the *average match* under 2, 3 and 4 TF gene regulatory architectures are given next in Figure 5.2. Here, the black bar indicates mTRIM and the grey bar indicates the proposed RIFT architecture.



2 TF regulatory architectures



3 TF regulatory architectures



4 TF regulatory architectures

Figure 5.2: Comparative bar charts showing average matching of 2, 3, and 4 TF gene regulatory architectures with published databases

The fruitfulness of the analysis is depicted on the basis of the AUROC plots dependent upon the specificity and sensitivity values of each DE gene present in individual clusters. For example, in ApTF_ApDE case, in each cluster of target DE genes, cluster specific plots are generated based on the interactions present between a target ApDE gene and all ApTFs. That is, for each DE gene, specificity and sensitivity needs to be computed. Accordingly, in this example there are two plots to compare the performances of mTRIM and RIFT from the AUROC (Area under Region of Characteristic). Similarly, there exist 2 plots for PTF_ApDE and 8 plots in each case of ApTF_PDE and PTF_PDE. The AUROC plots for ApTF_ApDE, ApTF_PDE, PTF_ApDE and PTF_PDE are given in Figures 5.3, 5.4, 5.5 and 5.6 respectively. In each plot, the dashed line corresponds to the proposed RIFT architecture and the solid line corresponds to mTRIM approach. In this regard, the specificity and sensitivity of any network can be defined as follows:

Specificity= 1- FPR, Sensitivity= TPR, where TPR stands for True Positive Rate and FPR stands for False Positive Rate.

This TPR and FPR are defined as:

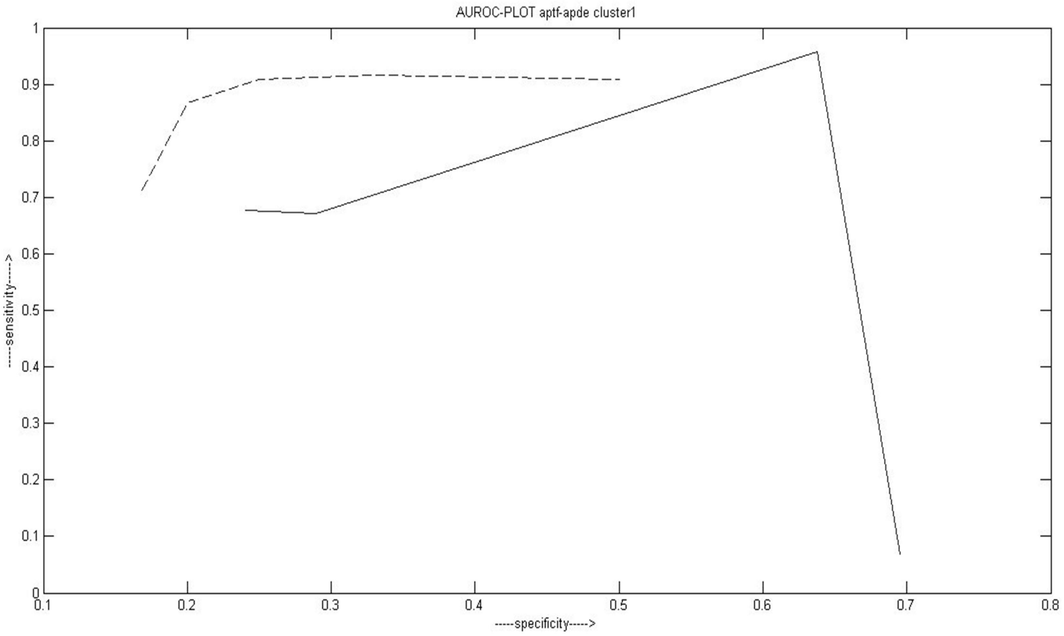
TPR = TP / (TP+FN) and FPR = FP / (FP+TN) where,

TP = True Positive or regulations present in the reported databases as well as in the obtained results.

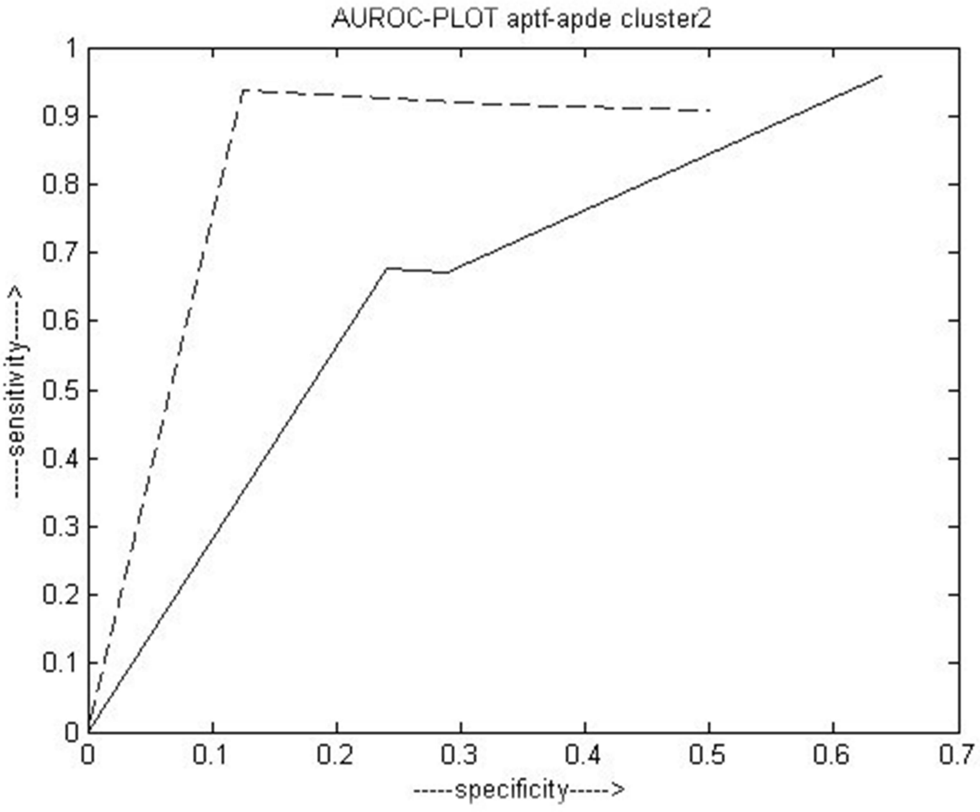
FP = False Positive or regulations absent in the reported databases but present in the obtained results.

TN = True Negative or regulations absent in both the databases as well as in the obtained results.

FN = False Negative or regulations present in the reported databases, but absent in the obtained results.

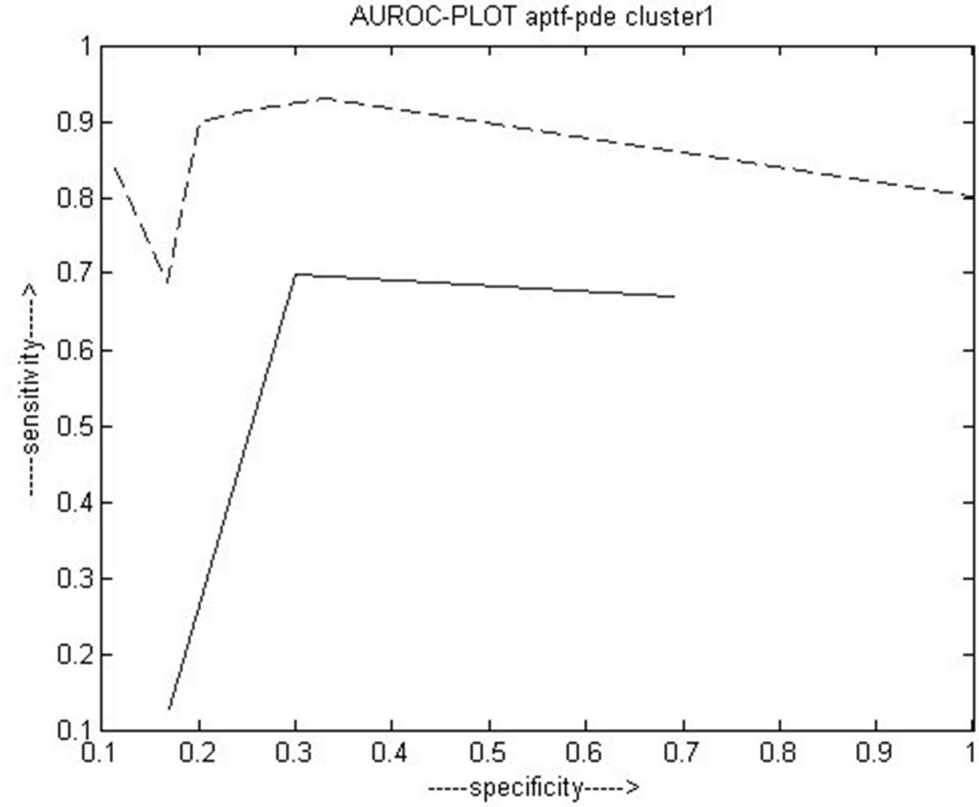


Cluster 1

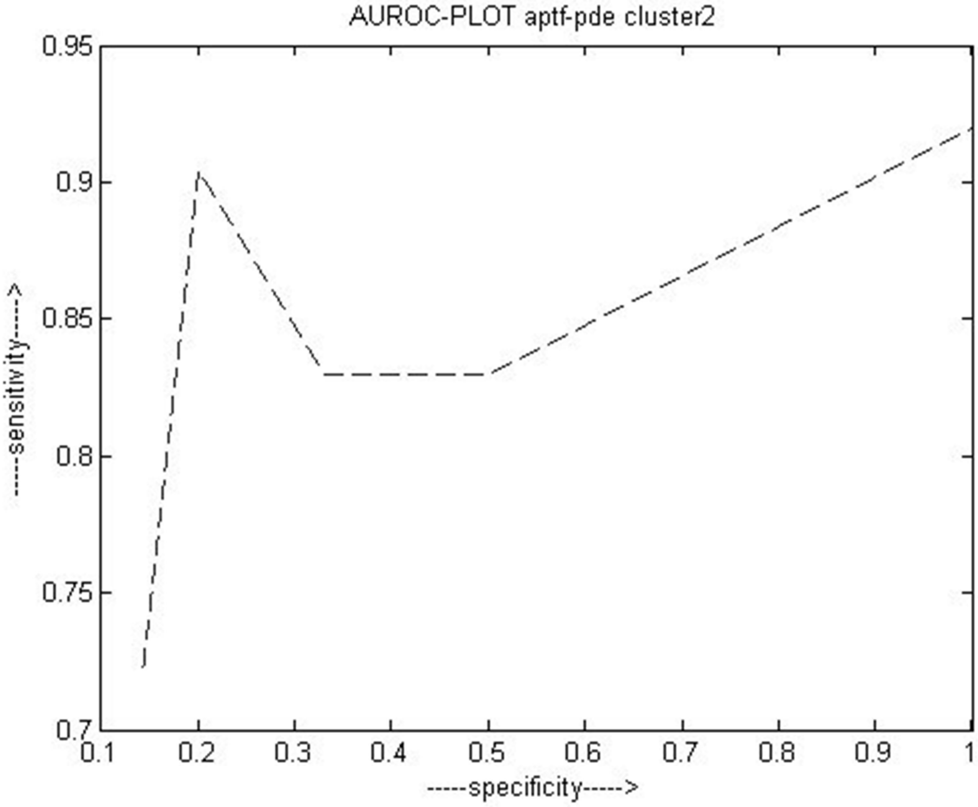


Cluster 2

Figure 5.3: Area under Region of Characteristic difference between RIFT and mTRIM (ApTF_ApDE)

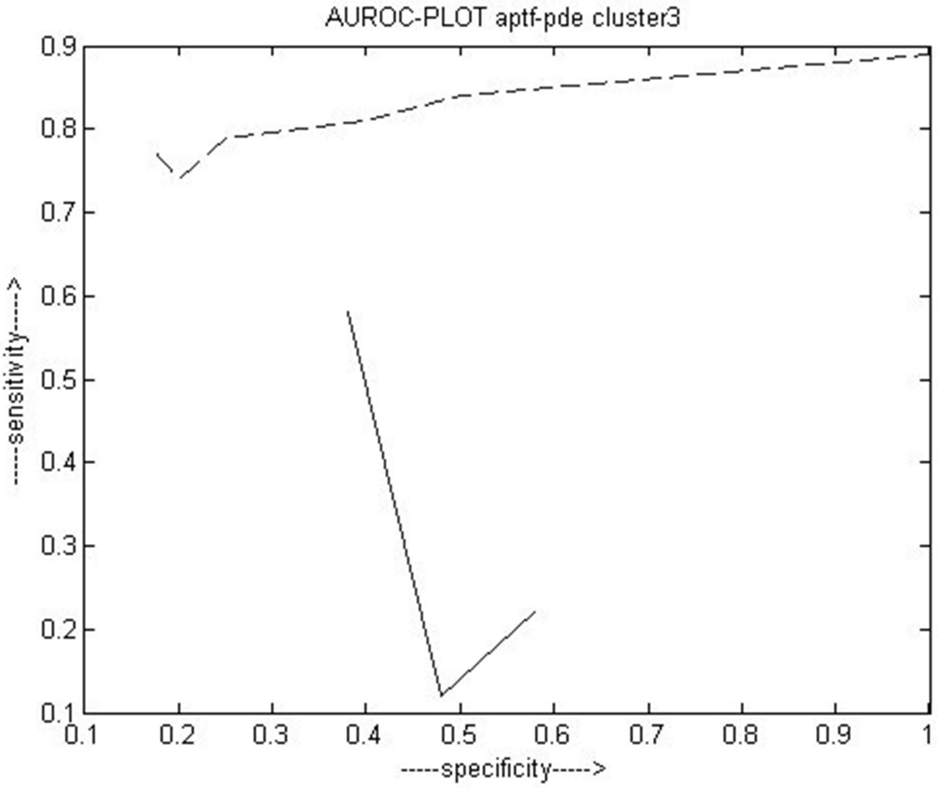


Cluster 1

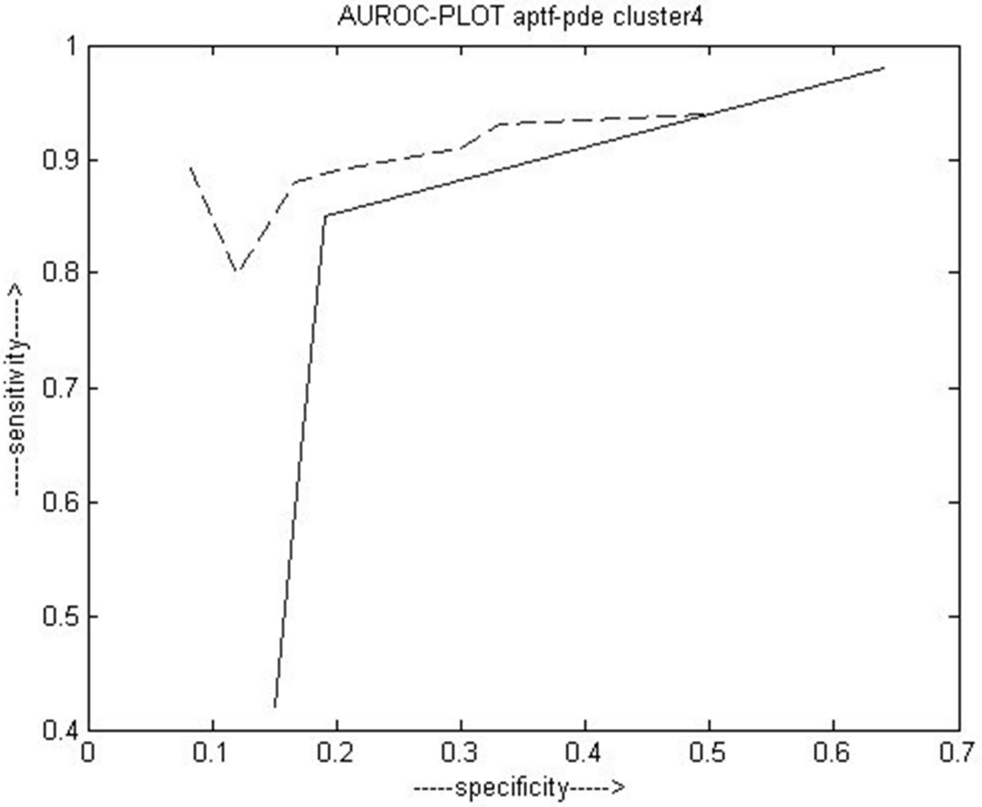


Cluster 2

Figure 5.4: Area under Region of Characteristic difference between RIFT and mTRIM (ApTF_PDE)..contd

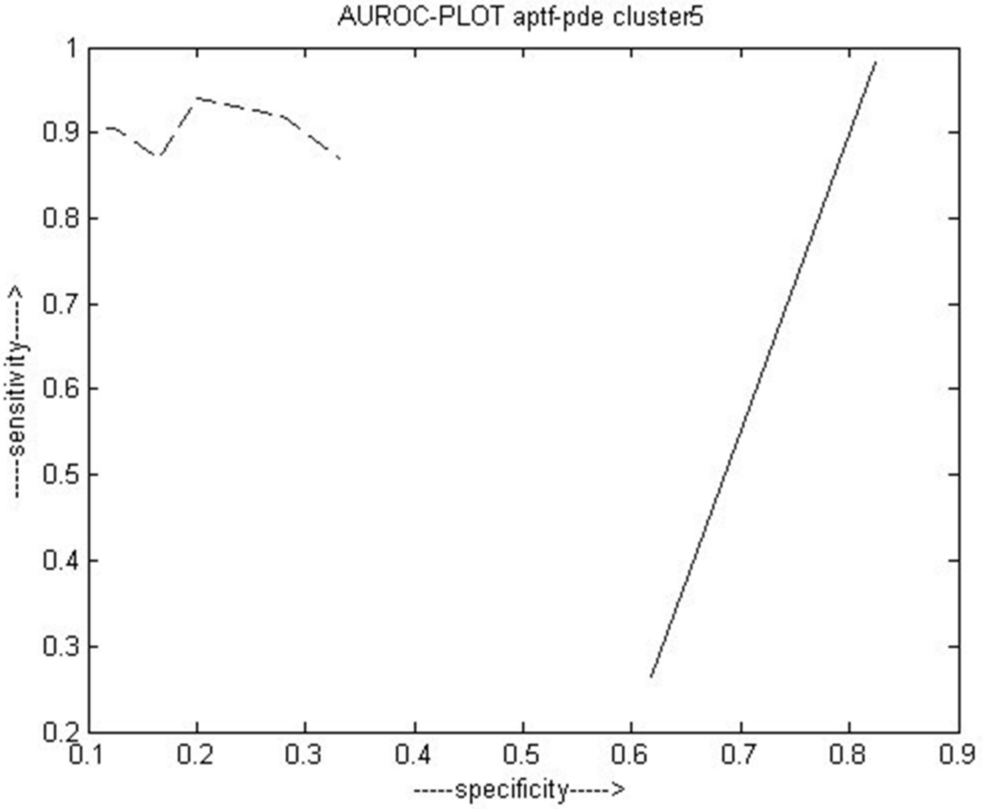


Cluster 3

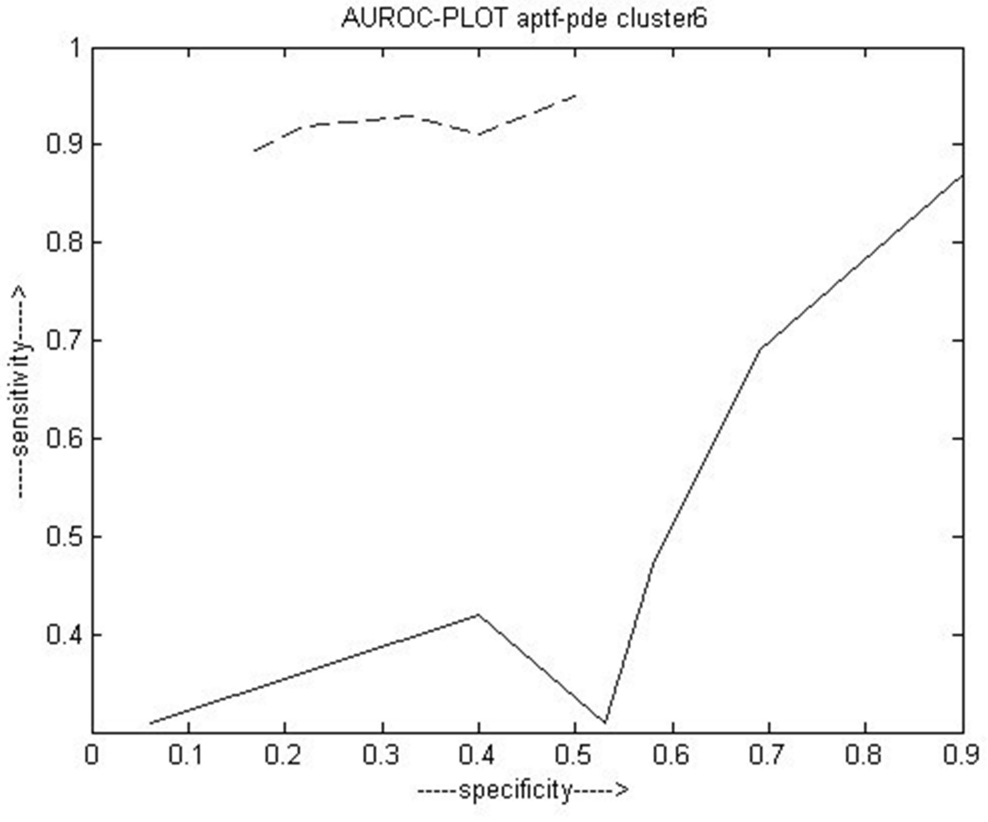


Cluster 4

Figure 5.4: Area under Region of Characteristic difference between RIFT and mTRIM (ApTF_PDE)..contd.

