AUTHOR: **Ritesh Krishna**

DEGREE: .....

TITLE: **From Gene-Expressions to Pathways**

DATE OF DEPOSIT: .....

I **agree** that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I **agree** that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

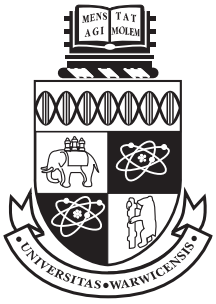
> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE	SIGNATURE	ADDRESS
.....	.....	.....
.....	.....	.....
.....	.....	.....
.....	.....	.....
.....	.....	.....



From Gene-Expressions to Pathways

by

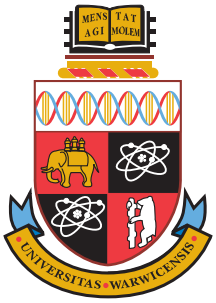
Ritesh Krishna

Thesis
Submitted to the University of Warwick
for the degree of
Doctor of Philosophy

Supervisors: Dr. C-T Li and Prof. J.F. Feng

Department of Computer Science
2009

THE UNIVERSITY OF
WARWICK



From Gene-Expressions to Pathways

by

Ritesh Krishna

Thesis
Submitted to the University of Warwick
for the degree of
Doctor of Philosophy

Supervisors: Dr. C-T Li and Prof. J.F. Feng

Department of Computer Science
2009

THE UNIVERSITY OF
WARWICK

In loving memories of my grandparents

Contents

List of Figures

Acknowledgements

Writing one's PhD research in a comprehensive document like this, rigorously and repeatedly tests the claim made by the famous American essayist Ralph W. Emerson: "sometimes a scream is better than a thesis". One can not bring the endeavor of PhD research to this stage and maintain the sanity without the precious support of supervisors, friends, and family.

The work reported in this thesis could never have been completed without the precious support of my supervisors Prof. Jianfeng Feng and Dr. Chang-Tsun Li. I am thankful to both of them not only to guide me throughout my research, but also to give me the confidence that they are ever present, and nothing could go wrong. More importantly, I am indebted to both of them to help me in the time when I really needed the help, both professionally and personally. I will always remember those gestures fondly and gratefully. I am also grateful to Dr. Vicky Buchanan-Wollaston for providing the biological data used in this study and explaining many valuable concepts. Very special thanks to Prof. Dongyun Yi for many stimulating discussions and contribution to my research.

I am thankful to the staff members of department of computer science for ensuring that the students face the least administrative problems during their research. In particular, I am grateful to our head of department Prof. Roland Wilson for his support to the research students.

A huge thanks to my friends in Oxford, Dr. Amol Patil and Mrs. Monica Mantri, for sheltering me in their home during the whole period of the thesis writing. The homely comforts and their warmth continuously challenged the above quoted claim by Mr. Emerson. This document would have taken much longer if they had not stepped in time to pick me up and bring to their lovely nest. My friends Ashutosh Trivedi and Tiziana Faveretto also deserve a grateful acknowledgment for the similar reasons. One can feel incredibly lucky to have two houses in Oxford, while still being a research student, that too the one who is writing his PhD thesis !!

I would like to thank my friends Rajesh Balakrishnan, Nikolaos Papanikolaou, Anil Sorathiya, Dr. Estelle Guyez, Marie Clucas, Michal Rutkowski, Agnieszka Rutkowska, Daniel Alejandro Valdes Amaro, Antony Holmes, Dr. Ashutosh, Jothi Philip, Tina Thomas and Mirela Domijan for continuously providing their support and encouragement throughout my PhD. I am also thankful to many of my other friends and colleagues at University of Warwick to make my life easier and enjoyable.

Last, but not the least, my sincere thanks to my ever supporting parents and my younger brother for many beautiful things in life.

Declaration

This thesis is presented in accordance with the regulations for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work in this thesis has been undertaken by myself under the joint supervision of Dr. C-T. Li and Prof. Jianfeng Feng.

Although the key findings of the thesis have been announced in an international journal and proceedings of some international conferences, this thesis presents the results with more details and complements them with illustrations. Chapter 2 and Chapter 5 are the extended versions of the paper "Listen to Genes: Dealing with Microarray Data in the Frequency Domain" [FYK+09], which appeared in the journal PLoS ONE 2009. Chapter 3 is an extended version of the paper "Interaction Based Functional Clustering of Genomic Data" [KLBW09], which appeared in the proceedings of the IEEE International Conference on Bioinformatics and Bioengineering(BIBE) 2009. Chapter 4 is an extended version of the paper "A partial Granger causality approach to explore causal networks derived from multi-parameter data" [KG08], which appeared in Lecture Notes in Computer Science (LCNS) 2008. Two more manuscripts have been prepared from the materials in Chapter 3 and Chapter 4 in this thesis and have been submitted to international journals.

Abstract

Rapid advancements in experimental techniques have benefited molecular biology in many ways. The experiments once considered impossible due to the lack of resources can now be performed with relative ease in an acceptable time-span; monitoring simultaneous expressions of thousands of genes at a given time point is one of them. Microarray technology is the most popular method in biological sciences to observe the simultaneous expression levels of a large number of genes. The large amount of data produced by a microarray experiment requires considerable computational analysis before some biologically meaningful hypothesis can be drawn. In contrast to a single time-point microarray experiment, the temporal microarray experiments enable us to understand the dynamics of the underlying system. Such information, if properly utilized, can provide vital clues about the structure and functioning of the system under study. This dissertation introduces some new computational techniques to process temporal microarray data. We focus on three broad stages of microarray data analysis - normalization, clustering and inference of gene-regulatory networks. We explain our methods using various synthesized datasets and a real biological dataset, produced in-house, to monitor the leaf senescence process in Arabidopsis thaliana.

Chapter 1
Introduction

1.1 Introduction

Over the last ten years or so genome sequencing has made rapid progress. Genome sequencing has facilitated transfer of information from DNA of a species to electronic computers. Identification and symbolic representation of correct genes are only the preliminary goals of genome sequencing, the holy grail of biology lies in understanding the functions of those genes. This has given birth to a new research field known as *functional genomics*.

Much of the success in genome sequencing can be accredited to high throughput DNA sequencing techniques. This led to what primarily used to be a wet science to become in larger part an information science [Qua07]. Similar high throughput techniques have been developed for functional genomics also. Most notable among them are DNA microarray technologies. Microarrays allow researchers to monitor simultaneous gene expression levels of thousands of genes in an organism in a single experiment. On one hand, advancing experimental techniques are producing tons of data which can provide clues about the functions of genes, on the other hand, it is becoming more complicated to extract mean-

ingful information from that data and build biological hypotheses which can be tested in laboratories. The development of computational methods and tools to analyse such massive data is the task of computational biology and bioinformatics.

This dissertation addresses some of the challenges found in analysis of microarray data and provides techniques to address them. We provide a complete pipeline for dealing with three broad stages in microarray data analysis, namely, normalization, clustering and inference of gene regulatory networks. Our goal is to automatically infer the meaningful signals using statistical techniques and build plausible biological hypotheses for further testing in the laboratory.

1.2 The Central Dogma of Molecular Biology

To understand how gene expression works, we need to understand the *Central Dogma* of biology. Cells are fundamental working units of every living organism. Cells are largely made of proteins which define their shapes and structures. Proteins are functional molecules essential for performing many life functions like catalysis, signalling etc. The central dogma of biology charts out the flow of information from DNA molecules to proteins. DNA is a stable molecule containing the complete genetic blueprint of living organisms. The information in DNA is stored as a code made up of four chemical bases known as nucleotides : adenine (A), guanine (G), cytosine (C), and thymine (T). Segments of DNA known as *genes* are *transcribed* into messenger RNA(mRNA) which are subsequently *translated* into proteins. This complete process is known as *gene expression*.

Figure 1.1 presents a pictorial representation of the stages involved in the central dogma. There are three broad stages in the inheritance of genetic information and its conversion from one form to another. In the first stage of *replication*, a

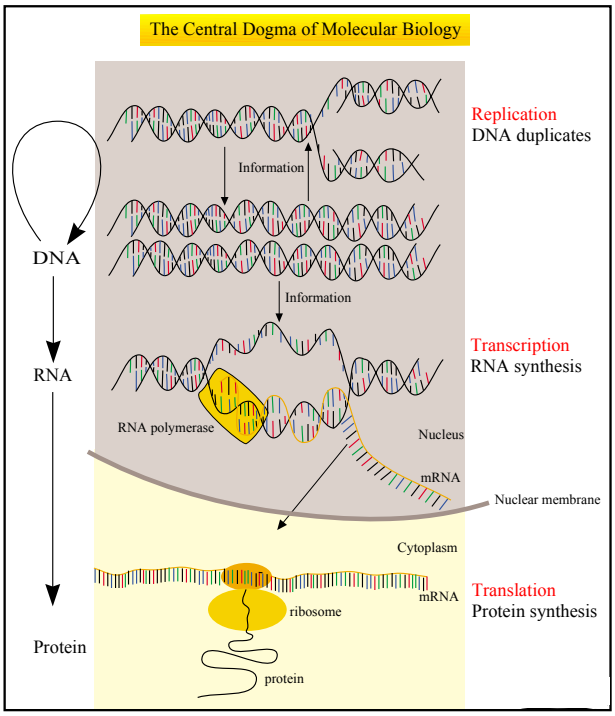


Figure 1.1: The central dogma of molecular biology. Figure from (Vie99).

DNA creates identical copies of itself and the genetic information is replicated. A segment of DNA called a *gene* contains both coding sequences that determine the function of the gene, and non-coding sequences that determine when the gene is active (expressed). When a gene is active, the second stage in the central dogma named as *transcription* takes place, where the coding and non-coding sequences of the gene are copied to produce a single stranded RNA copy of the gene's information. The RNA moves from the nucleus into the cytoplasm where the ribosomes are located. There are multiple types of RNAs in nature, but the one responsible for protein coding is known as messenger RNA or mRNA. The mRNAs are surrounded by ribosomes. In the final stage of the central dogma known as *trans-*

lation, the mRNA sequence is translated into a sequence of amino acids as the protein is formed. The translation of mRNA to protein is performed by ribosomes which read three bases (a codon) at a time in the mRNA sequence and translate them into one amino acid according to the rules specified by genetic code.

A key consideration is that all the cells in an organism's body contain the same copy of DNA molecule, yet all the cells are not same. The diversity is due to the difference in gene expressions across different cell types. Different gene subsets which eventually lead to different proteins synthesized, express themselves in different ways by reflecting both the cell types and their conditions. Microarrays quantify the gene expressions by monitoring the abundance of mRNA molecules during the transcription stage. The amount of each mRNA detected in the cell can provide information on the level of expression for the corresponding gene.

1.3 How Do Microarrays Work

Microarray technology is based on the principle of *DNA hybridization*, a process in which DNA strands bind to their unique complementary strands. A DNA molecule consists of two complementary strands, each strand containing the information to describe the other (adenine(A) bonds only to thymine(T), and cytosine(C) bonds only to guanine(G)). A microarray is typically a solid surface of either glass or silicon chip. A set of specific DNA oligonucleotides, or cDNA, or small fragments of PCR (polymerase chain reaction) products corresponding to mRNAs (these all are collectively called *known sequences*) are attached on the solid surface at fixed locations using covalent bonds. The choice of oligonucleotides, cDNA, or PCR products depends on the manufacturers of the arrays. The immobilized known sequences are also called *probes*. The fluorescently tagged targets (*unknown sequences*) bind by hybridization to the probes on the array with which they share

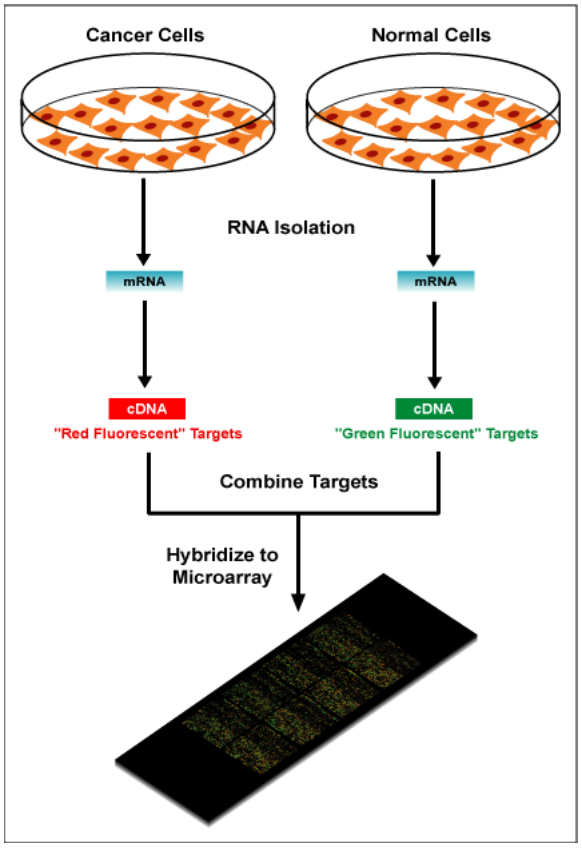


Figure 1.2: Overview of a typical microarray experiment with two samples. Figure obtained from (Com09).

significant sequence complementarity. After allowing sufficient time for the hybridization to take place, the excess sample is washed off the solid surface. The binding affinity of each probe with the labelled target reflects the proportion of the expression of the gene represented by that probe. Hybridized microarrays are excited by a laser and scanned at suitable wavelength for detection of fluorescent dyes to estimate the amount of intensity bound to each probe. The intensity measured at each probe is an indicator of the expression level of the gene on that array, which after adjustment for technical artefacts, should provide an estimate of the level of gene expression which can be used for further analysis.

Microarrays can be used to measure gene expressions in different ways. One of the most popular approaches is to compare gene expression levels in two different

samples representing the same cells or cell types under two different conditions. In this case, the mRNA extracted from each sample is labelled differently, for instance, a green label (using the fluorescent dye Cy3 having the fluorescence emission wavelength of approximately 570 nm) for the sample from condition 1, and a red label (using the fluorescent dye Cy5 having the fluorescence emission wavelength of approximately 670 nm) for the sample from condition 2. Both the Cy-labelled samples are mixed and hybridized on a single array. The hybridized array is scanned at the suitable wavelengths to produce separate expression profiles for Cy3 and Cy5 tags. Figure 1.2 shows steps in a typical microarray experiment involving two sample types. Such experiments typically rely on the ratio based analysis of relative intensities of samples to identify up-regulated or down-regulated genes.

Another popular variation of microarray experiment involves hybridization of single-labelled population of samples to each array. In this case, the experiments give estimations of the absolute level of gene expressions. If we want to compare gene-expressions for two different samples, we need to perform two separate single-label experiments and compare the absolute gene-expression levels. In this case, comparisons are primarily made between the data obtained from different arrays, as opposed to between the labelled populations hybridized to a single array. The experiments with two-colour labels are also known as two-channel experiments. On the same lines, the single-colour experiments are known as single-channel experiments. An advantage of single-channel experiments over two-channel experiments is that the absolute value of gene expressions may be easily compared between studies from different experiments conducted months or years apart. At the same time, it is possible to treat a two-channel experiment as a single-channel experiment by taking each channel intensity as the absolute expression level rather

than relying on the intensity ratios of spots between samples.

Microarrays are useful in a wide variety of studies to achieve wide variety of objectives. The objectives can be broadly divided into four categories -

1. Class comparison - Involves comparison of gene expression profiles among samples selected from predefined classes to identify the differentially expressed genes,

2. Class prediction - Similar as class comparison, but requires building a statistical model to predict the class of a new specimen based on its expression profile,

3. Class discovery - Involves the identification of novel subtypes of specimens within a population. In context of drug discovery, class discovery methods can be used to find putative (sub-)types of diseases and to identify informative subsets of genes with disease-type specific expression profile,

4. Pathway analysis - Involves identification of co-regulated genes, or the ones which belong to the same biochemical pathway.

Microarrays were first used to study global gene expressions in Saccharomyces cerevisiae in 1997 by DeRisi et al. [DIB97]. A genome-wide measurement of transcription is called an *expression profile* and provides us with a complete list of genes whose transcription level is affected in a given condition. In a biological sense, what we measure is how the gene expression of each gene changes to perform certain coordinated tasks.

1.4 Time Courses vs. Independent Data Point

There are two types of microarray datasets : time-independent (or single point steady-state), and time-series(time dependent) datasets. The majority of the microarray experiments are carried out for pair-wise comparison between different samples at a single time-point. It is relatively easy to compare and contrast single-point datasets belonging to different experimental conditions to identify the sets of differentially expressed genes across conditions. However, to verify that the obtained results are reliable and robust to variations in the experimental procedure, it is necessary to repeat a given experiment several times independently. Thus, a successful comparison for time-independent datasets requires several independent repetitions of the experiment in which the different conditions are tested in parallel. In general, the time-independent gene expression profiles are capable of recovering steady-state behaviour of the system, but fail to recover the temporal regulating relationships.

Time course experiments, on the other hand, can improve the inference greatly in contrast to time-independent data sets [ZSD06]. Time course experiments have been proven to be useful in a number of experimental systems, providing information about the difference in each transcript over different time points, reflecting information about the order of events and their trends. Another main advantage of the time-course experiment is that samples for a given experiment are all derived from a single relatively homogeneous population, making the results much less sensitive to the population-specific effects or the slight differences in the experimental or biological background. The time points at which mRNA samples are taken are usually determined by the investigator's judgement concerning the biological events of interest.

Time course experiments can be further classified into two categories : *periodic* and *developmental*. Periodic time-courses include natural biological processes whose temporal profiles follow regular patterns. Examples are cell cycles [SSZ+98], circadian rhythm [HHS+00, CCWN+01] etc. In developmental time course experiments, gene expression levels are measured at successive times, depending on the timing of phenomena of interest, during the developing phase, for example , natural growth or decay [TBW+02, HVV+04] in a cell type. Such experiments are also useful in understanding effect of controlled stimulus in a given system, for example, the effect of drug treatment on a cell type over a period of time.

1.5 Data Generation and Processing

Collection and analysis of data in any high-throughput experiment like microarray can be performed in two major stages :

1. Material processing and data collection stage, and
2. Information processing stage

The material processing and data collection stage is concerned with the lab based activities related to the biological samples and experimental instruments, and can be further broken down into the following steps -

1. preparation of biological samples to be studied;
2. extraction of RNAs from the samples;
3. labelling of RNAs using fluorescent dyes;
4. hybridization of labelled RNA extracts to the array;

5. excitation of arrays by laser at suitable wavelength to detect the hybridization intensity;

6. scanning the hybridized array to produce image files.

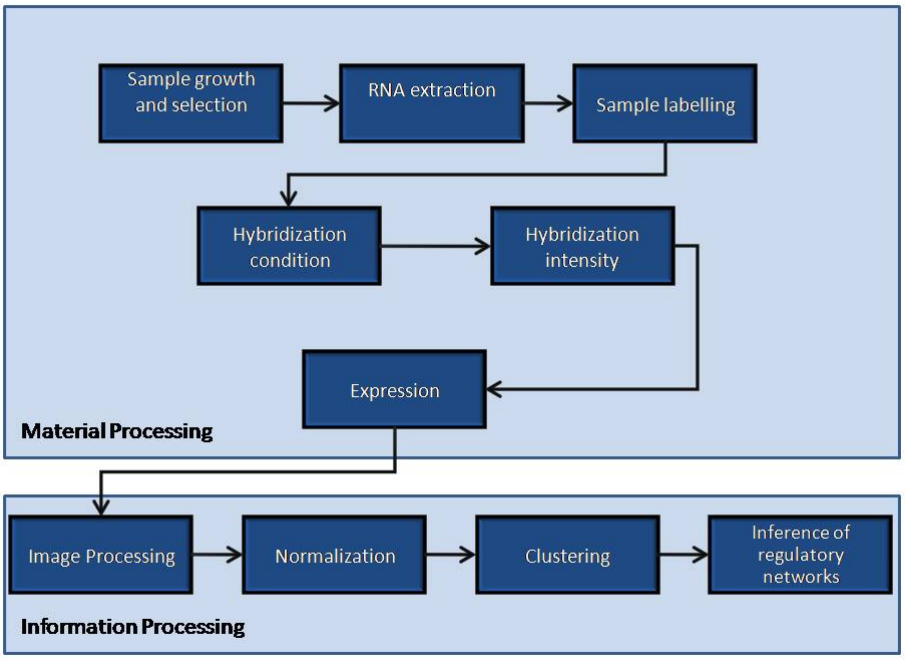


Figure 1.3: Processing pipeline for a typical microarray experiment.

The information processing stage is essentially computer based analysis of the data and can be broken down into four distinct steps -

1. image processing of the scanned images to extract gene expression levels;

2. normalization of gene expression values;

3. clustering of genes;

4. inference of regulatory networks.

A systematic diagram for the order of steps in each stage can be seen in Figure 1.3. The last three steps in the information processing stage are the focus

of this thesis. The steps need to be performed in a sequential order as shown in the pipeline in Figure 1.3 before some meaningful hypothesis from data can be derived. We present a brief description of the steps in the information processing stage -

- Image processing - The digital images obtained from microarray experiments need to be analysed in order to gain information about gene expression levels. Each spot on the array is identified and its intensity is measured and compared with the background. Image quantification is usually performed by image processing software which sometimes are provided by microarray manufacturers. The image quantification software extract the data from digitalized images and combine in a table commonly known as the *image quantization matrix*. Each row represents one spot on the array, and each column represents different quantitative characteristics like mean or median pixel intensity for that spot. Image quantification for microarray experiments is not a trivial task and can be regarded as an area for experts.

- Normalization - The data from multiple hybridizations or different arrays in the *image quantization matrix* must be further analysed and should be combined into a *gene expression matrix*. In the *gene expression matrix*, each row represents a gene and each column represents a particular biological sample or experimental condition. The combination of information from *image quantization matrix* to *gene expression matrix* is not a trivial task. There are several experimental artefacts which must be taken into account while doing the conversion. There are many biological and experimental variations which can affect the expression level of each spot. The biological samples and the experimental conditions may differ across different arrays, and obtaining a single value for each gene can require considerable attention. The data must be normalized to handle the variations in the experiments and the

values across spots. There are many methods for microarray normalization but there is no standard fitting all the cases. The data normalization is very much dependent on the platform, experimental setup and practitioner's hypothesis. Once the data has been normalized properly, we use values in the *gene expression matrix* for further analysis and data mining.

- Cluster Analysis - The next step after data normalization is to group the genes based on certain features which help in reduction of data dimension. The goal of clustering techniques is to discover the underlying gene pathways representing the biological processes. Genes lying in the same pathway are often activated or depressed simultaneously or sequentially upon receiving stimuli. Clustering can help in recognizing biologically relevant patterns among genes. The importance of clustering is more apparent while dealing with a large number of genes. Automated grouping of genes in several clusters on the basis of structural or functional similarity can substantially help in recognition of genes of interest, and thus, can reduce the amount of data to be analysed further.

- Inference of regulatory networks - The ultimate goal of microarray experiments is to understand the interactions among genes. Understanding the interactions can help unlock the functioning and behaviour of genes leading to development of potential therapeutic targets and drug discovery. Gene regulatory networks are indicators of networks among genes and are concerned with the control of transcription i.e., how genes are up or down regulated with respect to different signals. Inference or reverse-engineering of gene regulatory networks from data is a step downstream to cluster analysis in the microarray information processing pipeline. Reverse-engineering refers to an approach where one tries to design a model that fits the data. The

choice of computational methods for creating the models depends crucially on the kind of modelling techniques used. The models can produce further hypothesis which can be verified in additional laboratory experiments.

1.6 Overview of the Arabidopsis Experiment

Much of the work in this dissertation is explained and tested with the microarray data produced by scientists at Warwick Horticulture Research group at the University of Warwick, UK. The experiment was performed to study the process of senescence in leaves of *Arabidopsis thaliana* over a period of time.

Arabidopsis thaliana (also known as thale cress, mouse-ear cress or Arabidopsis) is a small garden weed type flowering plant. Although not an economically important plant, Arabidopsis has become popular as a model organism in plant biology due to its genome being one of the smallest among other plant genomes. It was also the first plant genome to be sequenced. Arabidopsis has several traits that make it a useful model for understanding the genetic, cellular, and molecular biology of flowering plants. Arabidopsis has a short life cycle of about 8-10 weeks and it can grow about 50 cm in height in as little as $1cm^3$ of soil. The small size and the rapid life cycle of the plant are advantageous for research. It can be grown in a small space and it produces many seeds. Each of these traits leads to Arabidopsis being a model plant organism for plant biologists.

The primary goal of the project undertaking the biological experiment explained in the following sections was to understand the senescence process in leaves of Arabidopsis. Senescence is a term for the collective process that leads to the ageing and death of a plant or a plant part, like a leaf. In the case of animals, ageing and senescence are used interchangeably, but, in case of plants,

Figure 1.4: Arabidopsis plant.

senescence is well differentiated from ageing which is a passive time-dependent degenerative process. Senescence in a plant, on the other hand, is an internally regulated developmental process based on an adoptive mechanism, and the death is its consequence. The basic molecular mechanism of senescence both in plant and animal systems may be the same. Senescence can take place due to natural reasons, or due to environmental stress factors. The process involves expression of specific genes. As for example, plants undergo the process of leaf senescence to prepare for winter and recycle some of the valuable and scarce mineral nutrients. Leaf senescence is also a mechanism to get rid of old and photosynthetically less efficient leaves in the evergreen plants.

In Arabidopsis, leaf senescence is a programmed cell event responding to wide range of external and internal signals. The leaf senescence in Arabidopsis is controlled by age in a predictable manner. Each individual leaf has a similar lifespan

and therefore, leaves that develop later in life, will senesce later. In addition to age, plant hormones and environmental conditions can modulate the progression of leaf senescence [Sma94, Pes05]. The process, however, is not only concerned with death alone, but involves several events associated with massive mobilization of nutrients in a highly ordered and regulated manner from senescing leaves to new leaves, seeds and buds, thus contributing to the nutrient cycling. Many different genes show enhanced expressions during senescence process, and can help elucidate the underlying signalling pathways. Identification of the key genes and pathways can result in understanding the mechanisms that occur during the senescence process. Although, the leaf senescence alone can not explain the senescence process in the whole plant, but can provide vital clues for understanding senescence as a whole process.

1.6.1 Material Processing and Data Collection

The experiment was performed over 40 days with the following steps involved in the material processing and data collection stage.

- *Plant growth and leaf material acquisition*: Arabidopsis plants (Columbia seed type also known as COL-0) were grown in a controlled environment at 20^oC temperature , 70% relative humidity and 250μ mol $m^{-2}s^{-1}$ light intensity. The plants were subjected to long days with 16 hours of sunlight. The seventh leaf (leaf 7) on its emergence during the development of each plant was tagged with a cotton around it. Figure 1.5 (a) shows a cotton tagged leaf. The cotton tags would act as identifiers later in the experiment. Four such leaves were selected for harvesting purposes. After 19 days from sowing, when the leaf 7 was fully developed, it indicated the beginning of the time course. The biological replicates were harvested both in morning and evening (7h and 14h into light period) on every other day for next 22

days. Figure 1.5(b) shows the plant growth on day 1,15 and 19 since the data collection started. Figure 1.6 shows the development of leaf 7 from fully developed until fully senescent. This resulted in total 22 time point samples for each leaf.

Figure 1.5: (a) Cotton tag around leaf 7 (b) The plant images on day 1, 15 and 19 since the data collection started. Leaf 7 is marked with an arrow.

Figure 1.6: Profiles of a leaf over 22 days during the senescence. The left most (first) profile shows a fully developed leaf and the profile was taken 19 days after sowing the plant.

- *RNA isolation and probe preparation*: RNA was isolated from 4 individual leaves as separate biological replicates using the Triazol method (Invitrogen) followed by RNeasy column purification (Qiagen). RNA was amplified using a MessageAmp II (Ambion) and then labelled with Cy3 or Cy5 using reverse transcriptase (SuperScript II, Invitrogen). Each amplified RNA sample was labelled twice with Cy3 and twice with Cy5 giving 4 technical replicates for each leaf sample. Two Cy3 and Cy5 labelled samples (in 25% formamide, 5x SSC, 0.1% SDS and 0.5 mg ml^{-1} yeast tRNA) were mixed in different combinations for hybridization to microarray slides.

- *Hybridization*: The microarrays used for analysis of the samples were Complete Arabidopsis Transcriptome Micro Arrays (CATMA). Each array contains 30,336 gene probes belonging to the genome of Arabidopsis. These arrays are produced by Warwick HRI using a sterile spotting machine. The description of the machine and the array can be found in [LKB+07]. These arrays were hybridized with labelled samples at 42°C overnight. Slides were washed and then scanned using an Affymetrix 428 array scanner at 532nm (Cy3) and 635nm (Cy5).

1.6.2 Information Processing

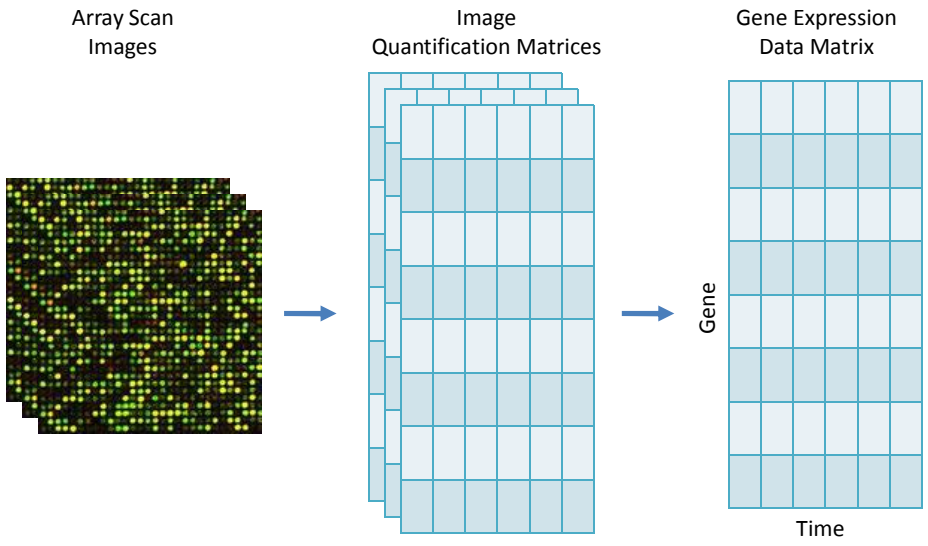


Figure 1.7: Scanned images are read using Imagene to produce text files with details of signal intensity and other statistics for each gene in the image file. The text files are read to produce quantization matrices after adjusting the gene intensity. The quantization matrices are further combined using normalization method to produce the final gene expression matrix for further data analysis.

The first step of the information processing stage, i.e. the image analysis

was performed using Imagene version 7 software (BioDiscovery, `http://www.biodiscovery.com/`). Figure 1.7 presents a schematic diagram of obtaining a final gene expression matrix from scanned image files. We have one scanned digital image file (.tiff format) for each replicate at each time-point. The image files were read using the Imagene software to produce text files with signal intensity values for the genes, along with other statistics like background mean, median etc. The quantified values for all the replicates were further adjusted and combined to produce a final gene expression matrix. The process to produce the gene expression matrix from quantification matrices is called *normalization*. The expression values in the final gene expression matrix will be used for the remaining stages in the information processing pipeline. The normalization step to produce the final gene expression matrix along with the other steps in the information processing stage are explained in greater details in later chapters.

1.7 Road-map of the Dissertation

This chapter presented an overview of the microarray technology and explained the experiment performed to understand the process of senescence in Arabidopsis leaves. The main focus of this dissertation is on the development of statistical techniques for the last three steps in the information processing stage in a microarray experiment, last three steps being normalization, clustering, and inference of regulatory networks.

The immediate step after the image analysis in the information processing stage of microarray is to clean the data from unwanted experimental variations, and combine the expressions collected from different arrays to produce a single gene expression matrix. Chapter 2 presents a normalization method to deal with different sources of experimental biases in the data. The data is normalized us-

ing a statistical error model and different sources of experimental variations are estimated. The method further uses an iterative algorithm to minimize the correlation among the residual terms across replicates.

The normalized data is clustered according to the temporal profiles of genes in Chapter 3. Clustering assigns genes into groups based on their related expression patterns; such groups contain functionally related genes or genes that are co-regulated. We adopt a Granger causality based temporal-precedence technique to cluster the Arabidopsis genes. The association graph showing the connectivity between various genes is analysed using a graph-theoretic method to detect dense regions in the graph. The dense regions could be indicators of the biologically relevant complexes. The genes in the subgraphs representing the dense regions are queried against publicly available Gene Ontology database to test for the functional similarities between genes.

In Chapter 4, we present a reverse-engineering technique to infer gene circuits from temporal microarray data. We extend the Granger causality technique presented in Chapter 3 to infer directional network structures from gene data in a multivariate context. We name this technique *Partial Granger Causality (PGC)*. Partial Granger Causality is tested using various artificial datasets representing different scenarios of connections among the participating entities. Partial Granger Causality is further applied to a publicly available dataset for Human T-cell activation. We apply Partial Granger Causality to infer gene circuits from the Arabidopsis data in chapter 5.

Chapter 5 brings the techniques discussed in previous chapters together and presents a complete pipeline for analysis of the Arabidopsis data; starting from

normalization to the inference of selected gene circuits. Chapter 5 also introduces a frequency based analysis for the Arabidopsis dataset to detect interesting patterns in the frequency domain. The Partial Granger Causality discussed in chapter 4 is extended to infer interactions between sets of genes, and the extended technique is called *Complex Granger Causality*. Three gene circuits, namely, circadian, ethylene, and a global gene circuit are inferred using Partial and Complex Granger Causality.

In the last chapter, we summarize the work done in each chapter. We conclude with a discussion on the possibilities of improvements in the proposed methods and some open questions.

Chapter 2

Normalization of Gene Expression Data

Microarray technologies provide a powerful mechanism to simultaneously detect and measure expression levels for tens of thousands of genes in a single experiment. The vast quantity of data if suitably analysed can help in understanding cellular processes, diagnosis of diseases and development of potential therapeutic targets. The effective analysis of the experimental results relies on the good quality of data. Experimental variations such as design of arrays, mRNA quality, labelling and dye effects, hybridization conditions, human and machine errors in the scanning process contribute towards obscuring variations found in a Microarray data. To overcome these obscure variations, and make the observations from different arrays comparable, an effective normalization technique is required.

The beginning point of any normalization method is reading the digital image files and forming the *image quantization matrices*. The gene-expression values in the matrices then need to be adjusted to produce cleaner data for further analysis. Image processing softwares like GeneSpring [gen09], Imagene [ima09] etc. analyze image files and generate a number of statistics for each gene intensity. The typ-

ical measurements reported are total intensity, mean, median and mode of pixel intensity distribution, as well as an estimate of these for the local background, and other statistics such as the standard deviation of both signal and background. Once the appropriate adjustments have been made during image analysis to produce the quantization matrices, the process of normalization begins to adjust the gene values for different systematic biases. Normalization techniques differ according to microarray platforms and designs of experiments. We discuss mainly the normalization techniques used for two-channel or two-colour (red and green) microarray platforms in this chapter.

The objective of normalization is to adjust the gene expression values of all genes on the array to remove experimental artefacts. All normalization methods are based on some underlying assumptions about the data and the experimental design, and consequently, the normalization approach used must be appropriate to the particular experiment in consideration. In a typical microarray experiment, there are two decisions to be made for normalization : which genes to use as normalization genes, and which normalization algorithm to use.

Normalization approaches typically use either the complete set of arrayed genes or a *control set* of genes, generally, either a set of *housekeeping genes* or a set of *spiked-in genes*. Housekeeping genes are the ones that are involved in essential activities of cell maintenance and survival, but that are not involved in cell function or proliferation. Because all cells need to express these genes to survive, it is reasonable to expect that such genes will be similarly expressed in all samples in the experiment. Unfortunately, it is often difficult to identify housekeeping genes, also, there is a accumulating evidence that many of these genes change in expression under some circumstances [LSGH02, TZL$^+$99]. Spiked-in genes on

the other hand are realized by including in all the samples some RNA which is generally not found in either sample. For example, adding some yeast RNA in human samples. By placing some relative amount of RNA into all the samples, it is possible to create an artificial set of housekeeping genes which will have same expression across both the channels in a two-colour microarray experiment. The problem with this method is that it is necessary that the proportion of sample RNA of spiked RNA must be the same in both the channels. This is a technically challenging problem.

Once a normalization gene set has been selected, a normalization factor is calculated for each gene and is used to adjust the data to compensate for the experimental variability. The key consideration in normalization methods involves determination of the amount by which the genes in the red channel change relative to the genes in the green channel. The bias can differ from array to array and gene to gene. Once the size of the bias is estimated, we obtain the final signal value by subtracting the normalization factor from the observed log ratio of genes between two channels. The normalization algorithms to estimate this bias can be categorized in three categories :

1. Linear or global normalization
2. Intensity based normalization, and
3. Location based normalization

Linear or global normalization assume that the red and green intensities have a linear relationship for the normalization genes on a given array. The slope of this linear relationship will determine the amount of normalization required. Since a single parameter is used to scale the whole data, this type of normalization technique is also known as *global* normalization. Some of the examples for the global

normalization technique are, log centering, rank invariant methods [TRLW01], quantile-based method [BIAS03], linear regression approach [CP91], and Chen's ratio based method [CDB97] etc. Yang et al.[YDLS01] summarized a number of normalization methods for global normalization in their publication.

Another group of normalization techniques called intensity based normalization methods consider that the overall magnitudes of the spot intensity may have an impact on the relative intensities between the channels. This claim is assessed using so-called M-A plots proposed by Yang et al. [YDLS01, YDL+02]. Locally weighted linear regression (LOWESS) analysis has been proposed as a normalization method that can remove such intensity-dependent effects [YDL+02, Cle79]. Kepler et al. [KCM02] proposed a local regression to estimate normalized intensities as well as intensity dependent error variance. Smyth et al. [SS03b] present a comparison of methods used for intensity based normalization methods.

An artefact that is sometimes observed is that the background subtracted log-ratios on an array varies in a predictable manner based on their position on the array. Such artefacts could appear due to the difference in print-tips used to create the slide. In order to perform such location based normalization, it is necessary that there be significant number of normalization genes within each grid. A non-linear LOWESS normalization for correcting spatial heterogeneity was proposed by Edward [Edw03]. Chen et al. proposed a normalization method to adjust for location based biases combined with intensity biases [CKS+03]. Fan et al. [FTVWY04, FHP05] have presented error model based iterative algorithms to estimate the print-tip effects on two-colour microarrays.

There are number of other techniques proposed to deal with the choice of nor-

malization techniques with reference samples. See [Qua01, Qua02, PYK+03] for a detailed review of the subject.

Much of the above techniques depend on the experimental designs where reference gene sets are available, which serve as a basis for comparison between samples. Another class of popular normalization techniques which do not utilize common reference samples have been studied using statistical error models. The experiments which do not compare two samples, or do not use reference samples, the log ratios of channel intensities do not need to be representative of expression level of genes on an array. Instead of taking log ratios, the analysis must be based on channel specific intensities. Models of this type have been suggested by Kerr and Churchill [KKC00, KC01], Wolfinger et al. [WGW+01], Dobbin et al. [DSS03] etc. Such models explicitly include various factors of variations like sample bias, dye bias, array bias etc. and are fitted for each gene on the array. As an example, we present a model presented by Kerr and Churchill [KKC00] to help understand this concept. Assume that a microarray experiment involves multiple arrays. Every measurement in the microarray experiment is associated with a particular combination of an array, a dye (Cy3 or Cy5), a variety type, and a gene. Let y_{ijkg} denote the log value of measurement from the i^{th} array, j^{th} dye, k^{th} variety type, and g^{th} gene. To account for the multiple sources of variation in a microarray experiment, they formulated the model in the following way :

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + AG_{ig} + VG_{kg} + \epsilon_{ijkg}$$

Here, μ denotes the overall average signal, A_i represents the effect of i^{th} array, D_j represents the effect of j^{th} dye, V_k represents the effect due to the k^{th} verity type, G_g represents the effect of the g^{th} gene, AG_{ig} represents the effect due to the combination of array i and gene g, and VG_{kg} represents the interaction between

k^{th} verity and g^{th} gene. The error terms ϵ_{ijkg} are assumed to be independent and identically distributed with mean 0. All of the above effects may not be of general interest but account for sources of variations in the microarray data. It is possible to include other effects as well in the model based on the requirement.

Although, there are many methods for microarray normalization but there is no standard fitting all the cases. The data normalization is very much dependent on the platform, experimental setup and practitioner's hypothesis. Also, normalization alone cannot control all systematic variations but plays an important role in microarray data analysis. The adjusted expression data can significantly vary with different normalization procedures. Subsequent analyses, such as detection of differentially expressed genes, clustering and inference of gene networks depend a lot on the quality of data obtained after normalization [PYK$^+$03].

2.1 Normalization Model for the Arabidopsis Experiment

To understand what could be a suitable model for normalization for the Arabidopsis data discussed in section 1.6, let us summarize the scenario again. We collected data from 4 leaves, each one provided us a biological sample at each time point. RNA samples were extracted from each leaf which were labelled twice with Cy3 and twice with Cy5, providing 4 technical replicates for each biological sample. The data has been collected over 22 days. The main aim of the experiment is to monitor the temporal activities in leaves during the process of senescence. The data from a two-colour experiment like ours (Cy3 and Cy5 dyes representing green and red colours respectively) can be analysed in two ways. One way is, when we are comparing fold change in two samples where each sample is labelled differently. We take log ratio of Cy3 and Cy5 for each gene to provide us one data point for every two spots on the array. The other way is, we read absolute level of gene expres-

sion for each spot on array and treat each Cy3 and Cy5 readings as separate data points. Since we are not interested in fold change or comparison of samples, nor do we use any reference sample in our experiment, we take absolute levels of gene expression for our analysis. We use background-subtracted gene intensities reported by Imagene software for preparing the *quantization matrix*. Using this approach, we have 4 leaves $\times$ 4 replicate $\times$ 22 time-points $\times$ 30336 genes $= 10,678,272$ values in our spot quantization matrix. The matrix can be arranged in a $30336 \times 16 \times 22$ form, where 30336, 16 and 22 are the number of genes, total replicates and total time points respectively. Considering that we know the independent biological and technical units in the experiment, and we have 16 values for each gene for each time point, we aim to estimate the influence of unwanted systematic variations and minimize them. Lack of reference samples and availability of multiple replicates allow us to construct a statistical error model for normalizing the gene expression values. As a test, whether our model satisfies the criteria of removing systematic biases, the residuals associated with genes in a replicate, standardized by the estimated gene-wise variances, should show a Gaussian distribution. Also, the correlation between residuals from one replicate to other replicate should be minimum.

In the following sections, we construct a statistical error model which can deal with our dataset. The model is generic in nature and can deal with any dataset adhering to similar experimental condition like ours. We test the model using the complete Arabidopsis dataset.

2.2 Methods

Let Y_{gbi} be the log of observed gene expression for gene g, on biological sample b, measured on the replicate i. To account for multiple sources of variations in a

microarray experiment, consider the following model

$$Y_{gbi} = \alpha_g + \beta_b + \gamma_{bi} + M_{gbi} + \zeta_{gbi} \tag{2.1}$$

where α_g is the assumed true value of the gene expression, β_b is the systematic variation associated with each biological sample,γ_{bi} is the systematic variation associated with the replicate i for the biological sample b. M_{gbi} is the confounding effect of dyes and other experimental conditions. Finally, ζ_{gbi} is the residual or error term in the model. Our goal is to -

1. Estimate all the above factors of variations
2. Estimate the error term ζ_{gbi} such that it is independent and identically distributed with zero mean and constant variance, and
3. Minimize the correlation between the error terms across replicates

Here we assume that each gene is spotted only once on each array and the replicates include both biological and experimental replicates.

For simplicity, the model in 2.1 can be expressed as

$$Y_{gbi} = \alpha_g + \beta_b + \gamma_{bi} + \epsilon_{gbi} \tag{2.2}$$

where

$$\epsilon_{gbi} = M_{gbi} + \zeta_{gbi} \tag{2.3}$$

For the further derivation, we fix $g = 1, 2, \ldots, G$, $b = 1, 2, \ldots, B$ and $i = 1, 2, \ldots, I$ associated with each b. We first present the following steps for estimation of γ_{bi}, β_b and α_g, and later we show how to process ϵ_{gbi} to achieve final goals.

With each biological sample b and the replicate i associated with it, the average gene expression for a gene g can be obtained as

$$\overline{Y}_{gb} = \frac{1}{I}\sum_{i=1}^{I} Y_{gbi} \tag{2.4}$$

The bias associated with each replicate i for a given b can be obtained by removing the effect of average gene expression $\overline{Y}_{gb}$ from each Y_{gbi}. Thus

$$Y_{gbi} - \overline{Y}_{gb} = \alpha_g + \beta_b + \gamma_{bi} - \alpha_g - \beta_b - \frac{1}{I}\sum_{i=1}^{I}\gamma_{bi} + \epsilon'_{gbi} \tag{2.5}$$

Using the Least-square estimates, the systematic variation γ_{bi} can be estimated as

$$\Rightarrow \widehat{\gamma_{bi}} = \frac{1}{G}\sum_{g=1}^{G}(Y_{gbi} - \overline{Y}_{gb}) \tag{2.6}$$

To estimate the variation β_b among biological samples we first remove the variation $\widehat{\gamma_{bi}}$ from the log of observed gene expression. Let

$$Y'_{gbi} \;=\; Y_{gbi} - \widehat{\gamma_{bi}}$$

hence according to our model in 2.2

$$Y'_{gbi} \;=\; \alpha_g + \beta_b + \epsilon_{gbi}$$

Now, for each b we have

$$Y''_{gbi} = Y'_{gbi} - \overline{Y}_{gi}$$

where $\overline{Y}_{gi}$ is the average gene expression for a given replicate i across all the biological samples $b = 1, 2, \ldots, B$.

$$Y''_{gbi} = \alpha_g + \beta_b - [\alpha_g + \frac{1}{B}\sum_{b=1}^{B}\beta_b] + \epsilon''_{gbi}$$

$\hat{\beta}_b$ for $b = 1, 2, \ldots, B$ can be estimated using Least-squares and by averaging over all the genes $g = 1, 2, \ldots, G$. In the next step, we can remove the biological variation β_b and the combined effect of variation of biological sample b and replicate i captured in $\widehat{\gamma_{bi}}$ from the Y_{gbi} to estimate the expected value of the gene expression α_g.

$$Y'''_{gbi} = Y_{gbi} - \hat{\beta}_b - \hat{\gamma_{bi}} = \alpha_g + \epsilon_{gbi}$$

$$\hat{\alpha_g} = \frac{1}{B \times I}\sum_{b,i} Y'''_{gbi} \tag{2.7}$$

After estimating $\hat{\alpha_g}$, $\hat{\beta}_b$ and $\hat{\gamma_{bi}}$, we can compute $\hat{\epsilon_{gbi}}$ from our model in Equation 2.2 as

$$\hat{\epsilon_{gbi}} = Y_{gbi} - \hat{\alpha_g} - \hat{\beta}_b - \hat{\gamma_{bi}} \tag{2.8}$$

In our model β_b and γ_{bi} are variations specific to biological samples and replicates. But there may be many other sources of variations in an experiment which may be confounded in various combinations and are captured in the expression ϵ_{gbi} which is specific to each gene, biological sample and replicate. In the ideal conditions, ϵ_{gbi} should be independent and identically distributed and should be uncorrelated across replicates. But this seldom is the case because of presence of many other unknown sources of experimental variations in a dataset. Thus ϵ_{gbi} demands a separate analysis and treatment.

2.2.1 Select-and-Reject Method

Recall that according to Equation 2.3, ϵ_{gbi} is composed of M_{gbi} and ζ_{gbi} which can be calculated separately. Consider a $G \times R$ matrix E having values of ϵ_{gbi}. G is the total number of genes and R are the total replicates present in the experiment. Let X be the $R \times R$ correlation matrix of E. In order to remove high degree of correlation among the values in E across different columns (corresponding to different replicates), we can apply an iterative procedure where the ϵ_{gbi} values for each replicate i denoted as the E_i column can be represented as a linear combination of highly correlated columns selected from the rest of the columns in E. The iterative process can be summarized in the following steps -

1. Set X to be the correlation matrix of E.

2. Pick a column i in the matrix X, corresponding to the column E_i in the matrix E. Find the entry having the lowest correlation coefficient in column i. The entry identifies the least correlated column E_j with the E_i column in E.

3. Estimate the S_k coefficient and ζ_i in the following equation with Least-square method

$$E_i = \Sigma S_k E_k + \zeta_i$$

   where $k \neq i, j$

4. Estimate the correlation of E_i with rest of the E_k columns and call the correlation matrix as X'. Also compute the variance ζ_i for the estimated $\hat{E}_i$.

5. If every entry in $X' < 0.1$ or ζ_i stabilizes across successive iterations, then store $\Sigma S_k E_k$ as M_{gbi} and ζ_i as ζ_{gbi}, and go to step 2 with next the i. Other-

wise, go to step 1 with E replaced by the E_k (where $k \neq i, j$) columns and iterate.

2.2.2 Equal Distribution of Negative Correlation

Upon inspection of the results on a small data sample, we found that though the correlation among replicates drops significantly, there is still a presence of more negatively correlated ζ_{gbi} terms compared to positively correlated ones. In order to deal with the skewness in the correlation terms, we distribute the correlation in system equally among all the replicates. Assuming ζ_{gbi} as -

$$\zeta_{gbi} = \zeta_{gbi} - a_l\zeta_{gbi} + a_l\zeta_{gbi}$$

and further

$$\theta_{gbi} = \zeta_{gbi} - a_l\zeta_{gbi}$$

So

$$\zeta_{gbi} = \theta_{gbi} + a_l\zeta_{gbi} \tag{2.9}$$

θ_{gbi} is the white noise element. By denoting the matrices having θ_{gbi} elements as θ, a_l elements as A and ζ_{gbi} matrix as ζ we have,

$$\zeta = \theta + A\zeta \tag{2.10}$$

The correlation matrix (ρ) of θ across all the replicates R has approximately the same correlation coefficient

$$\rho = \begin{bmatrix} 1 & \frac{-1}{R-1} & \dots \\ \frac{-1}{R-1} & 1 & \dots \\ \vdots & \vdots \ddots & \end{bmatrix}$$

The unknown matrix A can be computed from 2.10 where

$$\theta = (I - A)\zeta$$

Denoting the covariance matrix of θ and ζ as Σ_θ and Σ_ζ respectively, we have

$$A = I - \sqrt{(\Sigma_\theta)}(\sqrt{(\Sigma_\zeta)})^{-1} \tag{2.11}$$

So eventually, the final model in terms of Equation 2.1, after removing all the effects of variations and further breaking the error terms ζ_{gbj} in a way to equally distribute the remaining correlation in the system, can be expressed as

$$\hat{Y_{gbi}} = \hat{\alpha_g} + \theta_{gbi} \tag{2.12}$$

where $\hat{Y_{gbi}}$ is the adjusted value for the original expression value Y_{gbi} observed for gene g with biological sample b using the replicate i.

2.3 Results

The proposed normalization technique was applied to the Arabidopsis data by keeping the time fixed, i.e. for each time point, we took all the gene values from all the replicates and normalized them independent of the data obtained at other time points. Here we present the results obtained during normalization for the data obtained at time point $t = 1$. The process of normalization is same for all the time-points and similar results can be expected.

We first estimated the values of α_g,β_b and γ_{bi} according to Equation 2.2-2.7. Then, the residuals ϵ_{gbi} were calculated according to Equation 2.8. To check if the residuals between replicates were independent from each other, we computed

their correlation matrix. Figure 2.1 plots the correlation coefficients of ϵ_{gbi} values across all 16 replicates. The plot shows most of the coefficients to be distributed in the range of ± 0.3. There also exists a relatively high degree of correlation among residuals between some replicates, where the correlation coefficients are in the range of ± 0.4. The presence of high correlation among residuals between replicates is indicative of the fact that the residual values are not independent of each other. The high correlation between them must be minimized.

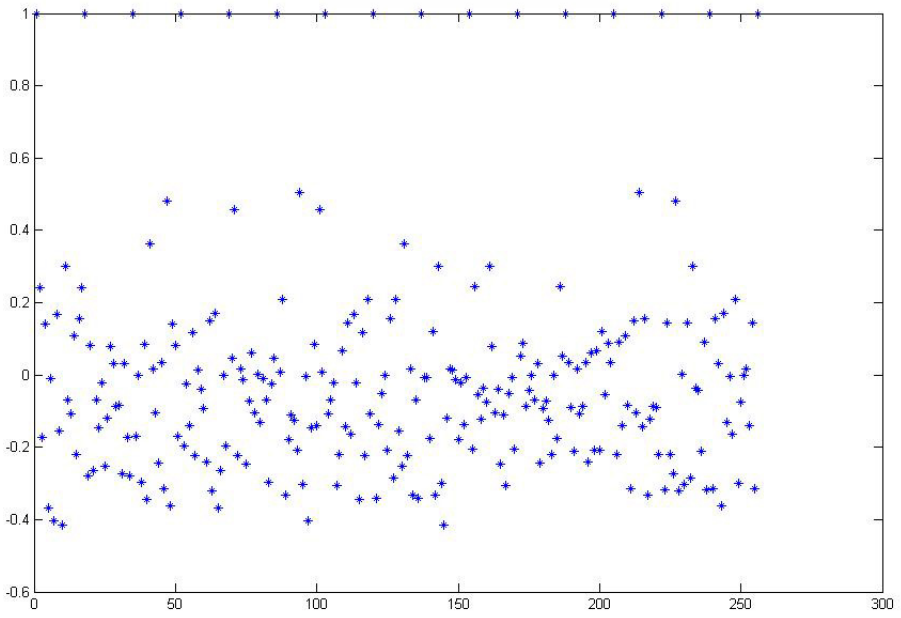


Figure 2.1: Correlation coefficients across replicates for ϵ_{gbi}

In order to minimize the correlation among residual terms between replicates, we applied our iterative Select-and-Reject method described in section 2.2.1. The Select-and-Reject procedure further divides the ϵ_{gbi} terms into M_{gbi} and ζ_{gbi}, leaving ζ_{gbi} as the final residual obtained by the method. Figure 2.2 plots the correlation coefficients for ζ_{gbi} terms between all the replicates. We can see that most of the correlation coefficients are confined within the region on ± 0.2 and very

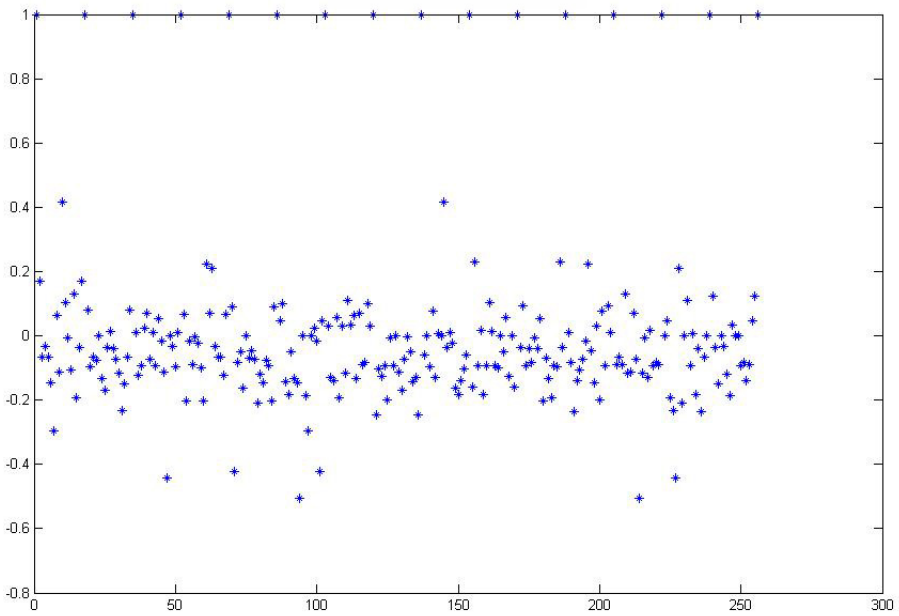


Figure 2.2: Correlation coefficients across replicates for ζ_{gbi}

few are near the value of ± 0.4. The plots shows a significant drop in correlation among residual terms across replicates compared to the ones shown in Figure 2.1, confirming the decline of correlation among residuals between replicates.

Furthermore, we can see that there is improvement in the estimation of $\hat{Y_{gbi}}$ after applying the Select-and-Reject algorithm. In an ideal situation, we would expect the estimated gene value vs. residual value plot to be completely flat, indicating that the estimated gene values and the residual terms are completely independent of each other. Figures 2.3 and 2.4 present the scatter plots of 1000 genes (randomly selected from dataset for better visibility of the plot) for corresponding $\hat{Y_{gbi}}$ values with ϵ_{gbi} and ζ_{gbi} respectively. We can see that there is some difference in the orientation of the scatter plot in Figure 2.4 due to the application of the Select-and-Reject procedure. The plot in Figure 2.4 is more balanced on both axes compared to the plot in the Figure 2.3.

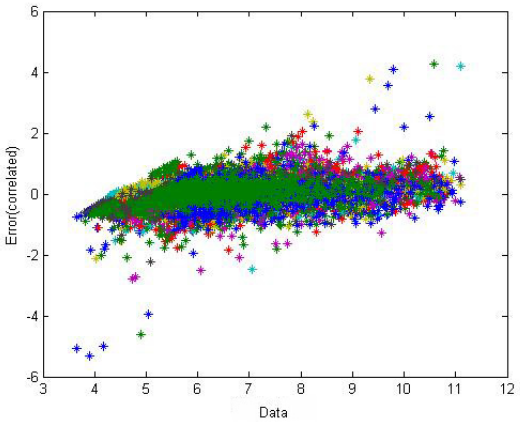


Figure 2.3: Scatter plot of $\hat{\epsilon_{gbi}}$ vs. $\hat{Y_{gbi}}$

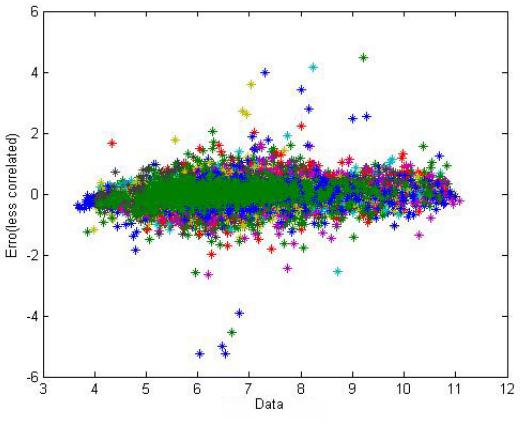


Figure 2.4: Scatter plot of $\hat{\zeta_{gbi}}$ vs. $\hat{Y_{gbi}}$

Even though the correlation among residuals between replicates had reduced significantly, there still existed some negative correlation values as seen in Figure 2.2. To adjust the residuals to be independently and identically distributed, we needed to disperse this correlation equally among all the replicates. The task of dispersing the negative correlation in the data was achieved by applying Equation 2.9 on the correlation matrix. Table 2.1 lists the correlation coefficients for ζ_{gbi} values, whereas, Table 2.2 lists the correlation coefficients for θ_{gbi} values after applying Equation 2.9. The values is Table 2.2 indicate the equal distribution of correlation between replicates for the θ_{gbi} values. A histogram plot of θ_{gbi} values for all the genes is shown in Figure 2.5. The plot shows a zero mean Gaussian distribution supporting the claim of independent identically distributed residuals for Equation 2.12.

Figure 2.6 shows the profile plots for 4 randomly selected genes (AT1G01160, AT1G01370, AT1G03900 and AT1G06460) from the dataset. The original profiles of the genes before normalization are plotted in blue. The adjusted values for those genes after applying the normalization procedure are plotted in red. We can

1	0.168	-0.066	-0.034	-0.067	-0.145	-0.297	0.062	-0.114	0.416	0.102	-0.008	-0.106	0.128	-0.192	-0.037
0.168	1	0.079	-0.095	-0.065	-0.077	-0.0005	-0.131	-0.169	-0.037	0.0148	-0.038	-0.0740	-0.115	-0.234	-0.150
-0.066	0.079	1	0.008	-0.121	-0.092	0.023	0.069	-0.074	0.009	-0.092	0.052	-0.016	-0.113	-0.442	0.000
-0.034	-0.095	0.008	1	0.067	-0.204	-0.015	-0.090	-0.003	-0.023	-0.098	-0.201	0.223	0.070	0.210	-0.032
-0.067	-0.065	-0.121	0.067	1	0.089	-0.423	-0.083	-0.051	-0.162	0.000	-0.0696	-0.047	-0.073	-0.211	-0.119
-0.145	-0.077	-0.092	-0.204	0.089	1	0.046	0.100	-0.142	-0.182	-0.050	-0.133	-0.145	-0.504	0.000	-0.184
-0.297	-0.0005	0.023	-0.015	-0.423	0.046	1	0.029	-0.128	-0.140	0.058	-0.192	0.030	-0.117	0.110	0.033
0.062	-0.131	0.069	-0.090	-0.083	0.100	0.029	1	-0.246	-0.102	-0.124	-0.092	-0.200	-0.0069	-0.092	0.000
-0.114	-0.169	-0.074	-0.003	-0.051	-0.142	-0.128	-0.246	1	-0.060	-0.0002	-0.097	0.077	-0.129	0.008	0.0000
0.416	-0.037	0.009	-0.023	-0.162	-0.182	-0.140	-0.102	-0.060	1	-0.160	0.229	-0.091	0.018	-0.183	-0.094
0.102	0.015	-0.092	-0.098	0.000	-0.050	0.058	-0.124	-0.0002	-0.160	1	-0.036	0.094	-0.092	-0.041	-0.081
-0.008	-0.038	0.052	-0.201	-0.069	-0.133	-0.192	-0.092	-0.097	0.229	-0.036	1	0.011	-0.082	-0.234	-0.138
-0.106	-0.074	-0.016	0.223	-0.047	-0.145	0.030	-0.200	0.077	-0.091	0.094	0.011	1	-0.088	-0.065	-0.089
0.128	-0.115	-0.113	0.070	-0.073	-0.504	-0.117	-0.006	-0.129	0.018	-0.092	-0.082	-0.088	1	0.000	0.045
-0.192	-0.234	-0.442	0.210	-0.211	0.000	0.110	-0.092	0.008	-0.183	-0.041	-0.234	-0.065	0.000	1	0.123
-0.037	-0.150	0.000	-0.032	-0.119	-0.184	0.033	0.000	0.000	-0.094	-0.081	-0.138	-0.089	0.045	0.123	1

Table 2.1: Correlation coefficients across replicates for ζ_{gbi} for all the genes

1	-0.062	-0.063	-0.106	-0.049	-0.045	-0.063	-0.070	-0.059	-0.068	-0.088	-0.062	-0.077	-0.040	-0.049	-0.077
-0.062	1	-0.067	-0.060	-0.070	-0.072	-0.067	-0.066	-0.068	-0.066	-0.062	-0.067	-0.064	-0.073	-0.070	-0.064
-0.063	-0.067	1	-0.063	-0.069	-0.070	-0.067	-0.066	-0.067	-0.066	-0.064	-0.067	-0.065	-0.071	-0.069	-0.065
-0.106	-0.060	-0.063	1	-0.042	-0.036	-0.063	-0.072	-0.057	-0.069	-0.097	-0.061	-0.082	-0.030	-0.043	-0.082
-0.049	-0.070	-0.069	-0.042	1	-0.083	-0.069	-0.065	-0.072	-0.066	-0.053	-0.070	-0.060	-0.087	-0.079	-0.060
-0.045	-0.072	-0.070	-0.036	-0.083	1	-0.070	-0.064	-0.074	-0.066	-0.050	-0.071	-0.059	-0.092	-0.083	-0.059
-0.063	-0.067	-0.067	-0.063	-0.069	-0.070	1	-0.066	-0.067	-0.066	-0.064	-0.067	-0.065	-0.071	-0.069	-0.065
-0.070	-0.066	-0.066	-0.072	-0.065	-0.064	-0.066	1	-0.065	-0.066	-0.069	-0.066	-0.067	-0.064	-0.065	-0.067
-0.059	-0.068	-0.067	-0.057	-0.072	-0.074	-0.067	-0.065	1	-0.066	-0.061	-0.068	-0.063	-0.076	-0.072	-0.063
-0.068	-0.066	-0.066	-0.069	-0.066	-0.066	-0.066	-0.066	-0.066	1	-0.067	-0.066	-0.067	-0.066	-0.066	-0.067
-0.088	-0.062	-0.064	-0.097	-0.053	-0.050	-0.064	-0.069	-0.061	-0.067	1	-0.063	-0.074	-0.046	-0.053	-0.074
-0.062	-0.067	-0.067	-0.061	-0.070	-0.071	-0.067	-0.066	-0.068	-0.066	-0.063	1	-0.064	-0.072	-0.070	-0.064
-0.077	-0.064	-0.065	-0.082	-0.060	-0.059	-0.065	-0.067	-0.063	-0.067	-0.074	-0.064	1	-0.057	-0.060	-0.070
-0.040	-0.073	-0.071	-0.030	-0.087	-0.092	-0.071	-0.064	-0.076	-0.066	-0.046	-0.072	-0.057	1	-0.086	-0.057
-0.049	-0.070	-0.069	-0.043	-0.079	-0.083	-0.069	-0.065	-0.072	-0.066	-0.053	-0.070	-0.060	-0.086	1	-0.060
-0.077	-0.064	-0.065	-0.082	-0.060	-0.059	-0.065	-0.067	-0.063	-0.067	-0.074	-0.064	-0.070	-0.057	-0.060	1

Table 2.2: Correlation coefficients across replicates for θ_{gbi} for all the genes

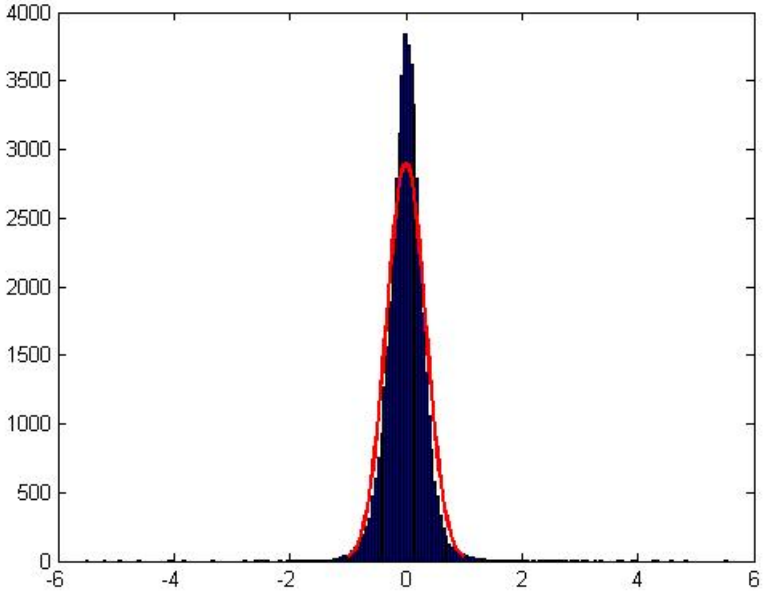
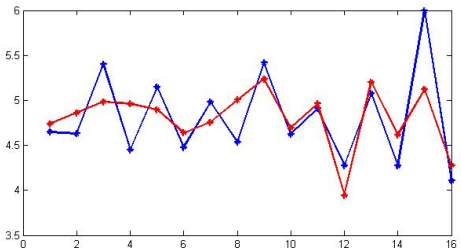
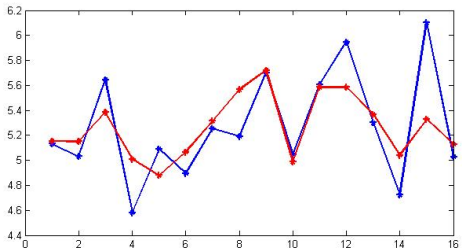


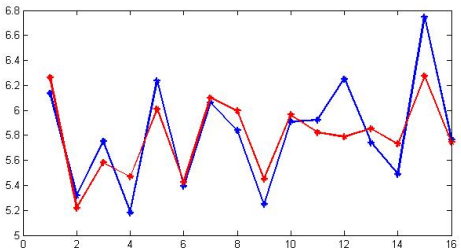
Figure 2.5: Histogram for θ_{gbi} for all the genes



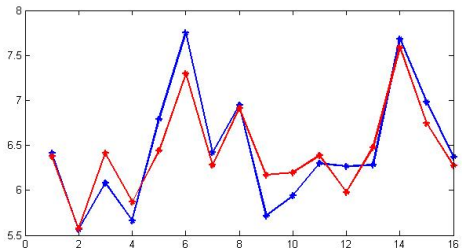
(a) Gene AT1G01160



(b) Gene AT1G01370



(c) Gene AT1G03900



(d) Gene AT1G06460

Figure 2.6: Original (in blue) and adjusted (in red) values for four genes across all replicates

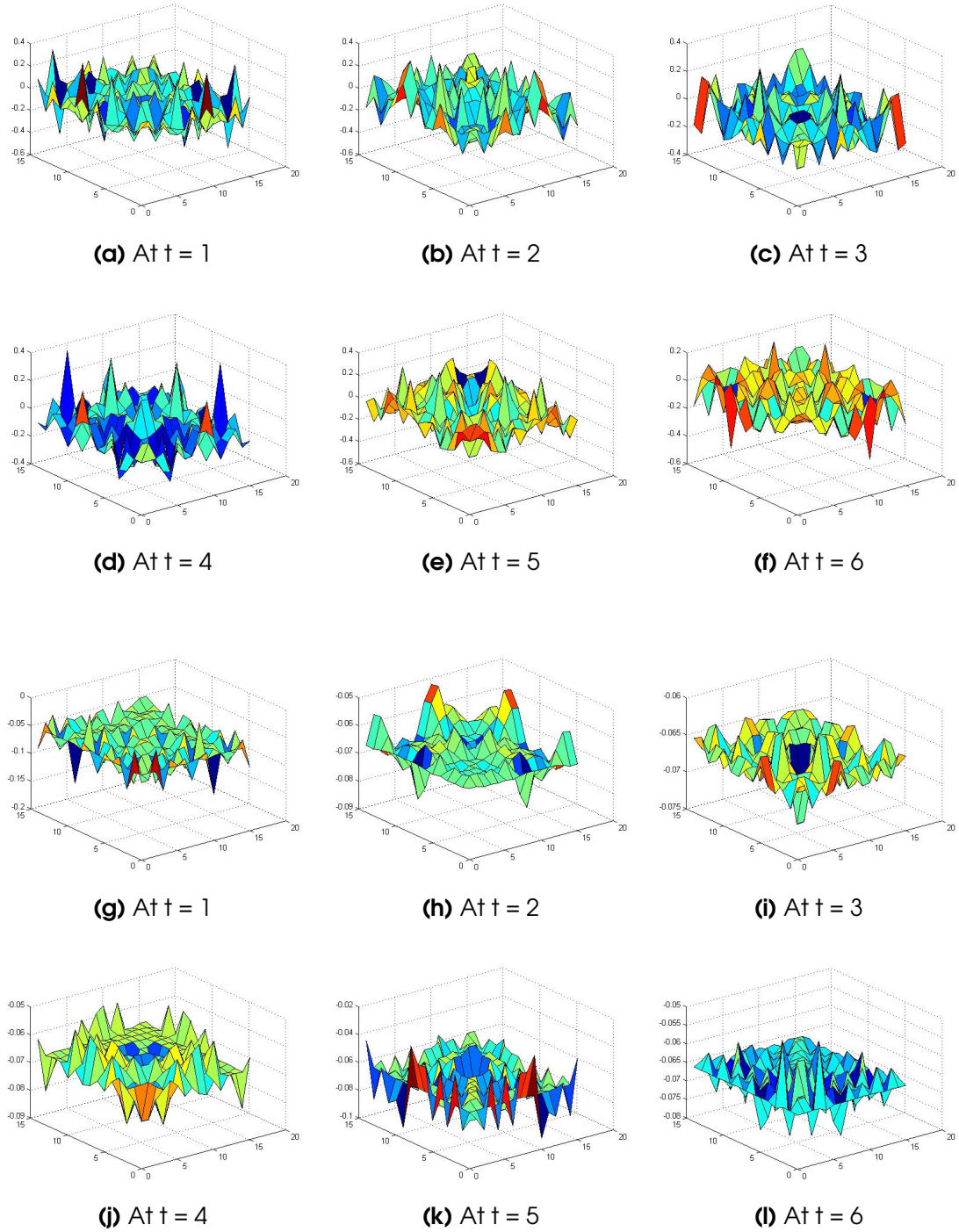
see that in all the cases, the profiles with red colour are much flatter than their corresponding blue profiles. The flatter profiles are desirable, because under ideal

conditions, when there are no systematic biases, we expect the values observed at all the replicates for any gene to be the same. But due to the lack of ideal conditions, the values observed for a gene at different replicates vary. An important purpose of a good normalization method is to adjust the observed values in such a way that it minimizes the difference observed at different replicates.