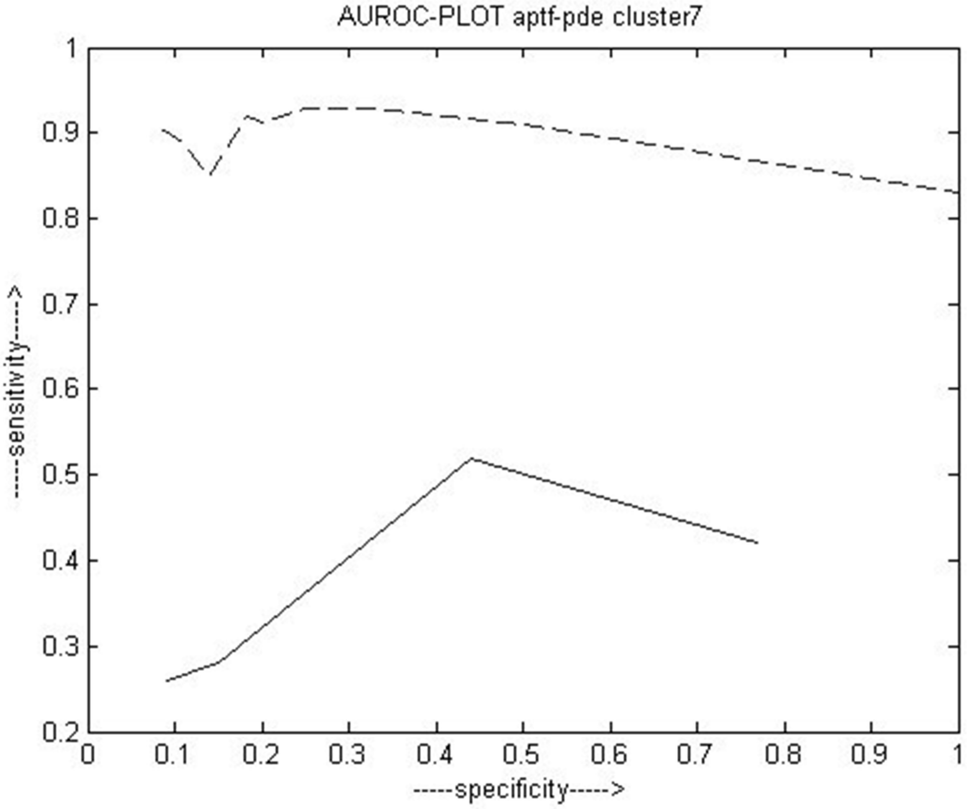


Cluster 5

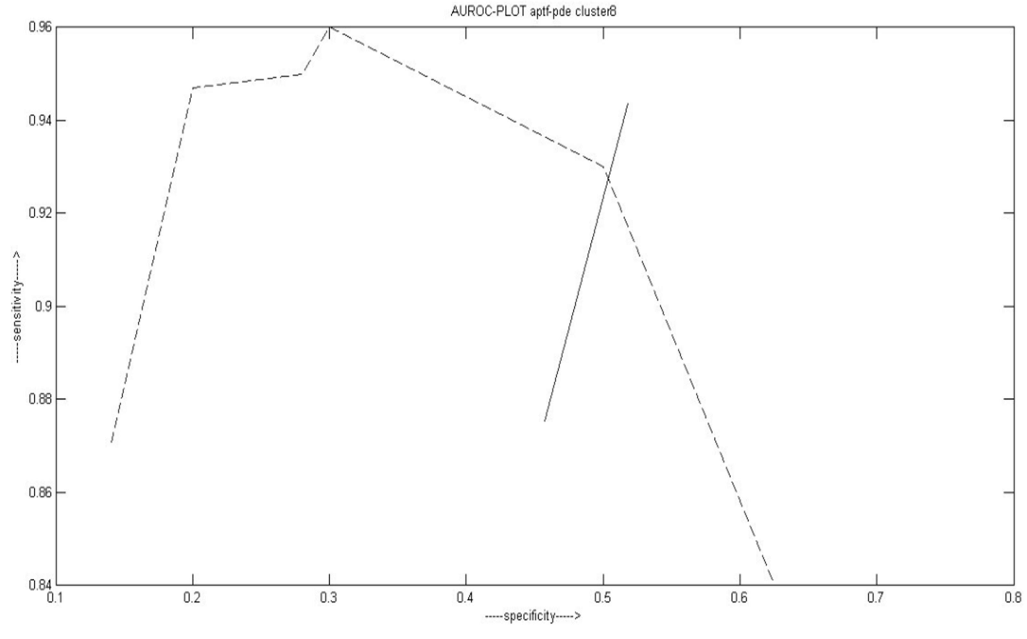


Cluster 6

Figure 5.4: Area under Region of Characteristic difference between RIFT and mTRIM (ApTF_PDE)..contd.

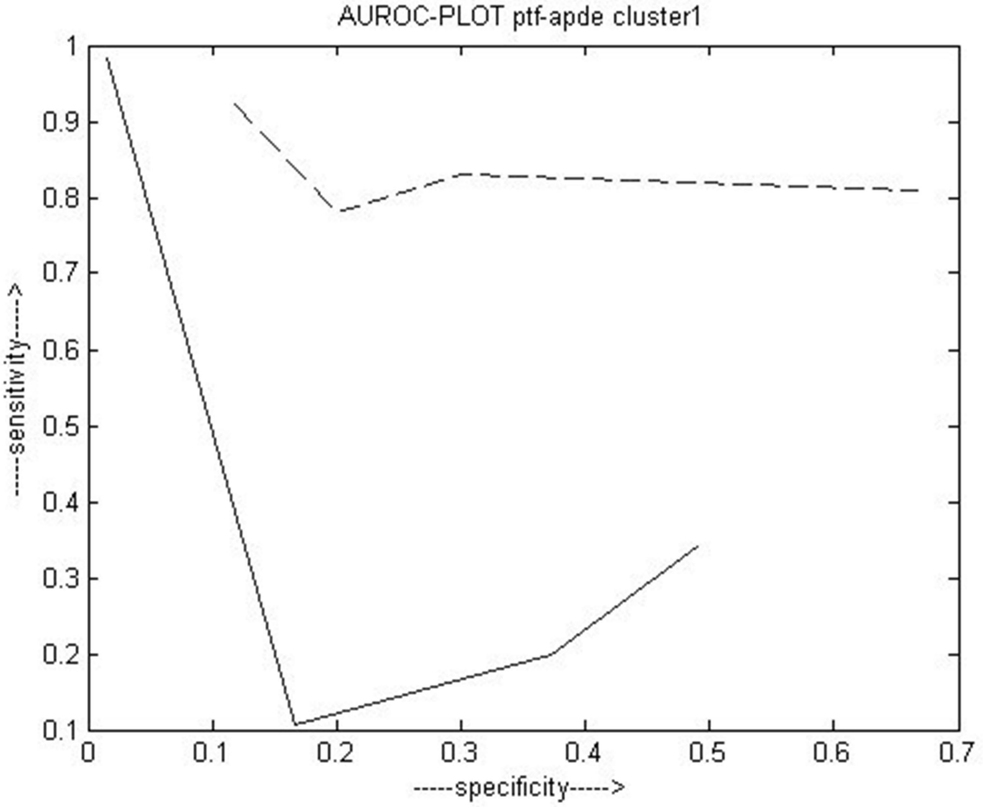


Cluster 7

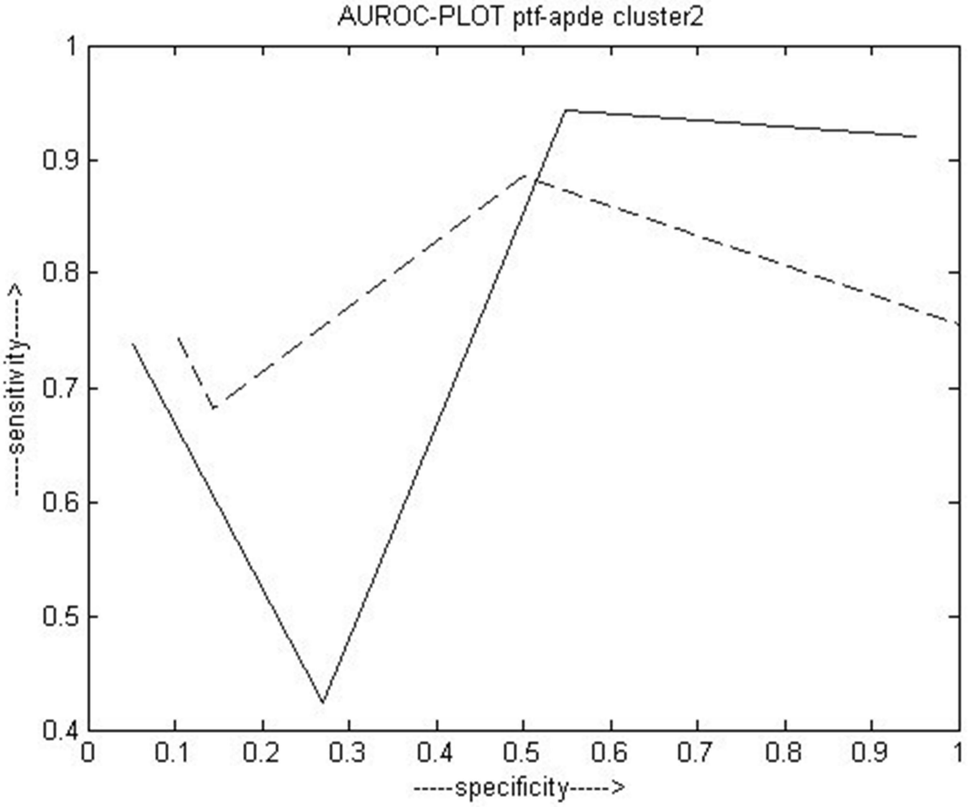


Cluster 8

Figure 5.4: Area under Region of Characteristic difference between RIFT and mTRIM (ApTF_PDE)

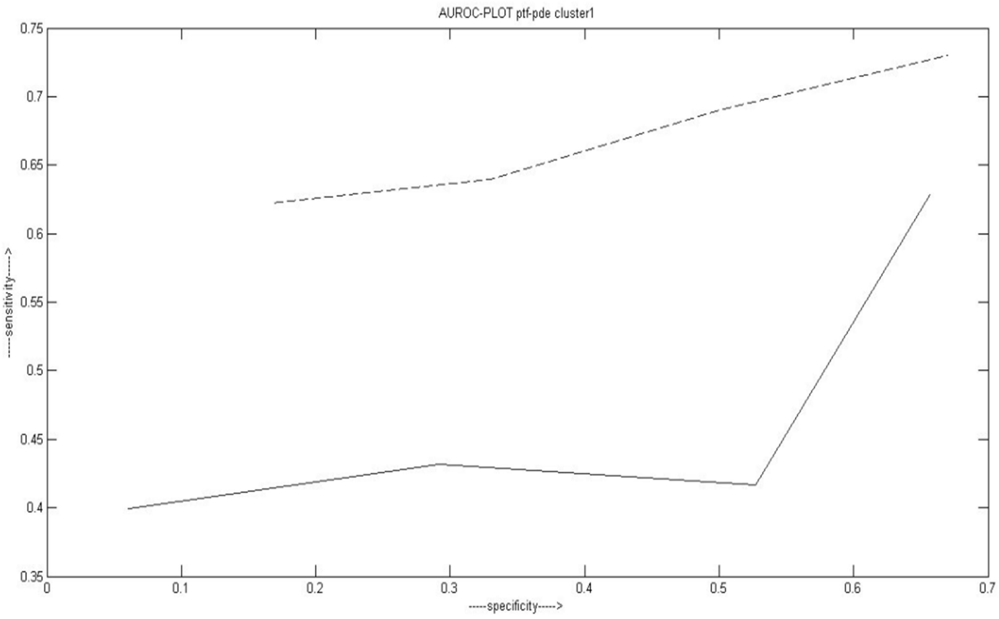


Cluster 1

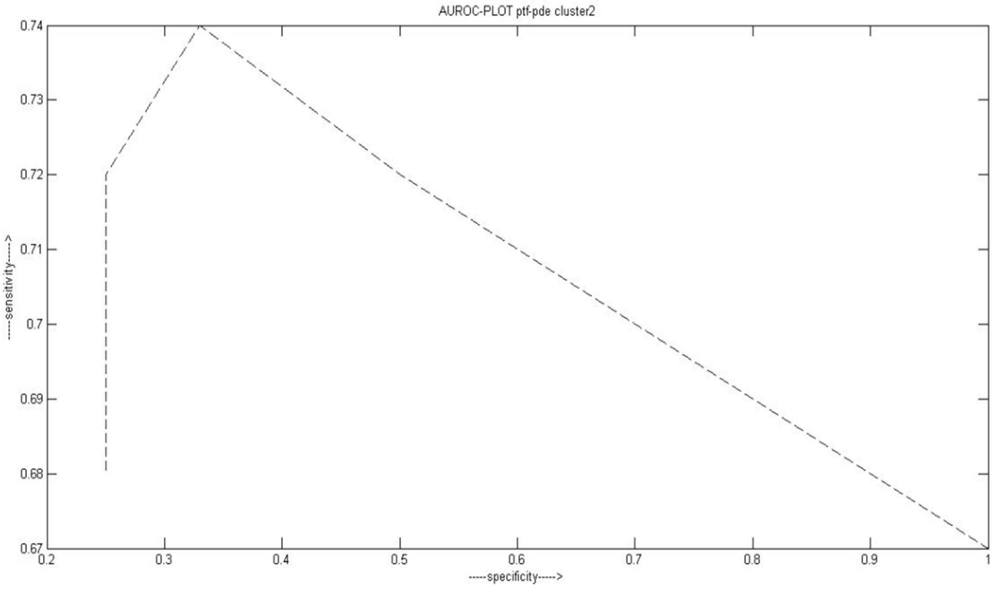


Cluster 2

Figure 5.5: Area under Region of Characteristic difference between RIFT and mTRIM (PTF_ApDE)

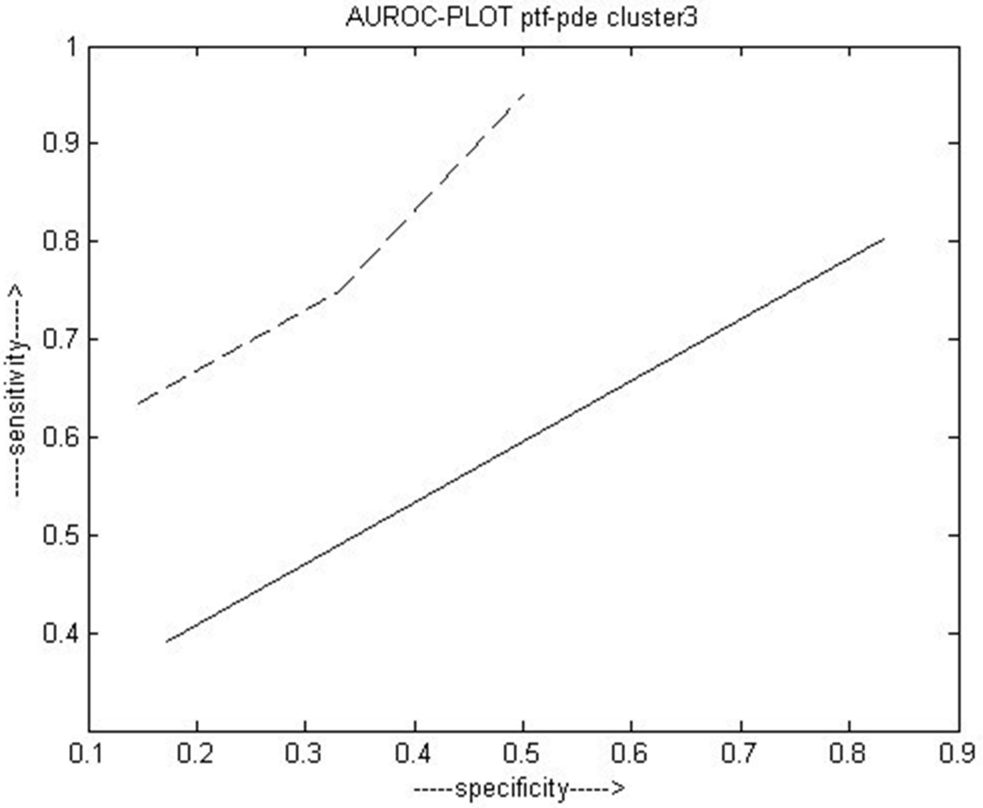


Cluster 1

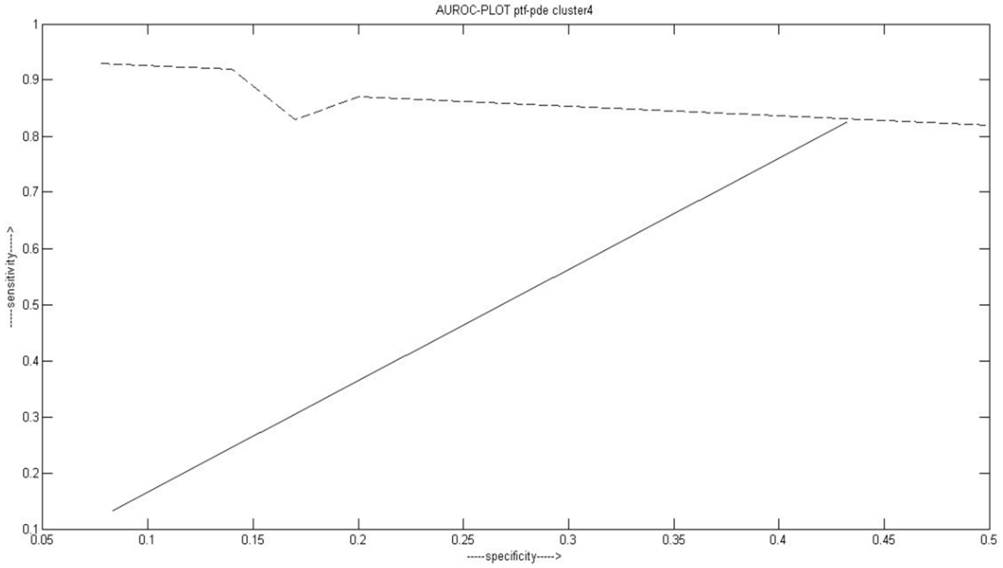


Cluster 2

Figure 5.6: Area under Region of Characteristic difference between RIFT and mTRIM (PTF_PDE)..contd