The same procedure for estimating biases and minimizing correlation among residual terms between all the replicates was applied for data at each time-point. The plots in Figures 2.7 denote the correlation coefficients of ζ_{gbi} and θ_{gbi} values for all the genes for time points $t = 1, 2, \ldots, 6$. We can see that the plots in Figure 2.7 (g)-(l) are much flatter compared to their counterparts in Figure 2.7(a)-(f). These results show the fulfilment of our targets that we set in the beginning of normalization.

2.4 Summary

Cleaning of data from unwanted experimental variations is an important step in microarray data analysis. Much of the the further investigation depends on the quality of data obtained at this stage. This chapter started with a discussion on the need for data normalization and provided a brief overview of various methods proposed for a two-colour microarray experiment. We provided a summary of the microarray experiment to collect the Arabidopsis data discussed in Chapter 1. This chapter proposed a statistical error model based normalization technique to deal with various unwanted artefacts in the data. After estimating various sources of biases in the data, the correlation among residual terms between replicates was minimized using an iterative procedure. We further proposed a way to deal with the remaining negative correlation in the data and correct the residual terms to show a zero mean Gaussian behaviour.



(a) At t = 1 **(b)** At t = 2 **(c)** At t = 3

(d) At t = 4 **(e)** At t = 5 **(f)** At t = 6

(g) At t = 1 **(h)** At t = 2 **(i)** At t = 3

(j) At t = 4 **(k)** At t = 5 **(l)** At t = 6

Figure 2.7: Plots in (a)-(f) show the correlation coefficients for ζ_{gbi} between the replicates for time points 1...6, whereas, plots in (g)-(l) show the correlation coefficients for θ_{gbi} between the replicates for the corresponding time points. We can see that the plots in (g)-(l) are much flatter compared to the ones in (a)-(f)

The normalization techniques available in the existing literature focus mainly on removing effects from certain predefined sources of technical variations. These predefined sources of variations like dye, sample, variety, gene-specific biases etc. are modelled as separate terms in the computational model for estimation. Though these methods are capable of dealing with individual terms, such specific designs can make the normalization models very specialized to certain types of microarray experiments, where the probability of those variations being present is large. Also, seldom is it the case that we have information about *all* types of biases present in the system. In the absence of the accurate distinction between the possible biases in the experiment, it is difficult to incorporate them as separate variables in the normalization model. Our proposed model, in contrast to the existing models, does not explicitly rely on distinction of individual sources of variations, but rather, only the biases associated with the biological samples and the replicates are modelled as individual terms. All other sources of variations are combined together in a separate term in the model to be analysed exclusively. Such a design makes the normalization model generic for any highly replicated microarray data. The model is traceable at each stage and the last stages requiring the application of the select-and-reject algorithm and the dispersion of negative correlation can be skipped, provided the model satisfies the statistical tests in the first stage itself. Our approach, however, cannot deal with location specific biases like print-tip effects etc. which need to be explicitly handled. The next step after normalization in the information processing pipeline of microarray data analysis is clustering of data, which we will discuss in the next chapter.

Chapter 3

Functional Clustering of Gene Expression Data

Microarray experiments have traditionally focused on measurement of gene expressions at a single time point; they are increasingly being applied to measure the expression-levels across multiple time points. Such time-course measurements can help in gaining insights into the dynamics of gene interactions [KLW06, HHS+00, WFS04]. The computational analysis of temporal microarray data requires three distinct stages to be performed before some meaningful hypothesis can be derived from data. The first stage is the normalization stage where data is cleaned from the effects of unwanted experimental biases [Qua01, Spe03]. The second stage requires the grouping of data based on certain features which helps in reduction of data dimensions. The third and final stage is the inference of the relationship between various genes of interest and understanding the functioning of smaller subsystems which comprise together to make a bigger system. Though these three stages have an ordered sequence of execution, the computational methods applied at these stages need not be dependent on each other. The normalization method solely relies on the experimental design of the microarray experiment [Spe03, KC01]. The clustering step can be performed using point-

based, model-based or feature based grouping of data [AYA07] depending on the hypothesis adopted by the practitioner. The final stage of relationship inference between genes is restricted to the sets of *selected* genes which can be studied as a system of bivariate or multivariate causal *interactions*. Keeping in mind, the final goal of microarray data analysis being identification of *interactions* between genes at the third level, the quest for this goal should ideally start when the data is being grouped together at the clustering stage.

There are plenty of clustering techniques which exist for clustering of temporal gene expression data and can be broadly classified as :

1. Point-wise distance based methods - group genes by minimizing an objective function based on a distance measure computed between gene pairs. The distance measure could be Euclidean distance, mutual information, correlation or its respective variants [DH05] etc. The point-wise methods can be further classified into two classes : (a) partitioning, and (b)hierarchical. Among partitioning methods, k-means [Seb84] and self-organizing maps (SOM) [EHI03] are widely used approaches. Hierarchical methods on the other hand create a hierarchy of relative distances and place multinomial points along a one-dimensional axis based on the relative distance between points. A typical representation of results obtained from hierarchy based methods is in form of a dendogram [JW88]. Point-wise distance based approaches are the most widely used clustering techniques for gene expression data due to their computational and conceptual simplicity. These methods are also popular due to their implementation in the large number of software packages designed for analysis of gene expression data. Some biological case studies using point-wise methods for clustering gene expression data can be found in [ESBB98, GSK+00, THC+99].

2. Feature based clustering methods - aim at detecting salient features and local or global shape characteristics of the expression profiles. As opposed to a distance based similarity measure, looking for general shape among the gene profiles can uncover more intricate relationships, such as time shifts and inversion in expression profiles. Ji and Tan [JT05] proposed a time-lagged based cluster identification technique which relied on the directional change of profiles across consecutive time points. Edge detection method by Chen et al. [CFS99] summed the number of edges of two gene expression curves where edges had the same direction within a time lag to generate a score. Directional changes were also used to compute the slope of expression values in Event method by Kwon et al. [KHN03] to cluster the gene profiles. Some of the feature based clustering methods transform the raw expression data to symbols which are further analysed to detect similarity between profiles [BHWK05, EBJ06]. Dominant Spectral Component Method by Yeung et al. [YSLY04] decomposes temporal expression sequences into spectral components using the autoregressive modelling technique to measure gene-gene relationship to form clusters. Graph-theoretic approaches studying the nature, properties, structure of the graph where the genes represent the nodes and the arcs representing association between genes also come under feature based clustering methods. Graph spectral clustering [NJW02] and minimal spanning tree method [GR69, XOX02] are other well-known feature based clustering methods.

3. Model based clustering methods - shift the similarity emphasis from the data to the unknown model that describes the data. Such methods are based on statistical mixture models which assume that data is generated by a finite mixture of underlying probability distributions, with each component corresponding to a distinct cluster [YLW08, PLL02]. Model based

clustering relies on the fundamental assumption that the observed expression profiles are clustered in functional space based on their characteristics. The focus of this approach is in functional decomposition of data, rather than the decomposition of raw data. The computational approach in model based clustering methods is based on maximizing the likelihood of data points. Expectation-maximization (EM) is a popular model based clustering approach to estimate unknown parameters (mean and standard deviation in case of Gaussian distribution) of underlying probability distribution for each cluster in order to maximize the likelihood of the observed expression profiles [DLR77]. Based on similar lines to the EM algorithm, Schliep et al. [SS03a, SSS05] suggested gene clustering based on a mixture of Hidden Markov Models (HMM). Along the similar thoughts that time-course gene dataset is a set of time series generated by stochastic processes, Ramoni et al. [RK02] suggested the use of autoregressive representation for each stochastic process defining a cluster. This method relies on regression and groups together genes whose dynamics can be expressed with roughly the same auto-regressive equation. Bar-Joseph et al. [BJGJ$^+$03] presented a clustering algorithm that uses splines to cluster the continuous representation of time series expression data. In some cases, prior knowledge has been used to fit the models to the expression profiles. For example, Zhao et al. [ZPB01] and Lu et al. [LZQ$^+$04] have used sinusoids to identify yeast genes with cyclic behaviour. Moller-Levet et al. [MLW03] presented a method based on a pre-defined comprehensive set of profiles to cluster genes according to their match with respective profiles.

One of the ultimate goals of all gene clustering algorithms is to discover the underlying gene pathways representing biological processes. Genes that are lying in the same pathway are often activated or depressed *simultaneously* or *se-*

quentially upon receiving stimuli. The biological signal is typically transmitted through intermediate gene interactions due to physical or chemical activities. The simultaneous or sequential activation or depression is delineated by the underlying network connection patterns. In this chapter, we present a novel approach for clustering of temporal microarray data. Our approach is a hybrid of both model-based and feature-based clustering methods. The temporal recording of gene expressions provides an excellent opportunity to view the gene profiles with respect to time and helps in understanding the underlying causal processes driving the behaviour of the genes and the system in turn. Like any dynamical system, in a system with a temporal expression profile, time plays a crucial role in the way the system behaves. The primary hypothesis behind the approach presented in this chapter is : *the observed effect on any gene is due to some cause propagated over time*. The observed expression of a gene could be due to the effect of other genes present in the system which may be activating or inhibiting the gene under observation with different time-lags. In other words, *we perceive the system as a set of interacting entities, where each entity is an independent process and the interactions between them are temporal activities taking place between a pair of processes*.

A system with such behaviour is a widely accepted concept in Economics and Neuroscience. Granger [Gra69] proposed a method to evaluate the influence of one time series on the other time series. Granger causality has recently been introduced in bioinformatics [MC07, NU08, KG08, GWDF08] to reverse engineer gene circuits from microarray data. We will utilize Granger causality in association with a graph-theoretic feature-based method to cluster the functional modules present in a large dataset. A *functional module* can be defined as a separate substructure of a network having a group of genes or their products that are related

by physical or genetic interactions. Biological networks are considered to have modular structures where the various substructures of a complex network behave as functional units [JMBO01, TSKS04]. Biological networks have been found to have certain architectural properties which distinguish them from randomly generated networks [Bar02].

This chapter is organized in the following way. We first present the method of bivariate Granger causality to quantify the associations between pairs of genes. We then present the graph theoretic method to detect the highly connected regions in the association network to find the modular complexes. We demonstrate our ideas using synthesized datasets and a small sample of our dataset for Arabidopsis thaliana. We finally apply our method on a larger dataset for Arabidopsis to extract interesting clusters. We also analyze the structural properties of the association graph obtained for the larger dataset.

3.1 Methods

In accordance to general equilibrium theory, economists assume that everything depends on everything else; and hence, the notion of causal relationship between different time-series arises. The idea of causality is related to the idea of succession in time and that the cause always precedes the effect. Consider two processes X and Y. If Y is causal to X, the current and lagged values of Y should contain information that can be used to improve the forecast of X, rather than considering only the past and present values of X alone. Granger [Gra69] proposed the definition of causality, widely known as Granger-causality in the literature to examine whether the forecast of future values of X can be improved if along with X's own values - the current and past values of Y are also taken into account. Another reason why lagged values are considered for corresponding variables is to

avoid spurious regressions between dependent and explanatory variables [GN86]. The inclusion of past values of both variables implies that the time-series are filtered. With respect to the causal relationship between two time-series, only the corresponding innovations matter [Sch79]. We assume that our time-series is stationary in nature. Let I_t be the total information present at time t. I_t contains two time series X and Y. Let $\bar{X_t}$ be the set of all current and past values of X_t i.e. $\bar{X_t} = \{x_t, x_{t-1}, \ldots\}$ and similarly $\bar{Y_t} = \{y_t, y_{t-1}, \ldots\}$. Let $\sigma^2(.)$ be the variance of the corresponding forecast error. Granger's definition of causality between X and Y included three scenarios.

1. Granger Causality : Y is Granger causal to X if and only if the future values of X can be predicted better i.e with a lower variance, if the current and past values of Y are used.

$$\sigma^2(x_{t+1}|I_t) < \sigma^2(x_{t+1}|I_t - \bar{Y_t})$$

2. Instantaneous Granger Causality : Y is instantaneously Granger causal to X if and only if the application of an optimal linear function leads to the better prediction of future value of X, x_{t+1} if the future value of Y, y_{t+1} is used in addition to the current and past values of Y.

$$\sigma^2(x_{t+1}|I_t, y_{t+1}) < \sigma^2(x_{t+1}|I_t)$$

3. Feedback : The feedback between X and Y exists if X is causal to Y and Y is causal to X.

Feedback is only defined for the case of simple causal relations because the direction of instantaneous causality cannot be determined without additional information or assumption.

The bidirectional Granger causality can be tested in the context of linear regressive models. For a pairwise interaction between two variables, we use autoregressive specification of a bivariate vector autoregression. Assume a particular autoregressive lag length p, and we can estimate the following unrestricted equation by ordinary least squares (OLS):

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{p} \beta_i Y_{t-i} + u_t \tag{3.1}$$

where X_t is the is the prediction of the X at time t based on its own past values as well as the past values of Y, α_i and β_i are the weighting factors, and u_t is the prediction error(residual) with a variance that measures the strength of the prediction error. If all the weighting factors β_i in Equation (3.1) are equal to zero then we can conclude that Y does not contribute towards the prediction of X, but in the case of any β_i being not equal to zero, we will say that the past values of Y are contributing towards the prediction of the current X. Therefore we can have two hypotheses as follows -

$$\text{Null Hypothesis} \quad H_0 : \forall i \in \{1, 2, 3, \ldots, p\}, \beta_i = 0 \tag{3.2}$$

$$\text{Alternate Hypothesis} \quad H_1 : \exists i \in \{1, 2, 3, \ldots, p\}, \beta_i \neq 0 \tag{3.3}$$

We can conduct a F-test of the hypotheses by estimating the following equation using Ordinary Least Squares

$$X_t = \sum_{i=1}^{p} \gamma_i X_{t-i} + \epsilon_t \tag{3.4}$$

where ϵ_t is the prediction error or residual.

Let RSS_1 and RSS_0 be the sum of squared residuals of Equation (3.1) and (3.4), respectively, i.e.

$$RSS_1 = \sum_{t=p+1}^{T} \widehat{u}_t^2 \tag{3.5}$$

$$RSS_0 = \sum_{t=p+1}^{T} \widehat{\epsilon}_t^2 \tag{3.6}$$

and

$$S_1 = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(T-2p-1)} \sim F_{p,T-2p-1} \tag{3.7}$$

If the test statistic S is greater than the specified critical value specified critical value, we reject the null hypothesis that Y does not Granger-cause X.

The results are strongly dependent on the number of lags of explanatory variables. To find a suitable lag value in Equations (3.1) and (3.4) we use Akaike Information Criteria(AIC, [Aka69]). Any value p which minimizes the AIC value is chosen as the lag order.

$$AIC(p) = 2log(|\sigma|) + \frac{2m^2 p}{n} \tag{3.8}$$

where σ is the estimated noise covariance, m is the dimension of the stochastic process and n is the length of the data window used to estimate the model.

We will use the test of Granger causality to establish association between gene pairs in our interaction network. If the test for causality passes in any direction, either from $X \rightarrow Y$ or from $Y \rightarrow X$, we add an edge in the network. We are not interested in the direction of the edge and the association network is not directional at all.

3.1.1 Network Analysis

Even though most of the biological networks are sparse in their connectivity, the complexity of connections increases with the increasing number of nodes. A network of interacting entities can be readily modelled as a graph where the entities are represented by nodes and the associations between them as edges. It is often argued [SMO+03, XSD+02] that graph theoretic approaches can help analyse large interacting networks to find clusters (highly dense regions) in a network. Clusters in a gene-gene interaction network are often biological complexes or part of biochemical pathways [DES08]. Algorithms for finding clusters or highly dense regions are an ongoing topic of research and are often based on network flow theory[Gol84] or spectral clustering[NJW02]. We use a clustering method proposed by Bader and Hogue [BH03] to detect the dense regions in the association network obtained by our Granger causality based method. The method weighs all the vertices based on their local network density to detect dense regions in the graph. The decision to use this algorithm to analyse our association matrix was based on two reasons: a) this is one of the earliest methods to use a clustering algorithm to identify molecular complexes in a biological network, and hence is widely known, and, b) it has a publicly available software plug-in for a widely

used network analysis platform called Cytoscape [SMO+03]. Thus, the method and its implementation are both widely used and tested. It should be noted that application of other clustering methods to detect dense regions can produce different clusters and some may have better performances but these are not tested here.

The functioning of the method by Bader and Hogue can be understood in the following way. Given a graph $G = (V, E)$, where V and E being the sets of vertices and edges respectively, the density of a graph is based on the connectivity level and is defined as $D_G = |E|/|E_{max}|$, where E_{max} is the total number of all possible edges in a complete graph G.

The vertex weighting in the graph starts by weighing all the vertices based on their local network density using the highest k-core of the vertex neighbourhood. A k-core is a graph of minimal degree, $\forall v \in V$ and the degree of $v \geq k$. The highest k-core of a graph is the central and most densely connected subgraph. The highest k-core component gives us the highest k-core level, k_{max} in the vertex neighbourhood. The final weight of the vertex is the product of k_{max} and the density of the corresponding highest k-core component. This type of weighting amplifies the weighting of heavily connected graph regions while removing the less connected graph regions which are present in abundance.

Once the vertex weighting is done, the algorithm seeds a subgraph(complex) with highest weighted vertex and moves outwards to include vertices in the neighbourhood whose weight is greater than a given threshold. The algorithm propagates through the included neighbours and recursively checks the subsequent nodes. The process stops when no more nodes can be added to the complex and is repeated for the next highest unseen weighted vertex in the network.

In the post-processing stage, the complexes which do not contain at least 2-core (graph with minimum degree 2) are filtered out. Finally, all the complexes in the network are scored and ranked. The complex score for a given subgraph $G_C = (V_c, E_c)$ is defined as the product of the density of the subgraph and the number of vertices ($D_c \times |V_c|$). Other scoring schemes are also possible but are not tested in the original algorithm.

3.2 Results

3.2.1 Illustrative Datasets

We test our method on three sets of synthetic multivariate datasets. Each set represents a collection of stochastic processes in the form of a time-series. We construct each set in such a way that the processes belonging to the set are inter-dependent, whereas the sets themselves are disjoint from each other.

Dataset 1:

$$
\begin{aligned}
x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \epsilon_1(t) \\
x_2(t) &= 0.5x_1(t-2) + \epsilon_2(t) \\
x_3(t) &= -0.4x_1(t-3) + \epsilon_3(t) \\
x_4(t) &= -0.5x_1(t-2) + 0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \epsilon_4(t) \\
x_5(t) &= -0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \epsilon_5(t)
\end{aligned}
$$

Dataset 2:

$$\begin{aligned}
x_1(t) &= 0.05\sqrt{2}x_1(t-1) + \epsilon_1(t) \\
x_2(t) &= 0.5x_1(t-3) + \epsilon_2(t) \\
x_3(t) &= 0.5x_2(t-2) + \epsilon_3(t) \\
x_4(t) &= -0.5x_3(t-3) + 0.25\sqrt{2}x_1(t-1) + \epsilon_4(t)
\end{aligned}$$

Dataset 3:

$$\begin{aligned}
x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \epsilon_1(t) \\
x_2(t) &= 0.2x_1(t-1) + \epsilon_2(t) \\
x_3(t) &= 0.15\sqrt{2}x_2(t-2) + \epsilon_3(t) \\
x_4(t) &= 0.25\sqrt{2}x_3(t-1) + \epsilon_4(t) \\
x_5(t) &= 0.5\sqrt{2}x_2(t-2) - 0.5\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \epsilon_5(t) \\
x_6(t) &= 0.25\sqrt{2}x_1(t-2) - 0.25\sqrt{2}x_5(t-1) + 0.25\sqrt{2}x_3(t-3) + \epsilon_6(t)
\end{aligned}$$

In the above datasets, $\epsilon_i \sim$ N(0,1) represents the uncorrelated random error associated with each process. In Dataset 1, x_1 is the driving force for x_2, x_3 and x_4 with time lags 2,3 and 2 respectively. x_4 further drives x_5 and they both share a feedback loop. Similarly, in Dataset 2, x_1 drives x_2 with time lag 3 and x_2 in turn drives x_3. x_1 and x_3 both together drive x_4. Similarly, in Dataset 3, we have x_1 driving x_2. x_2 drives x_3 with lag 2 and x_3 in turn drives x_4. The process x_5 is driven by x_2 and x_4 with time lag 2 and 1 respectively. In the end, x_6 receives the drives from x_1, x_5 and x_3 with time lags 2,1 and 3 respectively. The datasets are disjoint from each other due to different sources of initiation. The datasets show different arrangements of connections between the processes which include feedback loops, low and high coefficients of drive between processes, multiple pro-

cesses together driving a single process and all the processes interacting with other processes on a different time lag.

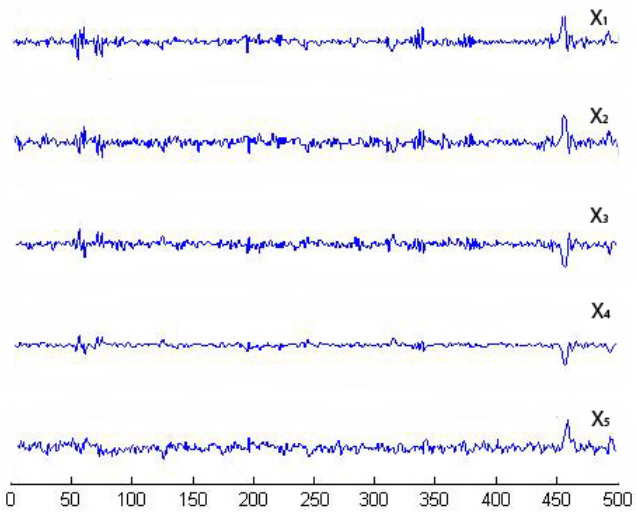


Figure 3.1: Plot of time-series for Dataset 1

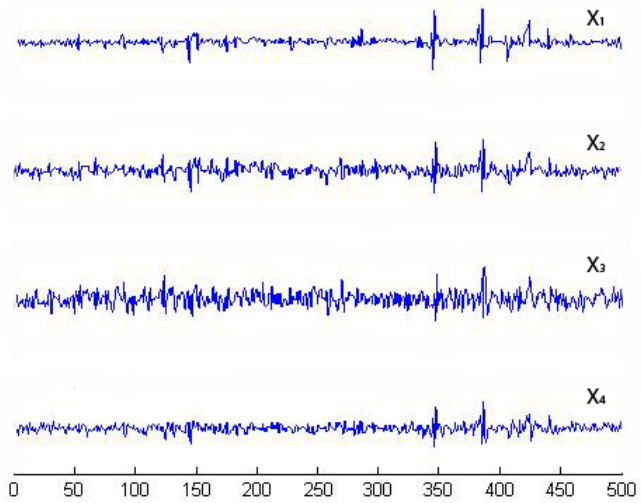


Figure 3.2: Plot of time-series for Dataset 2

Figures 3.1, 3.2 and 3.3 display the raw time series for the processes in Dataset 1, 2 and 3 respectively. We apply the Granger causality to infer the interactions between different entities in each dataset. A critical value of $\alpha = 0.05$ was chosen for the F-test to accept or reject the hypothesis. The causal hypothesis H_0 was

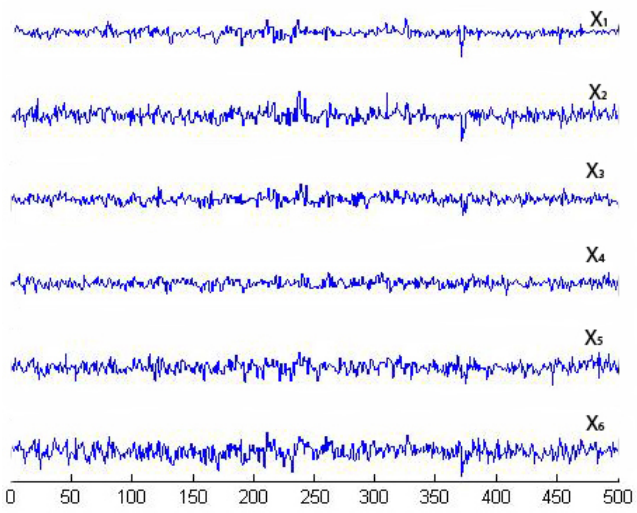


Figure 3.3: Plot of time-series for Dataset 3

tested for each pair of processes denoted by (X, Y) in both ways i.e, X causing Y, and Y causing X. Since we are only interested in the presence of interaction between (X, Y), we ignore the directionality of causal influence and quantify the association between the pair with the higher of causality value obtained from both directions. If there is no causal relationship between the pair, the association between X and Y is quantified as zero. The networks obtained after computing the Granger causality and weighing the edges for all the synthetic datasets are shown in Figures 3.4, 3.5 and 3.6. The true edges according to the equations describing the datasets are plotted with solid bold lines, whereas the extra detected edges are plotted with thin dashed lines.

We see in Figure 3.4 for (Dataset 1) that node 1 connects to nodes 2,3 and 4. Nodes 4 and 5 are also connected in the inferred network structure. The equations describing the Dataset 1 reflect these facts. One of the extra link present is the interaction of node 2 with node 3 showing the fact that nodes 2 and 3 are both driven by node 1. They exhibit an interaction according to the F-test criteria but their strength is very low compared to other interactions. Since node 1 is also a

driving force for node 4, so according to the previous argument, nodes 2 and 3 are also found to drive node 4. Node 4 and node 5 share a feedback loop, thus an interaction between them exists. There is a similar situation with nodes 1,2 and 3 interacting with node 5 due to node 1 being the common driving force behind nodes 2 and 3.

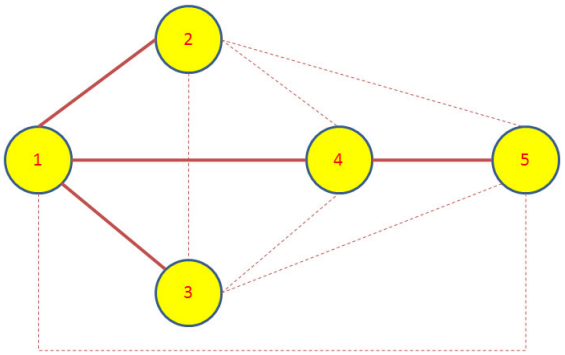


Figure 3.4: Inferred network for Dataset 1

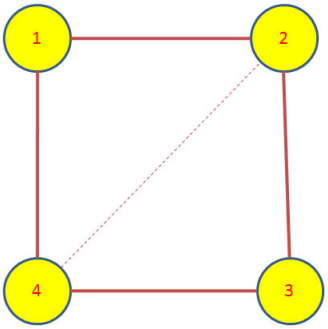


Figure 3.5: Inferred network for Dataset 2

The connections are simpler and more sparse in the case of Figure 3.5 for Dataset 2 where an extra edge not described by the system of equations is present in the inferred network. Similarly, in the network obtained for Dataset 3, the influence of node 1 on nodes 3 and 5 can be attributed to the fact that the influence is propagating through node 2 which is directly regulating nodes 3 and 5. The influence of node 2 on 6 is due to node 2 being the driving force for node 3 which in turn is directly influencing node 6. In the similar fashion, the dashed

line between nodes 4 and 6 can be explained due to node 4 driving node 5 which in turn is driving node 6.

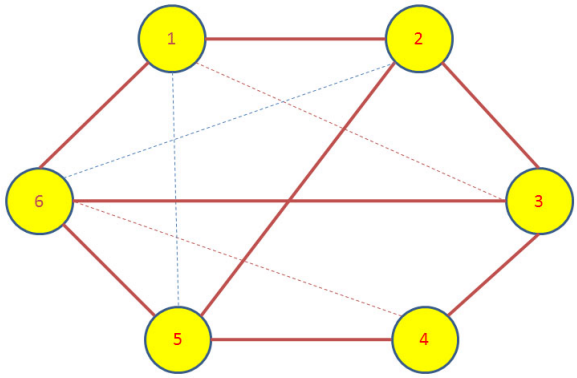


Figure 3.6: Inferred network for Dataset 3

Having analysed the individual datasets, we further investigate what happens when all the three datasets are put together to form a bigger system of processes and the pairwise interaction between the processes are computed. We create a system of 15 entities where the first 5 entities represented the processes in Dataset 1, the entities from 6 to 9 represented the processes from Dataset 2, and the last 6 entities represented the processes in Dataset 3. We then test for Granger causality for all possible pairs of processes (total 210 directional edges for a complete network with 15 nodes) in the system. We plot the interaction strength between the processes in Figure 3.7 where the x and y axes represent the 15×15 matrix of processes in the system. The interaction strengths between the processes are shown on the z-axis. We can clearly see three different island-like structures in the graph where entities 1 to 5 interact within themselves, 6 to 9 within themselves and 10 to 15 within themselves. The plot clearly shows that there is no cross talk between the entities across different sets even though they are present within the same system.

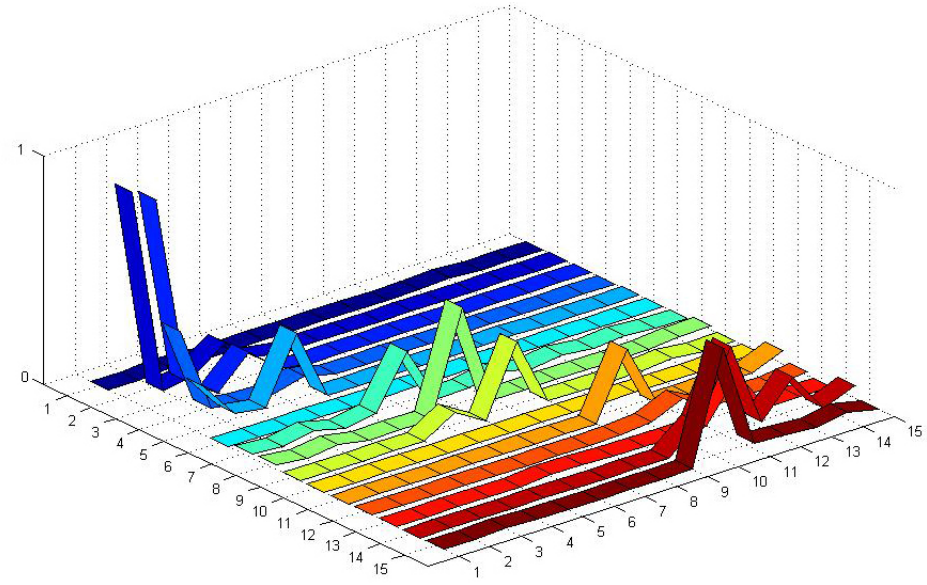


Figure 3.7: Simulation results with Dataset 1, 2 and 3 integrated into one system.

3.2.2 Arabidopsis Dataset : Small Example

After testing our method on the synthesized datasets, we test our method on the Arabidopsis discussed in Chapter 1. We test our method on two samples of different sizes of the same dataset. We first test our method on a smaller sample of 85 genes belonging to three different categories of biological processes. This smaller sample helps us mimic the scenario shown by our synthetic model. The primary advantages of choosing the smaller dataset is that it helps us in minimizing the search space for ontological validation of clusters by mining on-line repositories which may not be complete for all the genes. In the end, we apply our technique on a larger dataset of 1800 genes and study the clusters obtained and the general structural properties of the network.

For the smaller dataset, we selected 85 genes belonging to three different categories of biological processes according to Gene Ontology (GO) database [gen00].

The selected genes include genes which participate in maintaining the circadian rhythm of the plant, genes which are responsible for ageing and the genes involved in plant death. We use the gene ontology (GO) interface provided at the Arabidopsis repository at TAIR (http://www.arabidopsis.org/index.jsp) to find the names of the genes which are experimentally confirmed to perform above mentioned biological functions. It should be noted that this interface does not provide any p-value associated with the GO terms for the selected genes. This selection should be considered just as a weak indication of a gene performing the mentioned biological function. While verifying the results, we use another gene annotation tool (BinGO) [MHK05] which provides the statistical significance for the biological functions for the genes. We selected the time-series data for those genes from our microarray dataset described earlier. Some of the selected genes had flat profiles, i.e. the temporal expressions of the genes did not show much fluctuation across time. Such genes were filtered out using the 2σ technique and discarded. We finally had a set of 30 genes responsible for circadian rhythm, 34 genes involved in the ageing process and 21 genes participating in the cell death, leading in total to a set of 85 genes. Figure 3.8 shows the profiles of the selected genes.

The temporal profiles of genes were adjusted by taking the first difference of successive time points to obtain the stationary behaviour. We then applied the causality test to all the pairs of genes in the system. A complete network with 85 genes has total links equal to $2 \times \binom{85}{2} = 7140$. In the second stage, for each pair of nodes (X, Y), we selected the maximum of the causality values for directions $X \rightarrow Y$ and $Y \rightarrow X$ and assigned that value as the weight for the edge between X and Y. To further simplify the network, we applied a threshold corresponding to 0.975 quantile of all the edge value to ignore the edges with weights below that

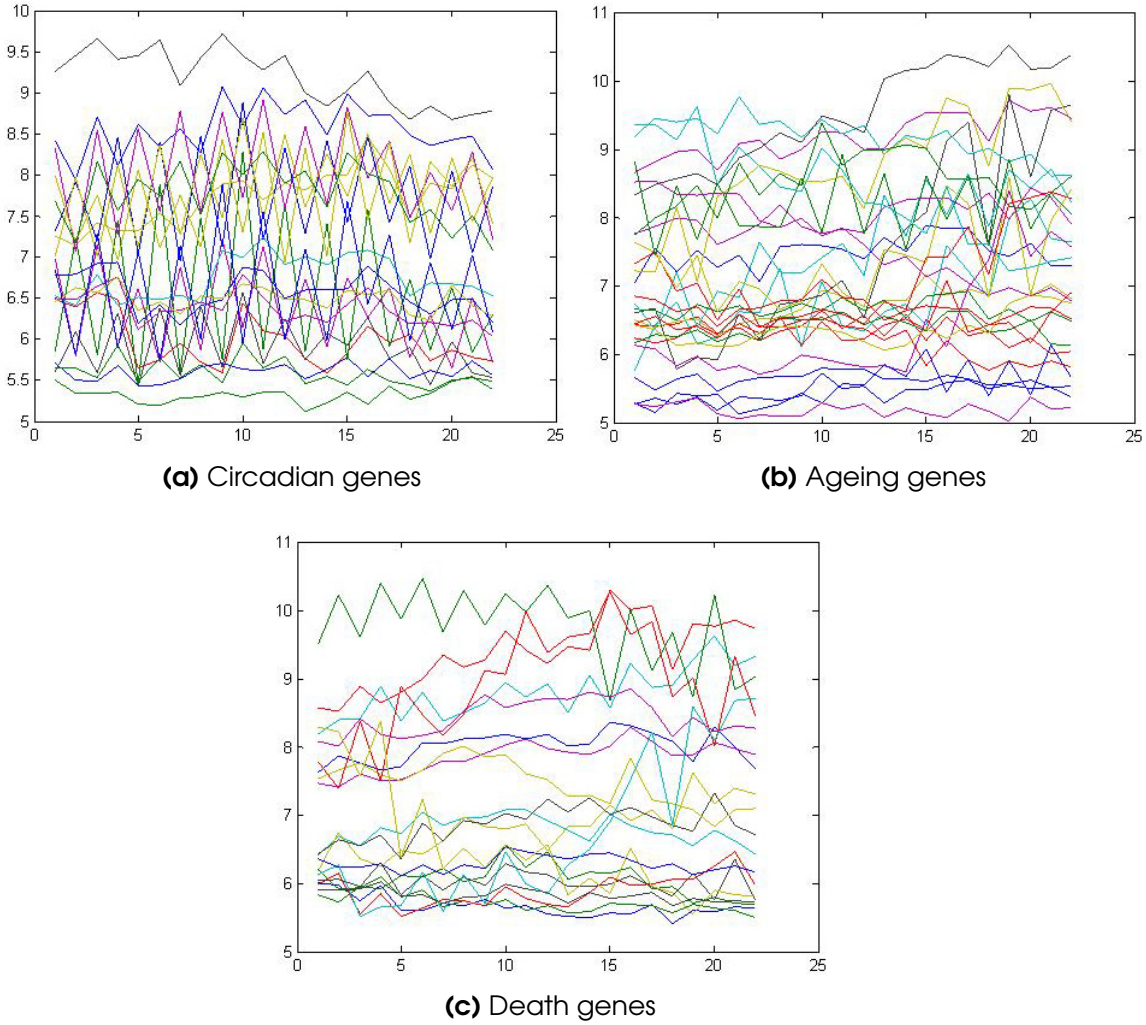
threshold. The final network is presented in Figure 3.9. The network is arranged in a degree sorted layout. The vertices with higher degree are bigger in size. The size of a vertex is decided according to the total degree associated with it. The biological relevance of the degree distribution of nodes in a biological network is discussed later in the chapter.

To find the modules in the network, we applied the graph-theoretic approach discussed in section 3.1. The approach detects densely connected regions in the network. Dense regions are the maximally connected sub-components in the graph and may be representative of the complexes in the context of biological networks. The graph-theoretic analysis gives us 4 subgraphs presented in Figure 3.10. These subgraphs are obtained by setting the k-core value = 2 and the results are presented after trimming the nodes with single degree.

To verify our hypothesis that these subgraphs represent functional modules, we use the functional information stored in the Gene Ontology (GO) database using the BinGO tool. Table 3.1 summarizes the information obtained for all the subgraphs. The first column in the table represents the GO-ID of the functional category stated in the *Functional Description* column. The genes in the table are grouped together to show the GO category they belong to, along with their statistical over-representations in columns 2 and 3. The p-values in column 2 are computed by the Hypergeometric Test which is exact and equivalent to an exact Fisher test. To reduce the False Discovery Rates (FDR), a multiple testing correction (Benjamini and Hochberg's FDR correction [BH95]) is applied and reported in column 3. The Functional Description column lists the biological functions the corresponding genes are associated with. The 'Known/Total' column represents the ratio of genes known to perform a certain biological function in the GO

GO-ID	p-value	corr p-value	Known/Total	Functional Description	Gene Names
Network 1					
48511	1.3744E-11	4.1921E-10	4/6	rhythmic process	AT5G02810, AT2G46830, AT1G68830, AT2G25930
7623	1.3744E-11	4.1921E-10	4/6	circadian rhythm	AT5G02810, AT2G46830, AT1G68830, AT2G25930
Network 2					
9814	4.5406E-11	7.6281E-9	5/8	defence response	AT1G55490, AT2G34690, AT5G03280, AT1G61560, AT4G14400
45087	2.5439E-10	2.1369E-8	5/8	innate immune response	AT1G55490, AT2G34690, AT5G03280, AT1G61560, AT4G14400
6955	3.8828E-10	2.1743E-8	5/8	immune response	AT1G55490, AT2G34690, AT5G03280, AT1G61560, AT4G14400
2376	5.7329E-10	2.4078E-8	5/8	immune system process	AT1G55490, AT2G34690, AT5G03280, AT1G61560, AT4G14400
8219	3.9627E-9	1.1096E-7	4/8	cell death	AT1G55490, AT2G34690, AT5G03280, AT4G14400
16265	3.9627E-9	1.1096E-7	4/8	death	AT1G55490, AT2G34690, AT5G03280, AT4G14400
Network 3					
7623	1.6563E-14	9.8551E-13	5/7	circadian rhythm	AT5G57360, AT2G46790, AT1G22770, AT5G61380, AT4G08920
48511	1.6563E-14	9.8551E-13	5/7	rhythmic process	AT5G57360, AT2G46790, AT1G22770, AT5G61380, AT4G08920
Network 4					
16280	3.0760E-13	1.4149E-11	5/6	ageing	AT3G12090, AT4G23410, AT5G14930, AT2G19580, AT2G21045
32502	1.1218E-8	2.5802E-7	6/6	developmental process	AT3G12090, AT4G23410, AT3G44880, AT5G14930, AT2G19580, AT2G21045

Table 3.1: Gene ontology details for the networks shown in Figure 3.10.



(a) Circadian genes

(b) Ageing genes

(c) Death genes

Figure 3.8: Temporal profiles of genes selected for smaller dataset for Arabidopsis

database with respect to the total number of genes having a reference in GO. We can see that the number of known genes in GO is less than the total number of genes submitted. This is due to the fact that the functional annotation of Arabidopsis genomes is incomplete and a particular type of annotation for a gene may differ. We may find a gene that has GO classification and no functional summary text, while other genes have functional summary text and no GO classification, while others have no classification whatsoever.

The subgraph in Figure 3.10(a) is composed of 8 genes (AT5G02810, AT1G68830,

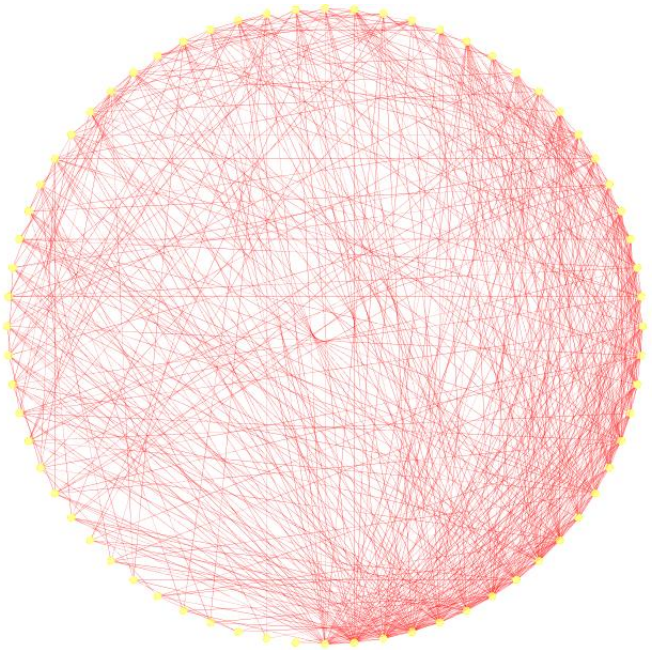


Figure 3.9: Degree sorted network structure

AT1G63260, AT2G46830, AT5G65970, AT5G15850, AT1G67070, AT2G25930). 6 out of the 8 genes are known in the GO database. No annotations could be obtained for the remaining 2 genes (AT5G65970 and AT1G67070). 4 out of the 6 known genes are clearly known as the genes participating in the circadian rhythm process. AT5G15850 is known to be associated with the regulation of flower development which is related to the circadian rhythm of the Arabidopsis plant. Gene AT1G63260 is wrongly classified as it is known to participate in the ageing process. Similarly, in the second network (Figure 3.10(b)), there are 13 genes in all(AT1G09530, AT4G14400, AT2G19450, AT2G02990, AT5G51810, AT5G20250, AT3G16770, AT2G29350, AT3G45290, AT1G55490, AT1G61560, AT2G34690, AT5G03280). 8 of the genes have entries in GO and no annotation could be found for the remaining 5 genes (AT1G09530, AT5G51810, AT5G20250, AT3G45290 and AT3G16770). 5 out of the 8 known genes are involved with the biological process of defence, immune response and cell death. AT2G19450 and AT2G02990 are known for 'response to stress'(GO process ID - 9651). Gene

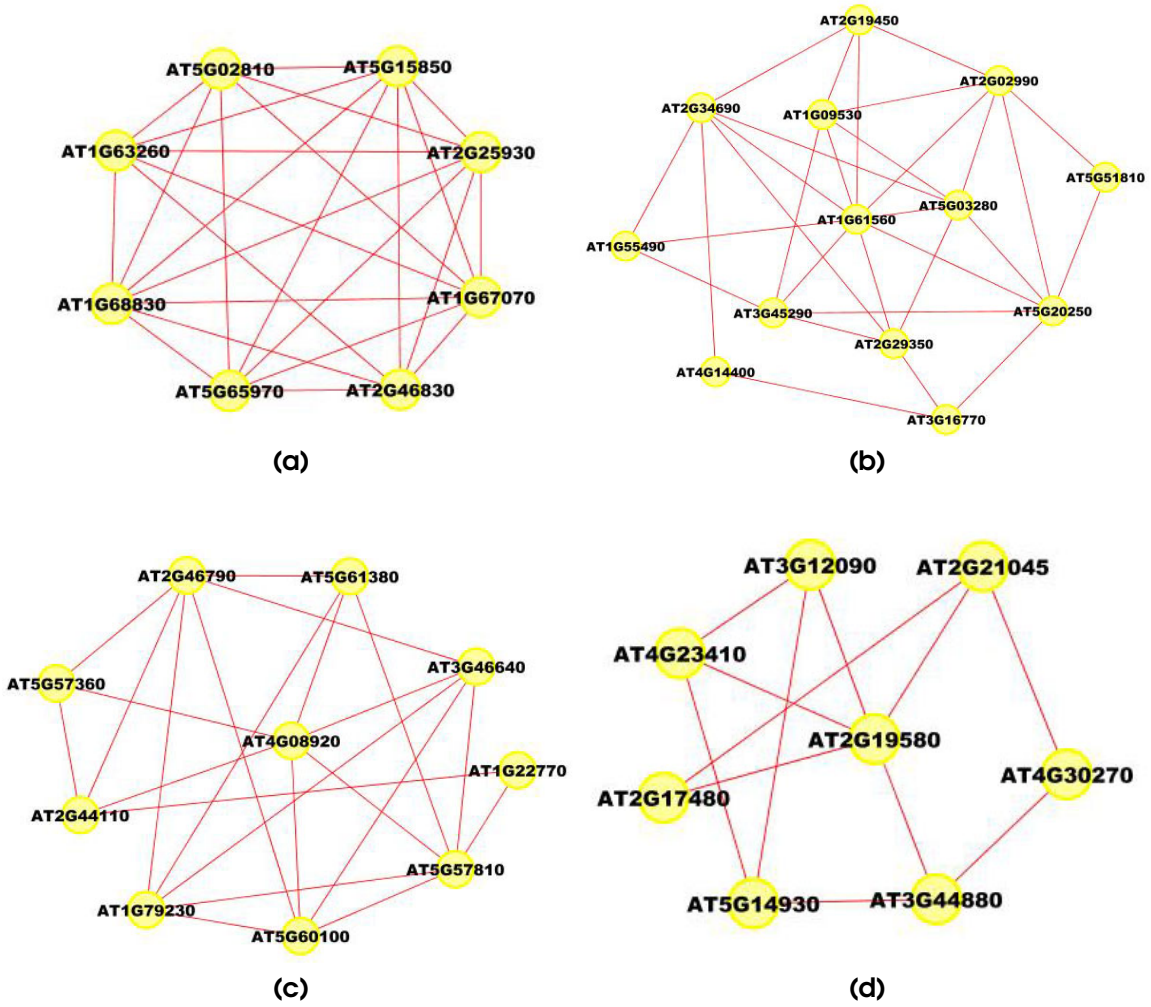


Figure 3.10: Extracted subgraphs indicating potential modules of interest in the smaller dataset. Biological functions performed by modules in respective figures are a.) Circadian rhythm b.) Immune and Defense response c.) Circadian rhythm d.) Ageing

AT2G29350 is classified for 'ageing' and is the odd member in the network. The third network shown in Figure 3.10(c) has 10 genes (AT2G44110, AT5G61380, AT4G08920, AT5G57360, AT2G46790, AT5G60100, AT1G79230, AT3G46640, AT5G57810, AT1G22770) with 7 of them known in the GO database and 3 (AT3G46640, AT1G79230, AT2G44110) are without any annotation. 5 out of the 7 annotated genes are known to participate in rhythmic activity. Gene AT5G60100 is known for regulation of circadian rhythm (GO process ID - 42752). Gene AT5G57810 is known for 'ageing' and is wrongly put in this network. The last sub-

network shown in Figure 3.10(d) is composed of 8 genes (AT4G23410, AT5G14930, AT3G44880, AT4G30270, AT2G17480, AT2G21045, AT2G19580, AT3G12090). 6 out of the 8 genes are known in the GO database. All the 6 genes are known to participate in ageing process of the plant. No annotations were found for genes AT2G17480 and AT4G30270.

3.2.3 Arabidopsis Dataset : Bigger Example

We next applied our method on a larger dataset of 1800 genes selected according to their frequency profile as described in Section 5.2.2 in Chapter 5. We ranked the genes according to their power spectrum in frequency domain and chose the top 1800 genes for analysis with our method. We constructed an association network for all the pairs of genes using the test for causality to detect the edges in the network. We applied a threshold corresponding to 0.99 percentile of all the edge values to select the most dominant edges in the network for further analysis. We applied the dense region finding method on the network using different combinations of k-core score which resulted in a number of different clusters. We present some of the clusters we found in Figures 3.12-3.15. The GO descriptions of selected genes in the shown clusters is summarized in Table 3.2. The table reports the information in the same manner as it did in case of the smaller sized data sample.

Simple network statistics : We computed certain network statistics to confirm that our network is not a randomly generated network and has the properties desired in a biological network. A total of 1353 nodes were present in the network after filtering out weaker edges. The total number of edges present in the network was 21,214 which is around 1.1% of the total possible directed edges in the net-

work, which is an indication of sparseness, a common characteristics of biological networks [BA99]. There is one connected component in the network indicating strong connectivity. The mean shortest path length is 2.6 which means that most genes are close to each other and the network diameter representing the maximum distance between two connected nodes is 6. Both the phenomenon have been described as small world properties of real networks [WS98]. We also compute and report the following widely used topological properties for our network.

Node degree distribution : We calculated the degree distribution $p(k)$ of the genes, measuring the probability that a given gene interacts with k other genes. Barabsi and Albert [BA99] used the node degree distribution to distinguish between the topologies of random and scale-free networks. Our network shows a power-law like distribution on log scale as shown in Figure 3.11(a). The plot shows that there are few nodes with large number of neighbours and they dominate the connectivity in the network. Also, the tail of power-law distribution on normal scale indicates that highly connected vertices have a large degree of occurring. Such networks exhibit preferential connectivity indicating that a new node will link to established nodes which are well connected, resulting in a structure where few hubs hold together numerous small nodes.

Shared neighbour distribution : Figure 3.11(b) shows the shared neighbour distribution for the network. $P(i, j)$ is the number of partners shared between nodes i and j, that is, nodes that are neighbours of both i and j. The shared neighbours distribution gives the number of node pairs (i, j) with $P(i, j) = k$ for $k = 1, 2, 3, \ldots$. The distribution again shows a power law like distribution indicating the presence of motifs with large numbers of connected components in the network.

Closeness centrality : Closeness centrality is a measure of how fast information flows from a given node to other reachable nodes in the network. Closeness centrality (C) of a network with n nodes is computed as the reciprocal of the average shortest path length is computed as follows : $C(n) = \frac{1}{mean(L(i,j))}$ where $L(i,j)$ is the length of the shortest path between two nodes i and j. Figure 3.11(c) plots the closeness centrality of all the nodes against number of neighbours. The isolated nodes have their closeness centrality equal to 0. An increasing trend of closeness centrality in our network further indicates strong connectivity and ability to form hubs.

Topological coefficient : Another characteristics of interaction networks can be captured by calculating the topological coefficients [GR03, RSM+02]. The topological coefficient,$TC(k)$, is a relative measure for the extent to which a gene in the network shares interaction partners with other genes. Also the topological coefficient as shown in Figure 3.11(d) decreases with the number of links (close to $\frac{1}{k}$), demonstrating that, relatively, in our network, hubs do not have more common neighbours than genes with fewer links. This indicates that genes with many links are not artificially clustered together. Moreover, it confirms the presence of modular structures in the network organization.

3.3 Comparison With Respect to Other Existing Methods

In order to have a comparison of our proposed method with some existing methods, we use the smaller Arabidopsis dataset of 85 genes discussed in Section 3.2.2. The small size and the knowledge about the functionality of genes are the main advantages of using the smaller dataset. The small size of dataset also allows us

GO-ID	*p*-value	corr *p*-value	Known/Total	Functional Description	Gene Names
Network 1					
6950	5.1715E-13	5.6542E-11	18/38	Response to stress	AT3G08730 AT5G27600 AT4G33030 AT1G53670 AT2G37220 AT4G34710 AT4G31550 AT5G54810 AT4G09650 AT4G29040 AT5G24770 AT2G14610 AT3G51780 AT3G53990 AT4G04020 AT1G16880 AT5G25610 AT5G02500
Network 2					
44444	3.5066E-4	2.0147E-2	7/11	Cytoplasmic part	AT5G42020 AT3G62030 AT1G27450 AT4G37910 AT2G45030 AT5G50950 AT1G69370
Network 3					
51869	5.5701E-12	2.3450E-9	20/41	Response to stimulus	AT5G20850 AT5G55120 AT3G08720 AT4G37680 AT5G26870 AT1G33560 AT2G47180 AT2G05520 AT1G48030 AT4G01060 AT5G37780 AT1G63840 AT2G14580 AT1G58220 AT3G26790 AT3G54320 AT5G10450 AT1G74310 AT5G45340 AT5G40350
Network 4					
9628	1.1048E-5	1.3147E-3	4/7	Response to abiotic stimulus	AT5G52310 AT3G17020 AT5G67030 AT5G63890
Network 5					
3824	9.0400E-4	4.2857E-2	10/15	Catalytic activity	AT2G17420 AT3G15020 AT5G04590 AT3G13235 AT1G23190 AT3G53160 AT3G48090 AT4G23600 AT4G08790 AT1G51680
Network 6					
6950	7.0271E-10	6.6055E-8	14/37	Response to stress	AT5G20230 AT5G61900 AT4G16845 AT3G22370 AT2G04030 AT1G55490 AT3G11820 AT4G12400 AT4G34990 AT4G23100 AT4G20260 AT3G49910 AT5G09810 AT5G05410
Network 7					
44464	2.6470E-3	1.8771E-2	25/36	Cell part	AT4G27670 AT5G59220 AT4G25100 AT3G58810 AT4G14630 AT3G53620 AT5G11520 AT3G27300 AT1G42970 AT5G43280 AT4G27430 AT1G49300 AT2G39460 AT2G37040 AT3G01480 AT5G24550 AT1G72140 AT5G62790 AT1G25540 AT1G02860 AT4G38970 AT2G43130 AT3G52960 AT3G01220 AT2G43750

Table 3.2: GO annotations for the highlighted genes shown in Figures 3.12-3.15

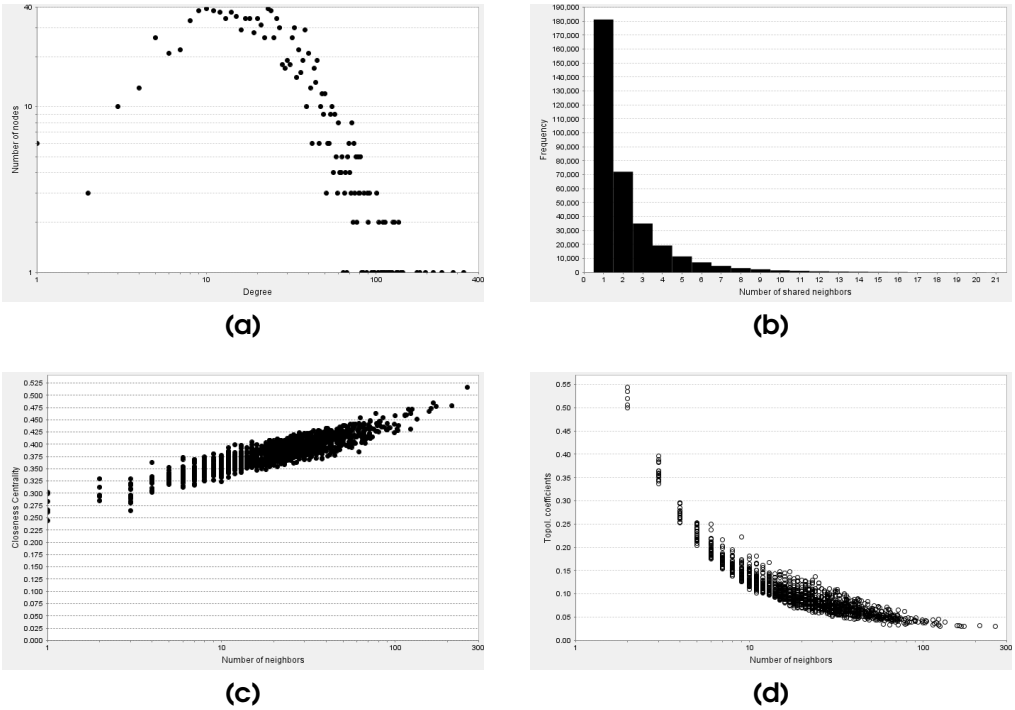


Figure 3.11: Structural properties of association network obtained for bigger dataset. a) A power-law like distribution obtained for the node degree distribution. b) A distribution of number of partners shared between a pair of nodes c) Closeness centrality of all the nodes d) Plot for topological coefficient.

to present the results in an easy-to-view graphical format. We apply two widely used techniques to establish association between the pairs of genes in the dataset. The association between genes are measured using a) Pearson correlation coefficient, and b)Euclidean distance. The genes in the dataset were arranged in an ordered fashion before computing the association between them, i.e., the first 30 in the dataset of 85 genes performed circadian rhythm related activity, the next 34 genes were associated with ageing, and the last 21 genes participated in cell death. Figure 3.17 and Figure 3.18 present the graphical representation of the association matrices obtained for the gene pairs using correlation coefficient and Euclidean distance respectively. Each cell in an association matrix is filled with a colour based on the quantitative entry in that cell. The mapping of colours with the magnitudes of cells is displayed by the colour-bars in the figures. We

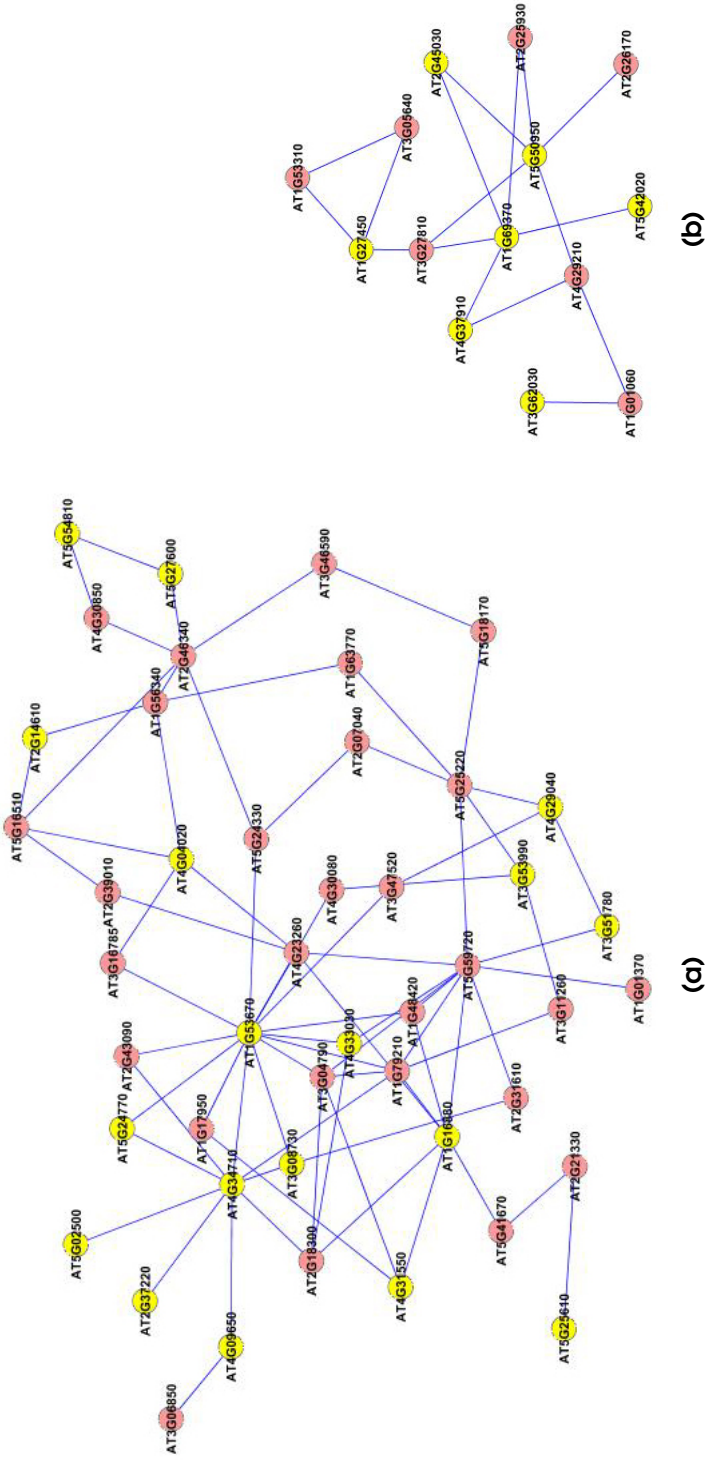


Figure 3.12: Modules 1-2 : Genes listed in Table 3.2 for network 1 (a. Response to stress) and network 2 (b. Cytoplast) are highlighted with yellow.

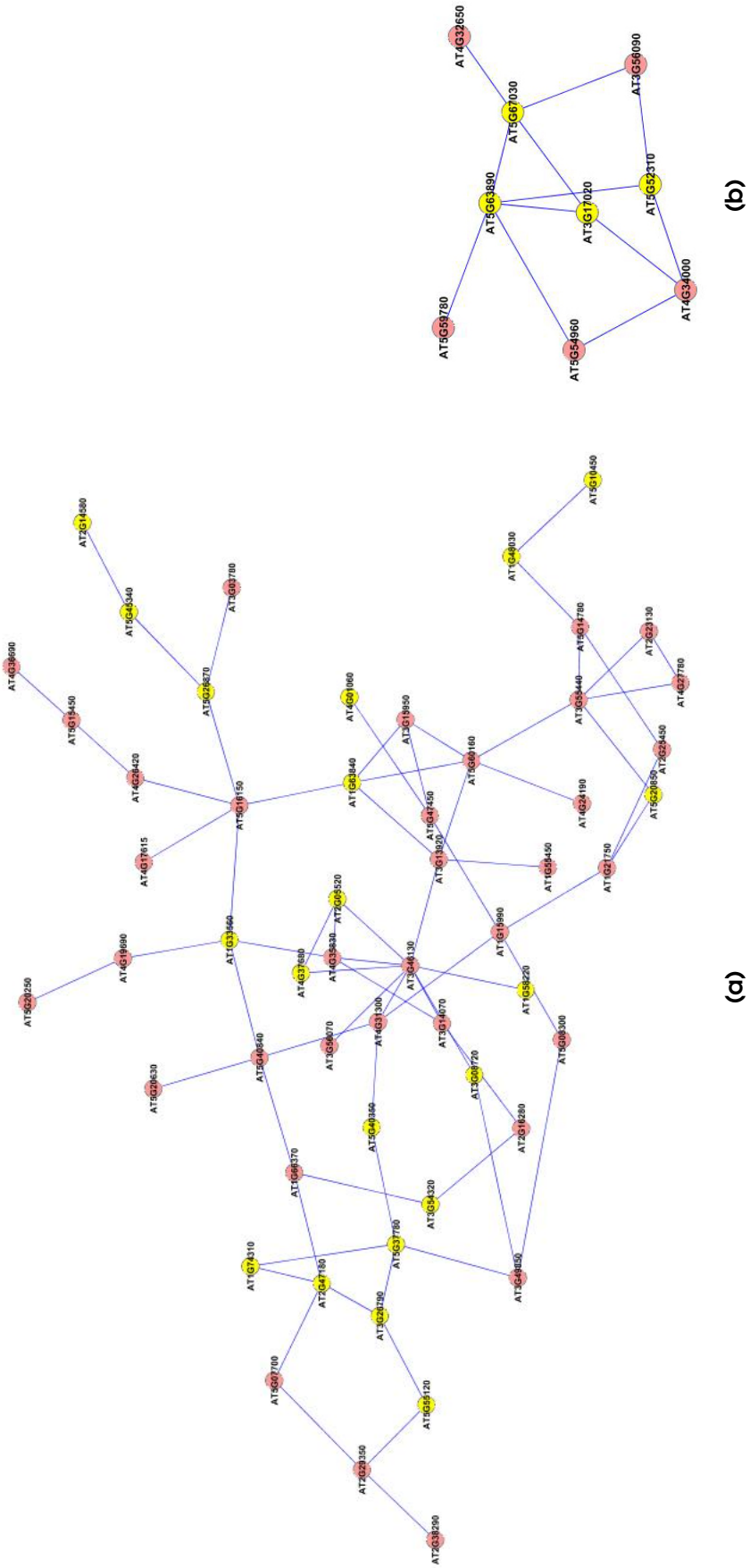


Figure 3.13: Modules 3-4 :Genes listed in Table 3.2 for network 3 (a. Response to stimulus) and network 4 (b. Response to abiotic stimulus) are highlighted with yellow.

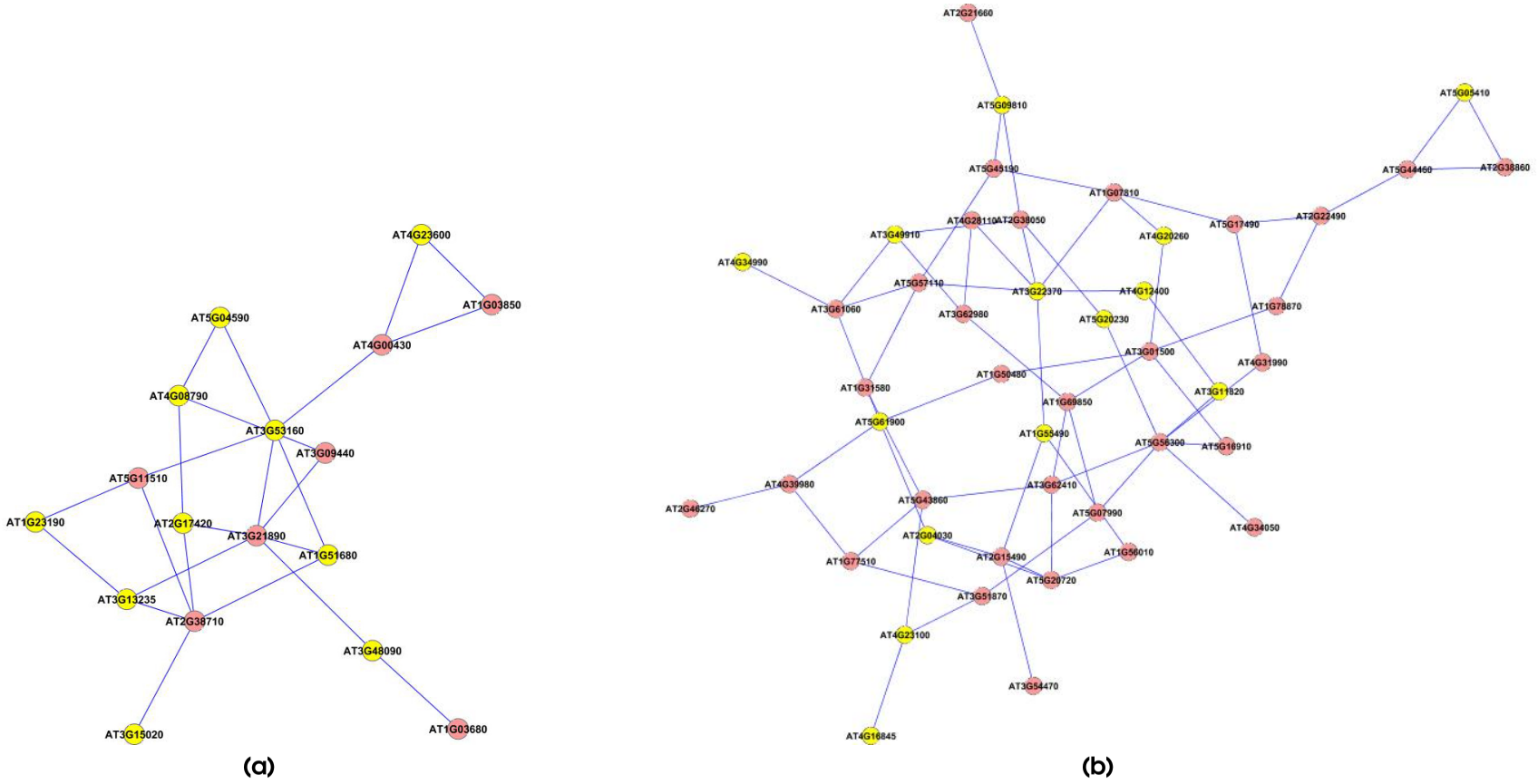
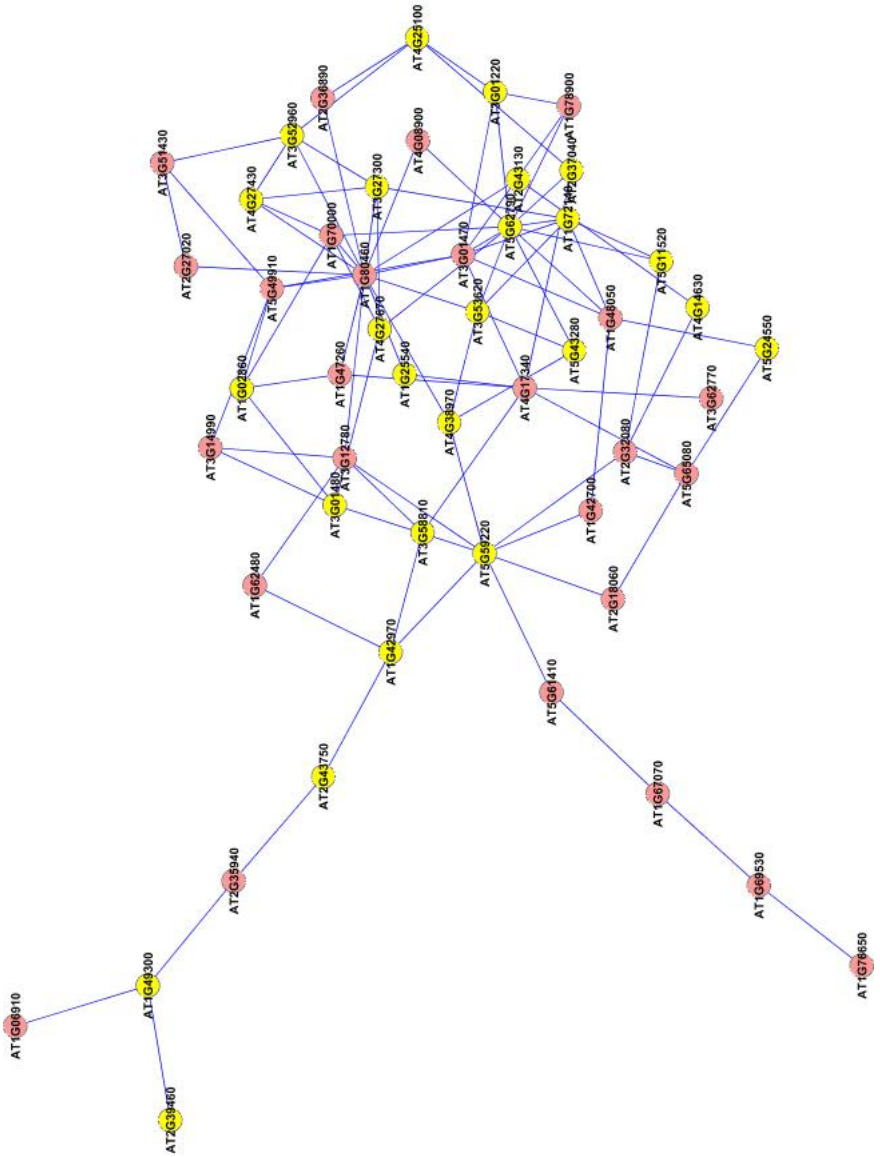


Figure 3.14: Modules 5-6 :Genes listed in Table 3.2 for network 5 (a. Catalytic Activity) and network 6 (b. Response to stress) are highlighted with yellow.