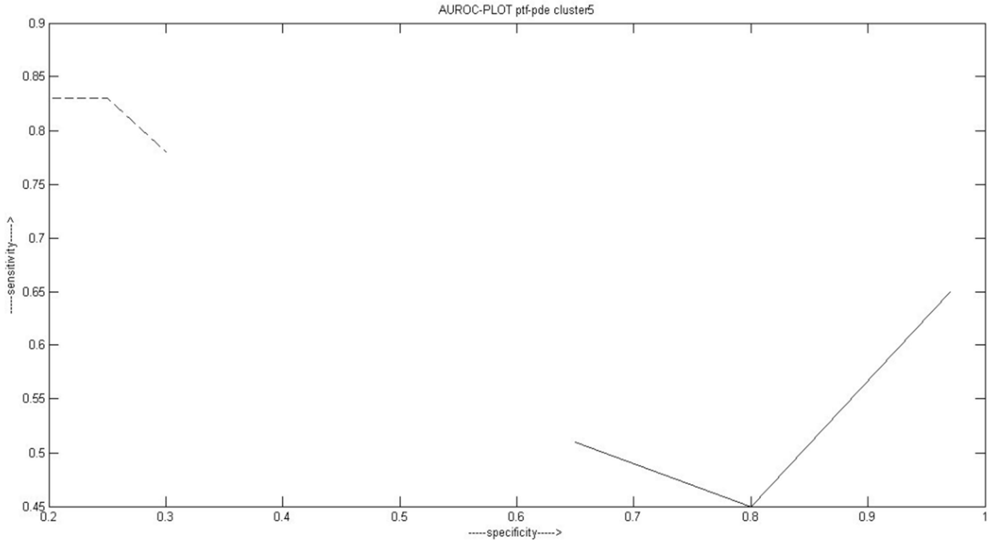


Cluster 3

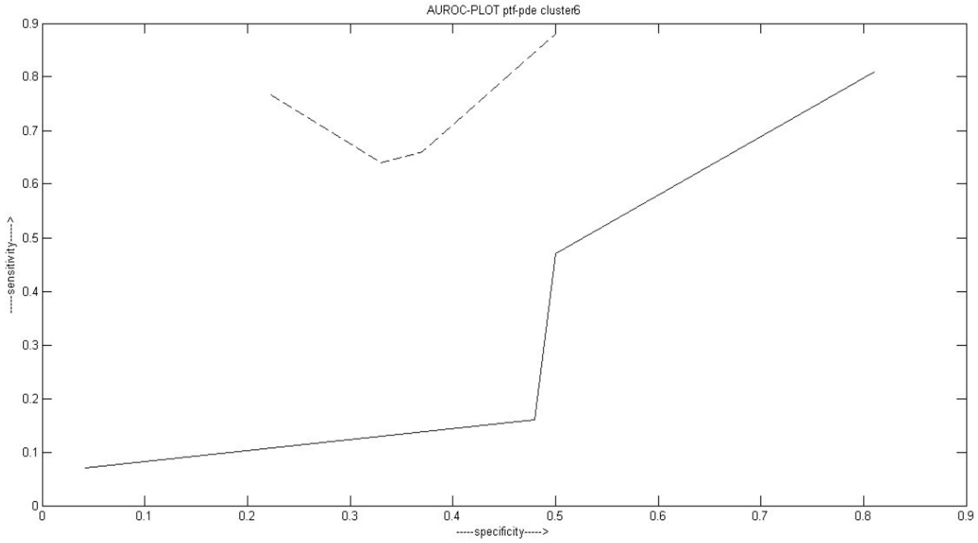


Cluster 4

Figure 5.6: Area under Region of Characteristic difference between RIFT and mTRIM (PTF_PDE)..contd

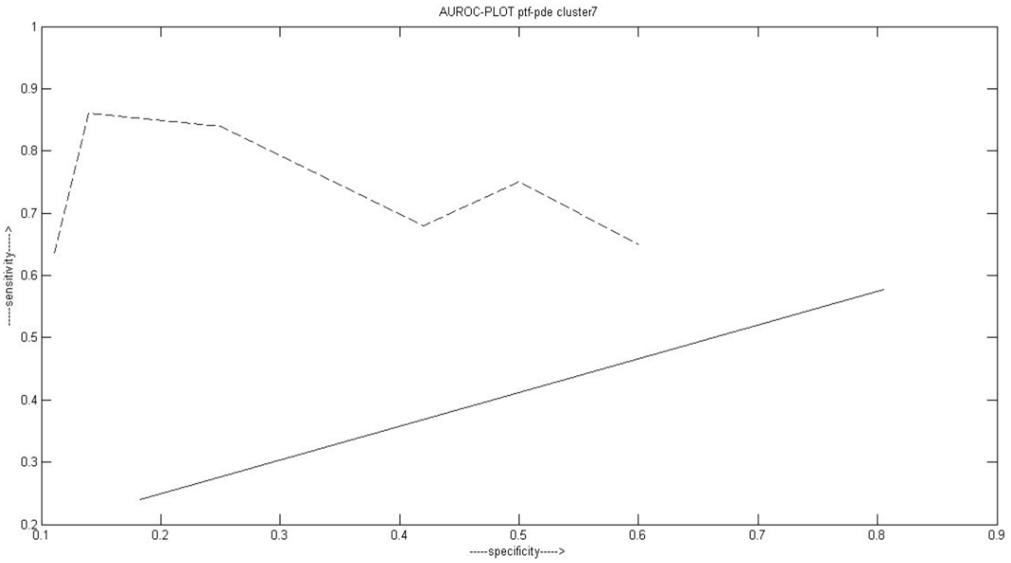


Cluster 5

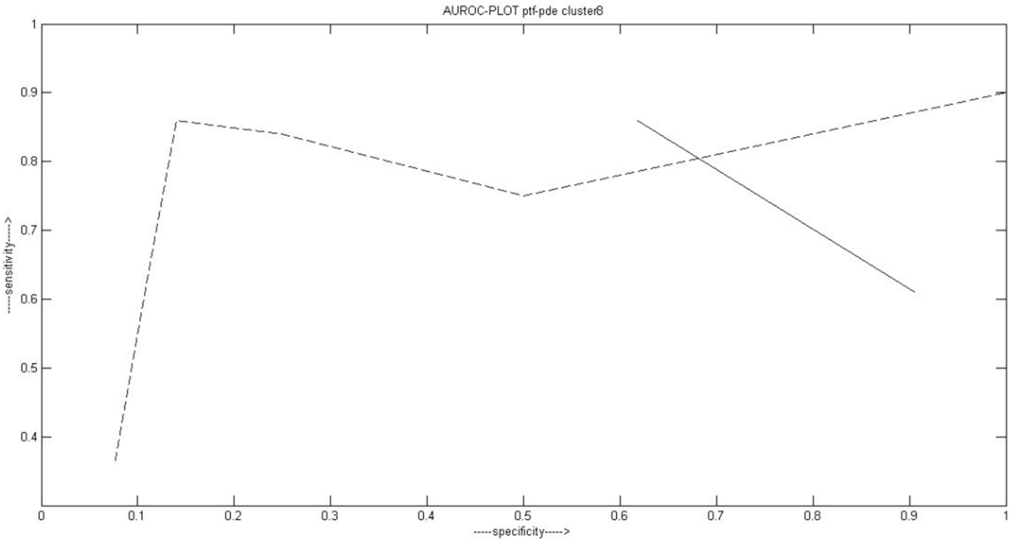


Cluster 6

Figure 5.6: Area under Region of Characteristic difference between RIFT and mTRIM (PTF_PDE)..contd



Cluster 7



Cluster 8

Figure 5.6: Area under Region of Characteristic difference between RIFT and mTRIM (PTF_PDE)

5.2.4 Discussion: From the involved computations, it has been observed that the number of TF genes present in a group for controlling a target DE gene is more in the simple *AfnScore* based mTRIM analysis compared to the proposed algorithm based on RIFT score. In the mTRIM case, collaborative groups consisting of maximum 5 TF genes are found, whereas the RIFT architecture yields groups of maximum 4 TF genes. However, this kind of result is not surprising as the proposed RIFT based network design follows a very stringent measure in order to choose regulator sets compared to mTRIM. As the temporal differential mode specific regulatory pattern is absent in the case of mTRIM, hence all TF genes can be potential regulators of each target DE gene. Accordingly, every TF to target DE gene interaction is a valid interaction suitable for significance assessment. But this novel RIFT based architecture initially filters out those TF-DE gene interactions showing irregularity or inconsistency in the composite temporal differential regulation between the phase pairs of HeLa cell cycle data, considering the middle phase to be the control or reference phase. This is followed by formation of TRN structures using the significantly assessed filtered TF genes. Hence, formation of higher order TF gene collaborations gets restricted. But even if lower sized groups are formed in this RIFT modelling, the specificity and sensitivity of the network designs found in every cluster of DE genes (corresponding to the various periodic and aperiodic regulatory combinations) happen to be far better or have outperformed the *AfnScore* based mTRIM analysis. This indicates the true power of the algorithm through understanding the specific temporal or time variant role of differential regulation present in a TRN.

Apart from above, there are certain significant observations related to the obtained results. One of these (already mentioned in Table 5.3) is about TF genes that are DE as well. It has been checked that there are 4 such genes participating in the network reconstruction process. Here, the TF to TF interactive regulations (interact amongst themselves to control some other targets) are being verified by the TRRUST [**20**] database. A typical example in this regard can be found in aperiodic TF-periodic DE gene regulation from the eighth cluster. Among the 72 DE targets present in the cluster, the gene *E2F1* is identified as a TF gene. Validation done using TRRUST shows TF genes *EP300*, *FOXO3*, *NFYA*, *NFYB* and *EGR1* interactively communicating with *E2F1*. In this regard, the result clarifies *EP300* is significantly associated with *E2F1* via single TF regulation in all the four composite temporal modes (AN_RS_AN, AS_RN_AS, RN_AS_RN and RS_AN_RS). Similarly *FOXO3*, *NFYA*, and *NFYB* show significant single TF gene composite temporal regulation with target *E2F1* in AS_RN_AS, AN_RS_AN and RN_AS_RN modes respectively. However for the TF regulator *EGR1*, it has been checked through the implementation that its single regulatory mechanism is insignificant, but together with other TF genes does interact significantly with the target regulatee *E2F1*. Another significant observation can be related to multiple TF based composite regulatory actions. To elucidate, an example considering three TF genes W, X, and Y controlling a target gene Z via a combination of composite temporal regulatory modes (A B C) is the matter of concern. Here, assuming that W and X are the significant regulators of Z verified through the interactive information present in the databases, the novel research architecture using the RIFT score confirms their coexistence with Y in the regulation of Z. This also clarifies that W and X together is not significant enough to regulate Z with the help of some other TF gene/s via other composite temporal mode differential regulations.

5.3 Conclusion

Given any time series gene expression data, the regulatory analysis can be done in two ways. In one case, instantaneous or first order regulatory system architecture such as [**25**] can be realized and in the latter through higher order regulations [**26-28**] involving time delayed networks. In the biological context it is mostly observed that a gene can regulate another gene by its products (RNAs or proteins). In this regard, a finite delay does come into existence which may include one or more of the following factors like the translation time of the source gene, protein folding time, translocation time, promoter binding time, and transcription time of the target gene. Although some significant algorithms [**29-31**] have been developed to analyze and process time series informational sequence verifying the time delayed effects in higher order gene regulatory networks, but most of these deal with short time series sequences. However, considering the contemporary research trend in this area, such higher order regulatory networks have minimal or trivial exploration of the functional and logical regulatory aspects at specific activation time points which happens to be the alma mater of this research contribution, assuming first order regulatory network.

5.4 References

[**1**] A. Fujita, P. Severino, K. Kojima, J.R. Sato, A.G. Patriota, S. Miyano, “Functional clustering of time series gene expression data by Granger causality”, BMC Systems Biology, 6, Article No.137, October 2012, https://doi.org/10.1186/1752-0509-6-137

[**2**] Y. Luan and H. Li, “Model-based methods for identifying periodically expressed genes based on timecourse microarray gene expression data”, Bioinformatics, 20(3), 332–339, February 2004, https://doi.org/10.1093/bioinformatics/btg413

[**3**] M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, D. Botstein, “Identification of genes periodically expressed in thehuman cell cycle and their expression in tumors”,

Molecular Biology of the Cell, 13(6), 1977–2000, June 2002, https://doi.org/10.1091/mbc.02-02-0030

[**4**] L.-Y. Lo, K.-S. Leung, and K.-H. Lee, “Inferring time-delayed causal gene network using time-series expression data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(5), 1169–1182, September/October 2015, https://doi.org/10.1109/tcbb.2015.2394442

[**5**] S. Awad and J. Chen, “Inferring transcription factor collaborations in gene regulatory networks”, BMC Systems Biology, 8, Article No.S1, January 2014, https://doi.org/10.1186/1752-0509-8-S1-S1

[**6**] C-H. Yeang and T. Jaakkola, “Modelling the combinational functions of multiple transcription factors”, Journal of Computational Biology, 13(2), 463–480, March 2006, https://doi.org/10.1089/cmb.2006.13.463

[**7**] S. Awad, N. Panchy, S.-K. Ng, and J. Chen, “Inferring the regulatory interaction types of transcription factors in transcriptional regulatory networks”, Journal of Bioinformatics and Computational Biology, 10(5), Article No.1250012, October 2012, https://doi.org/10.1142/s0219720012500126

[**8**] A. Majumder and M. Sarkar, “Simple transcriptional networks for differentially expressed genes”, In Proceedings of IEEE International Conference on Signal Propagation and Computer Technology, Ajmer, India, 642–647, July 2014, https://doi.org/10.1109/ICSPCT.2014.6885016

[**9**] A. Majumder and M. Sarkar, “Paired transcriptional regulatory system for differentially expressed genes”, Lecture Notes on Information Theory, 2(3), 266–272, September 2014, doi: 10.12720/lnit.2.3.266-272

[**10**] H. Yu, B-H. Liu, Z-Q. Ye, C. Li, Y-X. Li, and Y-Y. Li, “Link-based quantitative methods to identify differentially co-expressed genes and gene pairs”, BMC Bioinformatics, 12, Article No.315, August 2011, https://doi.org/10.1186/1471-2105-12-315

[**11**] J. Yang, H. Yu, B-H. Liu, Z. Zhao, L. Liu, L-X. Ma, Y-X. Li, and Y-Y. Li, “DCGL v2.0: An R package for unveiling differential regulation from differential co-expression”, PloS One, 8(11):e79729, November 2013, https://doi.org/10.1371/journal.pone.0079729

[**12**] J. Ernst, O. Vainas, C.T. Harbison, I. Simon, and Z. Bar-Joseph, “Reconstructing dynamic regulatory maps” Molecular Systems Biology, 3, Article No.74, January 2007, https://doi.org/10.1038/msb4100115

[**13**] R.O. Duda, P.E. Hart, and D. G. Stork, “Pattern Classifications”, Hoboken, NJ, USA: Wiley, vol. 2, 2001.

[**14**] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchinson, “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids”, Cambridge, United Kingdom: Cambridge Univ. Press, vol. 1, 1998.