(a)

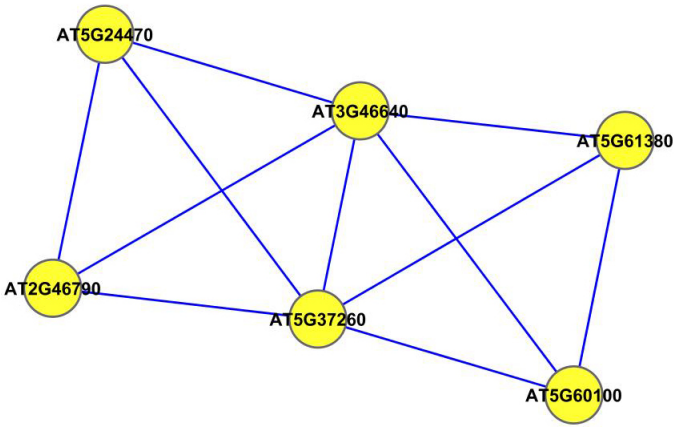
Figure 3.15: Modules 7: Genes listed in Table 3.2 for network 7 (Cell part) are highlighted with yellow.

can see that the colour coding starts from blue (for low magnitude of association) to red (for high magnitude of association). The strongly associated gene pairs are represented by shades of red in their respective cells. The diagonal entries of both the association matrices are drawn in dark red, indicating maximum degree of association between self-to-self pair. The association matrices are symmetric, thus, the inspection of only the lower diagonal entries should suffice in detection of strongly associated gene pairs.

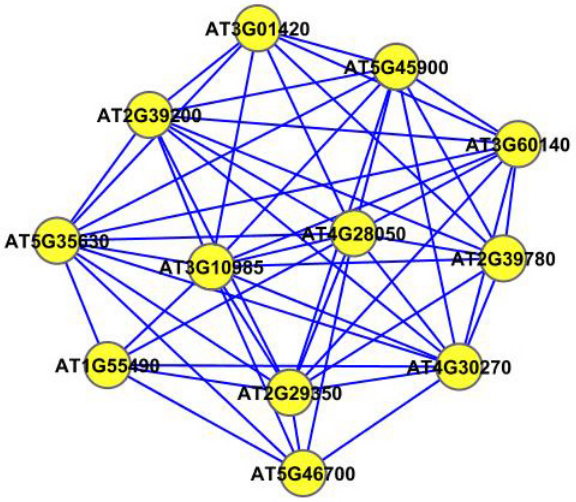
In an ideal scenario, where the genes performing similar activity group together, we expect three distinct regions in Figures 3.17 and 3.18. The lower diagonal blocks from cell 1 to 30, cell 31 to 64, and cell 65 to 85 should indicate a high degree of intra-block association, each block should be coloured in different shades of red according to the colour-magnitude mapping shown in the colour-bars. But, this is not the case in the figures obtained by us where we can see no clear blocks in the figures. The lack of any block-wise patterns in the colour coded cells of association matrices indicate the absence of strong associative information between genes based on the measures discussed above. This is the first indication that the measures like correlation and Euclidean distance may not be suitable for our dataset.

To investigate further, we applied a threshold to keep the strongest edges in the graphs obtained from the association matrices. The criteria to choose the threshold was same as the one used in section 3.2.2 for selecting the strong edges in the graphs. The filtered graphs were analysed using the graph-theoretic technique with the similar settings as used in case of smaller Arabidopsis dataset in section 3.2.2. The correlation based associative graph resulted in two subgraphs shown in Figure 3.16, whereas the euclidean distance based graph did not yield

any subgraph at all. The gene ontology analysis of the two subgraphs shown in Figure 3.16 is presented in Table 3.3. We can see that in Network 1, three out of total six genes belonged to rhythmic process related activity, whereas, in Network 2, five out of a total of twelve genes belonged to ageing process. These networks and their related biological relevance are much inferior compared to the subgraphs obtained in section 3.2.2 using our technique, where we obtained 4 distinct subgraphs with distinct biological functions and better gene ontology results.



(a) Network 1



(b) Network 2

Figure 3.16: Two subgraphs of potential interest were detected when correlation coefficient was used to establish association between genes in the smaller Arabidopsis dataset.

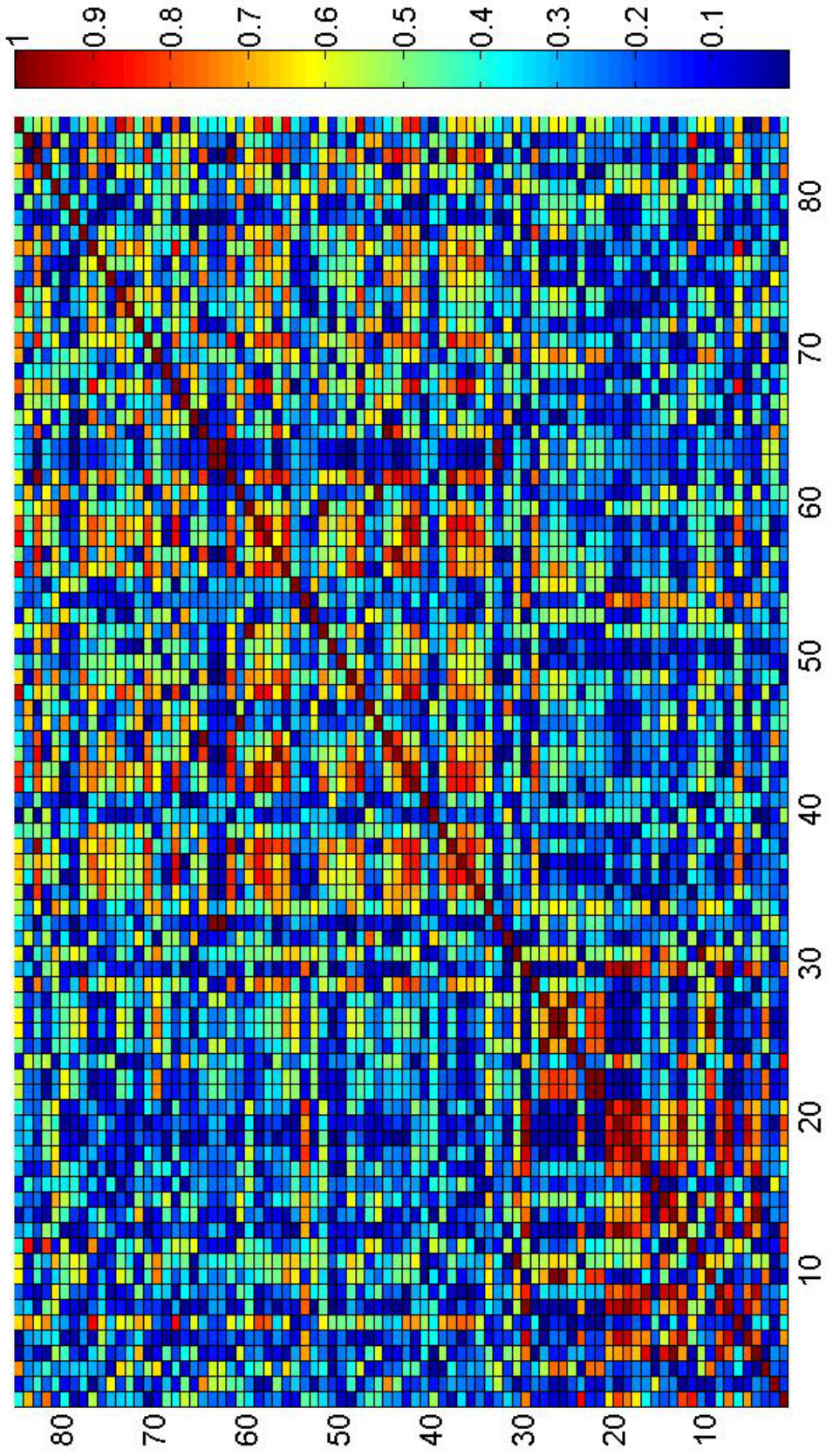


Figure 3.17: Correlation matrix for smaller Arabidopsis dataset

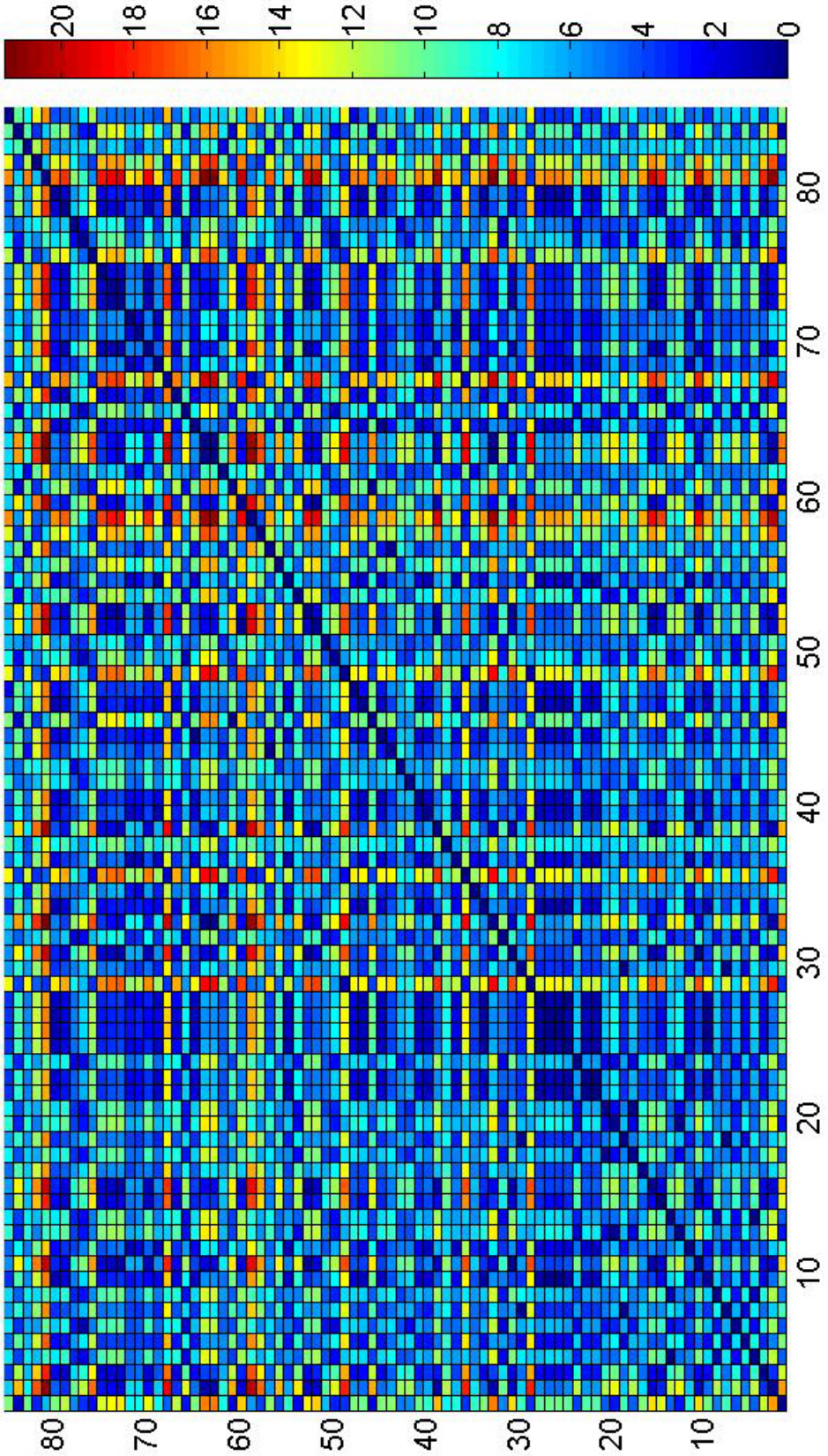


Figure 3.18: Distance matrix for smaller Arabidopsis dataset

GO-ID	*p*-value	corr *p*-value	Known/Total	Functional Description	Gene Names	
Network 1						
48511	4.0370E-9	8.2759E-8	3/4	Rhythmic process	AT5G24470 AT5G61380	AT2G46790
Network 2						
16280	2.8611E-12	2.0600E-10	5/8	Ageing	AT5G45900 AT5G35630 AT4G28050	AT2G29350 AT3G10985

Table 3.3: GO annotations for clusters found in the smaller Arabidopsis dataset using correlation as the measure of association between genes

We have used a fresh and distinct approach to cluster temporal microarray gene expression data. One of the key questions that we have tried to address using this method is that how some variables are useful for *forecasting* others. The proposed method facilitates a way to study such forecasting relationships between two variables. In other words, we are asking if a variable X can predict another variable Y. Equivalently, we can say if X is exogenous in time-series sense with respect to Y or not. Yet a third expression meaning the same thing is, if X is linearly informative about *future* Y. The basic idea behind this method is, if an event X causes another event Y, then X should *precede* Y in time. This is why our illustrative models are based on time, and within that time frame the lags like $t-1$,$t-2$,$\cdots$ etc. denote the temporal association within the processes.

While discussing widely used pairwise association methods for clustering, like any form of correlation or distance based methods, the time is static. In these methods, the time does not play any role. The core of these methods rely on association rather than prediction. So if we re-order the sequence of observations for any pair of variables (X, Y), the association measure between them does not change. As for example, let the original observation be $X = \{x_{t-1}, x_{t-2}, x_{t-3}\}$ and $Y = \{y_{t-1}, y_{t-2}, y_{t-3}\}$. The Association measure using correlation/distance for $(X, Y) = C$. After reordering of the observations, let $X' = \{x_{t-3}, x_{t-1}, x_{t-2}\}$ and $Y' = \{y_{t-3}, y_{t-1}, y_{t-2}\}$. The new association measure using correlation/distance

for $(X', Y') = C'$ where $C = C'$. Hence, this assumption is not suitable for dynamical systems. This is the reason why the usual pairwise association methods can give us less reliable results than the ones by our method. And hence a comparison between the two methods will not be fair. There has been some work in model based clustering methods based on Bayesian statistics where the dynamics of profiles (modelled as regressive processes) have been used to create clusters [RK02, ACDC$^+$08]. Such methods are different from our approach as, first, our approach is based on the frequentist approach rather than the Bayesian approach, and second, the essence of our approach lies in detecting the causal association between genes. Another important aspect to consider is the choice of time lag in our method which is decided using the AIC criteria. The lag value is not fixed, but is chosen iteratively for each individual pair (X, Y) according to what describes the variables best.

We have demonstrated the performance of the method using various artificial datasets and examples from real biological datasets. It is easy to see that the pair-wise association based techniques, like distance or correlation based measures, would not work as desired, when we are investigating a system where the interaction with respect to time is an important concept.

3.4 Summary

Clustering helps in reducing the data dimensions by grouping genes with similar profiles or similar functionalities. In this chapter, we proposed a clustering method to group functionally related genes in a temporal microarray dataset. Our method exploits the temporal interdependence between genes. The interdependence was determined using the test of Granger causality between two time series. The method is simple in its implementation, and testable at every stage. The as-

sociation graph representing the dependence between genes is further analysed using a graph-theoretic method. The graph-theoretic method detects the dense regions in the graph which could represent biological complexes or motifs. We test our approach using a set of artificial datasets and two datasets of different sizes belonging to the Arabidopsis experiment. The functional similarity between genes belonging to detected clusters was verified using the publicly available gene ontology database. We further analysed the structural properties of the association network obtained for the larger of the two datasets for Arabidopsis. We show using different network characteristics that the computed association network is not a random network in its structure, and has the properties expected in a real biological network.

There are a few considerations which should be taken into account while applying this approach. The data must be cleaned properly using appropriate normalization method to remove unwanted experimental biases. We discussed the normalization technique applied for our dataset in Chapter 2. For any time-series based statistical method, it is important that the data has been collected at intervals which capture the natural changes in the system. Selection of correct lag order using an information based criterion is also important as the test of Granger causality is strongly dependent on that decision. Most important of all, the experimental design should be able to support the hypothesis of the practitioner. Further care should be taken while discovering *directional* causal links using Granger causality. We will see more of this in the next chapter.

In this chapter, our effort was not to detect a *causal* network structure from gene data, but to find a suitable association matrix based on interactions between them. Once the interesting modules have been found, different reverse engineering

methods like Bayesian networks, Structural equations etc. can be applied to infer causal networks from selected genes of interest. It should be noted that in order to detect causal interactions between genes, the multivariate approach instead of pairwise one, can give better results while *re-engineering a causal network structure* from data. The next chapter discusses an extension of Granger causality presented in this chapter and demonstrates how it can be used in a multivariate context for reverse engineering of gene circuits. It should be noted that multivariate reverse engineering approaches are only possible for a small number of genes due to their massive computational requirements.

Chapter 4
Partial Granger Causality

Reconstructing gene-regulatory networks is one of the key problems of functional genomics [VSWBR02, Kit02]. A gene network can be visualized as a graph in which each node represents a gene and the interactions between them are represented by the edges in the graph. The edges can represent direct or indirect interactions between the genes. Large scale monitoring of gene expression is considered to be one of the most promising techniques for reconstructing gene networks [Ber01]. Microarrays generate abundant data which could be used for reconstructing gene networks. Inference of gene networks is also the last stage in the information processing pipeline of microarray data analysis as explained in section 1.5 of the Introduction chapter. Once the data has been properly normalized and genes of interest selected, a reverse engineering approach can be applied to understand the functional connectivity between genes. This computational approach can help us understand the underlying biological pathways and their patterns of connectivities. Mapping of gene pathways typically involves inferences arising from various studies performed on individual pathway components. Although pathways are often conceptualized as distinct entities, it is often understood that inter-pathway cross-talk and other properties of networks reflect underlying complexities that cannot by explained by consideration of individual

pathways in isolation. In order to consider interaction between individual paths, a global multivariate approach is required. A variety of approaches have been proposed to describe gene-regulatory networks, such as Boolean networks [Kau93], Difference equations [VSWBR02], Differential equations [YTC02] and Bayesian networks [FLNP00, PREF01] etc. While Boolean networks, Difference and Differential equations are based on prior biological understanding of the molecular mechanism, Bayesian networks on the other hand have been used to infer network structures directly from the data itself. The acyclicity constraint of the Bayesian networks are addressed by Dynamic Bayesian networks [DGM+06, KIM03] but computational and theoretical problems arise in the case of an incomplete dataset which is a common problem in gene expression measurements. Relevance networks and Gaussian graphical models [WGH06] are other commonly used methods to infer network structures from time-series data, both being simple but incapable of producing directed network structures. With each approach having its advantages and disadvantages, the field of inference of network structure from gene-expression data is still open to new techniques.

In this chapter, we present a gene network reconstruction technique based on the idea of Granger causality discussed in Chapter 3. We discuss an extension of the method which can be used to infer the interaction patterns among multiple time series representing a set of stochastic processes. As discussed in the earlier chapter, the proposed technique relies on the statistical interdependence among multiple simultaneous time series. The interdependence between a pair of time-series could be causal in nature and therefore symmetric measures may not be suitable for measuring it. Wiener [Wie56] proposed a new way to measure causal influence of one time series on another by conceiving the notion that the prediction of one time series could be improved by incorporating the knowledge

of the second one. Granger [Gra69] formalized this concept in the context of the linear autoregression model of causal influences which we discussed in Chapter 3. Granger causality was extended by Geweke [Gew82] who proposed a *measure* of interdependence between two sets of time series. We have seen a recent interest in biological community regarding application of Granger causality [MC07, NU08] for temporal Microarray data. But we realize that a straightforward application of Granger causality for biological data may not be suitable when the chances of latent and exogenous variables present in the system are high. Also, a pairwise detection of causal links can lead to discovery of redundant connections in the inferred network.

In this chapter, we introduce a definition of *Partial Granger Causality (PGC)*. Partial Granger causality computes the interdependence between two time series by eliminating the effect of all other variables in the system. The proposed idea of Partial Granger Causality is tested for various toy models representing different scenarios of interdependence between sets of time series. We then apply this approach to a highly replicated microarray time series data for T-cell activation to infer the gene network.

4.1 Methods

4.1.1 Measures of Linear Interdependence

Geweke proposed a measure to *quantify* the interdependence between two sets of stochastic processes based on the definition of causality proposed by Granger [Gew82]. To explain the method, we focus on a bivariate stochastic system with X and Y as its members. Expressing the processes in their autoregressive form

similar to Equation (3.4), we have

$$X_t = \sum_{i=1}^{\infty} a_{1i} X_{t-i} + \epsilon_{1t} \tag{4.1}$$

$$Y_t = \sum_{i=1}^{\infty} b_{1i} Y_{t-i} + \epsilon_{2t} \tag{4.2}$$

where ϵ_{1t} and ϵ_{2t} are the prediction errors. A joint autoregressive representation having information of the past measurements of both processes X and Y can be written as

$$X_t = \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} c_{2i} Y_{t-i} + \epsilon_{3t} \tag{4.3}$$

$$Y_t = \sum_{i=1}^{\infty} b_{2i} Y_{t-i} + \sum_{i=1}^{\infty} d_{2i} X_{t-i} + \epsilon_{4t} \tag{4.4}$$

Equation (4.3) represents the prediction of the current value of X based on its own past values as well as the past values of Y. The variance $\sigma^2(\epsilon_{3t})$ measures the strength of the prediction error. Granger causality suggests that if the prediction of one process is improved by incorporating its own past values and the past information of the other process, then the second process is said to Granger cause the first process. In other words, if the variance of the prediction error for the first process is reduced by the inclusion of past measurements of the second process, then a causal relationship between the two processes exist. According to Geweke's decomposition of causality measure, the causal influence from Y to X where $\sigma^2(\epsilon_{3t}) < \sigma^2(\epsilon_{1t})$ can be expressed as

$$F_{Y \to X} = \ln \left(\frac{|\sigma^2(\epsilon_{1t})|}{|\sigma^2(\epsilon_{3t})|} \right) \tag{4.5}$$

If $F_{Y \to X} > 0$, then $Y \to X$ exists. On the parallel lines, the causal influence

from X to Y is defined as

$$F_{X\to Y} = \ln\left(\frac{|\sigma^2(\epsilon_{2t})|}{|\sigma^2(\epsilon_{4t})|}\right) \tag{4.6}$$

The third kind of interdependence between X and Y is due to the factors possibly exogenous to (X, Y) system, and is termed as *instantaneous causality*, where $\gamma = \sigma(\epsilon_{3t}, \epsilon_{4t}) \neq 0$. The instantaneous causality can be expressed as

$$F_{X.Y} = \ln\left(\frac{|\sigma^2(\epsilon_{3t})|.|\sigma^2(\epsilon_{4t})|}{|L|}\right) \tag{4.7}$$

where

$$\mathrm{L} = \begin{bmatrix} \sigma^2(\epsilon_{3t}) & \gamma \\ \gamma & \sigma^2(\epsilon_{4t}) \end{bmatrix}$$

When $\gamma = 0$, $F_{X.Y} = 0$, no instantaneous causality exists. But when $\gamma^2 > 0$, then $F_{X.Y} > 0$ and the instantaneous causality exists.

The above definitions imply that the total interdependence between two time series X and Y can be defined as

$$F_{X,Y} = F_{X\to Y} + F_{Y\to X} + F_{X.Y} \tag{4.8}$$

Thus, the total interdependence between two time-series X and Y can be decomposed into three components: two directional causal influences between X and Y, and the instantaneous causality between the two.

4.1.2 Partial Granger Causality

In a network with multiple nodes, where each node represents a stochastic process, various possibilities of interdependence among them may arise. Such interdependences can be denoted by the edges in the network. Any two entities in the

network can be connected in either a direct way or in an indirect way. While inferring the network structure from data, it is important that only the direct influences of interactions are considered while drawing an edge between any two entities. This issue is of important concern for network inference in order to filter out redundant channels. Apart from that, there could be presence of exogenous inputs and latent variables in the system. Exogenous variables represent the common experimental drives present in any experimental setup, whereas, the latent variables account for the unobserved or hidden data which could not be captured during the experiment. The above definitions of directional causality apply to only two variables. But while considering a network of multiple variables, a multivariate approach is desirable. The benefit of multivariate model fitting is that it uses information from all the participating entities in the system, making it possible to verify whether two entities share direct causal influence while the effect of other entities are taken into account. Also, the pairwise analysis of two time series is not sufficient to reveal if the causal relationship between a pair is direct or not. We can assume that both exogenous and latent variables apply a common input in the current time as well as in past to all the observed variables. The above definitions of causality can be extended in a multivariate context to deal with the effects of common inputs and all other observable variables while measuring the directional influence between any two variables. We call this method *Partial Granger Causality (PGC)* . In the proposed definition of Partial Granger Causality, we compute the linear dependence between two entities by *eliminating* the effect of all other variables. As a result of this elimination, it is possible to compute the strength of direct interaction between two entities in a system. As an extension to the bivariate systems described in the previous section, we propose the definition of Partial Granger Causality in this section.

The basic idea of Partial Granger Causality comes from the definition of Partial Correlation [JW88]. Partial Correlation between two response variables Y_1 and Y_2 after eliminating the effects of other predictor variables $Z_1, \ldots, Z_r$ can be understood in the following way. Consider the grouping of variables such that the response variables are represented by vector $\boldsymbol{Y} = \{Y_1, Y_2\}$ and the predictor variables by the vector $\boldsymbol{Z} = \{Z_1, \ldots, Z_r\}$. The covariance matrix after performing the multivariate regression to predict $\boldsymbol{Y}$ from $\boldsymbol{Z}$ can be represented as $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ can be partitioned as

$$\left[\begin{array}{c|c} \boldsymbol{\Sigma_{YY}} & \boldsymbol{\Sigma_{YZ}} \\ \hline \boldsymbol{\Sigma_{ZY}} & \boldsymbol{\Sigma_{ZZ}} \end{array}\right]$$

Representing the identity matrix as $\boldsymbol{I}$, the matrix $\boldsymbol{\Sigma}$ can alternatively be represented by the following algebraic manipulation -

$$\boldsymbol{\Sigma} = \boldsymbol{I\Sigma I} = \left[\begin{array}{c|c} \boldsymbol{I} & -\boldsymbol{\Sigma_{YZ}}\boldsymbol{\Sigma_{ZZ}}^{-1} \\ \hline \boldsymbol{0} & \boldsymbol{I} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma_{YY}} & \boldsymbol{\Sigma_{YZ}} \\ \hline \boldsymbol{\Sigma_{ZY}} & \boldsymbol{\Sigma_{ZZ}} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{I} & \boldsymbol{0} \\ \hline -\boldsymbol{\Sigma_{YZ}}\boldsymbol{\Sigma_{ZZ}}^{-1} & \boldsymbol{I} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YZ}}\boldsymbol{\Sigma_{ZZ}}^{-1}\boldsymbol{\Sigma_{ZY}} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Sigma_{ZZ}} \end{array}\right]$$

where $\boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YZ}}\boldsymbol{\Sigma_{ZZ}}^{-1}\boldsymbol{\Sigma_{ZY}}$ represents the association between Y_1 and Y_2 after eliminating the effects of $Z_1, Z_2, \ldots, Z_r$. This formulation leads us to the definition of Partial Granger Causality, which can be understood in the following way

Consider two processes X and Z. The joint autoregressive representation for

X and Z can be written as

$$X_t = \sum_{i=1}^{\infty} a_{1i} X_{t-i} + \sum_{i=1}^{\infty} c_{1i} Z_{t-i} + \epsilon_{1t} \tag{4.9}$$

$$Z_t = \sum_{i=1}^{\infty} b_{1i} Z_{t-i} + \sum_{i=1}^{\infty} d_{1i} X_{t-i} + \epsilon_{2t} \tag{4.10}$$

The noise covariance matrix for the system can be represented as

$$\mathrm{S} = \begin{bmatrix} \sigma^2(\epsilon_{1t}) & \sigma(\epsilon_{1t}, \epsilon_{2t}) \\ \sigma(\epsilon_{1t}, \epsilon_{2t}) & \sigma^2(\epsilon_{2t}) \end{bmatrix}$$

Extending this concept further, the vector autoregressive (VAR) representation for a system involving three processes X,Y and Z can be written in the following way.

$$X_t = \sum_{i=1}^{\infty} a_{2i} X_{t-i} + \sum_{i=1}^{\infty} b_{2i} Y_{t-i} + \sum_{i=1}^{\infty} c_{2i} Z_{t-i} + \epsilon_{3t} \tag{4.11}$$

$$Y_t = \sum_{i=1}^{\infty} d_{2i} X_{t-i} + \sum_{i=1}^{\infty} e_{2i} Y_{t-i} + \sum_{i=1}^{\infty} f_{2i} Z_{t-i} + \epsilon_{4t} \tag{4.12}$$

$$Z_t = \sum_{i=1}^{\infty} g_{2i} X_{t-i} + \sum_{i=1}^{\infty} h_{2i} Y_{t-i} + \sum_{i=1}^{\infty} k_{2i} Z_{t-i} + \epsilon_{5t} \tag{4.13}$$

The noise covariance matrix for the above system can be represented as

$$\Sigma = \begin{bmatrix} \sigma^2(\epsilon_{3t}) & \sigma(\epsilon_{3t}, \epsilon_{4t}) & \sigma(\epsilon_{3t}, \epsilon_{5t}) \\ \sigma(\epsilon_{3t}, \epsilon_{4t}) & \sigma^2(\epsilon_{4t}) & \sigma(\epsilon_{4t}, \epsilon_{5t}) \\ \sigma(\epsilon_{3t}, \epsilon_{5t}) & \sigma(\epsilon_{4t}, \epsilon_{5t}) & \sigma^2(\epsilon_{5t}) \end{bmatrix}$$

The Partial Granger Causality between X and Y by eliminating all the effect of Z, can be calculated by partitioning the noise covariance matrices S and Σ in the following way -

$$\mathrm{S} = \left[\begin{array}{c|c} \sigma^2(\epsilon_{1t}) & \sigma(\epsilon_{1t}, \epsilon_{2t}) \\ \hline \sigma(\epsilon_{1t}, \epsilon_{2t}) & \sigma^2(\epsilon_{2t}) \end{array}\right] = \left[\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right]$$

$$\Sigma = \left[\begin{array}{c|c} \sigma^2(\epsilon_{3t}) & \sigma(\epsilon_{3t}, \epsilon_{5t}) \\ \hline \sigma(\epsilon_{3t}, \epsilon_{5t}) & \sigma^2(\epsilon_{5t)} \end{array}\right] = \left[\begin{array}{c|c} \Sigma_{XY} & \Sigma_{XYZ} \\ \hline \Sigma_{ZXY} & \Sigma_{ZZ} \end{array}\right]$$

The measure for Partial Granger Causality from Y to X by eliminating the effect of Z can be expressed as

$$F_{Y \to X|Z} = \ln \left(\frac{|S_{11} - S_{12} S_{22}^{-1} S_{21}|}{|\Sigma_{XY} - \Sigma_{XYZ} \Sigma_{ZZ}^{-1} \Sigma_{ZXY}|} \right) \tag{4.14}$$

Based on the above formulation, we demonstrate in the following sections that Partial Granger Causality is a good tool for inferring a network structure from a given set of time series data. An equivalent representation of Partial Granger Causality exists in the frequency domain as well, as proposed by Guo et al. [GWDF08] . The main results are presented in Appendix A.

4.1.3 Prerequisites For Causal Models

Stationary time series: Measurement of linear dependence between multiple time series assumes the time series to be stationary. We assume our time-series to be weakly stationary.

Linear independence among entities: Before fitting an autoregressive model on a set of processes, it is important to check that the processes are linearly independent. The check ensures that the fluctuation in the estimate of one parameter will not be compensated by the fluctuations in the estimate of other parameters. To check for linear independence among p variables, a sample variance-covariance matrix S can be calculated, which contains p variances and $\frac{1}{2}p(p-1)$ potentially different covariances. The determinant of S provides a generalized sample variance, and is equal to zero in case of linear dependence between the variables. In

the case of linear dependence among variables, some of the variables should be removed from the sample [JW88].

Selection of lag order: The above definitions are dependent on a choice of appropriate lag values for variables. A lag-value p which minimizes the Akaike Information Criterion (AIC, [Aka69]) according to Equation (3.8) can be used in estimation of OLS regressions.

4.1.4 Bootstrap Analysis

To construct a confidence interval for every edge present in the network, it is important to estimate the distribution of the PGC values between different pairs of entities in a network. The confidence interval can be used as a statistical measure to separate relevant edges from the pool of all possible edges in the network. The distribution of the PGC values in a network is determined by the bootstrap method. Consider a set of variables $Y = \{E_1, E_2, \ldots, E_N\}$, where each E_i is a time series of equal length. The PGC value between any two variables can be denoted as f_i which can be computed according to Equation (4.14). The set of all possible PGC values for all the possible p pairs of variables in Y can be denoted as $F = \{f_1, f_2, \ldots, f_p\}$. The following procedure can be applied to compute a bootstrap confidence interval for F using the 3σ method.

- Multiple samples of data for the system Y can be generated to create a bootstrap sample $B = \{Y_1^*, Y_2^*, \ldots, Y_L^*\}$.

- For each Y_i^* in B, compute PGC values F_i^*. This will give bootstrap estimates $F_1^*, F_2^*, \ldots, F_L^*$ for the PGC values obtained from the bootstrap sample B.