[**15**] A. Reverter, N.J. Hudson, S.H. Nagaraj, M. P-Enciso, and B.P. Dalrymple, “Regulatory impact factors: Unraveling the transcriptional regulation of complex traits

from expression data", Bioinformatics, 26(7), 896–904, April 2010, https://doi.org/10.1093/bioinformatics/btq051

[**16**] F. Xiao, L. Gao, Y. Ye, Y. Hu, and R. He, "Inferring gene regulatory networks using conditional regulation pattern to guide candidate genes", PLoS One, 11(5):e0154953, May 2016, https://doi.org/10.1371/journal.pone.0154953

[**17**] F. Liu, S-W. Zhang, W-F. Guo, Z-G. Wei, and L. Chen, "Inference of gene regulatory network based on local Bayesian networks," PLoS Computational Biology, 12(8):e1005024, August 2016, https://doi.org/10.1371/journal.pcbi.1005024

[**18**] X. Zhang, J. Zhao, J-K. Hao, X-M. Zhao, and L. Chen, "Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks", Nucleic Acids Research, 43(5):e31, March 2015, https://doi.org/10.1093/nar/gku1315

[**19**] G. Zheng, Y. Xu, X. Zhang, Z-P. Liu, Z. Wang, L. Chen, and X-G. Zhu, "CMIP: A software package capable of reconstructing genome-wide regulatory networks using gene expression data", BMC Bioinformatics, 17, Article No.535, December 2016, https://doi.org/10.1186/s12859-016-1324-y

[**20**] H. Han, H. Shim, D. Shin, J.E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, H. Kim, K. Kim, S. Yang, D. Bae, A. Yun, S. Kim, C. Y. Kim, H. J. Cho, B. Kang, S. Shin, and I. Lee, "TRRUST: a reference database of human transcriptional regulatory interactions", Scientific Reports, 5, Article No.11432, June 2015, https://doi.org/10.1038/srep11432

[**21**] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talon, "maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments", Bioinformatics, 22(9), 1096–1102, May 2006, https://doi.org/10.1093/bioinformatics/btl056

[**22**] M. Ahdesmaki, K. Fokianos, and K. Strimmer, "GeneCycle: Identification Periodically Expressed Genes", R software: Gene-Cycle_1.1.5, Available online at: https://cran.r-project.org/package=GeneCycle

[**23**] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien," R software: e1071_1.7-13, Install the latest version of this package by entering the following in R: install.packages("e1071", repos="http://R-Forge.R-project.org")

[**24**] Y. Rahmtallah, F. Emmert-Streib, and G. Glazko, "Gene Sets Net Correlations Analysis (GSNCA): A multivariate differential co-expression test for gene sets", Bioinformatics, 30(3), 360–368, February 2014, https://doi.org/10.1093/bioinformatics/btt687

[**25**] A. Wise and Z. Bar-Joseph, "SMARTS: Reconstructing disease response networks from multiple individuals using time series gene expression data", Bioinformatics, 31(8), 1250–1257, April 2015, https://doi.org/10.1093/bioinformatics/btu800

[**26**] J-R. Kim, S-M. Choo, H-S. Choi, and K-H. Cho, "Identification of gene networks with time delayed regulation based on temporal expression profiles", IEEE/ACM

Transactions on Computational Biology and Bioinformatics, 12(5), 1161–1168, September/October 2015, https://doi.org/10.1109/tcbb.2015.2394312

[**27**] L-Y. Lo, K-S. Leung, and K-H. Lee, "Inferring time-delayed causal gene network using time-series expression data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(5), 1169–1182, September/October 2015, https://doi.org/10.1109/tcbb.2015.2394442

[**28**] H. Chen, P.A. Mundra, L.N. Zhao, F. Lin, and J. Zheng, "Highly sensitive inference of time-delayed gene regulation by network deconvolution," BMC Systems Biology, 8, Article No. S6, December 2014, https://doi.org/10.1186/1752-0509-8-S4-S6

[**29**] S. Dejean, P.G.P. Martin, A. Baccini, and P. Besse, "Clustering time-series gene expression data using smoothing spline derivatives", EURASIP Journal on Bioinformatics and Systems Biology, 2007(1), Article No. 70561, May 2007, https://doi.org/10.1155/2007/70561

[**30**] J. Ernst, and Z. Bar-Joseph, "STEM: A tool for the analysis of short time series gene expression data", BMC Bioinformatics, 7, Article No.191, April 2006, https://doi.org/10.1186/1471-2105-7-191

[**31**] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data", Bioinformatics, 20(1), 5–20, January 2004, https://doi.org/10.1093/bioinformatics/btg364

Chapter 6

Confirming the Presence of Unknown Transcription Factor Genes on a Differential Gene Regulatory Link

Related Publication

[**1**] A. Majumder and P. Sharma, "Topologically Overlapped Fused LASSO Measure for Reconstructing Gene Regulation Networks", Published online: 13 Nov 2023, IETE Journal of Research, 2023.
https://doi.org/10.1080/03772063.2023.2280620

6.1 Introduction

Identifying target genes in Gene Regulation Network (GRN) models has always been an open challenge in Systems Biology [**1**]. Here, the coordinated action of different regulatory mechanisms affects either single independent or mutually dependent molecular activities. In the previous two chapters, contributions have been made in understanding this regulatory mechanism through single and collaborative scenarios via static (using identical and independently distributed gene expression profiles) and dynamic time variant (using highly auto-correlated time series gene expression profiles at specific activation time instants) network architecture formations. In this regard, the proposed and validated formations of Transcriptional Regulatory Networks (TRNs) or Gene Regulatory Networks (GRNs) which have been undertaken primarily incorporate direct causal links. However, in this regard, indirect gene regulatory architectures or some form of indirect causal effect on any target gene may be promising enough considering varied topological structures and unknown gene regulation factors [**2**]. Such forms of indirect causal regulatory aspects can be investigated, keeping in force all perturbation experiments of a dataset. It is noteworthy that contemporary state of research primarily highlights direct interaction networks which mostly forego the inevitable presence of a third entity, if any, towards varied forms of causal regulations. In this chapter, a contribution related to such complex indirect regulatory architecture has been discussed that helps in unveiling the genetic wiring through the Fused Least Absolute Shrinkage and Selection Operator (Fused-LASSO) technique with a Topological Overlap (TO) measure as the interaction structure. In this connection, the different statistically significant hierarchical regulation outcomes maintaining parity with the direct interaction structures, if any, to the target genes may throw new light on gene regulation statistics.

6.2 A Topological Fused LASSO framework discovering unknown gene regulation or transcriptional factors

In Transcriptional Regulatory Networks (TRNs), considered as a subset of the Gene Regulatory Networks (GRNs), the Transcription Factors (TFs) or the protein-building causal factors of any target gene generally act as concertmasters of the orchestrated biological actions [**3**]. The inference from any TRN or GRN is primarily based on the application of different kinds of similarity measure [**4-8**] (like correlation, mutual information, spline regression, and polynomial regression) defined direct and indirect causal regulations [**9-11**]. In this regard, most of the direct causal regulatory network formations lack the complete co-expression structure and the connectivity pattern between genes.

It is a well-established thought present in existing literature [**12,13**] that not only Differential Expression (DE) but Differential Connectivity (DC) across conditions leads to varied functionalities of a gene set. In this regard, a Topological Overlap (TO) metric based on a common set of neighbours had been developed to decide the level of connectivity between two genes [**12**]. From various TO measures, the far less investigated Generalized Topological Overlap Measure (GTOM) can be selected to portray a complete picture of indirect association among genes. In this regard, the flexibility of assessing indirect regulatory architecture depends on the number of intermediate nodes/genes. The focus on the indirect regulatory architecture stems from the fact that driver genes in carcinogenic studies may bring a phenotypic change in a cell considering the overall co-expression pattern and the mutation frequency in some sets of localized genes [**14-17**]. In each localized set or pathway measuring the indispensability of the driver gene followed by its distinct mutational change compared to neighbouring genes helps in understanding the heterogeneous role of the driver in the presence of the associated genes in cancer cells [**18**]. To understand such localized pathways, the above mentioned GTOM measure can be applied on a regression-based method to extract one to many associations among genes. In this regard, a regularized regression model called least absolute shrinkage and selection operator (LASSO) is chosen in this research because of its ability to infer correctly in high dimensional data sets [**19**]. Fundamentally the LASSO operator shrinks some coefficients and makes others zero, and thus tries to hold the functional attributes of both subset selection and ridge regression [**20**]. In this specific contribution towards understanding the indirect roles of one or more regulator genes, a fused LASSO based framework is developed using the Generalized Topological Overlap Measure (GTOM) [**21**] as the fundamental measure in the LASSO structure instead of using the gene expression values [**22**]. In this perspective, mTOM (multiple Topological Overlap Measure) signifying the regulatory application in the nth order variant of GTOM [**21**] maintains a crucial role in finding the unknown or hidden factors of regulation.

6.2.1 The basic findings: The performance of the proposed topologically overlapped fused LASSO measure for reconstructing regulatory networks has been checked in four real datasets; YEAST 5 ON/OFF data [**23**], YEAST 11, YEAST Cell Cycle [**24**], and Human Cancer data (HeLa) [**1,25**]. In the context of YEAST 5 and YEAST 11 regulation network, the regulatory performance outcomes has been checked and compared with some benchmark algorithms like Time Delay Network De-convolution (TD_ND) [**26**], Time Delay ARACNE [**24**], Correlation with Lasso (XCorr+Lasso) [**27**], Delay Detection Lasso (DD-Lasso) [**28**], and Group Lasso [**29**]. In each case, the improvement is judged in terms of precision, recall, and F score. The above comparative analysis has been made complete through inclusion of novel interactions obtained from YAGM [**30**] and Regulator DB [**31**] in addition to available benchmark networks given in [**24**]. An unmitigated investigation has also been carried out involving datasets possessing a large number of gene entities like the YEAST cell cycle and HeLa. In the context of YEAST cell cycle data, YEASTRACT [**32**] along with YAGM and Regulator DB is utilized to acquire all biologically relevant TF to TF and TF to target interaction particulars. In the case of HeLa, the same is done with the help of the TF2target [**33**] and TRRUST [**34**] database.

6.2.2 Methodology: The algorithm, mentioned below, is implemented in a LASSO framework. It is documented in [**20**] that LASSO relies on the combination of L2 and L1 norms, introducing sparsity by reducing the coefficients to zero. Moreover, the fusion in LASSO is meant to deal with problems with a reasonable order of features. Another added characteristic of fused LASSO happens to be the successive differences of regression coefficients in ordered fashion over and above minimization constraint on the regression coefficients.

The prime requirement of the algorithm is about assimilating information on the difference between semi-direct and higher-order indirect connectivities, between the target and controller genes. In the cases of YEAST 5 and YEAST 11, every gene is considered a target, with all other genes being the candidate regressors or potential regulators. However, to indulge and understand some significant portions of the varied forms of gene regulations present in complex pathways involving immense gene to gene associations, the analysis is conducted verifying the biological regulation performance, in the context of the YEAST Cell Cycle and HeLa Cancer Data.

ALGORITHM: Gene Regulatory Network using mTOM based LASSO (TO-LASSO)

Input:

1. *Dataset* with multiple conditions, having N TF and T target genes.
2. Known biological database for validation.

Output:

Gene regulatory pathways with significant contribution of unknown or hidden regulators on multiple differential regulatory links.

Description

```
1: Initialize the set of Regulatory Pathways RP = $\phi$
2: for each Dataset
3:   initialize responce vector Y=$\phi$, predictor matrix X=$\phi$, weight matrix W=$\phi$
4:     for each condition
5:       for each target t (t$\in$T)
6:         calculate Topological Overlap (TO) between the target gene t and each TF
7:           for each reference TF gene n (n$\in$N)
8:               calculate multiple Topological Overlap Measure (mTOM) between the
                 reference TF gene n and target gene t considering all possible
                 combinations of other TF genes
9:           end for
10:      end for
11:    end for
12:    for each target t (t$\in$T)
13:        Append the vectors of (N-1) TO values obtained in each condition to get the
           single response vector Y
14:        Append the matrices of mTOM values obtained in each condition to get the
           predictor matrix X
15:        Calculate the matrix comprising probability of differential connectivities
           (pDC) between the vector of (N-1) TO values and the corresponding mTOM
           matrix in each condition
16:        Generate weight matrix W putting in place the pDC matrices in a condition
           specific manner
17.        Apply fused LASSO modeling to compute optimized regressor vector ($\beta$)
18.        Use $\beta$ to find most similar indirect regulatory pathways between a TF and the
           target gene considering all conditions of interest
19:        Set of Regulatory Pathways RP is appended with statistically significant
           indirect regulatory pathways to the target gene
20:    end for
21: end for
```

The above algorithm works on any dataset having two or more experimental conditions. Lines 4 through 6 along with lines 12 and 13 of the developed algorithm highlight the 1st order topological overlap (TO) or semi direct connectivity between a target gene ‘t’ and all other genes in each experimental condition. In this regard, a TO valued response vector ‘Y’ can be obtained for each target gene ‘t’. Hence, assuming N-1 regulator or TF genes, the structure of ‘Y’ in the presence of two experimental conditions will be:

$$Y = \begin{bmatrix} C_1 \begin{cases} m_{1.t} \\ m_{2,t} \\ . \\ . \\ m_{N-1,t} \end{cases} \\ \\ C_2 \begin{cases} m_{1,t} \\ m_{2,t} \\ . \\ . \\ m_{N-1,t} \end{cases} \end{bmatrix}_{[2(N-1) \times 1]}$$

Here, C_n stands for n^{th} condition, and $m_{x,t}$ represents the TO or semi direct connectivity between x^{th} regressor and ‘t’. The concept of $m_{x,t}$ can be elucidated as given in [**21**]:

$$m_{x,t} = \frac{|N(x,t)|+a_{xt}}{min\{|N(x,-t)|,|N(-x,t)|\}+\binom{V}{2}} \quad \text{where, } |N(x,t)| = \sum_{u \neq t,x} a_{tu} a_{xu}$$

$$\& \quad |N(x,-t)| = \sum_{u \neq x} a_{xu} - a_{tx}$$

Above, a_{tu} stands for the correlation between t and u. V stands for the number of participating nodes required to calculate mTOM between source and target, including both. In the above context, two nodes (x and t) are in consideration. Thus, V=2. The binomial coefficient given in the denominator gives us the upper bound of a_{xt}. In other words, only one connection is possible between ‘x’ and ‘t’. The 1st order topological overlap states the physical dependence between two genes considering the effect of direct and other intermediate genes, one at a time. In other words, the direct dependency is considered or studied in the presence of individual involvement of the rest (N-2) genes. Putting n=2, the size of the vector ‘Y’ is [2(N-1)×1].

Next thing is to calculate the predictor matrix ‘X’ for each target gene following lines 7 through 9 and 14 of the designed algorithm. It comprises of mTOM values between a target gene ‘t’ and each reference TF or regulator gene considering all possible combinations of the other regressors (TFs or regulator genes). In this regard, the increasing degree of connectivity in each condition is considered. Hence, with n=2 experimental conditions, the structure of ‘X’ is:

$$X = \begin{bmatrix} V_{C_1} & 0 \\ 0 & V_{C_2} \end{bmatrix}_{2(N-1)\times 2(N-2)} \quad \text{where,}$$

$$V_{C_n} = \begin{bmatrix} 1_{TOM}2_{TOM}.......(N-2)_{TOM} \ \ (for \ \ 1^{st} \ \ gene \ ) \\ 1_{TOM}2_{TOM}.......(N-2)_{TOM} \ \ (for \ \ 2^{nd} \ \ gene \ ) \\ . \\ 1_{TOM}2_{TOM}.......(N-2)_{TOM} \ \ (for \ \ (N-1)^{th} \ \ gene \ ) \end{bmatrix}$$

For the xth TF gene acting as source, 1_{TOM} can be written as $m_{x,y,t}$ indicating the overlap between x and t via y, i.e. depicting full indirect connectivity of order one. As given in [**21**], the mathematical representation of mTOM is:

$$m_{x,y,t} = \frac{|N(x,y,t)| + a_{xy} + a_{xt} + a_{yt}}{\min\{|N(x,y,-t)|, |N(x,-y,t)|, |N(-x,y,t)|\} + \binom{V}{2}}$$

where, $|N(x, y, t)| = \sum_{u \neq t,x,y} a_{tu} a_{xu} a_{yu}$

and $|N(x, y, -t)| = \sum_{u \neq x,y} a_{xu} a_{yu} - a_{xt} a_{yt}$

The case considered in the above expression has three different nodes in alignment (x, y, and t). Hence, the value of V= 3. In this context, the complete set of 1_{TOM} can be framed considering individually (i.e. y) all the remaining N-2 genes; finally adding the individual results. A similar analogy is applicable for all higher order mTOM analysis. In the case of higher order mTOM analysis, more than one intermediate TF regressor or regulator gene comes in consideration, i.e. delving into indirect connectivity of order more than one. In other words, mTOM analysis with m >1 depicts proper indirect connectivity to a target gene. Here, the size of V_{C_n} is (N-1) × (N-2). Accordingly, the dimension of matrix 'X' is 2(N-1) × 2 (N-2).

For large datasets like YEAST cell cycle and HeLa, only those genes are taken in the higher orders (greater than one), which have some biological evidence of co-existence with the source gene 'x'. For example, in the context of HeLa, the two databases

TF2target [**33**] and TRRUST [**34**] are considered to reconstruct TF to target based GRN accurately. At the initial stage, these two databases have been used to form a regulatory pathway that states the possible number of regulators that may be present in a hierarchy (indirect mode of interaction to a target gene) for a reference gene ‘t1’. For example, it is assumed that genes 'a1', 'b1', and 'c1' are the first set of regulatees of ‘t1’. Further exploration of the database reveals that gene 'a1' controls 'm1' and 'n1'; gene 'b1' controls gene 'o1' and gene 'c1' controls gene 'p1' with no additional regulatee for any one of 'm1', 'n1', 'o1' or 'p1'. With this assumption, mTOM computation corresponding to 't1' in sparse matrix X is completed. However, for small datasets like YEAST 5 or YEAST 11, all possible regulator combinations are considered in the mTOM computation discussed above.

After calculating ‘X’ and ‘Y’, the weight matrix W is framed for any pair. It is done by calculating the mean probability of differential connectivity between semi-direct and higher-order indirect associations, as given in line 15 of algorithm. Finally, as per line 16 of the same algorithm, the weight matrix is calculated as:

$$W_{Y,j} = \begin{bmatrix} P_{Y,j}^{C_1} & 0 \\ 0 & P_{Y,j}^{C_2} \end{bmatrix} \; where \;\; , P_{Y,j}^{C} = pvalue \; _{Y,j}^{C}$$

Here, above ‘j’ indicates a particular degree of connectivity; effectively counting the number of regulators that might be present in the regulation path to the target gene. Effect of all possible combination of regulators can be obtained simply by varying the degree ‘j’. In the above equation, W portrays the statistical significance of the similarity of differential connectivity between target and regulators, considering semi-direct and fully indirect perspective with varying degrees of connectivity. In other words, corresponding to a degree ‘j’, if any entry of W is quite low, the similarity between semi-direct and fully indirect situation with degree ‘j’ is statistically significant for the concerned target and the source gene under consideration. Hence, dimensionally $P_{Y,j}^{C}$ is 1×(N-2). Thus, for (N-1) regulators, the size of

$P^{C}_{Y,j}$ becomes (N-1) × (N-2). Hence, identical to 'X', the total dimension of 'W' is [2(N-1) × 2(N-2)].

At this point, comes the regression vector β between 'Y' and 'X' across each condition. Fetching the concept from [**22**], the lasso term $\left\|\beta^{C_2} - \beta^{C_1}\right\|_1$ is added to ensure that the two reconstructed networks are as close as possible. In this regard, the following action is taken as per line 17 of the algorithm.

$$\hat{\beta} = \arg_{\beta} \min[\left\|Y - X\beta\right\|_2^2 + \lambda_1\left\|W\beta\right\|_1 + \lambda_2\left\|\beta^{C_2} - \beta^{C_1}\right\|_1]$$

Here, the best possible estimate of the regression vector is the motto, keeping in frame simultaneous optimization of tuning parameters λ_1 and λ_2. R package 'GLMNET' [**35**] has been used to obtain the optimized tuning parameters. Following line 18 of the algorithm, the vector β is used to extract out interactions to the target gene from a source gene that shows a similar type of connectivity between semi-direct and completely indirect connectivity of all possible degrees till (N-2) or as the case may be depending on biological pathway evidence.

At the ultimate level, before carrying any comparison, the values $\hat{\beta}$, λ_1 and λ_2 are applied in the equation $f = \left\|Y - X\beta\right\|_2^2 + \lambda_1\left\|W\beta\right\|_1 + \lambda_2\left\|\beta^{C_2} - \beta^{C_1}\right\|_1$ separately in each condition. A pair is considered accurate, provided the source gene yields a significantly low score in 'f' vector (fused LASSO vector) per condition of interest. This approach can suitably be extended to more than two conditions of interest, provided statistical significance analysis of the 'f' vector shows a low p-score in each condition of interest. However, in such a case, higher number of added LASSO terms toward computation of the regression vector will come into the application. It means with four conditions of interest with C1 indicating the control state, the added lasso terms including the tuning parameters will be like $\lambda_2\left\|\beta^{C_2} - \beta^{C_1}\right\|_1$, $\lambda_3\left\|\beta^{C_3} - \beta^{C_1}\right\|_1$, and $\lambda_4\left\|\beta^{C_4} - \beta^{C_1}\right\|_1$.

6.2.3 Results: The algorithm discussed above has been applied on four datasets (individual application outcomes are shown below) to reconstruct GRNs or TRNs, wherever applicable, followed by checking the robustness of the approach through the network decisive parameters, namely Precision, Recall, and/or F score. These parameters can be defined as follows:

Precision (i.e. P) = True Positive / (True Positive + False Positive), i.e. TP/(TP+FP)

Recall (i.e. R) = True Positive / (True Positive + False Negative), i.e. TP/(TP+FN)

& F score = 2×Precision×Recall / (Precision + Recall), i.e. 2PR/(P+R)

Above, TP = True Positive or regulations present in the reported databases as well as in the obtained results, FP = False Positive or regulations absent in the reported databases but present in the obtained results, TN = True Negative or regulations absent in both the databases as well as in the obtained results, FN = False Negative or regulations present in the reported databases, but absent in the obtained results.

The various datasets used to prove the robustness of the research work presented in this chapter through comparative analysis with standard benchmark algorithms and/or checking the level of statistical similarity with biological evidence of hierarchical gene regulatory pathways are presented below. Henceforth, the abbreviation TO-LASSO (mentioned in the algorithm nomenclature as well) is used to present the research conducted with the designed algorithm.

<u>At first the results from YEAST 5 ON/OFF data</u>: Here switching between ON and OFF states can be done by formulating glucose and galactose in cells. The 5 genes under consideration in this process are SWI5, ASH1, CBF1, GAL4, and GAL80. Culturing of cells and corresponding detailed mechanisms can be found in [**23**]. GRN obtained from the data by applying TO-LASSO is given in Figure 6.1. In the figure, solid lines represent True Positives and dashed lines signify False Positives.

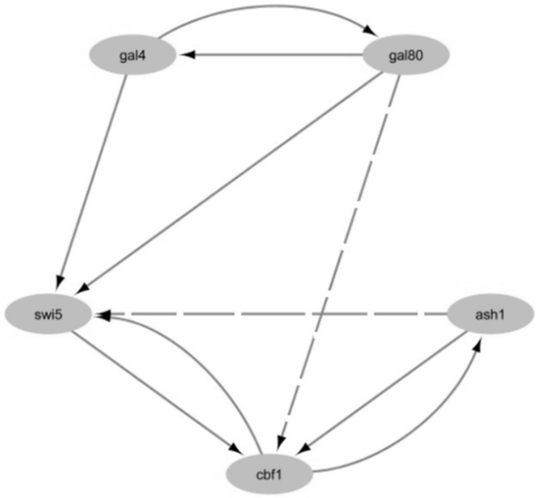


Figure 6.1: TO-LASSO based GRN for YEAST ON/OFF Data

The results of Precision, Recall, and F score after implementing TO-LASSO and comparing with the standard benchmark algorithms (already mentioned in section 6.2.1) are listed in Table 6.1.

Table 6.1: Comparative results from YEAST ON/OFF data

Method	Dataset	Precision	Recall	F Score
TO-LASSO	Yeast 5	0.8	0.727	0.762
	Yeast 5 ON	0.75	0.818	0.78
	Yeast 5 OFF	0.66	0.909	0.77
DD LASSO	Yeast 5 ON	0.58	0.63	0.61
	Yeast 5 OFF	0.571	0.6	0.545
Group LASSO	Yeast 5 ON	0.308	0.444	0.364
	Yeast 5 OFF	0.371	0.436	0.4
TD_ND	Yeast 5 ON	0.643	0.818	0.72
	Yeast 5 OFF	0.625	0.909	0.741
TD_ARACNE	Yeast 5 ON	0.667	0.182	0.286
	Yeast 5 OFF	0.5	0.091	0.154
XCorr+LASSO	Yeast 5 ON	0.8	0.364	0.5
	Yeast 5 OFF	0.5	0.182	0.267

The existent benchmark algorithms which have worked on this dataset have applied reverse engineering methodologies separately in ON and OFF data. In TO-LASSO (an extended analysis of the fused lasso framework [**22**], the result is obtained considering a conjoined version of ON and OFF data followed by similarity analysis with the benchmark network given in [**23**]. However, maintaining the line of research followed by the benchmark reverse engineered algorithms depicted in Table 6.1, individual outcomes of ON and OFF counterparts are given along with the conjoined result stated above. It is clear from Table 6.1 that TO-LASSO outperforms all the benchmark algorithms in terms of the final F score. The results also depict that the overall scores in TD_ND are very close to TO-LASSO. In this regard, it is essential to mention that TO-LASSO highlights on fully indirect regulation (between a source-target gene pair) possessing statistical significant similarity with the semi direct regulatory effect. In other words, this depicts the inevitable presence of one or more intermediary entities important to be considered in the regulation process for the corresponding source-target gene pair. Hence, the same is not similar to direct regulation models explored in

TD_ND and other compared algorithms. To date, there is no existent architecture following a similar kind of concept.

Figure 6.1 depicts the reconstructed network showing the TO-LASSO regulations between any pair of genes. For example, GAL80 to GAL4 is not a direct regulation. In other words, this source-target pair does have the explicit presence of one or more intermediary genes as per the proposed algorithm. This interaction happens to yield statistically similar causal effect with respect to direct regulation context in between GAL80 and GAL4, if any. Thus, the true and false positive outcomes shown in the figure relate our predicted outputs to the benchmark network [**23**]. In this regard, to highlight the fully indirect process of order one (could not be higher in the implementation because of very few genes under consideration from the parent dataset), the most significant indirect pathways are enlisted in Table 6.2 for each target gene. In this table, any row gives us the regulatory pathway for each target, shown in the first column. The gene present in the second column is the intermediary gene with the source gene given in the third or last column of the table. TO-LASSO analysis elucidates that all the other four genes indirectly regulate SWI5, but the best regulation in terms of β score is listed here. From the table, it is observed, the best regulator gene of SWI5 is GAL80 via intermediary gene GAL4. The same is repeated for each target gene.

Table 6.2: Best Regulatory Pathways of YEAST ON/OFF Data

Target Gene	Intermediate (associate) Gene	Final Regulator
CBF1	SWI5	GAL80
GAL4	***CBF1***	**GAL80**
SWI5	GAL4	GAL80
GAL80	**CBF1**	**GAL4**
ASH1	**SWI5**	CBF1

Here, corresponding to a pathway, any italicized or bold pair of genes present in adjacent columns indicates the absence of fully indirect regulation as per TO-LASSO algorithm. This indicates the gene pair possesses statistically dissimilar fully indirect and semi-direct connectivities. A deeper analysis clarifies an italicized pair signifying a false negative, i.e. the pair is biologically valid in existing literature like DAVID or Gene Trail (depicting by default direct biological connectivity). However, the bold pair highlights a true negative connection, i.e. the pair is not biologically validated either.

For example, we inspect the regulation of GAL4 by GAL80 via CBF1. In this simple hierarchical regulation, neither GAL4 nor CBF1 has been found to possess full indirect regulation by CBF1 and GAL80 respectively. As per the findings of TO-LASSO approach, CBF1 to GAL4 is a false negative (association can be found in [**37**]) and GAL80 to CBF1 is a true negative connection. Similar true negative connections have been discovered between CBF1 to GAL80, GAL4 to CBF1, and SWI5 to ASH1.

Next, exploring the results from YEAST 11 data: Here, from a total of 6178 genes, a subset of 11 genes is extracted as given in [**24**]. The 11 genes are CLN1, CLN2, CLN3, SWI4, SWI6, MBP1, CLB5, CLB6, SIC1, CDC28, and CDC6.

For this dataset, the performance of TO-LASSO analysis is compared with the same set of standard algorithms considered earlier. The comparative results in terms of Precision, Recall, and F score are listed in Table 6.3. From this table, it is evident that TO-LASSO analysis performs better than any of the state-of-the-art techniques analyzing this dataset. In this case of 11 genes extracted from the YEAST Cell Cycle dataset, each of the compared benchmark algorithms show the performance level with respect to a single gene regulation network corresponding to any one condition of the G1 cycle. However, TO-LASSO performs the analysis considering all the four experimental conditions (alpha, cdc15, cdc28, and elu) distributed over 18, 24, 17, and 14 time points, respectively. In this algorithm, any two of the four time-series information sequences are taken at a time. This yields six pairs of differential data from where six gene networks are designed. The regulatory links in all these six networks including the necessary documentation on the significantly best pathway obtained for each of these 11 genes, similar to the one depicted for YEAST ON/OFF data, are given in Figures 6.2 to 6.7 with the corresponding Tables 6.4 to 6.9. In all the above mentioned figures solid lines represent True Positives and dashed lines represent False Positives. In each of the significant best regulatory pathway matrices presented in the above tables, the first column signifies final regulated gene and the last column shows the controller gene able to maintain statistically similar semi-direct and fully indirect connectivity to the regulated gene, present in the first column, with the intervention of the intermediate genes, which state the order of the statistically significant indirect connectivity. Here pairs, represented in bold signifies True Negative and by italic signifies False Negative.

Table 6.3: Comparative results from Yeast Cell-Cycle Data (11 genes; YEAST 11)

Method	Data Set	Precision	Recall	F Score
TO-LASSO	Yeast Cell Cycle (alpha-cdc15)	0.648	0.676	0.662
	Yeast Cell Cycle (alpha-cdc28)	0.606	0.642	0.623
	Yeast Cell Cycle (alpha-elu)	0.591	0.56	0.573
	Yeast Cell Cycle (cdc15-cdc28)	0.732	0.703	0.717
	Yeast Cell Cycle (cdc15-elu)	0.549	0.582	0.565
	Yeast Cell Cycle (cdc28-elu)	0.633	0.662	0.647
DD LASSO	Yeast Cell Cycle	0.309	0.815	0.449
Group LASSO	Yeast Cell Cycle	0.253	0.621	0.36
TD_ND	Yeast Cell Cycle	0.5778	0.3662	0.4483
TD_ARACNE	Yeast Cell Cycle	1	0.239	0.386
XCorr+LASSO	Yeast Cell Cycle	0.714	0.14	0.235

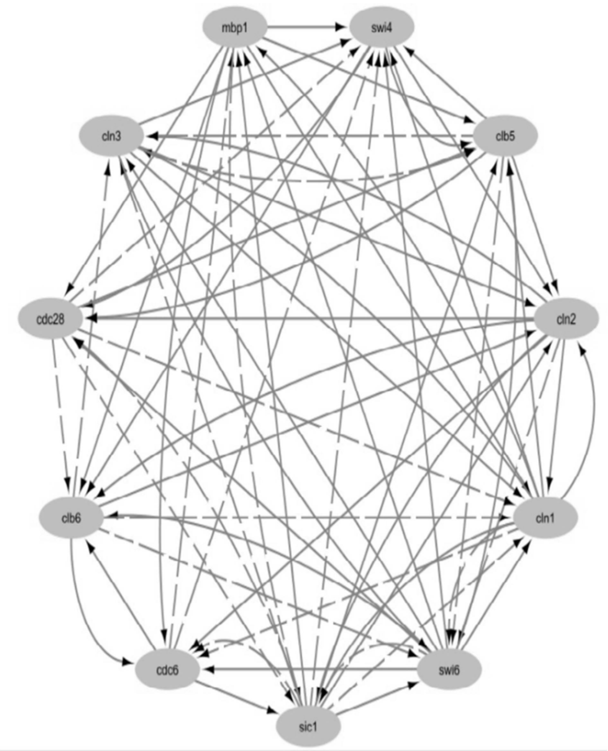
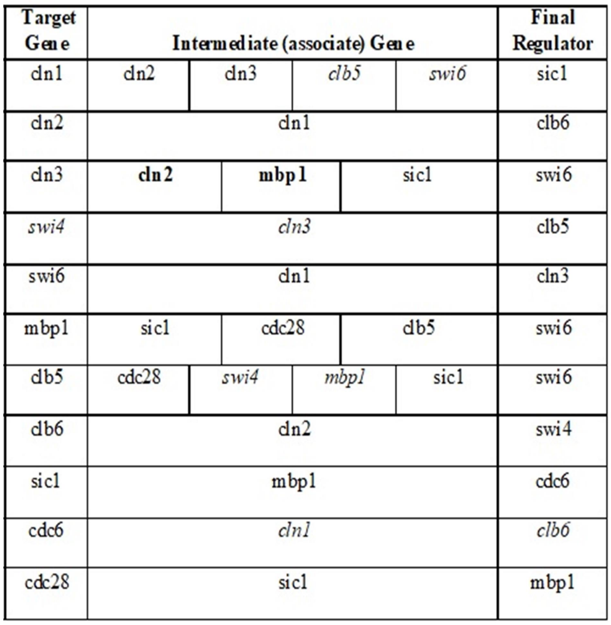


Figure 6.2: TO-LASSO based GRN for YEAST 11 (alpha_cdc15)

Table 6.4: Best Regulatory pathways (YEAST 11 data; alpha_cdc15)



Target Gene	Intermediate (associate) Gene				Final Regulator
dn1	dn2	dn3	*clb5*	*swi6*	sic1
dn2	dn1				clb6
dn3	**dn2**	**mbp1**	sic1		swi6
swi4	*cln3*				clb5
swi6	dn1				cln3
mbp1	sic1	cdc28	db5		swi6
db5	cdc28	*swi4*	*mbp1*	sic1	swi6
db6	dn2				swi4
sic1	mbp1				cdc6
cdc6	*cln1*				*clb6*
cdc28	sic1				mbp1

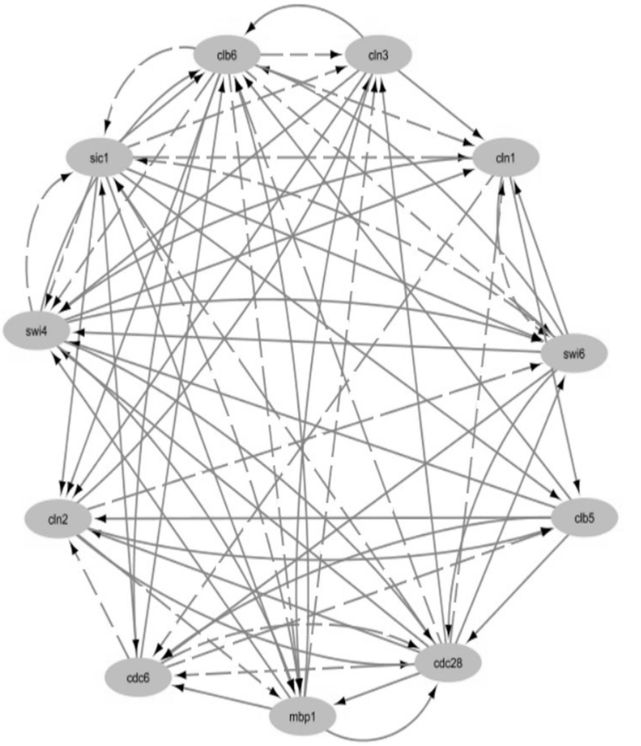
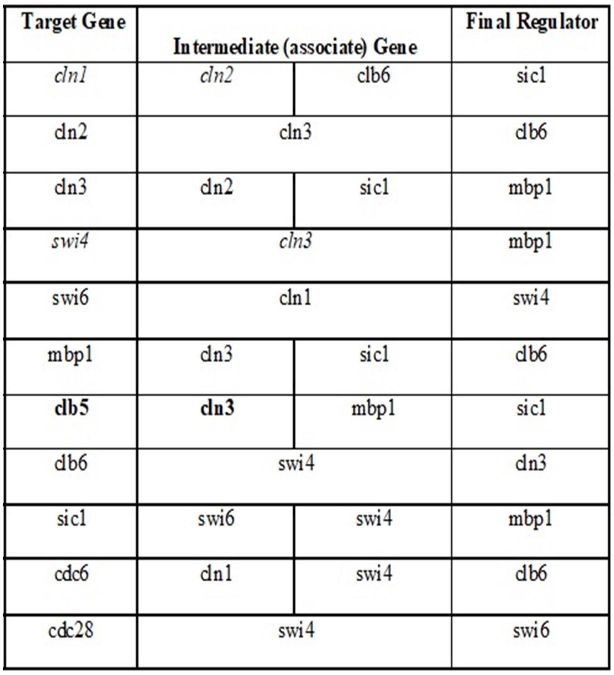


Figure 6.3: TO-LASSO based GRN for YEAST 11 (alpha_cdc28)

Table 6.5: Best Regulatory pathways (YEAST 11 data; alpha _cdc28)



Target Gene	Intermediate (associate) Gene		Final Regulator
cln1	*cln2*	clb6	sic1
cln2	cln3		clb6
cln3	cln2	sic1	mbp1
swi4	*cln3*		mbp1
swi6	cln1		swi4
mbp1	cln3	sic1	clb6
clb5	**cln3**	mbp1	sic1
clb6	swi4		cln3
sic1	swi6	swi4	mbp1
cdc6	cln1	swi4	clb6
cdc28	swi4		swi6

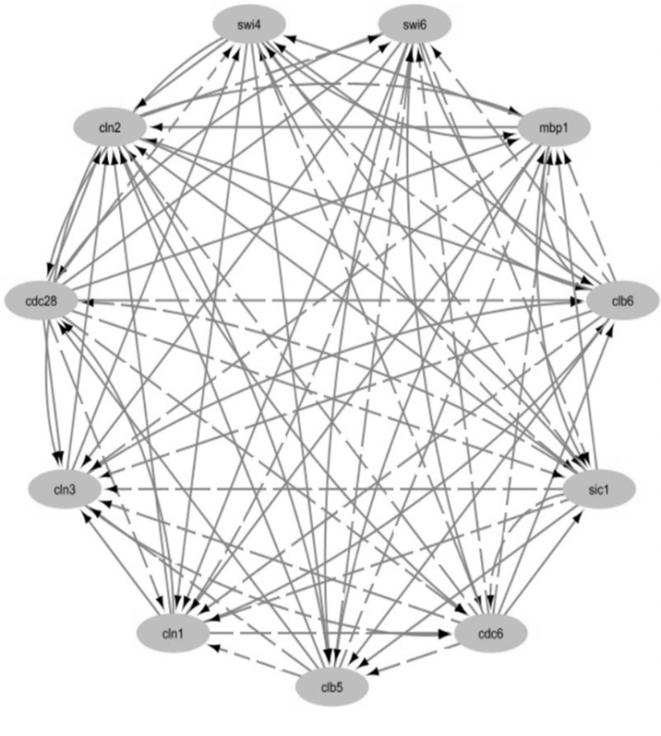
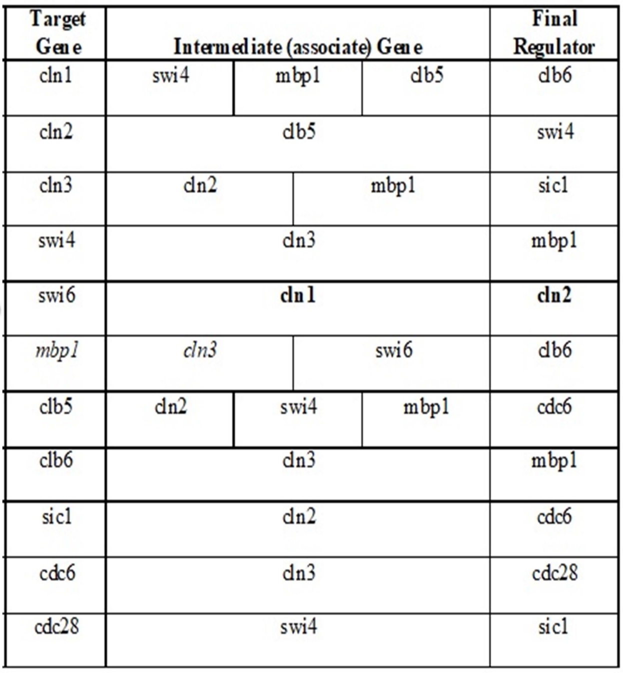