- A standard deviation σ_i^* for each f_i in F can be computed by the distribution of corresponding f_i^* values in $F_1^*, F_2^*, \ldots, F_L^*$.

- For 99.7% confidence level, obtain lower bound(lb) and upper bound(ub) for each f_i.

$$(lb, ub) = \{f_i - 3 \times \sigma_i^*, f_i + 3 \times \sigma_i^*\}$$

- Test the null hypothesis that the f_i value is significant by rejecting the null hypothesis if the confidence interval does not contain the value 0. So, the edges having their f_i values in F are accepted to appear in the network whose lb >0. The rest of the edges are supposed to be absent.

4.2 Results

4.2.1 Illustrative Examples

We demonstrate the concept of Partial Granger Causality for network inference with the following toy models. These models have been used earlier in the literature [BS01]. A Matlab routine was developed to compute the PGC values and was tested on the following examples.

Each example has a set of 5 time-series where each time-series represents a node in the interaction network. For a complete graph of 5 nodes, where each node is connected to every other node in the network, there are at the most 20 possible *directed* edges. See Table 4.1 for edge enumeration for all the directed edges. We computed PGC values for all the node pairs (X, Y) forming an edge in the complete graph for both the directions ($X \rightarrow Y$ and $Y \rightarrow X$) by eliminating the effects due to all other nodes and common input present in the system. The magnitude of PGC for each directed edge represents the weight associated with that edge. The confidence interval for each edge was constructed using the bootstrap criteria for 2000 generated datasets for each example.

Example 1: Consider a set of 5 simultaneously generated time-series according to the following equations:

$$
\begin{aligned}
x_1(n) &= 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + w_1(n) \\
x_2(n) &= 0.5x_1(n-2) + w_2(n) \\
x_3(n) &= -0.4x_1(n-3) + w_3(n) \\
x_4(n) &= -0.5x_1(n-2) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\
x_5(n) &= -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5(n)
\end{aligned}
$$

where $w_i(n)$ are zero-mean uncorrelated white processes with identical variance. One can see that $x_1(n)$ is a direct source to $x_2(n)$,$x_3(n)$, and $x_4(n)$. $x_4(n)$ and $x_5(n)$ share a feedback loop and there is no direct connection from $x_1(n)$ to $x_5(n)$. The final network structure obtained after computing the PGC values and the confidence intervals on each edge can be seen in Figure 4.1a. Figure 4.2a plots the magnitude of PGC values computed for each edge in the network for a dataset representing Example 1. We can see that the PGC values for edges 10 ($4 \leftarrow 5$), 11 ($1 \rightarrow 2$), 12 ($1 \rightarrow 3$) , 13 ($1 \rightarrow 4$) and 20($4 \rightarrow 5$) are comparatively higher than the PGC values for the other edges in the network. Figure 4.3a represents the inferred edges for 20 different datasets representing the above set of equations. The x-axis in Figure 4.3a represents the edge numbers according to Table 4.1 and the y-axis represents the dataset numbers. The bright boxes in the figure represent the selected edges after computing confidence intervals for PGC values. We can see that for most of the datasets, the selected edges are edge numbers 10,11,12,13 and 20.

Example 2: The system in Example 1 is modified where $x_1(n)$ influences $x_2(n)$,

which in turn affects $x_3(n)$ and then finally couples to $x_4(n)$. $x_4(n)$ and $x_5(n)$ share a feedback loop in same way as in the previous example. The modified system of equations can be represented by the following equations.

$$\begin{aligned}
x_1(n) &= 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + w_1(n) \\
x_2(n) &= -0.5x_1(n-2) + w_2(n) \\
x_3(n) &= 0.4x_2(n-3) + w_3(n) \\
x_4(n) &= -0.5x_3(n-2) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\
x_5(n) &= -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5(n)
\end{aligned}$$

$w_i(n)$ are again zero-mean uncorrelated white processes with identical variance. The network structure obtained after applying our method is shown in Figure 4.1b. Figure 4.2b plots the magnitude of PGC values for each edge in the network for a dataset for this example. Figure 4.3b shows the selected edges for 20 such datasets generated for this example.

Example 3: The system in Example 2 is further modified and a direct connection from $x_5(n)$ to $x_1(n)$ is formed. Following set of equations represent the modified system.

$$\begin{aligned}
x_1(n) &= 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + 0.5x_5(n-2) + w_1(n) \\
x_2(n) &= -0.5x_1(n-2) + w_2(n) \\
x_3(n) &= 0.4x_2(n-3) + w_3(n) \\
x_4(n) &= -0.5x_3(n-2) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\
x_5(n) &= -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5(n)
\end{aligned}$$

The network structure found after computing PGC values for Example 3 is

Edge Number	Edge	Edge Number	Edge
1	$(1 \leftarrow 2)$	11	$(1 \rightarrow 2)$
2	$(1 \leftarrow 3)$	12	$(1 \rightarrow 3)$
3	$(1 \leftarrow 4)$	13	$(1 \rightarrow 4)$
4	$(1 \leftarrow 5)$	14	$(1 \rightarrow 5)$
5	$(2 \leftarrow 3)$	15	$(2 \rightarrow 3)$
6	$(2 \leftarrow 4)$	16	$(2 \rightarrow 4)$
7	$(2 \leftarrow 5)$	17	$(2 \rightarrow 5)$
8	$(3 \leftarrow 4)$	18	$(3 \rightarrow 4)$
9	$(3 \leftarrow 5)$	19	$(3 \rightarrow 5)$
10	$(4 \leftarrow 5)$	20	$(4 \rightarrow 5)$

Table 4.1: Enumeration of all the directed edges in the toy examples.

shown in Figure 4.1c. The PGC values for inferred edges obtained for a dataset are shown in Figure 4.2c. Figure 4.3c shows the selected edges for 20 such datasets generated for this example.

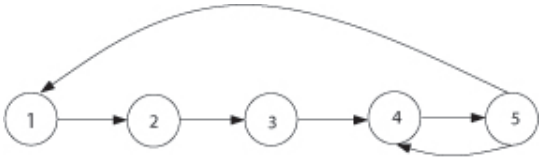
Example 4: The system in Example 2 is modified where $x_1(n)$ connects to $x_4(n)$ via two distinct pathways, through $x_2(n)$ and $x_3(n)$ respectively.

$$
\begin{aligned}
x_1(n) &= 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + w_1(n) \\
x_2(n) &= -0.5x_1(n-2) + w_2(n) \\
x_3(n) &= 0.5x_1(n-3) - 0.4x_2(n-2) + w_3(n) \\
x_4(n) &= -0.5x_3(n-1) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\
x_5(n) &= -0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_5(n)
\end{aligned}
$$

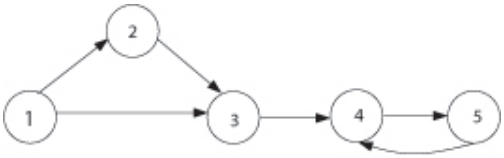
The network structure found after computing PGC values for Example 4 is shown in Figure 4.1d. The magnitudes of PGC values computed for a dataset for this example are plotted in Figure 4.2d, and the inferred edges for multiple datasets are shown in Figure 4.3d.

(a) Example 1

(b) Example 2

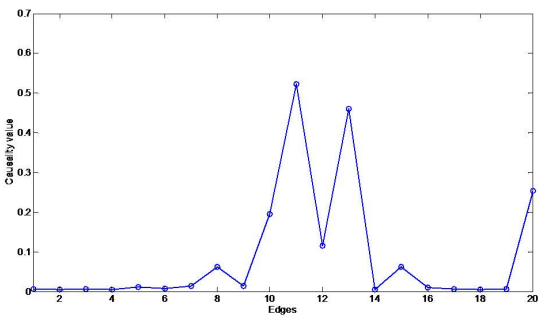


(c) Example 3

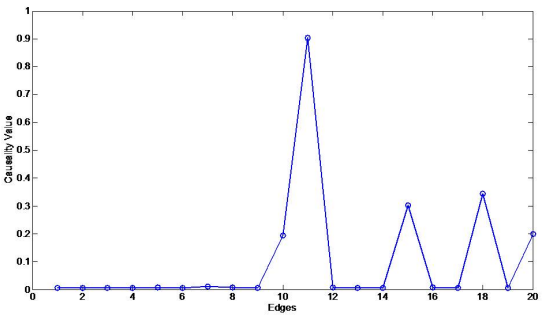


(d) Example 4

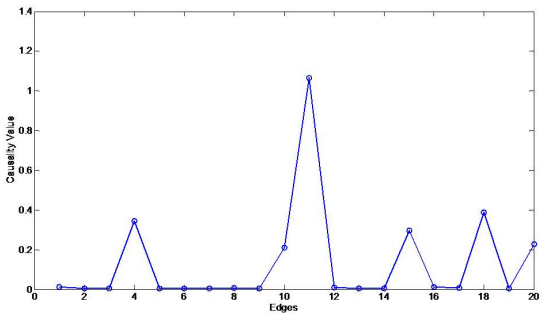
Figure 4.1: Network structures for the discussed examples.



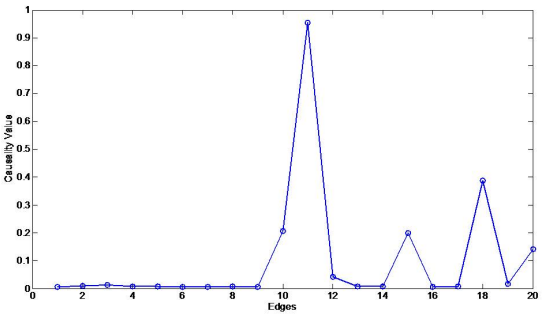
(a) Example 1



(b) Example 2



(c) Example 3



(d) Example 4

Figure 4.2: Plot of PGC values for edges in the discussed examples. See Table 4.1 for edge enumeration.

	Example 1			Example 4		
Edge number	Mean - 3σ	Mean + 3σ	Mean	Mean - 3σ	Mean + 3σ	Mean
1	-4.03e-003	8.03e-003	2.00e-003	-4.48e-003	1.05e-002	3.02e-003
2	-3.94e-003	7.92e-003	1.99e-003	-4.48e-003	1.05e-002	3.03e-003
3	-4.16e-003	8.36e-003	2.10e-003	-4.37e-003	1.04e-002	3.03e-003
4	-3.94e-003	8.00e-003	2.03e-003	-4.44e-003	1.08e-002	3.18e-003
5	-4.25e-003	8.47e-003	2.11e-003	-4.31e-003	1.03e-002	3.04e-003
6	-4.17e-003	8.29e-003	2.05e-003	-4.64e-003	1.08e-002	3.08e-003
7	-4.00e-003	7.87e-003	1.93e-003	-4.64e-003	1.06e-002	2.97e-003
8	-6.85e-003	3.70e-002	1.51e-002	-4.39e-003	1.04e-002	3.03e-003
9	-4.84e-003	9.84e-003	2.49e-003	-4.60e-003	1.08e-002	3.12e-003
10	6.95e-002	2.04e-001	1.36e-001	6.76e-002	2.01e-001	1.34e-001
11	2.50e-001	4.50e-001	3.50e-001	4.02e-001	6.25e-001	5.13e-001
12	2.35e-002	1.24e-001	7.41e-002	1.95e-001	3.74e-001	2.84e-001
13	2.51e-001	4.50e-001	3.51e-001	-4.34e-003	1.03e-002	3.00e-003
14	-4.02e-003	8.00e-003	1.99e-003	-4.39e-003	1.05e-002	3.08e-003
15	-6.26e-003	4.19e-002	1.78e-002	9.49e-002	2.47e-001	1.71e-001
16	-3.95e-003	7.97e-003	2.00e-003	-4.53e-003	1.07e-002	3.08e-003
17	-3.96e-003	7.96e-003	1.99e-003	-4.71e-003	1.10e-002	3.15e-003
18	-4.00e-003	8.02e-003	2.00e-003	1.64e-001	3.46e-001	2.55e-001
19	-3.95e-003	8.02e-003	2.03e-003	-4.44e-003	1.05e-002	3.06e-003
20	8.72e-002	2.37e-001	1.62e-001	6.88e-002	2.01e-001	1.35e-001

Table 4.2: Confidence interval bounds for Examples 1 and 4. The edge numbers correspond to the edges enumerated in Table 4.1

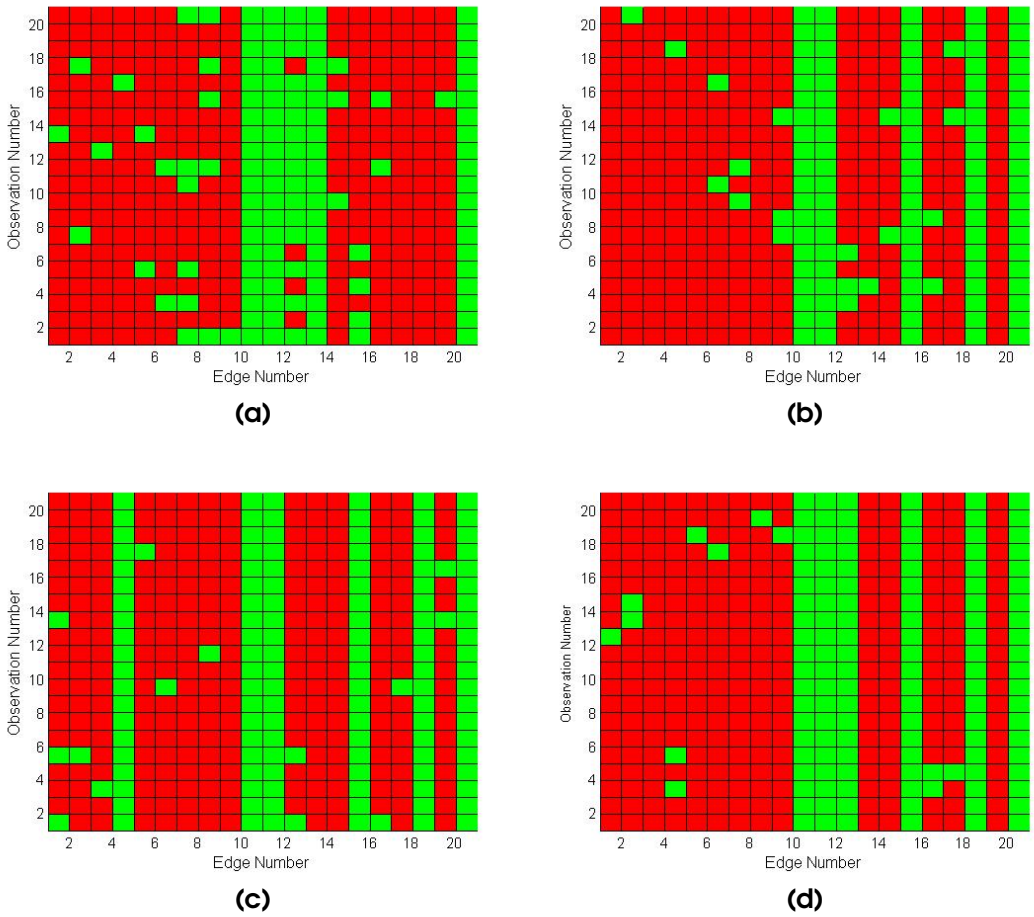


Figure 4.3: Detection of edges on multiple datasets. The x-axis represents the edges which were expressed for the corresponding dataset on the y-axis. (a) The network in Example 1 has edge number 10,11,12,13 and 20 expressed for most of the datasets. See Table 4.1 for relationship between edge numbers and the edges. (b) Example 2 has edges 10,11,15,18 and 20 expressed for most of the datasets. (c) The network in Example 3 has edge number 4,10,11,15,18 and 20 expressed for most of the datasets and (d) Example 4 has edges 10,11,12,15,18 and 20 expressed for most of the datasets.

The bootstrap approach helped us in identifying the statistically relevant edges for each dataset. We performed the confidence interval test using 2000 datasets for each example. Figure 4.3 shows the selected edges for 20 datasets for each example. Table 4.2 presents the confidence interval limits and the mean of causality values for the bootstrap samples for toy Examples 1 and 4 The plots in Figure 4.2 show PGC values for 4 datasets, each dataset belonging to a particular example. Figures 4.4 represents the Q-Q plot for actual and fitted values for a dataset for

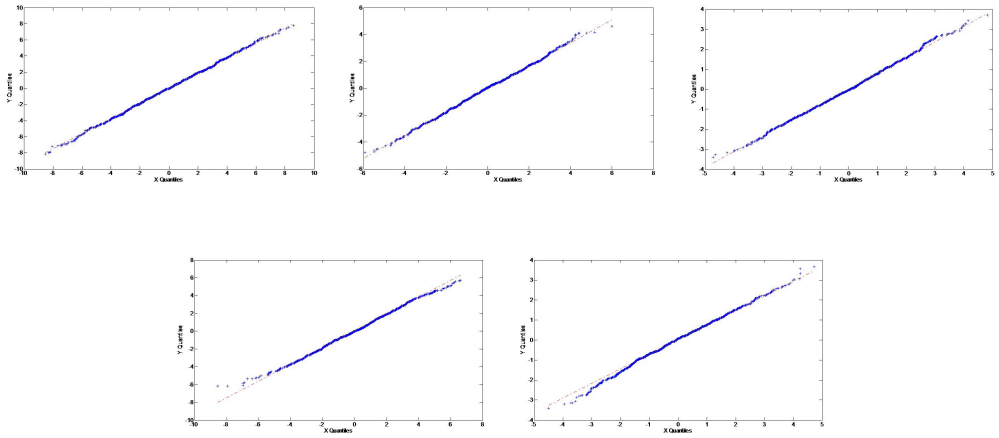


Figure 4.4: Q-Q plots for the variables in Example 1.

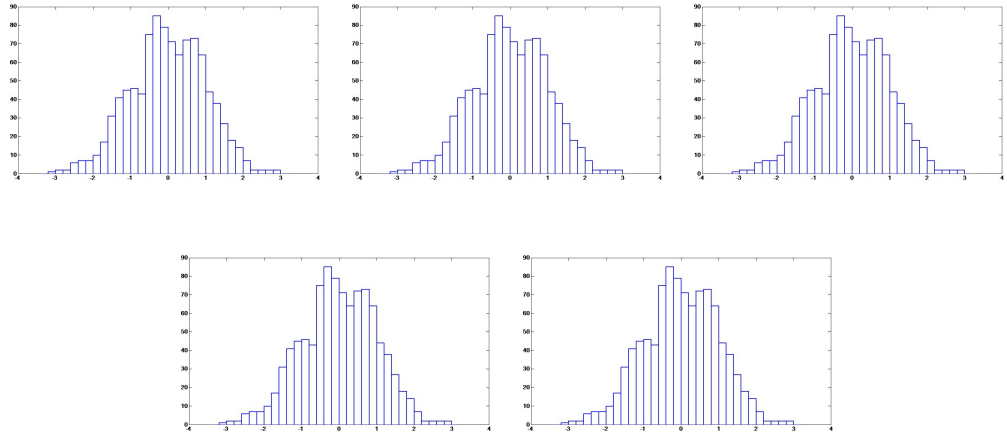


Figure 4.5: Residual plots for the variables in Example 1.

Example 1. Figure 4.5 plots the distribution of residuals obtained after fitting the model to data. Linear Q-Q plots and normal distribution of residuals are desirable for a good model fit. Such plots can be used as diagnostic tools to access fit of model to data.

A careful consideration at the PGC values for each example indicated that only the edges with comparatively higher PGC values passed the bootstrap criterion for edge selection. This can be seen by considering the confidence interval limits and the mean causality values in Table 4.2 for Example 1 and 4. The mean causality values for the edges passing the bootstrap criteria is significantly higher than the mean causality values of the other edges in that network. Similar observations can be made by looking at the Figure 4.3 which plots the selected edges for 20 different datasets for each example. As can be seen in the figure, the majority of those datasets represent the expected network structures. The filled bright squares in the figure denote the edges which passed the bootstrap confidence interval criteria. These are also the edges having considerably higher PGC values than other edges in the network. This phenomena was observed for all the toy models indicating that the edges with higher magnitude have a more significant role in detection of network structure. This is an important observation considering that bootstrap is a computationally expensive and time-consuming process. The VAR (Vector Auto Regressive) modeling of a process with q entities requires $O(q^2)$ parameters and is suitable for modelling small networks but time consuming for bigger networks. Performing bootstrapping on a bigger network using this technique will require considerable time and computational resources.

The toy models mimic different scenarios of connections between the participating entities. We saw that PGC was correctly able to infer the network structure

from data for each example. The visual matrix in Figure 4.3 showed that similar edges were expressed for most of the datasets for a given example. There were a few extra edges for some of the datasets which could be attributed to the property of data, some signals in a particular dataset were more dominant due to the introduced noise. Finally, the entries in Table 4.2 show that only the most dominant edges, i.e. edges with higher PGC value pass the confidence interval criteria for edge selection. The positive lower bound for the relevant edges supports the hypothesis in the bootstrap section.

4.2.2 Application to T-cell Data

T-cells are part of the adaptive immune response and are of two types : Helper T-cells and Killer T-cells. Activation of T-cells is a central event in initiation of immune response. T-cell activation is initiated by the interaction between the T-cell receptor (TCR) complex and the antigenic peptide present on the surface of an antigen-presenting cell. T-cells act by releasing certain proteins known as cytokines which are responsible for activating cells, triggering them to grow and divide or to die. A complex network of cytokines is secreted by the helper T-cells that determine the course of the immune response. Killer T-cells, on the other hand, release cell-damaging enzymes that create holes in the membranes of the target cells and trigger them to undergo programmed cell death. The proposed methodology was applied to a publicly available microarray gene expression data obtained from a well-established model of T-cell activation by Rangel et al.[RAG$^+$04]. The data was collected from 2 experiments characterizing the response of a human T-cell line (Jurkat) to PMA and ionomycin treatment. The dataset comprises recorded expression levels of 58 genes observed after 0, 2, 4, 6, 8, 18, 24, 32, 48 and 72 hours. The dataset can be downloaded from the web-

site http://public.kgi.edu/∼wild/LDS/index.htm mentioned in the publication by Rangel et al.

Fitting the VAR model on the data: The VAR model was fitted on the transformed dataset with the lag selection performed according to the AIC criterion mentioned in Equation 3.8. A lag value between 2 to 6 was chosen which minimized the AIC value for the system. Figure 4.6 represents the Q-Q plot for four genes. The plots were obtained after fitting the VAR model on the whole dataset. The linearity of the plots indicates that the actual time series values for a gene were in accordance with the predicted series. Plots in Figure 4.7 represent the histograms and cross-correlation measures for the standard innovations obtained for those four genes. The innovations exhibit Normal distribution. A similar pattern was observed for other genes as well after fitting the VAR model. The coefficient of determination for all 58 equations, each one representing a gene, is also presented in Figure 4.8. After the model fitting was done, a variance-covariance matrix for the residuals was obtained for the whole system. PGC values were computed for each pair of genes in the dataset according to Equation 4.14. The distribution of calculated PGC values can be found in the Figure 4.8.

Detection of the network structure: The total number of possible edges in a system of 58 entities is $\binom{58}{2} \times 2 = 3306$. Performing a bootstrap on such a big system is extremely time-consuming and computationally demanding due to the complexity of VAR models. We then relied on the observation that we made while studying the toy models, which revealed only the edges with higher PGC values compared to the rest of the edges. This was confirmed by the confidence interval tests performed on those models. Figure 4.3 and Table 4.2 support this theory for toy models. A threshold to select the most dominant edges was chosen from the

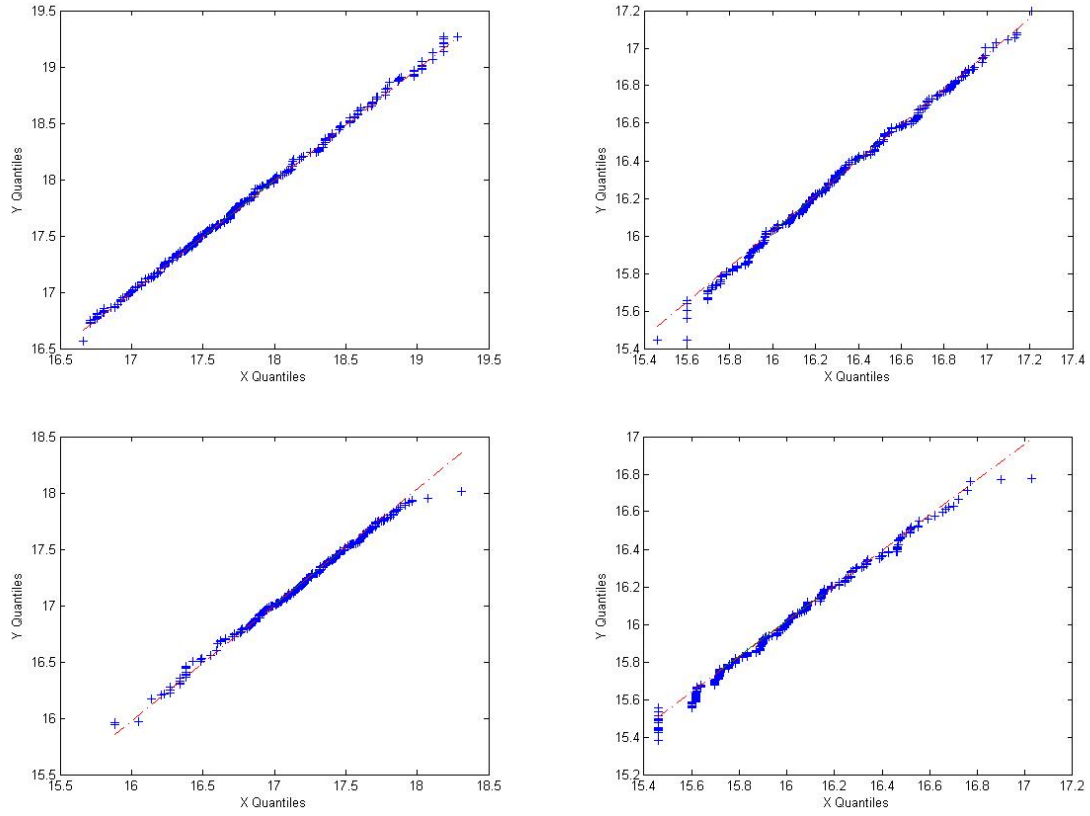


Figure 4.6: Q-Q plots of actual data versus predicted data after fitting the autoregressive model.

tail of the empirical distribution of PGC values for the T-cell data. Though the choice of threshold is user dependent and can vary from case to case, we use the value of 0.5743 which corresponds to the 97.5 percentile as the threshold to detect the relevant edges. A total of 83 edges were found to have PGC value higher than the threshold. The obtained network can be seen in Figure 4.9.

Analysis of the inferred network structure: The threshold criteria for inference of network structure resulted in the elimination of 11 genes from the final network obtained. The elimination of nodes doesn't imply that they don't play an active role in the T-cell system, but only indicates that the interaction caused by them in our inferred causal network is weaker than the interactions caused by the entities present in the network. A complete list of the genes shown in Figure 4.9 along

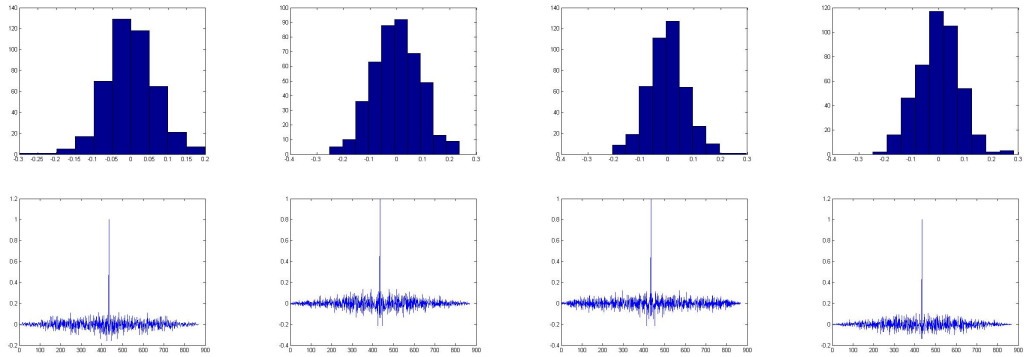


Figure 4.7: Histogram and cross-correlation plot for innovations after fitting the autoregressive model.

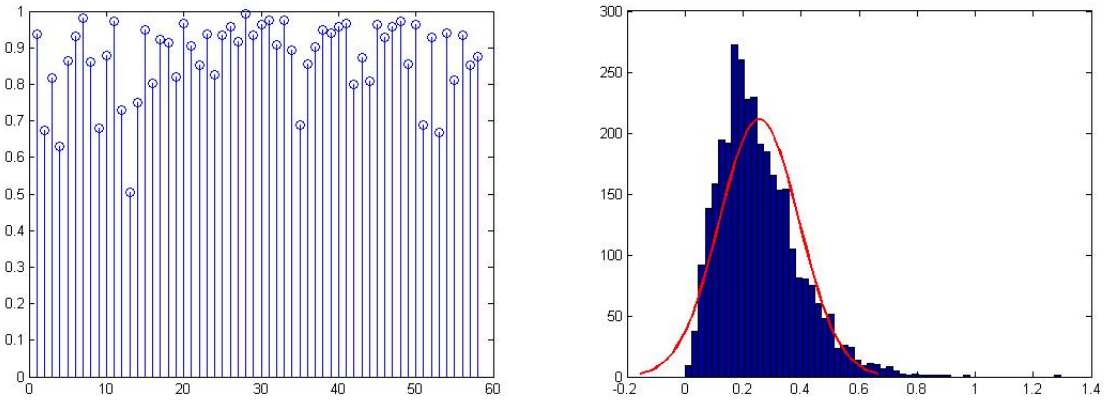


Figure 4.8: (a) Plot of coefficient of determination after fitting the VAR model on T-cell data.(b) Histogram plot for the PGC values between all pairs of genes in the dataset.

Figure 4.9: The full names and complete list of all the genes are provided in Table 4.3. Key genes mentioned in the discussion are CD69 (gene 7),LAT (gene 57), FYB (gene 45),integrin-αM (gene 15),IL-2Rγ (gene 46),NF-κB (gene 56) and IL-16 (gene 23).

with the missing ones can be found in Table 4.3. Some key genes are listed in the caption of Figure 4.9.

From a purely computational point of view, the network has two remarkable properties which are commonly found in most of the biological networks. The first property is the sparseness of connections in the network, and the second is the existence of hub-and-spoke structure in the network. There are several edges emanating from nodes 32 (superoxide dismutase 1), 57 (LAT) and 58 (v-akt murine thymoma viral oncogene homolog 1), and several edges terminating at nodes 7 (CD69), 11 (jun D proto-oncogene) and 24 (adenomatosis polyposis coli). Barabasi argues that such structures are natural for the biological systems and knocking out a hub can break down the network [Bar02]. Among the links found in the network, we obtained a few gene-gene interactions that have been documented earlier. Zhang et al.[ZIT$^+$99] showed that LAT is required for up-regulation of CD69 in T-cells, whereas the role of IL-2Rγ for regulation of CD69 was discussed by Cheng et al.[CONG02]. Pasquet et al. reported the activation of integrin-αM by LAT [PGQ$^+$99]. Influence of FYB on CD69 has been reported by Cambiaggi et al. [CSC$^+$92]. A significant correlation between NF-κB activation and level of IL-16 was discovered by Takeno et al.[THU$^+$02] and also reported by Hidi et al. [HRAA$^+$00].

4.3 Comparison With Respect to Other Methods

We summarize the main advantages of using this technique and compare this technique with other commonly used approaches for inferring network structure from biological data. The main benefits of using our technique are the following:

- Non-availability of prior knowledge about the system is not a limitation and does not restrict us from studying large systems.

Gene number	Gene name
1	retinoblastoma 1 (including osteosarcoma)
2	cyclin G1
3	TNF receptor-associated factor 5
4	clusterin (complement lysis inhibitor, SP-40,40, sulfated glycoprotein 2, apolipoprotein J)
5	mitogen-activated protein kinase 9
6	CD27-binding (Siva) protein
7	CD69 antigen (p60, early T-cell activation antigen)
8	zinc finger protein, subfamily 1A, 1 (Ikaros)
9	interleukin 4 receptor
10	mitogen-activated protein kinase kinase 4
11	jun D proto-oncogene
12	lymphocyte-specific protein tyrosine kinase
13	small inducible cytokine A2 (monocyte chemotactic protein 1, homologous to mouse Sig-je)
14	ribosomal protein S6 kinase, 70kD, polypeptide 1
15	integrin, alpha M (complement component receptor 3, alpha; also known as CD11b (p170)
16	catenin (cadherin-associated protein), beta 1 (88kD)
17	survival of motor neuron 1, telomeric
18	caspase 8, apoptosis-related cysteine protease
19	E2F transcription factor 4, p107/p130-binding
20	proliferating cell nuclear antigen
21	cyclin C
22	phosphodiesterase 4B, cAMP-specific (dunce (Drosophila)-homolog phosphodiesterase E4)
23	interleukin 16 (lymphocyte chemoattractant factor)
24	adenomatosis polyposis coli
25	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein
26	Src-like-adapter
27	cyclin-dependent kinase 4
28	early growth response 1
29	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
30	myeloid cell leukemia sequence 1 (BCL2-related)
31	cell division cycle 2, G1 to S and G2 to M
32	superoxide dismutase 1, soluble (amyotrophic lateral sclerosis 1 (adult))
33	cyclin A2
34	quinone oxidoreductase homolog
35	interleukin-1 receptor-associated kinase 1
36	SKI-INTERACTING PROTEIN
37	myeloid differentiation primary response gene (88)
38	caspase 4, apoptosis-related cysteine protease
39	transcription factor 8 (represses interleukin 2 expression)
40	apoptosis inhibitor 2
41	GATA-binding protein 3
42	retinoblastoma-like 2 (p130)
43	chemokine (C-X3-C) receptor 1
44	interferon (alpha, beta and omega) receptor 1
45	FYN-binding protein (FYB-120/130)
46	interleukin 2 receptor, gamma (severe combined immunodeficiency)
47	colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)
48	myeloperoxidase
49	apoptosis inhibitor 1
50	cytochrome P450, subfamily XIX (aromatization of androgens)
51	CBF1 interacting corepressor
52	caspase 7, apoptosis-related cysteine protease
53	mitogen-activated protein kinase 8
54	jun B proto-oncogene
55	interleukin 3 receptor, alpha (low affinity)
56	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
57	linker for activation of T cells
58	v-akt murine thymoma viral oncogene homolog 1

Table 4.3: List of genes in the T-cell data

- Our model is inherently able to capture feedback cycles in a system which is a common feature for biological systems.
- The computation in its simplest case is very straightforward and as a result the outcome is very reliable and robust.
- In this chapter we only deal with linear causality cases, but we can readily extend the concept to non-linear cases.
- An equivalent representation of this method exists in frequency domain which can be useful while dealing with frequency domain datasets.

Among the commonly used methods for inference of network structure from time-series data, relevance networks, Gaussian graphical models and Bayesian networks are the prominent and widely used ones. In the following text we present a brief overview of these methods and the problems associated with them.

Relevance networks (RN): Relevance networks are based on pairwise association score which is a correlation based method. The principle disadvantage of this method is, that inference of an interaction between two nodes is not performed in the context of the whole system. Correlation based methods are incapable of distinguishing between direct and indirect interactions and are unable to capture feedback loops.

Gaussian graphical models (GGM): GGMs - also known as undirected probabilistic graphical models are inferred by calculating partial correlation between two nodes conditional on all the other nodes in the network. Though the direct interaction between two nodes is computed in the context of the whole network, the method still suffers from some of the problems of relevance networks, namely

lack of direction and feedback cycles in the inferred network.