Figure 6.4: TO-LASSO based GRN for YEAST 11 (alpha_elu)

Table 6.6: Best Regulatory pathways (YEAST 11 data; alpha _elu)



Target Gene	Intermediate (associate) Gene			Final Regulator
cln1	swi4	mbp1	clb5	clb6
cln2	clb5			swi4
cln3	cln2	mbp1		sic1
swi4	cln3			mbp1
swi6	**cln1**			**cln2**
mbp1	*cln3*	swi6		clb6
clb5	cln2	swi4	mbp1	cdc6
clb6	cln3			mbp1
sic1	cln2			cdc6
cdc6	cln3			cdc28
cdc28	swi4			sic1

Table 6.7: Best Regulatory pathways (YEAST 11 data; cdc15_cdc28)

Target Gene	Intermediate (associate) Gene				Final Regulator
cln1	cln2	cln3	mbp1	clb5	sic1
cln2	swi6				swi4
cln3	mbp1				sic1
swi4	**cln2**	**mbp1**	clb5		sic1
swi6	cln1				cln2
mbp1	cln3	swi4	swi6		sic1
clb5	*cln2*	swi4	swi6		cln3
clb6	**cln1**				swi4
sic1	cln3	swi4			mbp1
cdc6	**cln1**	cln2			swi4
cdc28	cln1	swi4			mbp1

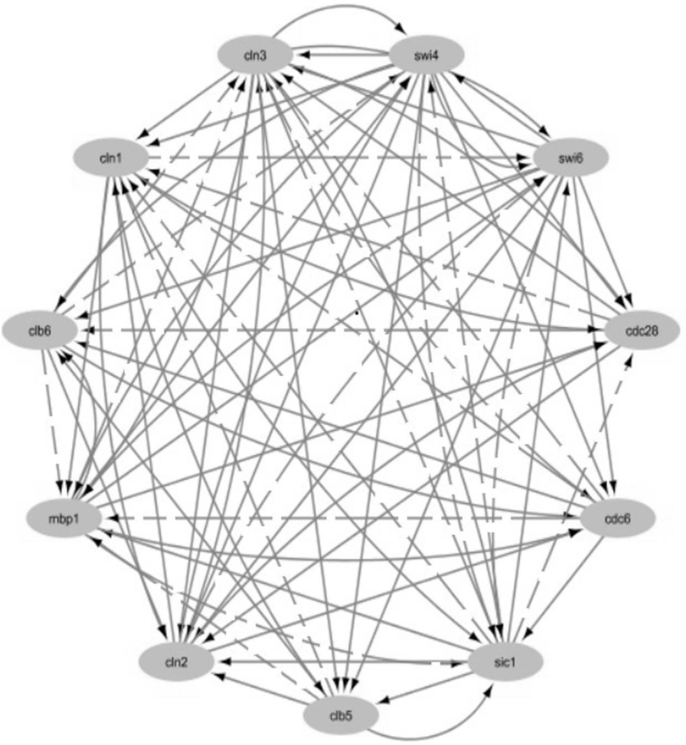


Figure 6.5: TO-LASSO based GRN for YEAST 11 (cdc15_cdc28)

Table 6.8: Best Regulatory pathways (YEAST 11 data; cdc15_elu)

Target Gene	Intermediate (associate) Gene			Final Regulator
cln1	**cln3**	mbp1		clb6
cln2	**mbp1**			sic1
cln3	mbp1			cdc6
swi4	clb6	**mbp1**	**clb5**	sic1
swi6	cln1			clb6
mbp1	cln3	swi6		clb6
clb5	cln2	swi4	swi6	cln1
clb6	**cln3**			mbp1
sic1	cln3	swi4		clb6
cdc6	cln1	cdc28		cln3
cdc28	sic1	clb5		cln3

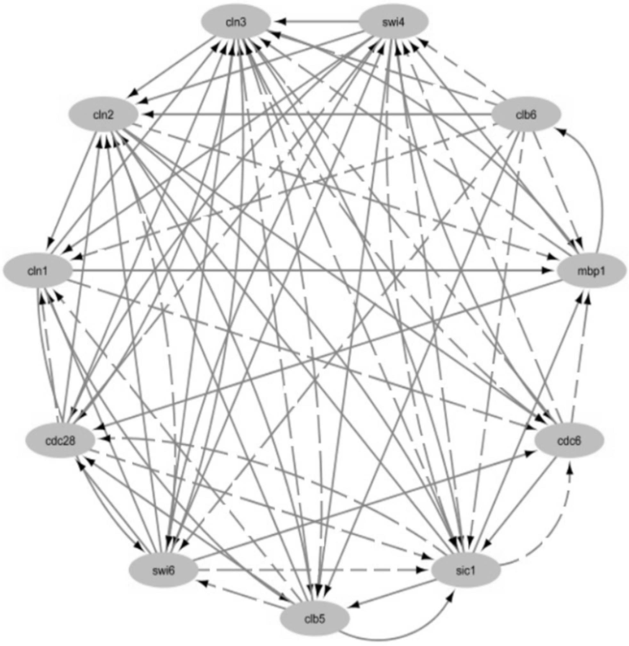
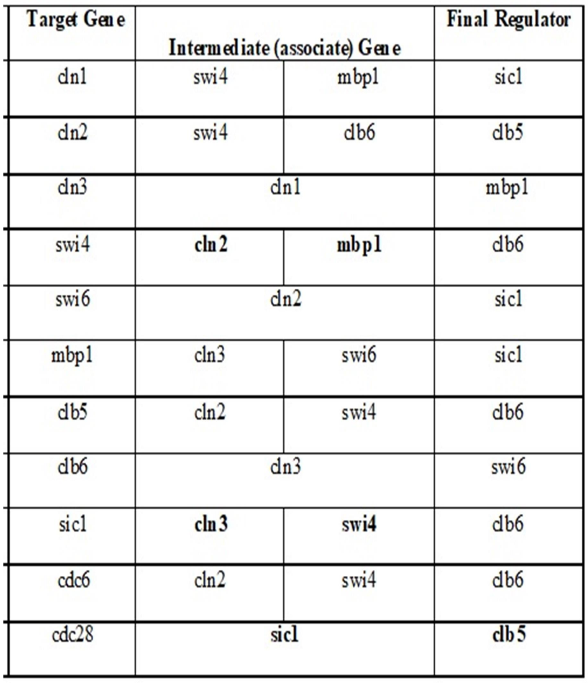


Figure 6.6: TO-LASSO based GRN for YEAST 11 (cdc15_elu)

Table 6.9: Best Regulatory pathways (YEAST 11data; cdc28_elu)



Target Gene	Intermediate (associate) Gene		Final Regulator
cln1	swi4	mbp1	sic1
cln2	swi4	clb6	clb5
cln3	cln1		mbp1
swi4	**cln2**	**mbp1**	clb6
swi6	cln2		sic1
mbp1	cln3	swi6	sic1
clb5	cln2	swi4	clb6
clb6	cln3		swi6
sic1	**cln3**	**swi4**	clb6
cdc6	cln2	swi4	clb6
cdc28	**sic1**		**clb5**

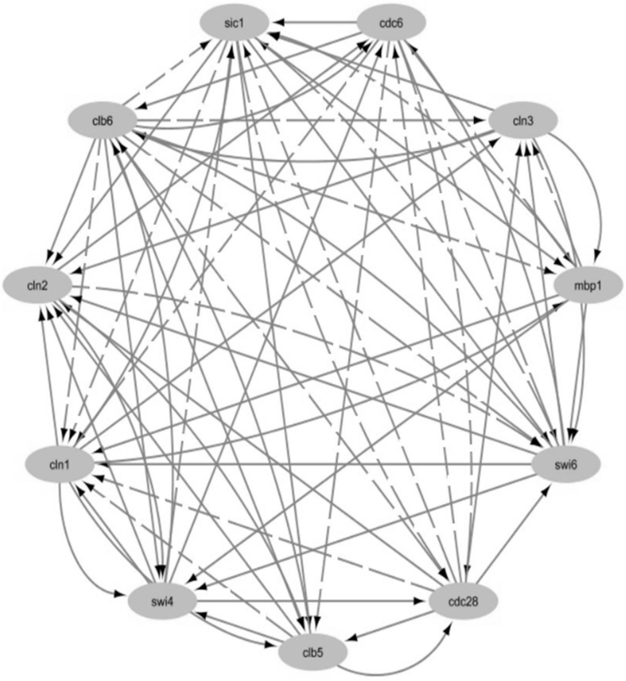


Figure 6.7: TO-LASSO based GRN for YEAST 11 (cdc28_elu)

<u>Presenting the results from YEAST Cell Cycle Data having more than 200 genes</u>: Here the concern is about reconstructing gene regulation networks through framing of TF to target regulatory pathways via the TO-LASSO method.

Initially, the TF and the DE genes are found from the given data comprising of 6178 genes [**24**]. This yields 17 TF and 235 DE genes [**37**] . Here, the DE genes act as final targets. The regulation study from YEASTRACT [**32**], YAGM [**30**], and Regulator DB [**31**] databases helps to understand that 197 out of 235 DE genes are controlled by at least 1 TF. The remaining 38 DE genes are not controlled by any of the TF genes obtained above. Hence, the focus is restricted on the 197 DE genes only.

To understand the problem, the regulation of a DE gene 'd' by 3 TFs, 't1', 't2', and 't3' via some intermediate TFs can be taken as an example. Extending the discussion on the example of a regulation study given in section 6.2.2 (Methodology section) by assuming that ‘d’ is directly regulated via ‘m1’, ‘n1’, ‘o1’ and ‘p1’, it can be stated that 't1' can control ‘d’ via 4 hierarchical paths comprising 'a1', 'b1', 'c1', 'm1', 'n1', 'o1', and 'p1' from the regulation databases. As a consequence of TO-LASSO application, if it is assumed that in the above hierarchical regulation perspective by 't1' for target DE gene 'd', the best regulatory action (both in terms of prediction as well as significance analysis) is formed including 'a1', 'b1', 'm1', and 'o1', then the thought can portray the existence of 2 regulatory paths involving 't1', 'a1', 'm1' and 't1', 'b1', 'o1' working cohesively towards the regulation of target DE gene 'd'. A similar analogy can be repeated for 't2' and 't3'. The principal regulatory pathway of 'd' is formed by checking the best significance level, and F score among the three pathways separately obtained from 't1', 't2', and 't3', respectively. Following this thought the number of statistically significant F score judged TF gene regulatory pathways for the target DE genes is shown in Figure 6.8. In this regard, it is to be noted that the TF gene regulatory pathways are found for every DE gene considering all possible pairwise combinations of the four conditions (alpha, cdc15, cdc28, and elu) of YEAST Cell Cycle time course information. From the figure it is possible to apprehend that the combinational pair of conditions, alpha_cdc28, yields on an average better F score judged gene regulatory pathways compared to other combinational pairs of conditions. Thus the TF gene regulatory links to target DE genes are on an average more significant in the context the combinational pair, alpha_cdc28, considering the inherent presence of unknown TF genes computed through the TO-LASSO technique. The presence of these unknown TF genes of a certain order (order of mTOM association present in the TO-LASSO method) hence can be considered as the indispensable hidden factors of the so called direct regulatory links.

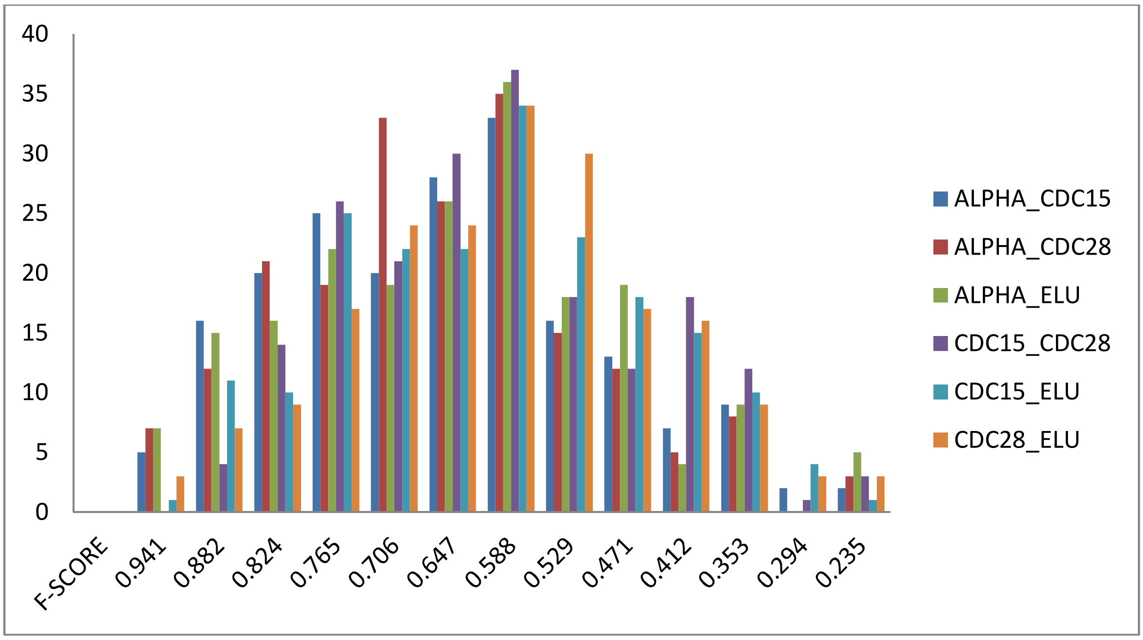


Figure 6.8: Number of statistically significant F score judged gene regulatory pathways (YEAST Cell Cycle)

One crucial observation noted in this regard is the indirect association of same average number of TF genes corresponding to any F score irrespective of the pair of conditions except cases where there is no evidence of any regulatory pathway corresponding to an F score. Such examples come in the context of (F score, pair of conditions) combinations like (0.941, cdc15_cdc28), (0.294, alpha_cdc28), and (0.294, alpha_elu) respectively.

Finally, coming up with the results from HeLa Cancer Data: Fetching the required data from [**25**] followed by extracting the necessary information from the databases (TF2target and TRRUST) and applying the R packages (maSigPro and GeneCycle) used in the last chapter, 137 TF and 619 DE genes are obtained. Out of this, as per the results presented in the last chapter, 37 TF genes and 496 DE genes happen to be periodic in nature. Thus at the very beginning the required combinations taken forward for implementation of TO-LASSO are the same as those presented in the last chapter, i.e. Aperiodic TF-Aperiodic DE (ApTF_ApDE), Aperiodic TF-Periodic DE (ApTF_PDE), Periodic TF-Aperiodic DE (PTF_ApDE), and Periodic TF-Periodic DE (PTF_PDE) respectively.

TO-LASSO is applied on each cluster [**38**] of DE genes to avoid computational complexity and to check the necessary performance of the outputs obtained per cluster. In this regard, similar to the last chapter, the 496 periodic DE genes are distributed in 8 clusters with 51, 68, 48, 60, 61, 62, 74, and 72 genes respectively. The remaining 123 aperiodic DE genes are accordingly distributed between 2 clusters with 69 and 54 genes respectively. Now the information is ready for TO-LASSO execution on the four different periodic and aperiodic combinations of TF to DE gene regulations.

The following results depicted in Figures 6.9 and 6.10 correspond to the number of statistically significant F score judged regulatory pathways for the aperiodic DE genes. Similar comment is applicable for interpreting the information pertaining to regulatory pathways shown in Figures 6.11 and 6.12 in relation to periodic DE genes. In these figures, the two conditions of interest are Phase 1 and Phase 2 on one hand and Phase 2 and Phase 3 on the other. TO-LASSO architecture demands a combinational pair of phases/stages for execution. As per the above two conditions following the TO-LASSO algorithm, Phase 2 of HeLa data can be treated here as the control state.

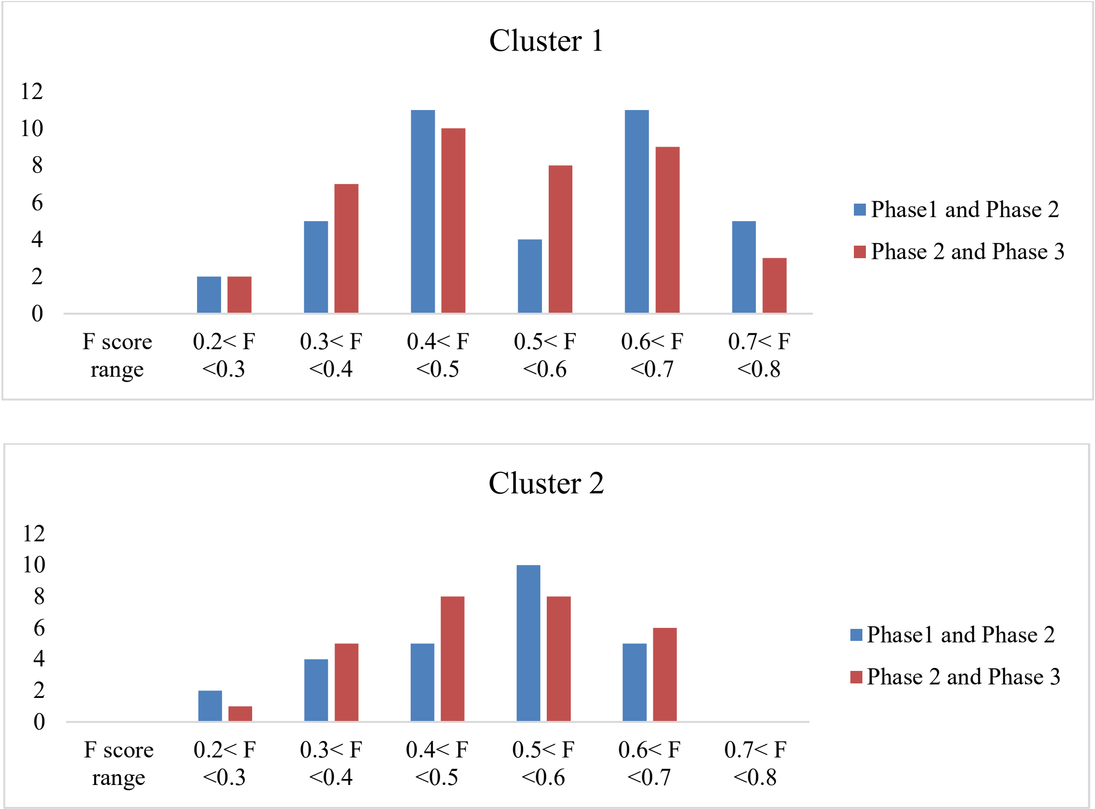


Figure 6.9: Number of statistically significant F score judged gene regulatory pathways (ApTF_ApDE)

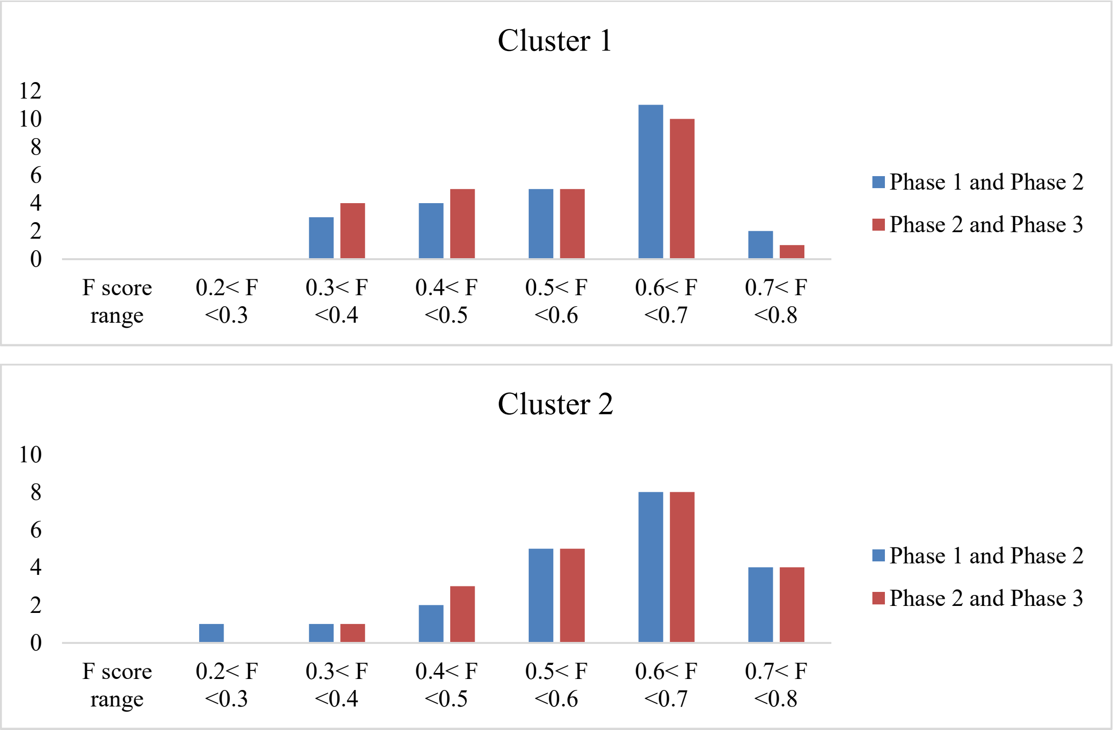


Figure 6.10: Number of statistically significant F score judged gene regulatory pathways (PTF_ApDE)

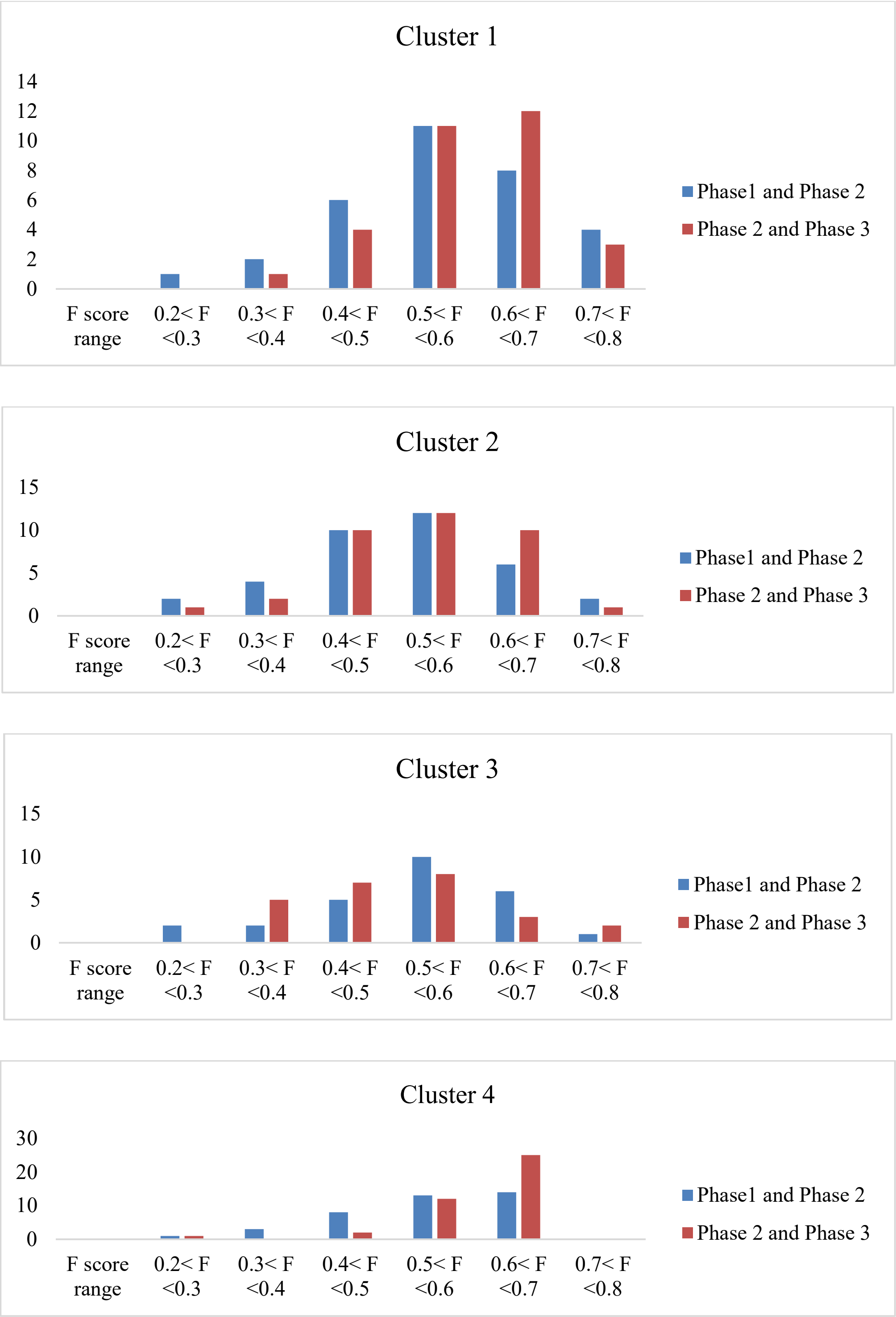


Figure 6.11: Number of statistically significant F score judged gene regulatory pathways (ApTF_PDE)…Contd

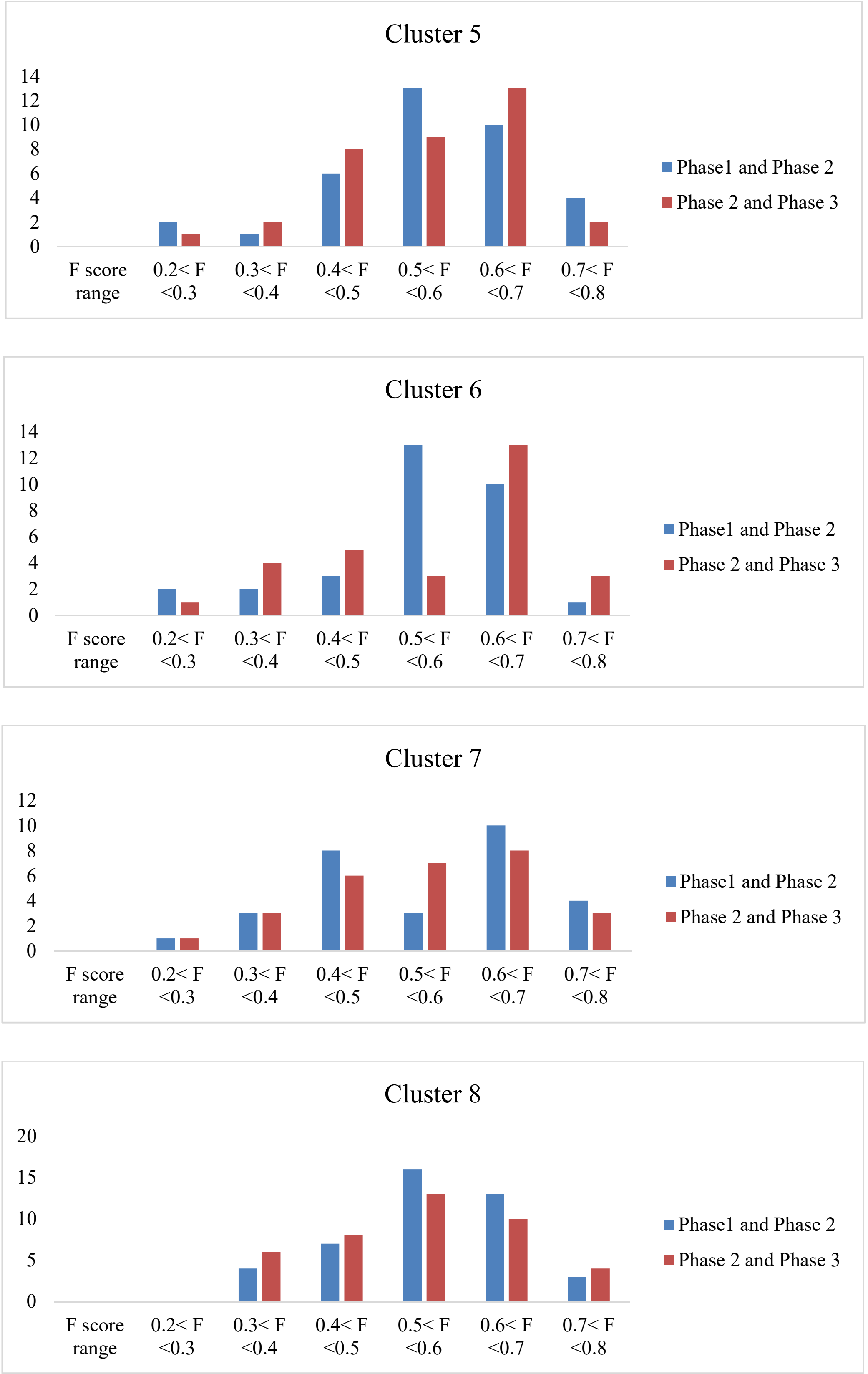


Figure 6.11: Number of statistically significant F score judged gene regulatory pathways (ApTF_PDE)

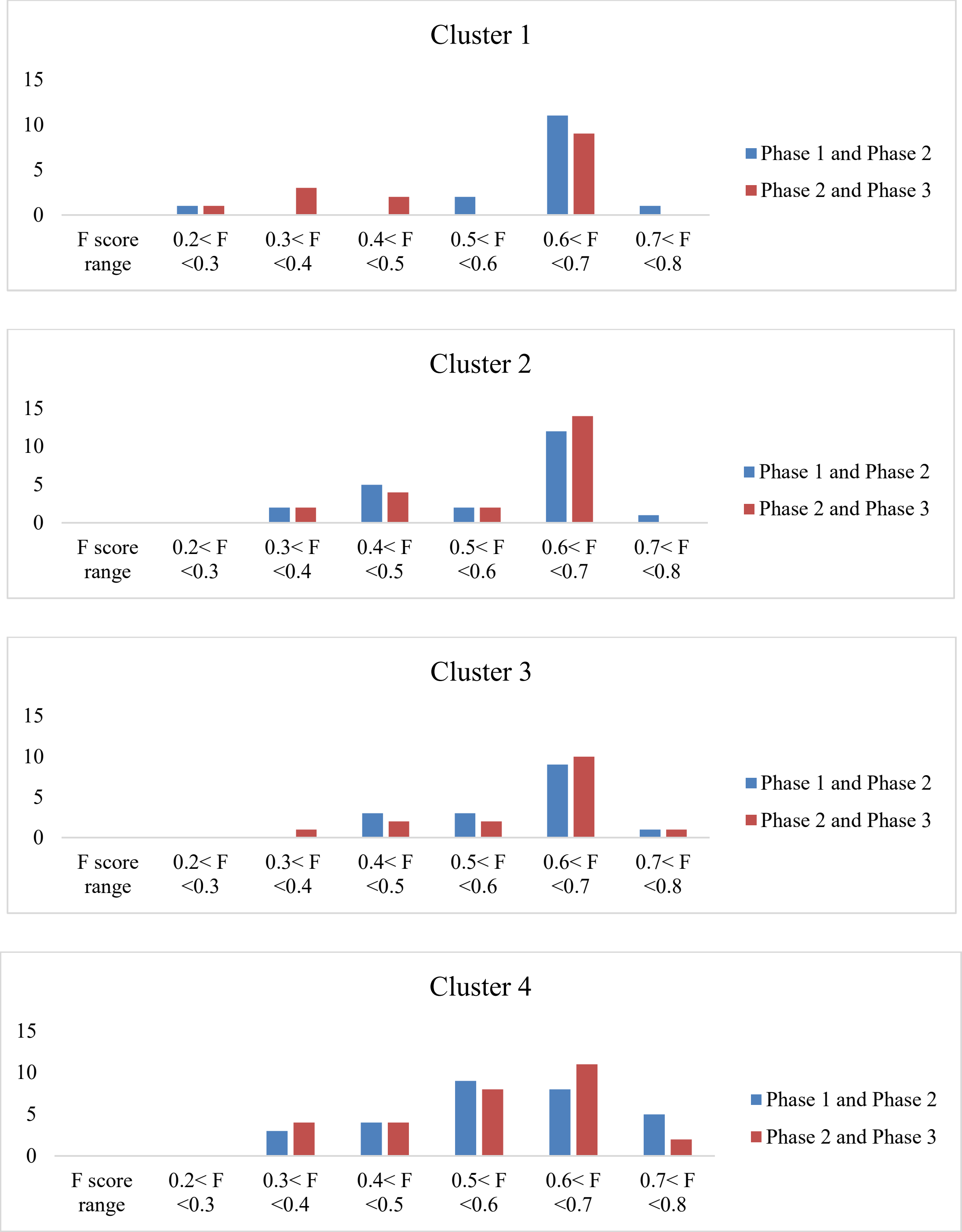


Figure 6.12: Number of statistically significant F score judged gene regulatory pathways (PTF_PDE)…Contd

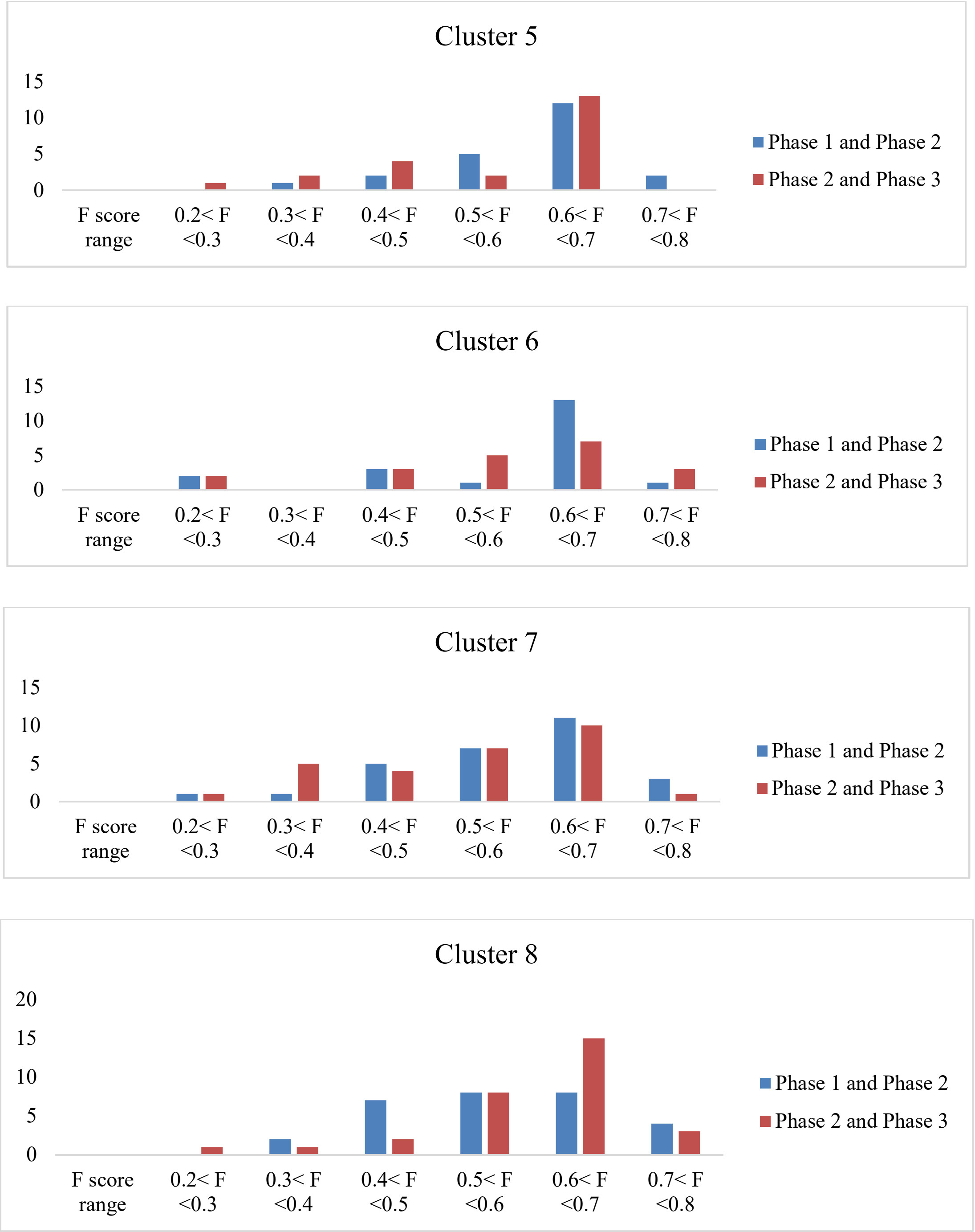


Figure 6.12: Number of statistically significant F score judged gene regulatory pathways (PTF_PDE)

From the above figures, salient observations related to the various cluster of DE genes in the aperiodic and periodic context is given as follows:

Cluster 1 and Cluster 2 for Aperiodic DE genes: On an average higher contribution of F score judged pathways are observed for the Cluster 1 DE genes with respect to aperiodic TF gene control. However for Cluster 2 DE genes, the regulatory matter remains more or less same for both aperiodic and periodic TF genes, except a better or larger contribution of periodic TF gene control with respect to higher F score (i.e. for 0.6<F<0.7 and 0.7<F<0.8) judged pathways. This implies prediction of hidden aperiodic TF genes maintaining indirect control over the target aperiodic DE genes in Cluster 1 is close to direct regulatory controls maintaining the TO-LASSO framework. Similar comment is applicable for a selected set of target aperiodic DE genes present in Cluster 2.