Bayesian networks (BN): Bayesian networks are probabilistic graphical models that represent a set of variables and their probabilistic independencies. Temporal observations of variables are modeled using Dynamic Bayesian networks which can overcome the limitation of acyclicity constraint associated with Bayesian networks. In a fully observable dataset, the parameter estimation and structure learning should not be overly complicated, but in case of missing variables, the computational and theoretical problems arise. The structure learning generally requires use of optimization algorithms which can be computationally costly and suffer from their own problems like getting stuck in a local minima. The final inferred structure is also dependent on the choice of prior and the scoring function used to evaluate the generated structures. Bayesian networks themselves are not carriers of causal connections between variables. However, causal connections can be derived by means of interventions in the dataset by fixing the values of a variable to a constant and assuming that values of none other variables are affected.

The ability of Bayesian networks to infer a causal network from intervention data distinguishes it from other computational techniques discussed above including ours. This leads to an interesting observation and takes us back to the basic definition of causality. Bayesian networks use intervention data to detect *effective causality* by modelling the hidden states of variables causing the observed data, whereas, our technique relies on *functional causality* resting on the temporal dependencies among the data themselves without reference to how they were caused by the underlying processes. Causality in the context of Granger causality is strictly based on temporal precedence and assume that data reflects the states that cause data. Other techniques like Differential equations, Petri nets, Process

calculi etc. use extensive biological understanding to model the system and are better suited to studying effective causality between variables. Granger causality based models on the other hand are applicable to any time-series dataset without requiring any prior knowledge about the system. Our method contains number of implicit assumptions like weak stationarity of time-series and Gaussian behaviour of uncorrelated residuals. A number of transformation methods have proposed to deal with such issues and can be applied according to the data. The assumption of linearity in our models can be overcome by extending the models to their non-linear forms. Some non-linear extensions of Granger causality can be found in the publications by Ancona et al. [AMS04] and Marinazzo et al. [MPS06].

4.4 Summary

Advances in experimental techniques in molecular biology have enabled researchers to perform high-throughput experiments and simultaneously monitor activities of numerous biological entities at different time points. Quantitative analysis of experimental data helps researchers to build hypothesis about the system and design new experiments. In this chapter, we proposed the use of Partial Granger Causality to quantitatively infer the causal network structure based on microarray data. The application of this technique was first studied for various toy models and then later applied to the T-cell microarray data for inference of the network structure. The multivariate nature of this technique makes it useful for the systems having large number of entities engaged in cross-communication with each other. The technique is simple in nature and can be easily applied to small as well as bigger systems. The proposed causality model can be most useful when experimental conditions are chosen in such a way that they activate the measured network strongly and there is minimum error in data recording.

Chapter 5
Listening to Genes

Uncovering the biological meaning embedded in time-series gene expression data is one of the most challenging problems in the post-genomic era. The expression patterns observed over multiple time points provide us with a rich set of information detailing the temporal profiles of the genes. Such profiles when studied at the genome wide level can help us understand the functional mechanism of the underlying cellular processes.

Temporal analysis of microarray data has not only helped in identifying functional categories of genes but also in understanding the behaviour of various gene circuits. Techniques like Fourier estimation have been used to detect periodic signals in various organisms including yeast and human cells [SSZ+98, WFS04, KLW06]. Claridge-Chang et al. [CCWN+01] used Fourier components to determine a set of genes expressed with a robust circadian rhythm in adult Drosoplhila heads. Similar microarray study on circadian rhythm in Arabidopsis was carried out by Harmen et al.[HHS+00] which empirically tested for statistically significant cross-correlation between the temporal profile of each gene and a cosine wave of definite period and phase. Temporal microarray data has also been helpful in understanding the gene circuits using methods like Ordinary Differential Equa-

tions [Alo07] and Dynamic Bayesian networks [DGM+06, KIM03]. This chapter presents a complete pipeline for analysing the Arabidopsis data discussed in Chapter 1. We apply the techniques discussed in the previous chapters on the complete dataset. We supplement those techniques with a frequency domain analysis.

The first step in dealing with the microarray dataset is to process the data using the normalization technique developed in Chapter 2. The normalization can help us deal with unwanted systematic variations associated with each biological sample and experimental conditions. The gene expressions were corrected for various unwanted biases and negative correlation across replicates was minimized.

The next step after normalization is clustering of data to reduce the data dimension. Three popular approaches [AYA07] for clustering microarray data are : a) Point-wise distance based methods [DH05] b) model-based clustering methods [PLL02] and c) feature-based clustering methods [BHWK05]. We presented a technique to find functional clusters from temporal data in Chapter 3. In this chapter, we introduce another approach to cluster data based on its frequency profile. This approach certainly belongs to the feature-based clustering methods, but differs from the idea of *visual clustering techniques* where genes are classified according to their distances from class centres. Frequency is an important feature in any temporal data and there are many advantages of dealing with temporal data in frequency domain. We call our method *auditory clustering* approach.

After applying our frequency based clustering approach to the data, we use Partial Granger Causality discussed in Chapter 4 to infer network structures for interactions among selected genes. We present three gene circuits in this chapter. The first circuit is the Circadian circuit comprising of 7 genes (ELF4, TOC1,

CCA1, LHY, PRR7, PRR9 and GI). The second circuit is the Ethylene signalling circuit comprising of 16 genes, and the third circuit is a global gene profile circuit of 9 genes. For all the circuits, we present the causality analysis in both the time and frequency domain. We further introduce the idea of Complex Granger Causality to show the interactions among *groups of genes* in the Circadian and Ethylene circuits. In both the circuit, we find that the complex Granger causality plays an important role in reconciling experimental and computational results. Interactions in the global gene profile circuit present more interesting results to answer questions like, if there is a global picture of interactions among genes. To create the global profiles for the genes, we first simply cluster the genes using the k-mean clustering algorithm. We then use the cluster centres (means) as representatives of each cluster and apply the Partial Granger Causality to infer the interaction pattern. We see a clear hierarchical structure of interactions among the representative genes. At the top of the hierarchy are the genes with a peak in the middle, at the middle level there are genes with a decreasing trend, and at the bottom level the genes exhibit an increasing trend.

5.1 Methods

5.1.1 Data Generation : Overview of the Dataset

An overview of the microarray experiment to obtain the dataset was presented in Section 1.6 of Chapter 1. We provide here a brief recap of the dataset obtained. Gene expressions for a total of 30,336 genes were collected. The data was obtained over 22 days during the leaf senescence process. The biological replicates were harvested both in morning and evening at every alternate day, thus giving total 22 time points. There were four biological replicates collected at each time point where each biological replicate resulted in four technical replicates. The final spot

quantization matrix obtained after scanning the hybridized arrays resulted in a $30336 \times 16 \times 22$ matrix where 30336, 16 and 22 are the number of genes, total replicates and total time points respectively.

5.1.2 Normalization

The data needs to be cleaned from various unwanted experimental biases and the expression values obtained at different replicates need to be comparable. We proposed a normalization technique to deal with our data in Chapter 2. We first estimated the various sources of experimental variations in the data using an error model. After removing the sources of unwanted variations in the dataset, the residuals associated with genes in a replicate, standardized by the estimated gene-wise variances showed a Gaussian distribution. Also, the correlation between residuals from one replicate to other replicate was minimized using an iterative algorithm. The normalized data is used for further analysis.

5.1.3 Clustering : Auditory Clustering

The genes were clustered according to their frequency profiles, thus the name auditory clustering. Availability of temporal data allows us to analyse it in both time and frequency domains; we take advantage of this fact and investigate the behaviour of the data in frequency domain. We use a toy example to illustrate the method. We randomly select 3000 genes from our dataset to build the toy example. Analysis of toy example is presented in Figure 5.1. A power spectrum of selected 3000 genes was computed using discrete Fourier transformation. An analysis of the power spectrum indicated presence of two major frequencies present in the system. The major frequencies were found for day 1 and day 22. We used

this information to generate the simulation data using the following equations :

$$x_{gt} = a_g + k \times cos(\frac{2\pi t(1)}{N}) + k \times cos(\frac{2\pi t(11)}{N}) + t \times \epsilon_t \tag{5.1}$$

where

$$k = \begin{cases} 0.1 \times U(0,1) & \text{if } 0 < g \leq 1000 \\ 0.5 \times U(0,1) + 0.1 & g > 1000 \end{cases}$$

The term a_g is the DC term computed for the gene g from selected dataset after taking the Fourier transform. ϵ_t is the uniform error associated with simulated x_{gt}.

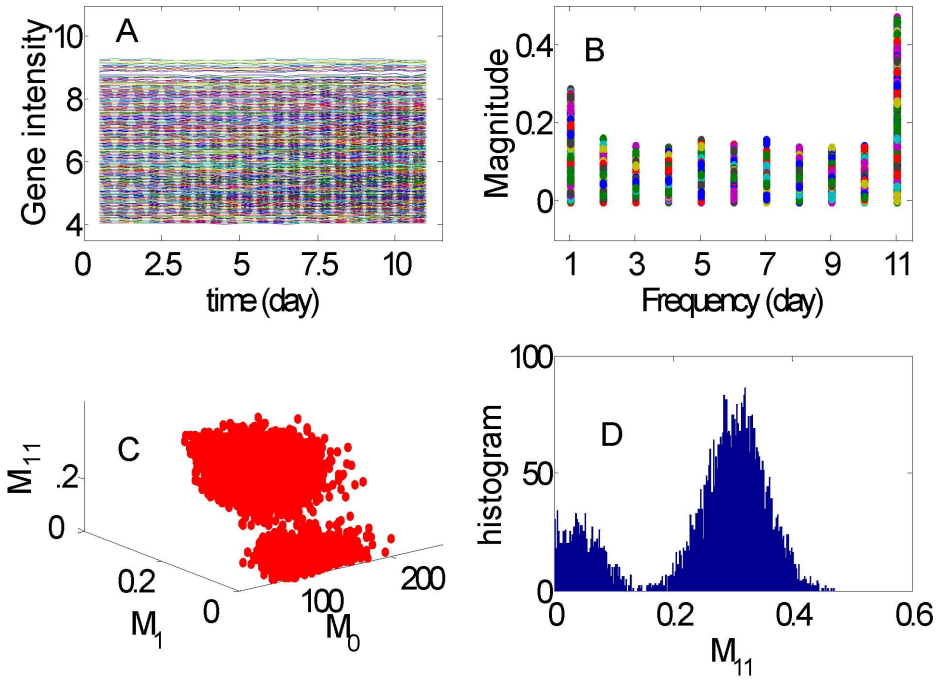


Figure 5.1: Synthesized data. A. Gene intensity vs. time. B. The magnitude of discrete Fourier transform of the data in A. The DC term is not shown. C. M_0 (DC term), M_1 (corresponding to the first column in B) and M_{11} (the 11th column in B).A clear structure of two clusters is shown. D. The histogram of the magnitude of M_{11}.

The panel (A) in Figure 5.1 plots the time domain representation of the generated 3,000 genes. Though it is difficult to see the grouping of genes in the time

domain representation, we transform the data to the frequency domain and the results are shown in Figure 5.1 (B). Two dominant frequencies corresponding to the 1^{st} (named M_1) and the 11^{th} (named M_{11}) components of the discrete Fourier transformation can be seen in the figure. Figure 5.1 (C) also confirms that two different frequencies are present in the data, one in the high frequency (M_{11}) and the other being in the low frequency (M_1). The behaviour can also be seen in the Figure 5.1(D). The functional meaning of clusters is obvious in terms of frequency. Each frequency has its own physiological meaning. The genes with higher M_{11} values are most sensitive to (controller of) faster changes, whereas the genes with higher M_1 values are responsible for slower changes. In general, we can deal with a data set which has multiple frequencies.

5.1.4 Network Analysis : Complex Granger Causality

In Chapter 4, we introduced the concept of Partial Granger Causality to infer network structure from temporal gene expression data. Based on the same principles, we introduce here a system of complex interactions. A *complex* essentially means a set of multiple time series. Complex interactions are considerably different from interactions observed at pairwise level. For example, a pair of variables may not have individual interaction with the third variable, but when in combination with each other, they may interact with the third variable. On the other hand, when two variables are negatively correlated, each of them can interact with the third variable, but when they are grouped together, the interaction may disappear. The complex interaction between sets of time series can be explained in the following way.

Consider three *multiple* stationary time series $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ with k, l and m dimensions respectively. We consider the relationship from $\boldsymbol{Y}$ to $\boldsymbol{X}$ by conditioning

on $\boldsymbol{Z}$. The joint autoregressive representation for $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be written as

$$\begin{cases} \boldsymbol{X}_t &= \sum_{i=1}^{\infty} \boldsymbol{a}_{1i}\boldsymbol{X}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{c}_{1i}\boldsymbol{Z}_{t-i} + \boldsymbol{\epsilon}_{1t} \\ \boldsymbol{Z}_t &= \sum_{i=1}^{\infty} \boldsymbol{b}_{1i}\boldsymbol{Z}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{d}_{1i}\boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_{2t} \end{cases} \tag{5.2}$$

The noise covariance matrix for the system can be represented as

$$\Gamma = \begin{pmatrix} \sigma^2(\boldsymbol{\epsilon}_{1t}) & \sigma(\boldsymbol{\epsilon}_{1t}, \boldsymbol{\epsilon}_{2t}) \\ \sigma(\boldsymbol{\epsilon}_{2t}, \boldsymbol{\epsilon}_{1t}) & \sigma^2(\boldsymbol{\epsilon}_{2t}) \end{pmatrix} = \begin{pmatrix} \Gamma_{xx} & \Gamma_{xz} \\ \Gamma_{zx} & \Gamma_{zz} \end{pmatrix}$$

where σ^2 and σ represent variance and co-variance respectively. Extending this representation, the vector autoregressive representation for a system involving all the three time series $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ can be written in the following way.

$$\begin{cases} \boldsymbol{X}_t &= \sum_{i=1}^{\infty} \boldsymbol{a}_{2i}\boldsymbol{X}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{b}_{2i}\boldsymbol{Y}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{c}_{2i}\boldsymbol{Z}_{t-i} + \boldsymbol{\epsilon}_{3t} \\ \boldsymbol{Y}_t &= \sum_{i=1}^{\infty} \boldsymbol{d}_{2i}\boldsymbol{X}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{e}_{2i}\boldsymbol{Y}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{f}_{2i}\boldsymbol{Z}_{t-i} + \boldsymbol{\epsilon}_{4t} \\ \boldsymbol{Z}_t &= \sum_{i=1}^{\infty} \boldsymbol{g}_{2i}\boldsymbol{X}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{h}_{2i}\boldsymbol{Y}_{t-i} + \sum_{i=1}^{\infty} \boldsymbol{k}_{2i}\boldsymbol{Z}_{t-i} + \boldsymbol{\epsilon}_{5t} \end{cases} \tag{5.3}$$

The noise covariance matrix for the above system can be represented as

$$\Sigma = \begin{pmatrix} \sigma^2(\boldsymbol{\epsilon}_{3t}) & \sigma(\boldsymbol{\epsilon}_{3t}, \boldsymbol{\epsilon}_{4t}) & \sigma(\boldsymbol{\epsilon}_{3t}, \boldsymbol{\epsilon}_{5t}) \\ \sigma(\boldsymbol{\epsilon}_{4t}, \boldsymbol{\epsilon}_{3t}) & \sigma^2(\boldsymbol{\epsilon}_{4t}) & \sigma(\boldsymbol{\epsilon}_{4t}, \boldsymbol{\epsilon}_{5t}) \\ \sigma(\boldsymbol{\epsilon}_{5t}, \boldsymbol{\epsilon}_{3t}) & \sigma(\boldsymbol{\epsilon}_{5t}, \boldsymbol{\epsilon}_{4t}) & \sigma^2(\boldsymbol{\epsilon}_{5t}) \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{pmatrix}$$

where $\boldsymbol{\epsilon}_{it}, i = 1, \cdots, 5$ are the uncorrelated prediction errors over time. The conditional variance $\Gamma_{xx} - \Gamma_{xz}\Gamma_{zz}^{-1}\Gamma_{zx}$ measures the accuracy of the autoregressive prediction of $\boldsymbol{X}$ based on its previous values conditioned on $\boldsymbol{Z}$ whereas the conditional variance $\Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}$ measures the accuracy of the autoregressive prediction of $\boldsymbol{X}$ based on its previous values of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ conditioned on

$\boldsymbol{Z}$. The traces of matrix $\Gamma_{xx} - \Gamma_{xz}\Gamma_{zz}^{-1}\Gamma_{zx}$ and the matrix $\Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}$ are denoted by $T_{x|z}$ and $T_{xy|z}$ respectively. We define the Complex Granger causality from vector $\boldsymbol{Y}$ to vector $\boldsymbol{X}$ conditioned on vector $\boldsymbol{Z}$ to be

$$F_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{Z}} = \ln\left(\frac{T_{x|z}}{T_{xy|z}}\right) \tag{5.4}$$

The 99.7% confidence interval can be constructed using the bootstrap method. An interaction between two genes or two group of genes is significant if and only if the low bound of the confidence interval is greater than zero.

5.2 Results

5.2.1 Normalization

The correlation matrix of residuals from 16 replicates for time point 1 is shown in Figure 5.2. The result obtained after removing the different biases is shown in $x \in [1, 16] \times y \in [1, 16]$. The existence of negative correlation among the replicates can be seen in Figure 5.2 (more downward spikes than upward). After applying our method to the data, the negative correlation is evenly distributed over all replicates $x \in [21, 36] \times y \in [1, 16]$. This considerably improves the outcome of the normalization. A detailed description of the method can be found in Chapter 2.

5.2.2 Frequency Analysis

After normalizing the data, we turn our attention to create clusters in frequency domain. The representation of a time dependent signal as a weighted sum of sine and cosine functions can be studied using Fourier analysis. It is important to note that a successful Fourier analysis depends on a careful design of experiment and data collection method. Too short data or a collection on data points on irregular

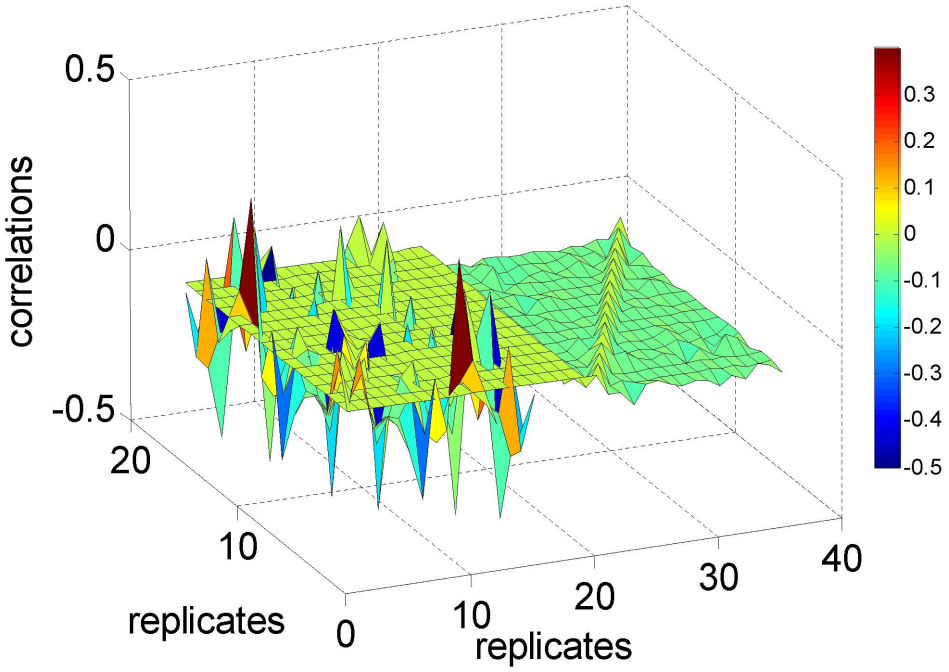


Figure 5.2: Correlation matrix of residuals before and after the application of select-and-reject algorithm during normalization (see Chapter 2 for details). For $x = 1, 2, \cdots, 16$ is the correlation matrix before applying the algorithm. For $x = 21, 22, \cdots, 36$ is the correlation matrix after applying the algorithm. The diagonal elements of two matrices are all set to 0.

intervals can miss the natural cycles present in the system and the Fourier analysis may not be fruitful. We took such important issues in consideration while collecting the data. First, our data is long enough and collected over 22 days which allows to capture changes in gene expression profiles during the senescence process. Second, our data was collected to capture the cyclic behaviour due to daily activity (24 hour period) in the plant. Twice a day data collection also allowed us to monitor the gene expressions due to day and night effect. Though our data was not collected on smaller intervals which meant that we missed the smaller frequencies but the larger frequencies could still be captured and utilized for our purpose.

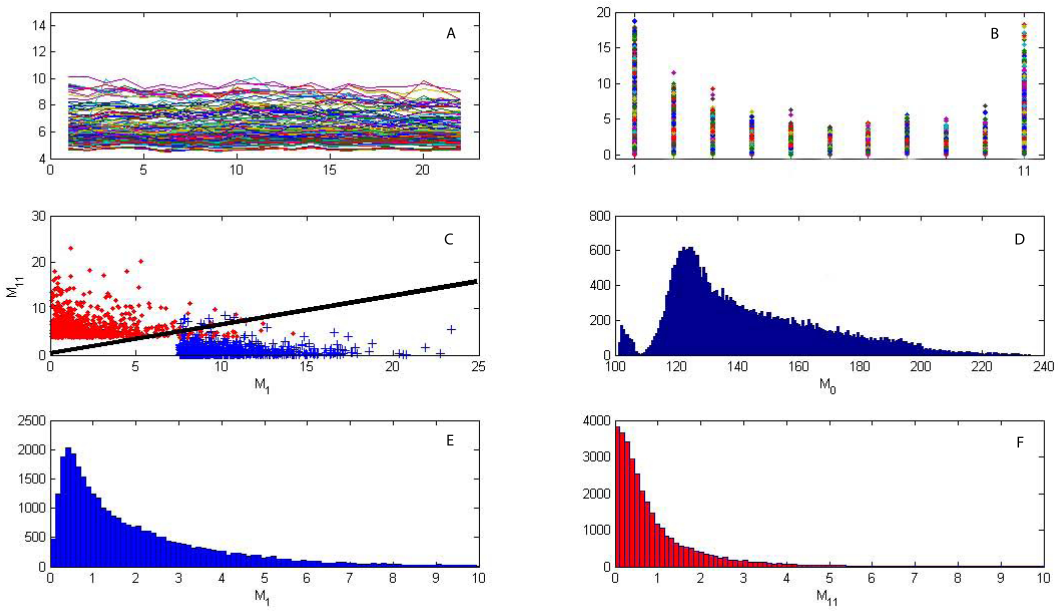


Figure 5.3: (A) Gene intensity vs. time. Only 200 genes are shown. (B) Magnitude of all genes vs. frequency. It is clear to see that there are two main frequencies in the data, i.e. the one of one day period (M_{11}, the 11th column) and the other of 22 days period (M_1, the first column). The DC term M_0 is not shown. (C) Two dimensional plot of M_{11} vs. M_1. (D) The histogram of the DC term. There are two peaks in the histogram. (E) The histogram of M_1, it is a Weibull distribution. (F) The histogram of M_{11}, it is an exponential distribution.

We show the results of frequency domain analysis in Figures 5.3 and 5.4. Figure 5.3 (A) plots the time-domain representation of 200 randomly selected genes. We can see that it is difficult to visualize any grouping in this small dataset. We then perform the Fourier transformation on the complete dataset and plot the power spectrum in Figure 5.3(B). We can see the presence of two dominant frequencies in the spectrum; one with a period of 22 days (M_1, the 1^{st} column) and other with a period of 1 day (M_{11}, the 11^{th} column).

The column corresponding to the DC term (M_0) is not shown in the figure. A histogram plot of the DC terms in shown in Figure 5.3(D). The plot indicates a bimodal distribution, indicating two groups of overall signalling strength in the gene data. To visualize the relationship between the two dominant frequencies at

M_1 and M_{11}, we plot them opposite to each other in Figure 5.3(C). For visualization purposes, we select only the top values of M_1 (all values > 8) and M_{11} (all values > 5). The thick black line in the plot indicates that genes with strong 1 day rhythm are separate from genes with 22 day rhythm. Figures 5.3(E) and (F) plot histograms of M_1 and M_{11} values respectively. We can see that distributions of values in M_1 and M_{11} resemble Weibull and Exponential distributions respectively.

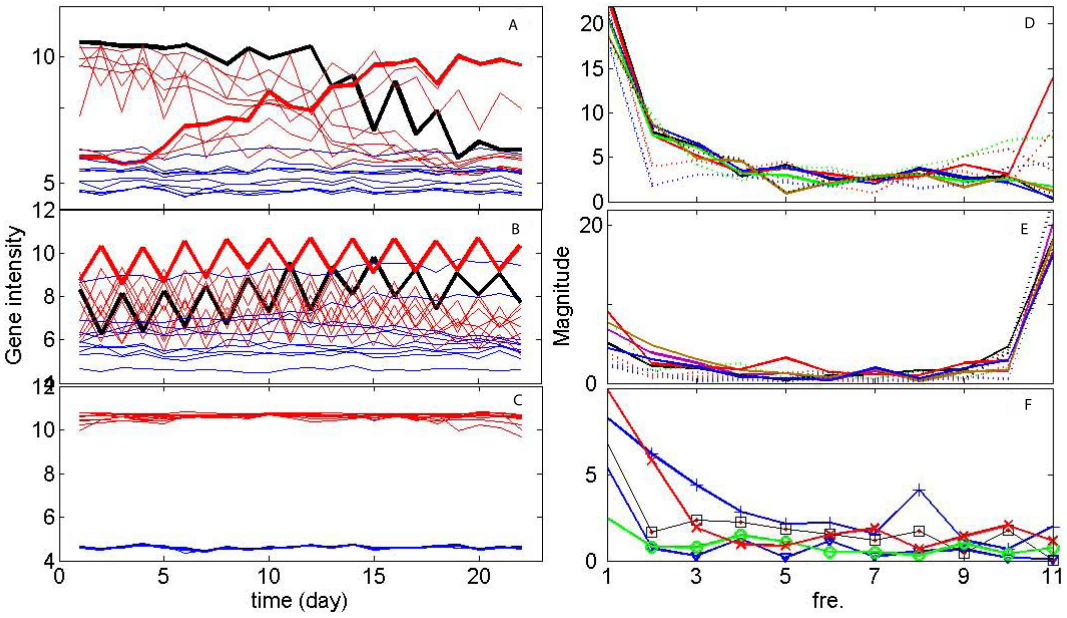


Figure 5.4: (A) Time trace of the top (in red and black) and bottom (in blue) ten genes with the strongest amplitude of the period of 22 days. There are two classes: one is up-regulated (red thick line), the other is down-regulated (black thick lines). (B) Time trace of the top (in red and black) and bottom (in blue) ten genes with the strongest amplitude of period of 1 day. There are two classes: one is on-phase (red thick line), the other is off-phase (black thick line). (C) Time trace of the first top (in red) and bottom (in blue) ten genes without rhythms. Plots in (D), (E) and (F) plot the frequency representation of top genes in (A), (B) and (C) respectively.

In order to have a clear understanding of the power spectrum, we perform further analysis and present the results in Figure 5.4. We select top ten and bottom ten genes ranked according to their magnitudes in M_1 and plot their time-domain representation in Figure 5.4(A). In Figure 5.4(B), we present the top ten

and bottom ten genes according their corresponding magnitudes in M_{11}. Similarly, top ten and bottom ten genes were selected according to their ranking in the DC term (M_0) and plotted in Figure 5.4(C). Figures 5.4 (D),(E) and (F) are frequency representations of top ten genes in Figures 5.4(A),(B) and (C) respectively. Figure 5.4(A) has time plot of genes with a period of 22 days. The bottom ranked ten genes are plotted in blue and have flat profiles. The top ten ones, drawn in red and black show big fluctuations across time. All the top ten genes can be divided into two classes (a) down-regulated (6 genes) and (b) up-regulated (4 genes). We plot one of the up-regulated genes with a thick red line, and one of the down-regulated ones with a thick black line for clear visualization. One could infer that genes are related to leaf senescence. We next turn our attention to the genes having a period of 1 day in Figure 5.4(B). Again the bottom ranked ten genes are plotted in blue and have flat profiles. The top ten genes, drawn in red and black are oscillatory genes and could qualify as circadian genes. The top ten genes can again be classified as (a) In-phase genes (6 genes) and (b) Out-phase genes (4 genes). We have drawn examples of an in-phase gene and an out-phase gene using thick red and thick black lines respectively. Finally, in Figures 5.4(D),(E) and (F), we plot the frequency domain representations of top ten genes plotted in Figures 5.4(A),(B) and (C) respectively.

5.2.3 A Circadian Circuit

The frequency plot of top ten genes with one day period in Figure 5.4(B) shows an interesting mix of in-phase and out-phase genes. A look at the gene annotation database identifies the gene with highest M_{11} value as ELF4. The plot indicates a strong oscillatory rhythm for ELF4 for each day. ELF4 plays an important role in maintaining the circadian rhythm of the plant as reported by Doyle et al.[DDB$^+$02] and McWatters et al.[MKH$^+$07]. The strong oscillatory behaviour

of circadian genes is also reported by Harmen et al. [HHS+00]. From the gene annotation database, we found that ELF4 is related to two other genes CCA1 and LHY. ELF4 is necessary for light-induced expressions of both CCA1 and LHY. Yonovsky et al. [YK03] reported a Circadian circuit involving ELF4, CCA1, TOC1 and LHY. See Figure 5.5(A). We plot the expression of these genes in Figure 5.6(A).

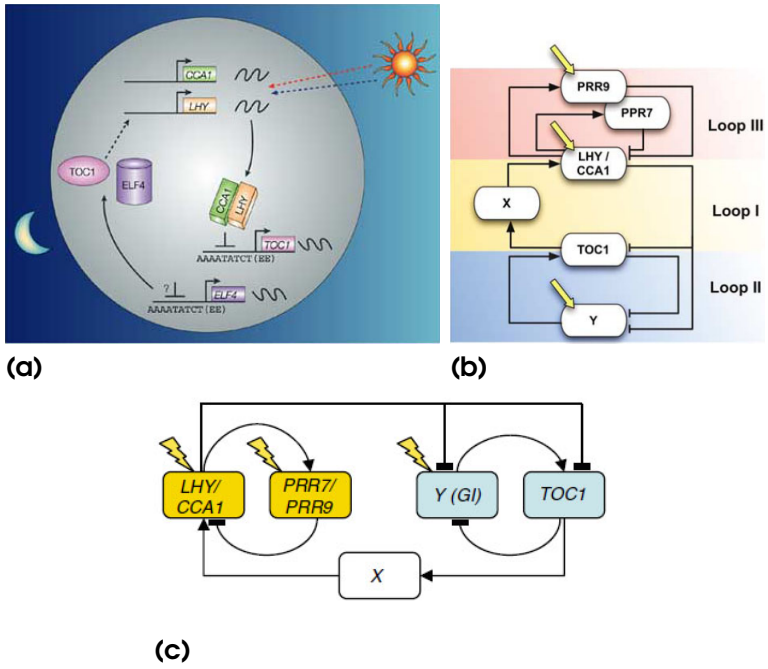


Figure 5.5: Circadian circuits reported in literature. (a)Morning and evening loop in Arabidopsis. From Yonovsky et al. (YK03). (b)Morning, evening and an unknown loop by Ueda (Ued06) (c) Inclusion of GI gene in the circuit by Locke et al.(LKBG+06)

Ueda [Ued06] and Locke et al. [LKBG+06] present circadian circuit with three loops: PRR9, PRR7 and LHY/CCA1 are in one loop (morning loop or loop III), TOC1 and GI as another loop (night loop or loop II), and TOC1,LHY/CCA1 and a unknown gene making the third loop (loop I). See Figure 5.5(B) and (C).

We therefore consider a circuit of 7 genes for our computational analysis. A complete gene annotation for selected genes is provided in Appendix B. We applied partial Granger causality discussed in Chapter 4 on these genes and the resulting

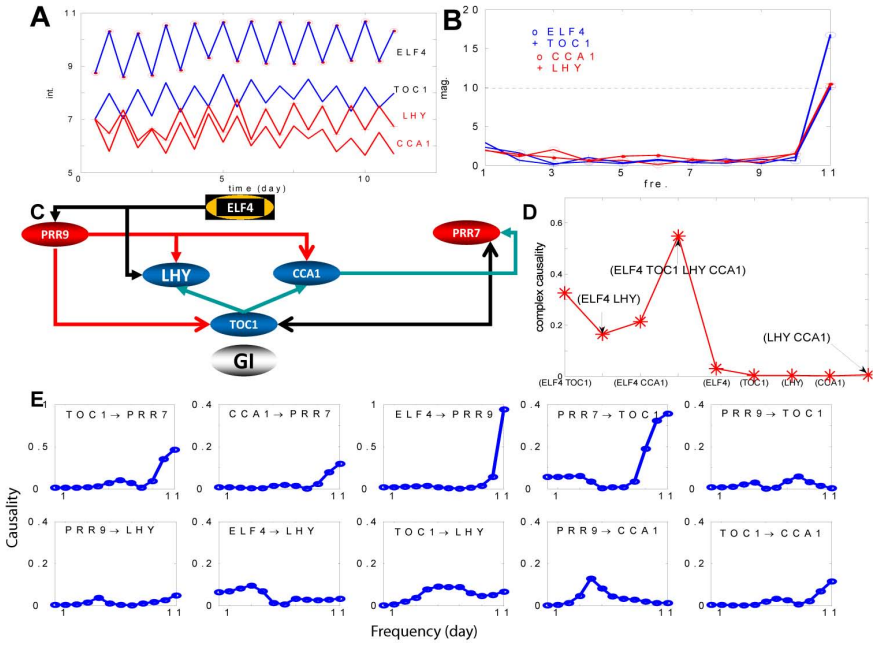


Figure 5.6: One gene circuit controlling circadian activity. A. Time trace of four genes, ELF4, TOC1, LFY and CCA1. ELF4 and TOC1 are in-phase oscillators, LFY and CCA1 are in-phase oscillators, but they are off-phase oscillators with respect to ELF4 and TOC1. B. Magnitudes vs. frequency for the four genes. They have highest magnitude at the frequency of one-day period. C. The gene circuit obtained in terms of PGC (see annotation in Supplemental material II). D. Complex interactions between different group of genes and GI. D. Gene interactions in the frequency domain.

network is shown in Figure 5.6(C). ELF4 plays an important role in regulating the circadian rhythm and is the most upstream gene. It interacts with both loop III and loop I. Loop III genes are closely interconnected via interactions between PRR9, LHY and CCA1, and the interaction between CCA1 and PRR7. Similarly, in loop I, TOC1 modulates LHY and CCA1. There are also links between loop III and loop I. PRR9 exerts influence on TOC1. TOC1 and PRR7 have a feedback loop. GI is an isolated gene on our structure without having any interaction with other genes. The observation of GI being an isolated gene in our structure also coincides with the experimental findings. On page 4 of [LKBG+06], it is mentioned that *The GI single mutant had a relatively weak phenotype, whereas our assays of the triple GI; lhy;cca1 mutant demonstrate GI's importance.* This led us to introduce the notion of interaction between complexes as presented in Section

5.1.4. We plot the interaction results of individual and grouped genes with GI in Figure 5.6(D). We see that interaction of single genes ELF4, TOC1, LHY and CCA1 with GI is almost negligible. The complex interactions of GI with pairs like (ELF4,TOC1), (ELF4,LHY) and (ELF4,CCA1) is also low. But when a complex of four genes (ELF4, TOC1, LHY and CCA1)is formed and their collective influence on GI is computed, we see a high peak in the interaction graph. This leads to an indication that there may be a possibility of strong phenotype of GI when complex interactions are taken into account.