Cluster 1 for Periodic DE genes: The indirect regulatory contribution of aperiodic TF genes is definitely on the better side compared to periodic TF genes. In other words, the F score of the mTOM based regulatory pathways in the TO-LASSO approach, considering every target DE gene, is higher with respect to aperiodic TF gene regulation or the hidden effect of indirect aperiodic TF genes is more prominent.

Cluster 2 for Periodic DE genes: Here too, the contribution of aperiodic TF genes is having a slight edge over periodic TF genes, except for the F score segment (0.6<F<0.7) in the context of periodic TF gene control. Thus the indirect contribution of unknown TF genes (aperiodic or periodic, as the case may be; judged by the F score) can be considered close to the almost direct regulatory effect on the target DE genes exerted by the reference TF gene (aperiodic or periodic, as the case may be; judged by the F score).

Cluster 3 for Periodic DE genes: Same comments as given above for Cluster 2 is applicable in this context.

Cluster 4 for Periodic DE genes: Here, the significant observation can be related to the F score ranges, 0.6<F<0.7 and 0.7<F<0.8. In the former one, indirect aperiodic TF gene control is far more effective. However, for the latter range, it is a bit better corresponding to indirect periodic TF gene control with almost nil contribution to the F score with respect to aperiodic TF control. Accordingly, the significance of the indirect mTOM based pathways controlling the target DE genes in these two cases can be clarified.

Cluster 5 for Periodic DE genes: For the F score range, 0.5<F<0.6, the indirect contribution of aperiodic TF genes are far more prominent compared to periodic TF genes. For higher F scores, the desired content of any design, the matter is more or less similar.

Cluster 6 for Periodic DE genes: Aperiodic TF gene indirect contribution on direct regulatory links to target DE genes is prominent in the range 0.5<F<0.6 for Phase 1 and Phase 2 pair combination.

Cluster 7 for Periodic DE genes: No significant difference is observed between aperiodic and periodic TF gene controls in the indirect regulation perspective on the basis of the number of F score judged regulatory pathways.

Cluster 8 for Periodic DE genes: On an average the indirect aperiodic TF gene regulation is slightly better.

In all the above cases, the DE gene control executed by the two kinds of TF genes is being considered or compared, as the case may be. Apart from this inference, the graphical plots may signify one more important aspect. The invaluable presence of one or more TF genes guided by the mTOM structure in any regulatory pathway can be clarified by the number of such pathways present with respect to any F score. In this regard, there are some number of DE gene regulations which, via the application of TO-LASSO approach, experience far higher F score in one condition compared to the other (for example, F score in Phase 1 and Phase 2 is quite higher compared to Phase 2 and Phase 3 or just the reverse). In other words, higher F score judged regulatory pathway predicts the indispensible presence of one or more indirect TF genes because in this case the pathway in consideration is significantly close to the performance of direct DE gene regulation, as per the framework of fused LASSO that has been worked out on a topological perspective. Hence, existence of differential regulatory link can be claimed to the concerned target DE gene. In one condition, there is the confirmed presence of a direct regulatory link (here the F score of the regulatory pathways, if any, will be close to zero) and on the other there is a high chance of the presence of indirect TF gene regulations via one or more pathways.

6.2.4 Discussion: All the dataset specific results have been explored based on the performance parameter, F score. As per the definition of F score given in section 6.2.3 under Results, it is essentially a parameter reflecting the biological significance of any reconstructed GRN or TRN. In the context of YEAST 5 ON/OFF data, only 5 genes are considered in the reconstruction process. The network connectivities that get generated as an outcome of TO-LASSO application can be perfectly judged by the F score. The same comments do hold for YEAST 11 data as well where all possible connectivities are explored via TO-LASSO.

However, the YEAST Cell Cycle and HeLa time course information corresponding to the respective TF and the target DE genes, have a different process of implementation. Here, in both cases, each and every DE gene specific regulatory pathway comprising of indirect contribution from one or more TF genes is considered. Thus instead of a fully connective GRN, the matter concentrates on specific TRNs that may be defined for target DE genes. In this regard, an example can be helpful to understand the F score of any such regulatory pathway. Considering N TF genes involved in the so called regulation of a particular DE gene and the presence of biological interactive or regulatory information from various renowned databases, the F score can be computed for any mTOM generated pathway in the TO-LASSO approach. For a particular TF and DE gene pair, the matter lies in understanding the statistical similarity between almost direct and completely indirect regulation, if any, present between these entities. In the process, it can be assumed that the biological information in hand states about interactive regulations between M TF genes and the concerned TF gene referred above. In other words, among the N-1 TF genes, M TF genes are known to be associated with the referred TF gene. Now, on reconstructing a TF gene regulatory pathway following the mTOM approach in TO-LASSO, which highlights statistical similarity between direct and indirect associations between the referred TF gene and the particular DE gene, the necessary TF genes out of N-1 TF genes that may be associated indirectly, can be found. This forms a TRN representing a regulatory pathway for the specific DE gene. This reconstructed TRN can then be validated with the known set of interactions between M TF and the referred TF gene using the F score analysis.

Coming onto the precision and recall components of the F score, yielding a higher value of precision indicates greater proportion of true positive outcomes designed for reconstructing the GRN or TRN, as the case may be. In TO-LASSO, it is to be noted that higher precision comes with the benefit of predicting the obvious presence of one or more intermediary factors crucial to gene regulation statistics which normally are kept in abeyance as per biologically validated networks or databases. Again, lower level of false negative, hence higher recall, depict the fact that fully indirect connectivity design using TO-LASSO algorithm is capable of understanding the true potential of direct connectivity (as per biological validated networks) in between two genes. In other words, there is an appreciable chance of exploring the presence of any hidden causal factor responsible in the regulation process.

6.3 Conclusion

The research content of this chapter centres around GRN mining using high dimension gene throughput data manipulated using a mTOM metric proposed novel measure, TO-LASSO. Here, the design keeps in frame possible indirect gene association based hierarchical regulations, assigning equal weightage to all operational combinations in each order. The sole intention of this research work is about exploring the potential hierarchical or fully indirect regulations of a target gene from any source gene possessing statistically close regulatory performance with respect to semi-direct regulation, if any, between the source-target gene pair. The novelty lies in realizing the importance of the degree of indirect connectivity in a hierarchical model representing a biological regulatory pathway to the target gene. In this regard, the performance of the connectivity measure (in this case mTOM) between genes could help us probe further into regulation networks compared to direct regulations considering gene expression values in the LASSO based prediction framework. The connectivity measure is developed keeping at par the contribution of all possible genes according to the degree of consideration in mTOM metric. The regulatory cascade indicating indirect connectivity to any target gene plays a crucial role in determining the physical constraints required to ascertain different time delays in the regulation procedure [**3,39,40**]. In other words, there may be an intermediary presence of a physical gene entity or any unknown biological event that leads to the cascaded or hierarchical regulation possessing statistically similar causal effect to the target gene through direct interaction, if any.

6.4 References

[**1**] A. Fujita, P. Severino, K. Kojima, J.R. Sato, A.G. Patriota, and S. Miyano, “Functional clustering of time series gene expression data by Granger causality”, BMC Systems Biology, 6, Article No.137, October 2012, https://doi.org/10.1186/1752-0509-6-137

[**2**] L-Y. Lo, M-L. Wong, K-H. Lee, and K-S. Leung, “Time Delayed Causal Gene Regulatory Network Inference with Hidden Common Causes”, 10(9):e0138596, September 2015, https://doi.org/10.1371/journal.pone.0138596

[**3**] L-Y. Lo, K-S. Leung and K-H. Lee, "Inferring Time-Delayed Causal Gene Network Using Time-Series Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(5), 1169-1182, September-October 2015, https://doi.org/10.1109/tcbb.2015.2394442

[**4**] M.J. Mason, G. Fan, K. Plath, *et al.*, "Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells", BMC Genomics, 10, Article No.327, July 2009, https://doi.org/10.1186/1471-2164-10-327

[**5**] M. Bockmayr, F. Klauschen, B. Györffy, *et al.*, "New network topology approaches reveal differential correlation patterns in breast cancer", BMC Systems Biology, 7, Article No.78, August 2013, https://doi.org/10.1186/1752-0509-7-78

[**6**] X. Tu, Y. Wang, M. Zhang and J. Wu, "Using Formal Concept Analysis to Identify Negative Correlations in Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(2), 380-391, March-April 2016, https://doi.org/10.1109/tcbb.2015.2443805

[**7**] L. Song, P. Langfelder, and S. Horvath, "Comparison of coexpression measures: Mutual information, correlation, and model based indices", BMC Bioinformatics, 13, Article No.328, December 2012, https://doi.org/10.1186/1471-2105-13-328

[**8**] X. Zhang, J. Zhao, J.K. Hao, X.M. Zhao, L. Chen, "Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks", Nucleic Acids Research, 43(5):e31, March 2015, https://doi.org/10.1093/nar/gku1315

[**9**] J. Schäfer, and K. Strimmer, "An empirical bayes approach to inferring large-scale gene association networks", Bioinformatics, 21(6), 754–764, March 2005, https://doi.org/10.1093/bioinformatics/bti062

[**10**] L. Han, and J. Zhu, "Using matrix of thresholding partial correlation coefficients to infer regulatory network", Biosystems, 91(1), 158–165, January 2008, https://doi.org/10.1016/j.biosystems.2007.08.008

[**11**] S. Hempel, A. Koseska, Z. Nikoloski, and J. Kurths, "Unraveling gene regulatory networks from time-resolved gene expression data- a measures comparison study". BMC Bioinformatics, 12, Article No.292, July 2011, https://doi.org/10.1186/1471-2105-12-292

[**12**] M. Ray and W. Zhang, "Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks", BMC Systems Biology, 4, Article No.136, October 2010, https://doi.org/10.1186/1752-0509-4-136

[**13**] A. de la Fuente, "From 'differential expression' to 'differential networking' identification of dysfunctional regulatory networks in diseases", Trends in Genetics, 26(7), 326–333, July 2010, https://doi.org/10.1016/j.tig.2010.05.001

[**14**] W. Peng, M. Li, L. Chen, and L. Wang, "Predicting protein functions by using unbalanced random walk algorithm on three biological networks", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(2), 360-369, March-April 2017, https://doi.org/10.1109/TCBB.2015.2394314

[**15**] P. Creixill, E.M. Schoof, J.T. Erler, and R. Linding, “Navigating cancer network attractors for tumor-specific therapy”, Nature Biotechnology, 30(9), 842-848, September 2012, https://doi.org/10.1038/nbt.2345

[**16**] A. Cho, J.E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee, “MUFFINN: Cancer gene discovery via network analysis of somatic mutation data”, Genome Biology, 17, Article No.129, June 2016, https://doi.org/10.1186/s13059-016-0989-x

[**17**] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D.G. Huntsman, C. Caldas, S.A. Aparicio, and S.P. Shah, “DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer”, Genome Biology, 13, Article No.R124, December 2012, https://doi.org/10.1186/gb-2012-13-12-r124

[**18**] J. Song, W. Peng, and F. Wang, “An entropy based method for identifying mutually exclusive driver genes in cancer”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(3), 758-768, May-June 2020, https://doi.org/10.1109/TCBB.2019.2897931

[**19**] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer Series in Statistics, 2009

[**20**] R. Tibshirani, “Regression Shrinkage and Selection via the LASSO”, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288, 1996, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[**21**] A. Li and S. Horvarth, “Network neighborhood analysis with the multi-node topological overlap measure”, Bioinformatics, 23(2), 222-231, January 2007, https://doi.org/10.1093/bioinformatics/btl581

[**22**] N. Omranian, J. Eloundou-Mbebi, and B. Mueller-Roeber *et al.*, “Gene regulatory network inference using fused LASSO on multiple data sets”, Scientific reports, 6, Article No.20533, February 2016, https://doi.org/10.1038/srep20533

[**23**] I. Cantone *et al.*, “A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches”, Cell, 137(1), 172-181, April 2009, https://doi.org/10.1016/j.cell.2009.01.055

[**24**] P. Zoppoli, S. Morganella, and M. Ceccarelli, “TimeDelay- ARACNE: Reverse Engineering of Gene Networks from Time-Course Data by an Information Theoretic Approach”, BMC Bioinformatics, 11, Article No.154, March 2010, https://doi.org/10.1186/1471-2105-11-154

[**25**] M.L. Whitfield *et al.*, “Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors”, Molecular Biology of the Cell, 13(6), 1977-2000, June 2002, https://doi.org/10.1091/mbc.02-02-0030

[**26**] H. Chen *et al.*, “Highly sensitive inference of time-delayed gene regulation by network deconvolution”, BMC Systems Biology, 8, Article No.S6, December 2014, https://doi.org/10.1186/1752-0509-8-S4-S6

[**27**] P.A. Mundra *et al.*, “Inferring time delayed gene regulatory networks using cross-correlation and sparse regression”, In Bioinformatics Research and Applications, Part of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 64-75, 2013, https://doi.org/10.1007/978-3-642-38036-5_10

[**28**] O. ElBakry, M.O. Ahmad and M.N.S. Swamy, "Inference of Gene Regulatory Networks with Variable Time Delay from Time-Series Microarray Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(3), 671-687, May-June 2013, https://doi.org/10.1109/tcbb.2013.73

[**29**] A.C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped Graphical Granger Modeling for Gene Expression Regulatory Network Discovery", Bioinformatics, 25(12), i110-i118, June 2009, https://doi.org/10.1093/bioinformatics/btp199

[**30**] W-S. Wu *et al.*, "YAGM: a web tool for mining associated genes in yeast based on diverse biological associations", BMC Systems Biology, 9, Article No.S1, 2015, https://doi.org/10.1186/1752-0509-9-S6-S1

[**31**] J.A. Choi, and J.J. Wyrick, "RegulatorDB: a resource for the analysis of yeast transcriptional regulation", Database (Oxford), 2017:bax058, August 2017, https://doi.org/10.1093/database/bax058

[**32**] P.T. Monteiro *et al.*, "YEASTRACT-DISCOVERER new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae", Nucleic Acids Research, 36 (Database issue): D132-136, January 2008, https://doi.org/10.1093/nar/gkm976

[**33**] J. Yang *et al.*, "DCGLv2.0: An R package for Unveiling Differential Regulation from Differential Co-expression", PLoS One, 8(11):e79729, November 2013, https://doi.org/10.1371/journal.pone.0079729

[**34**] H. Han *et al.*, "TRRUST: a reference database of human transcriptional regulatory interactions", Scientific reports, 5, Article No.11432, June 2015, https://doi.org/10.1038/srep11432

[**35**] J. Friedman, T. Hastie and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent", Journal of Statistical Software, 33(1), 1-22, February 2010, https://doi.org/10.18637/jss.v033.i01

[**36**] J. J. Hsieh *et al.*, "CIR, a corepressor linking the DNA binding factor CBF1 to the histone deacetylase complex", PNAS, 96(1), 23-28, January 1999, https://doi.org/10.1073/pnas.96.1.23

[**37**] A. Majumder *et al.*, "A Composite Entropy Model in a Multiobjective Framework for Gene Regulatory Networks", Current Bioinformatics, 13(1), 85-94, 2018, http://dx.doi.org/10.2174/1574893611666161202104422

[**38**] E. Dimitriado *et al.*, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly:E1071), TU Wien", Version 1.7-13, URL-https://rdrr.io/rforge/e1071/

[**39**] B. Yang *et al.*, "HSCVFNT: Inference of Time-Delayed Gene Regulatory Network Based on Complex-Valued Flexible Neural Tree Model", International Journal of Molecular Sciences, 19(10), 3178, October 2018, https://doi.org/10.3390/ijms19103178

[**40**] M.M. Kordmahalleh *et al.*, "Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network", BioData Mining, 10, Article No. 29, August 2017, https://doi.org/10.1186/s13040-017-0146-4

Chapter 7 Conclusion and Future Work

7.1 Conclusion

The research works presented in this thesis report are in line with differential gene regulatory network design. Design and significant reconstruction of such regulatory networks can be important in deciding the therapeutic targets, and hence contributing to the course of development of regulation networks that may be beneficial in rectifying the states of the concerned targets. In the process, the differential perspective can be fully topological based, i.e. the different kinds of interaction present among the concerned genes under varied conditions of interest, or based on the kind of regulations a differential gene may get involved, or a combination of the two above.

In this regard, at the initial part of this thesis the statistical property of the differentially expressed (DE) genes are been looked into, followed by devising biologically significant interactive structures between these DE genes (can be considered to be a subset of standard gene regulatory networks or GRNs) maintaining differential topology. As the transcription factor (TF) genes are the primary motivating entities for the DE genes, the suitable transcriptional regulatory networks (TRNs) and the development of the same has also been analyzed with statistically and biological significant regulatory outcomes. In the context of development of these DE specific GRNs or TRNs, analysis has been done utilizing static (independent and identically distributed gene expression profiles) and time course (highly auto-correlated time series gene expression profiles) information obtained from Microarray and/or RNA-seq experiments. The static profiles are used to reconstruct the statistically and biologically significant differential network structure under varied conditions of any living cell; thus helping in understanding the different states or types of any disease or different phenotypic versions of genetically related organisms. However, the time course profiles are more into understanding the pattern of differential connectivity between genes with respect to time. In this context, first order (no time delay) and higher order GRNs and/or TRNs is mostly time invariant in nature disregarding the specific contributory presence of certain activation time instants at different conditions of any living cell. Hence, time variant TF to DE gene architectures in single and collaborative regulations can help to understand the differential course of development of any disease between varied conditions, like having an idea on the differential development of cancer progression between different stages. Be a time dependent dynamic or static differential design, the importance in understanding the crucial or inevitable role of unknown gene factors in target DE gene regulations cannot be neglected in the process of development of these networks. In this regard, concerned network regulatory pathways can help to judge such external unknown or hidden factors that may work in cohesion with a protein forming gene complex towards regulating therapeutic DE gene targets.

7.2 Future Scope of Research

Differential regulatory network design based on time variant regulatory property of specific activation time instants across conditions can be explored further incorporating the concept of hierarchical or cascaded regulatory pathways. The prime reason to indulge in this network development lies in finding the necessary delays in time variant differential regulations between gene pairs which can be extended through cascaded analysis involving more gene pairs in succession. This thought can also be appended with the inclusion of unknown gene factors or hidden factors of regulation at any stage of the cascaded pathway following the time variant differential regulatory property.

This humongous domain of exploration can be addressed directly utilizing the available time course expression data with necessary incorporation of predicted gene information corresponding to hidden factors, wherever applicable. In this perspective, some kind differential wavelet analysis based on the phase difference may also be informative enough to apprehend the nature and the time delay of the interactive regulatory mechanisms. At the very end, it is all about understanding the rate of recovery of any diseased cell through therapeutic drug monitoring.

Aurpan Majumder