We also analyse the interaction between genes in the frequency domain using partial Granger causality and present the results in Figure 5.6(E). We can see that most of the interactions show a 24 hour periodic behaviour by exhibiting a peak at one day period.

5.2.4 Ethylene Circuit

The Ethylene signalling pathway [NJ00, HE04] is one of the most well studied circuits in the literature due to its importance in developmental processes and fitness responses. We selected a group of 16 genes which have been reported in the literature to play a central role in the pathway. See Appendix B for annotation of selected genes. Ethylene is perceived by a family of integral membrane receptors. In Arabidopsis, at least five family members are involved: ETR1, ETR2, ERS1, ERS2 and EIN4. ETR1 and ERS1 belong to type 1 receptors whereas EIN4, ETR2 and ERS2 are type 2 receptors. The receptors are hypothesized to be in a functionally active form that constitutively activates CTR1. It is reported that the interaction of type 1 receptors with CTR1 is stronger than type 2 receptors. CTR1 is an upstream gene, and has been reported as the regulator of the pathway [HE04]. In our inferred circuit, we obtain interactions of CTR1 with ERS2 and

CTR1 with ETR1. Both are biologically verified [HE04]. Though EIN2 is an important component in the Ethylene circuit, its function is not completely understood [SA05]. It has been suggested that in the downstream of CTR1 and on the upstream of EIN2, a SIMKK-MPK6 pathway exists which may be regulated by CTR1, but this is yet to be verified biologically. So, we directly focus on the interactions between CTR1 and EIN2 and check whether CTR1 regulates EIN2 or not. To understand the interactions between CTR1 and EIN2, we chose to use complex causality by grouping together CTR1, ETR1 and ERS2; and analysing the interaction of the group with EIN2. We found that CTR1 does have a relationship with EIN2 and this is shown in the Figure 5.7A (thick arrow). EIN3 is most closely related to EIL1 [HE04] and this interaction can be found in the inferred network. Except two genes (EIN4 and EIL2) which are isolated and have no interactions with the rest of the genes, we see that the pathway shows a clear hierarchical structure. Interactions in the frequency are shown in Figure 5.7(B). Some interactions, for example ETR2 $\rightarrow$ ERS1, EIN6 $\rightarrow$ EIL4 etc., exhibit a strong daily rhythm.

5.2.5 A Global Circuit

Finally, we turn our attention to a global picture involving all the genes in the dataset to see if there is any interaction network of interest showing how leaf senescence is turned on. All genes were clustered using k-mean clustering algorithm provided by Matlab [Seb84, mat]. Initially, we clustered all the genes by pre-specifying the number of clusters to be 32,28,24,...etc. We then removed all the clusters with flat profiles using visual inspection and selected the ones showing clear trends of high activity. We then picked up the one gene from each cluster representing the centre of that cluster for further analysis. The time trace of representative genes is shown in Figure 5.8(A).

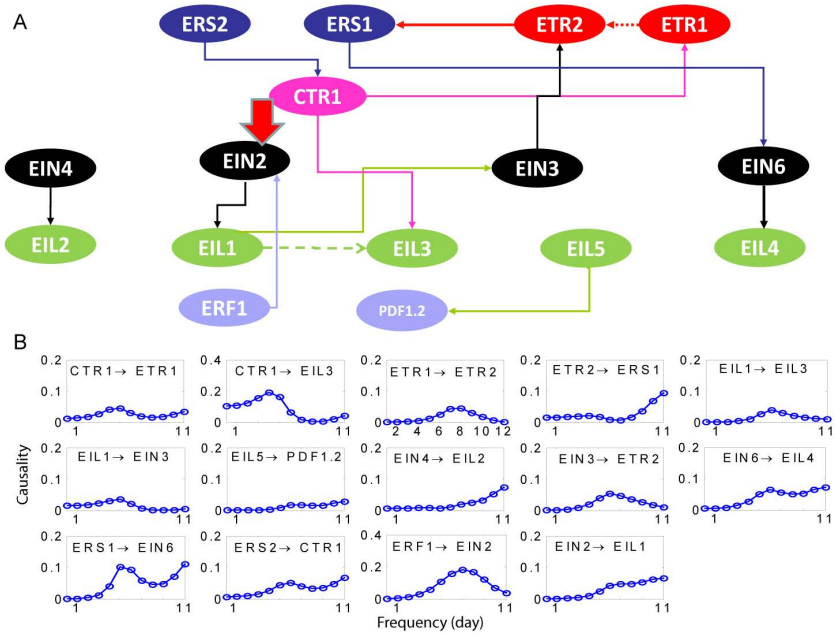


Figure 5.7: A. An ethylene gene circuit with 16 genes. Only genes with interactions are shown here. The thick arrow is the complex interaction between CTR1, ETR1 and ERS2 and EIN2. B. Interactions in the frequency domain calculated in terms of PGC. Only 14 significant interactions are shown.

A clear hierarchical structure can be seen in Figure5.8(A). The upstream genes show a typical concave shape. All genes in the middle layer in hierarchy have a peak at the beginning and then they decrease. Finally, the bottom layer genes increase their intensity as they approach 22 days. This behaviour of fits our intuition while thinking about the senescence process in leaf. During the life time of a leaf, senescence associated genes are first expressed at a relatively low intensity. Their intensity increases to peak level as an indication of the initiation of leaf senescence. The hierarchical and stable global circuit indicates that senescence is a stable process and is independent of a single or even a group of genes. Whether this is true for other genomes (as for example ageing in mammals) as well is an interesting and challenging problem. On the other hand, network inference results show the power of our Granger causality approach. Intuitively, one would not expect that gene 1, for example in Figure 5.8, would cause gene 3, since the down-regulation of gene 3 starts at an early time (day one).

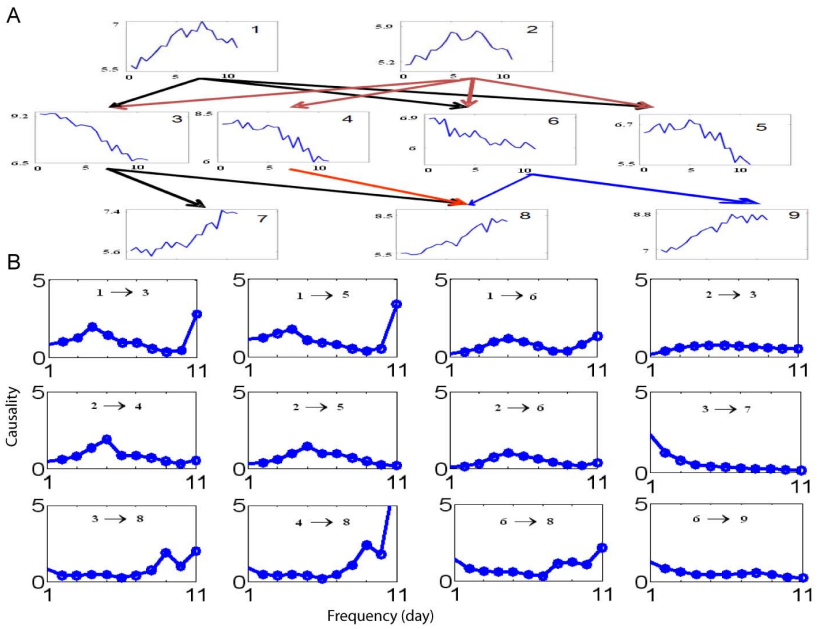


Figure 5.8: Causal relationship between genes: a global circuit. A. A total of 11 genes are shown and a clear hierarchy structure is demonstrated. B. The interactions in the frequency domain.

In Fig. 5.8B, the interactions in the frequency domain are plotted. We can see that the interactions at the frequency domain are different between the top and middle layer and the middle layer and bottom layer. In general, we have a peak in the middle frequency between the top and the middle layer: see for example, $1 \rightarrow 3$. But the interactions between the middle and the bottom layer are concentrated on either the high or the low frequencies.

5.3 Summary

In this chapter, we presented a complete work-flow for temporal microarray data processing. A fresh approach has been taken to accomplish each step in the work flow; from processing of raw data to gene network inference. The normalization method allows each gene to be represented as identical and independent stochastic process, and the auditory clustering reduces the data dimension by applying sim-

ple but powerful frequency based approach. We have shown in this chapter that the clustering method not only categorizes the genes according to their functionality but also allows a purely data driven natural ranking of genes based on their power spectrum profile. Two important concerns, namely, the optimal number of clusters in a dataset and ranking of each gene within each cluster are naturally handled using our method.

We describe in Section 5.2.3 how the natural ranking of genes allowed us to select key genes involved in the circadian circuit of Arabidopsis. Encouraged by these results, we decided to study the circadian circuit in more detail and analyse it. We used our method of complex and partial Granger causality to infer the gene interaction network for the circuit. Our time and frequency based analysis show that the computationally inferred network structure is in agreement with the experimental findings. We further applied Partial Granger Causality in time as well as frequency domain to selected genes involved in the Ethylene pathway. In the end, we clustered the complete dataset of 30,336 genes with a standard k-mean clustering method to detect any pattern among the genes. After selecting a representative gene from each cluster, we applied Partial Granger causality to obtain a global interaction circuit. A clear hierarchical interaction pattern emerged for the genes involved in the global circuit.

These are the first steps in applying a frequency domain approach to deal with temporal microarray data. There remain many issues to be further explored on the lines of frequency domain analysis. Is there any random gene (white signal) or a group of random genes having a flat power spectral density? Is there an link between the life-span distribution of genes and power spectral density distribution? Here we only checked for the frequency domain interactions at an

identical frequency. For a complex system, we expect that interactions at different frequencies exist. In the frequency domain analysis, it is known that the most efficient way to nullify an input signal at a given frequency is by applying a filter. Can we develop biological *filter* to fulfil certain purposes, for example, to prolong the life span of a leaf? An approach like this on datasets collected at shorter intervals can lead us to detection of other interesting frequencies hidden in the data.

Chapter 6
Summary and Future Work

This thesis attempts to introduce some new computational techniques to analyse temporal microarray data. We concentrated on the information processing part of microarray data analysis, essentially on normalization, clustering and reverse engineering of gene circuits from data.

Chapter 1 presented an overview of microarray techniques and the experimental details to produce the data used in this thesis. Chapter 2 introduced a normalization technique to clean the data from various unwanted artefacts. The effect of normalization was confirmed by verifying the distribution plot of residuals associated with genes, and by minimizing correlation coefficients between residual terms across replicates. In Chapter 3, we discussed a temporal-precedence based clustering technique to group functionally similar genes. The core method in the proposed clustering technique was based on the definition of causality provided by Clive Granger in Economics. Chapter 4 extended the idea of Granger causality to introduce a definition of Partial Granger Causality to infer causal network structures from gene-expression data. Finally in Chapter 5, we provided a complete analysis of Arabidopsis data using the techniques developed in the previous chapters. We also presented the analysis of data in the frequency domain and

discussed three gene circuits of potential interest to Arabidopsis researchers. In this chapter, we first summarize the important steps covered in each chapter, and then conclude with a discussion on future work.

6.1 Recapitulation

Chapter 1- Introduction

The chapter starts with a brief discussion of the central dogma of molecular biology and the process of gene expression. We provided an overview of the microarray technology which allows simultaneous measurement of thousands of genes. The chapter discusses the two broad stages in the processing of microarray data : a) material processing stage, which is mainly performed in wet labs and involves growth of biological samples, RNA extraction, hybridization experiments and collection of gene expression data, and b) information processing stage, which involves computational analysis of the collected data to extract meaningful signals and build plausible biological hypotheses. The focus of this thesis is on the development of new computational techniques for each step within the information processing stage of microarray data analysis. The chapter provides an overview of the obstacles faced at each step, and prepares a road-map for the techniques proposed in this thesis. In the final sections, we discussed the case study of the microarray experiment to understand the senescence process in leaves of Arabidopsis Thaliana.

Chapter 2 - Normalization of Gene Expression Data

The data must be normalized to minimize the systematic biases before some meaningful analysis can be carried out. The chapter introduces a novel method to normalize highly replicated microarray data using a statistical error model. The existing error models in literature have traditionally focused on modelling different

sources of variations as individual terms in their analysis. A careful application of the existing models can lead to minimization of the error sources included in the models, but may miss out the effects due to other sources of unwanted variations which were not explicitly handled. Our model, instead of distinguishing and incorporating individual sources of experimental variations, rather relies on the information present in terms of available biological and technical replicates, and groups all other sources of experimental variations in a single term for a separate treatment. A design of this nature makes the proposed normalization technique generic for processing any highly replicated microarray data. The execution steps in the model are modular in design, and at each step, statistical tests are defined to check for the behaviour of the estimated data. Such checks at each step can determine if the data meets the required standard, or, whether there is a need to execute the subsequent steps for further processing. The final aim of the normalization method is to have a zero mean Gaussian behaviour for the residuals within a replicate, and minimize the correlation between residuals across the replicates.

Chapter 3 - Functional Clustering of Gene Expression Data

Clustering is the next step in information processing stage of the microarray data analysis. The chapter starts with an overview of various clustering techniques applied in microarray data analysis. The chapter then introduces a new clustering technique based on the Granger test of causality to group temporal gene expressions in a microarray data. The approach further utilizes a graph-theoretic method to detect hubs and modules in the connection graph obtained from the clustering method. This is the first study based on the concept of Granger Causality which uses interdependence between two time series to construct a gene connection matrix. The Granger Causality test allows for the connection between genes to be determined based on the prediction, rather than based on associative measures

like point-wise, or, global or local shape-wise similarities. The graph-theoretic method helps in analysis of large connection graphs to automatically detect the hubs and modules in the network which could be potential candidates for functionally related biological modules. The concepts in the chapter are demonstrated with synthetic and real datasets of different sizes. The gene in the inferred modules obtained from the analysis of the real biological datasets, were queried against Gene Ontology (GO) databases to check for their biological functions. Our results show that the majority of the genes in the inferred modules were related in terms of their GO identifiers. The chapter further investigates the topological properties of the connection graph obtained for the larger dataset for Arabidopsis. Our analysis shows that the computed topological properties clearly distinguished our inferred graph as having the properties expected from of a real network compared to a random one.

Chapter 4 - Partial Granger Causality

This chapter introduces a new computational technique called Partial Granger Causality to infer causal network structures from temporal gene expression data. Partial Granger Causality extends the idea of pair-wise Granger causality in the context of multivariate systems. In a biological system, where genes are engaged in cross communication with each other in direct or indirect ways, a pairwise analysis to infer gene networks from data will not produce a desirable result. The concept of Partial Granger Causality builds on the idea of pairwise Granger Causality, and uses partial correlation to restrict the effects from other variables while computing the interdependence between two variables. Partial Granger Causality is capable of producing directed networks which are not possible with other widely used correlation based approaches like Relevance Networks or Gaussian Graphical Models. Partial Granger Causality also offers an advantage in terms of distinction

between direct and indirect links, which is an important concern while inferring network structures of biological entities. The chapter presents examples for testing Partial Granger Causality on various artificial datasets, and also on a real biological dataset for Human T-Cell activation. The implementation of Partial Granger Causality is simple and the models produced are statistically traceable at each stage. Partial Granger Causality does not require prior knowledge for inference of network structures from data. The linear form of Partial Granger Causality can easily be extended to non-linear forms when needed. Partial Granger Causality also has an equivalent representation in frequency domain which can be useful while understanding interactions at the frequencies of interest.

Chapter 5 - Listening to Genes

The chapter presents a complete pipeline for processing the Arabidopsis data using the techniques discussed in the previous chapters. It also presents a novel approach to cluster and rank genes according to their frequency profiles. A clustering of this type, naturally allows genes with similar dynamics (potentially indicative of similar biological functions) to rank close to each other in a frequency range. The chapter presents an extension of Partial Granger Causality introduced in previous chapter to infer interactions between sets of genes. The set-wise interaction method is named as Complex Granger Causality. The Complex Granger Causality results in a directed network where the interactions between different sets of genes, or influence of a set of genes on an individual gene is being estimated. Inference of set based interactions is a new concept in microarray data analysis. We have shown the usefulness of this concept with two examples of gene interactions belonging to different pathways of Arabidopsis. The chapter also presents three inferred gene circuits of potential interest to Arabidopsis researchers : a circadian gene circuit, an ethylene circuit, and a global gene circuit. The interactions are

presented both in time as well as frequency domain.

6.2 Future Work

While most of the work in this thesis is based on the same datasets, the computational techniques introduced in each chapter are quite generic in nature and a straightforward application of these methods to other datasets is possible. Although, the hierarchy in the information processing pipeline of microarray data analysis is fixed, the methods at each stage are independent of each other. While the technique at each stage attempts to solve a problem, it also raises a series of other questions and creates opportunities for further developments.

The normalization method proposed in Chapter 2 includes several factors of variations in its formulation, but there are many other factors which can be included explicitly and would depend on the design of experiment and the platform used. A particular issue is removal of location based print-tip effects. Since print-tip effects are specific to different arrays, and can also have variations within arrays, a global approach of having a single term like M_{gbi} in our model to handle such effects may not be sufficient. Similar issues can arise for other local experimental artefacts like gene-specific dye bias etc. Our model is inherently linear in its formulation at *all* the stages. The assumption of linear relationships needs to be overcome while handling the cases where the linearity between variable or replicates is not viable. This leads to another question, if there are other suitable alternatives to the select-and-reject algorithm, and under what conditions those alternatives can be applied for faster and better results. Our normalization method does not include the effect of temporal effects or correlations which may exist in form of *common cause* due to biological or non-biological(experimental) reasons. Much effort has been put into dealing with such temporal effects while

normalizing expression data. See for examples [ACDC+08, BJGS+03, LL04].

The association graph in Chapter 3 is built on the linear formulation of the Granger causality test presented in section 3.1. The relationship between all the pairs of genes need not be linear in the system, and thus the linear assumption in the Granger causality need not hold true for all the cases. Several non-linear formulations of Granger causality tests have been proposed in literature [AMS04, MPS06], but with the restriction of a limited number of temporal recordings for our data, a better strategy for performing the Granger causality test can be adopted. It will also be useful to have a comparison of network structures inferred from datasets of different sizes and see how the dataset size impacts the behaviour of the the results. We apply just one of the many techniques [BH03] proposed to detect dense regions in an association graph. Application of different techniques will lead to detection of different clusters and it will be interesting to see how the biological meanings of those cluster differ with the currently detected ones. The complexity of connections in the association graph can be reduced by avoiding the application of a global threshold to detect the dominant edges in the graph. An alternative technique like shortest-path measure [MC07, FJ09] between two genes in the association graph can be applied to keep the dominant connections. Such selections will rely more on the local connectivity patterns to decide on the edges to keep rather than a global threshold. Finally, the biological verification of detected subgraphs can be extended to other public ontology databases as well, and need not be dependent on the information provided by the Gene Ontology (GO) Consortium alone. Such information can help us answer questions like which development stages a set of genes are expressed at, or, whether they are involved in a certain disease or not; such questions are beyond the scope of the present GO system and a list of specific ontology databases can

be found at [SAR+09, web09].

Chapter 4 proposes an extension of Granger causality discussed in Chapter 3 to infer network structures from data in a multivariate context. Improvement steps like linear to non-linear models and comparisons using different dataset sizes as proposed in Chapter 3 apply to the technique in Chapter 4 as well. The unrestricted vector autoregressive model, as the one used in Chapter 4 is a theory-free method but can be computationally exhaustive. The unrestricted form can be simplified to restricted form at times when some information about the system is already known. Also, a Bayesian approach to fit vector autoregressive models can be applied in case of prior information about variables in the system. The Bayesian approach can also overcome the problems of over-parametrization faced by VAR models. It can happen that the residual terms are sensitive to normal distribution in some cases, especially with small sized datasets. Hacker and Hatemi-J [HHJ06] have recently proposed a bootstrap based approach in context of Granger causality that is not sensitive to the normal distribution of the error term, and can be studied further to improve our proposed method.

Our frequency based approach in Chapter 5 opens further possibilities for analysis of temporal data in frequency domain. A deeper understanding of physiological meaning of frequencies present in dataset need to be developed. As an example, Guo et al. [GWDF08] have reported the importance of three dominant frequencies in HeLa cell cycle. Though, the power spectral density of genes provided us important clues in analysing expression data, it lacked the variations at smaller frequencies due to absence of data points at smaller intervals. Collection of data at smaller intervals can increase the analysis power of the method. We also need to understand the distribution of individual spectral profiles at each

frequency. We saw that the data at frequencies of 1 day period and 22 day period in the power spectral density presented in section 5.2.2 had Weibull and Exponential distributions respectively. Can we related the life span distribution of genes with their corresponding power spectral density distribution ? In our analysis, we have only considered interactions at an identical frequency. The method can be extended to infer interactions at different frequency giving a dynamic profile of the activities taking place in the system. A widely used approach while dealing with frequency domain data is to develop filters to amplify or abbreviate a signal, a filter-based approach to understand the behaviour of the system can be helpful in gaining some useful insights about the functioning of the system. Last but not the least, the overall computational analysis of any experimental dataset depends on the high quality of data, and the choice of best experimental conditions to captures the biological variations in the best way.

Appendix A
Partial Granger Causality

Frequency Domain Formulation of Partial Granger Causality.

We discussed the time domain formulation of Partial Granger Causality (PGC) in Chapter 4. In this section, we present the frequency domain formulation of Partial Granger Causality. The formulation of PGC in frequency domain was proposed by Guo et al. [GWDF08]. We present the main results in this section. Before we start, we briefly discuss the following concepts which will be useful for understanding the further concepts.

The autoregressive model of time series assumes that x_t, the value of the process at a time t depends on its p previous values weighted by coefficients a plus a random white noise residual ϵ:

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \epsilon_t \tag{A.1}$$

In a multivariate case $\boldsymbol{X}(t)$ having k channels, the process value at time t is a vector of size k, the model coefficients $\boldsymbol{A}(t)$ are $k \times k$ matrices, and the residual component $\boldsymbol{E}$ is a vector of size k:

$$\boldsymbol{X}(t) = (X_1(t), X_2(t), \ldots, X_k(t))^T \quad (A.2)$$

$$\boldsymbol{X}(t) = \sum_{i=1}^{p} \boldsymbol{A}(i)\boldsymbol{X}(t-i) = \boldsymbol{E}(t) \quad (A.3)$$

Assuming that $\boldsymbol{A}(0) = \boldsymbol{I}$ (the identity matrix), and changing the signs of $\boldsymbol{A}(i)$ changed, the Equation (A.1) can be rewritten as

$$\boldsymbol{E}(t) = \sum_{i=1}^{p} \boldsymbol{A}(i)\boldsymbol{X}(t-i) \quad (A.4)$$

Equation (A.4) can be transformed to the frequency domain by applying the Z-transformation [OS89, Mar87], and can be written as

$$\boldsymbol{E}(f) = \boldsymbol{A}(f)\boldsymbol{X}(f) \quad (A.5)$$

$$\boldsymbol{X}(f) = \boldsymbol{A}^{-1}(f)\boldsymbol{E}(f) = \boldsymbol{H}(f)\boldsymbol{E}(f) \quad (A.6)$$

$$\boldsymbol{H}(f) = \left(\sum_{m=0}^{p} \boldsymbol{A}(m) exp(2\pi i m f \Delta t) \right) \quad (A.7)$$

where Δt is the sampling interval. The matrix $\boldsymbol{H}$ is called the transfer matrix of the system. The power spectrum of the signal is then computed as

$$\boldsymbol{S}(f) = \boldsymbol{X}(f)\boldsymbol{X}^*(f) = \boldsymbol{H}(f)\boldsymbol{E}(f)\boldsymbol{E}^*(f)\boldsymbol{H}^*(f) = \boldsymbol{H}(f)\boldsymbol{\Sigma}\boldsymbol{H}^*(f) \quad (A.8)$$

where $\boldsymbol{\Sigma}$ is the noise covariance matrix and is not dependent on the frequency. Based on these concepts, we explain how Partial Granger Causality can be formulated in frequency domain.

Consider two processes, X and Z in their time domain autoregressive representation using the lag polynomial L (where $L(X_t) = X_{t-1}, L^2(X_t) = X_{t-2}$ etc.)

$$\begin{pmatrix} D_{11}(L) & D_{12}(L) \\ D_{21}(L) & D_{22}(L) \end{pmatrix} \begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \begin{pmatrix} \phi_t \\ \psi_t \end{pmatrix} \tag{A.9}$$

with $D_{11}(0) = 1, D_{22}(0) = 1, D_{12}(0) = 0, D_{21}(0) = 0$ and $\sigma\phi_t, \psi_t = 0$. Let $\Sigma^{(1)}$ be the noise covariance matrix for the above system, where

$$\Sigma^{(1)} = \begin{pmatrix} \Sigma^{(1)}_{xx} & \Sigma^{(1)}_{xz} \\ \Sigma^{(1)}_{zx} & \Sigma^{(1)}_{zz} \end{pmatrix}$$

Similarly, for a system with three variables X, Y and Z, we have their similar representation as

$$\begin{pmatrix} B_{11}(L) & B_{12}(L) & B_{13}(L) \\ B_{21}(L) & B_{22}(L) & B_{23}(L) \\ B_{31}(L) & B_{32}(L) & B_{33}(L) \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} \epsilon_{xt} \\ \epsilon_{yt} \\ \epsilon_{zt} \end{pmatrix} \tag{A.10}$$

and the noise covariance matrix for the system can be denoted as $\Sigma^{(2)}$ where

$$\Sigma^{(2)} = \begin{pmatrix} \Sigma^{(2)}_{xx} & \Sigma^{(2)}_{xy} & \Sigma^{(2)}_{xz} \\ \Sigma^{(2)}_{yx} & \Sigma^{(2)}_{yy} & \Sigma^{(2)}_{yz} \\ \Sigma^{(2)}_{zx} & \Sigma^{(2)}_{zy} & \Sigma^{(2)}_{zz} \end{pmatrix}$$

Taking the Z-transformation of Equations (A.9) and (A.10), and applying the transformation matrix as shown in Equation (A.5), we have

$$\begin{pmatrix} X(\lambda) \\ Z(\lambda) \end{pmatrix} = \begin{pmatrix} G_{xx}(\lambda) & G_{xz}(\lambda) \\ G_{zx}(\lambda) & G_{zz}(\lambda) \end{pmatrix} \begin{pmatrix} \phi(\lambda) \\ \psi(\lambda) \end{pmatrix} \tag{A.11}$$

for Equation (A.9) and

$$\begin{pmatrix} X(\lambda) \\ Y(\lambda) \\ Z(\lambda) \end{pmatrix} = \begin{pmatrix} H_{xx}(\lambda) & H_{xy}(\lambda) & H_{xz}(\lambda) \\ H_{yx}(\lambda) & H_{yy}(\lambda) & H_{yz}(\lambda) \\ H_{zx}(\lambda) & H_{zy}(\lambda) & H_{zz}(\lambda) \end{pmatrix} \begin{pmatrix} \epsilon_x(\lambda) \\ \epsilon_y(\lambda) \\ \epsilon_z(\lambda) \end{pmatrix} \tag{A.12}$$

for Equation(A.10), where λ represents frequency. We make an assumption that the spectra of $X(\lambda)$ and $Z(\lambda)$ from Equation (A.11) remain identical to spectra from Equation (A.12), we can perform the following substitution,

$$\begin{pmatrix} \phi(\lambda) \\ Y(\lambda) \\ \psi(\lambda) \end{pmatrix} = \begin{pmatrix} G_{xx}(\lambda) & 0 & G_{xz}(\lambda) \\ 0 & I & 0 \\ G_{zx}(\lambda) & 0 & G_{zz}(\lambda) \end{pmatrix}^{-1} \begin{pmatrix} H_{xx}(\lambda) & H_{xy}(\lambda) & H_{xz}(\lambda) \\ H_{yx}(\lambda) & H_{yy}(\lambda) & H_{yz}(\lambda) \\ H_{zx}(\lambda) & H_{zy}(\lambda) & H_{zz}(\lambda) \end{pmatrix} \begin{pmatrix} \epsilon_x(\lambda) \\ \epsilon_y(\lambda) \\ \epsilon_z(\lambda) \end{pmatrix}$$

$$= \begin{pmatrix} Q_{xx}(\lambda) & Q_{xy}(\lambda) & Q_{xz}(\lambda) \\ Q_{yx}(\lambda) & Q_{yy}(\lambda) & Q_{yz}(\lambda) \\ Q_{zx}(\lambda) & Q_{zy}(\lambda) & Q_{zz}(\lambda) \end{pmatrix} \begin{pmatrix} \epsilon_x(\lambda) \\ \epsilon_y(\lambda) \\ \epsilon_z(\lambda) \end{pmatrix} \tag{A.13}$$

Now, consider $\Sigma^{(1)}$ and $\Sigma^{(2)}$ to be the noise covariance matrices for Equations (A.9) and (A.10) respectively. The spectral decomposition of the spectral density of X from Equation (A.9) can be denoted as

$$S_X(\lambda) = G_{xx}\hat{\Sigma^{(1)}_{xx}}G^*_{xx} + G_{xz}\hat{\Sigma^{(1)}_{zz}}G^*_{xz} \tag{A.14}$$

and the spectral decomposition of the spectral density of X from Equation (A.10) can be denoted as

$$S_X{}'(\lambda) = Q_{xx}\hat{\Sigma^{(2)}_{xx}}Q^*_{xx} + Q_{xy}\hat{\Sigma^{(2)}_{yy}}Q^*_{xy} + Q_{xz}\hat{\Sigma^{(2)}_{zz}}Q^*_{xz} \tag{A.15}$$

where $\hat{\Sigma^{(1)}_{xx}}$ and $\hat{\Sigma^{(2)}_{xx}}$ denote the conditional variances of residuals for X for

systems shown in Equations (A.9) and (A.10) respectively by partitioning the $\Sigma^{(1)}$ and $\Sigma^{(2)}$ matrices in way we explained in section 4.1.2 of Chapter 4. We have

$$\begin{cases} \hat{\Sigma^{(1)}_{xx}} &= \Sigma^{(1)}_{xx} - \Sigma^{(1)}_{xz}{\Sigma^{(1)}_{zz}}^{-1}\Sigma^{(1)}_{zx} \\ \hat{\Sigma^{(1)}_{zz}} &= \Sigma^{(1)}_{zz} \\ \hat{\Sigma^{(2)}_{xx}} &= \Sigma^{(2)}_{xx} - \Sigma^{(2)}_{xz}{\Sigma^{(2)}_{zz}}^{-1}\Sigma^{(2)}_{zx} \\ \hat{\Sigma^{(2)}_{zz}} &= \Sigma^{(2)}_{zz} \\ \hat{\Sigma^{(2)}_{yy}} &= \Sigma^{(2)}_{yy} - \Sigma^{(2)}_{yz}{\Sigma^{(2)}_{zz}}^{-1}\Sigma^{(2)}_{zy} - \frac{(\Sigma^{(2)}_{yx} - \Sigma^{(2)}_{yz}{\Sigma^{(2)}_{zz}}^{-1}\Sigma^{(2)}_{zx})(\Sigma^{(2)}_{xy} - \Sigma^{(2)}_{xz}{\Sigma^{(2)}_{zz}}^{-1}\Sigma^{(2)}_{zy})}{\Sigma^{(2)}_{xx} - \Sigma^{(2)}_{xz}{\Sigma^{(2)}_{zz}}^{-1}\Sigma^{(2)}_{zx}} \end{cases}$$

We require only the first terms from Equation (A.14) and (A.15) for our purpose, which can be thought of as intrinsic power after removing the effect of other variables in the system. This leads to the following definition of PGC at frequency λ:

$$f_{Y\to X|Z}(\lambda) = \ln \frac{|G_{xx}\hat{\Sigma^{(1)}_{xx}}G^*_{xx}|}{|Q_{xx}\hat{\Sigma^{(2)}_{xx}}Q^*_{xx}|} \tag{A.16}$$

By the Kolmogorov formula [Kol33] for spectral decompositions and under some mild conditions, the Granger causality in the frequency domain and in the time domain measures satisfies

$$F_{Y\to X|Z} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_{Y\to X|Z}(\lambda)d\lambda \tag{A.17}$$

Appendix B

Gene Annotations

B.1 Gene Annotations

Gene Number	Gene Name	Description
At2g40080.1	ELF4 (EARLY FLOWERING 4)	Encodes a novel nuclear 111 amino-acid phytochrome-regulated component of a negative feedback loop involving the circadian clock central oscillator components CCA1 and LHY. ELF4 is necessary for light-induced expression of both CCA1 and LHY, and conversely, CCA1 and LHY act negatively on light-induced ELF4 expression. ELF4 promotes clock accuracy and is required for sustained rhythms in the absence of daily light/dark cycles. It is involved in the phyB-mediated constant red light induced seedling deetiolation process and may function to coregulate the expression of a subset of phyB-regulated genes.
At1g01060.1, At1g01060.2, At1g01060.3, At1g01060.4	LHY (LATE ELONGATED HYPOCOTYL)	myb-related putative transcription factor involved in circadian rhythm along with another myb transcription factor CCA1
At2g46830.1, At2g46830.2	CCA1 (CIRCADIAN CLOCK ASSOCIATED 1)	Transcription factor; encodes a transcriptional repressor that performs overlapping functions with LHY in a regulatory feedback loop that is closely associated with the circadian oscillator of ARABIDOPSIS.
At5g61380.1	TOC1 (TIMING OF CAB1 1)	Transcription regulator; pseudo response regulator involved in the generation of circadian rhythms. TOC1 appears to shorten the period of circumnutation speed. TOC1 contributes to the plant fitness (carbon fixation, biomass) by influencing the circadian clock period.
At5g02810.1	PRR7 (PSEUDO-RESPONSE REGULATOR 7)	Transcription regulator; PRR7 and PRR9 are partially redundant essential components of a temperature-sensitive circadian system. CCA1 and LHY had a positive effect on PRR7 expression levels.
At2g46790.1, At2g46790.2	PRR9 (PSEUDO-RESPONSE REGULATOR 9)	Pseudo-response regulator PRR9. Involved in clock function. PRR7 and PRR9 are partially redundant essential components of a temperature-sensitive circadian system. CCA1 and LHY had a positive effect on PRR9. Interact with TOC1 in a yeast two-hybrid assay.
At1g22770.1	GI (GIGANTEA)	Together with CONSTANTS (CO) and FLOWERING LOCUS T (FT), GIGANTEA promotes flowering under long days in a circadian clock-controlled flowering pathway. GI acts earlier than CO and FT in the pathway by increasing CO and FT mRNA abundance. Located in the nucleus. Regulates several developmental processes, including photoperiod-mediated flowering, phytochrome B signaling, circadian clock, carbohydrate metabolism, and cold stress response. The gene's transcription is controlled by the circadian clock and it is post-transcriptionally regulated by light and dark.

Table B.1: Gene names and descriptions in circadian circuit.

B.2 Gene Annotations

Gene Number	Gene Name	Description
At5g03730.1 ,At5g03730.2	CTR1 (CONSTITUTIVE TRIPLE RESPONSE 1)	Kinase; Homologous to the RAF family of serine/threonine protein kinases. Negative regulator in the ethylene signal transduction pathway. Interacts with the putative ethylene receptors ETR1 and ERS. Constitutively expressed.
At1g66340.1	ETR1 (ETHYLENE RESPONSE 1)	Two-component response regulator; Similar to prokaryote sensory transduction proteins. Contains a histidine kinase and a response regulator domain. Homodimer. Membrane component. Binds ethylene. Mutations affect ethylene binding and metabolism of other plant hormones such as auxin, cytokinins, ABA and gibberellic acid. Ethylene receptor. Has histidine kinase activity.
At4g20880.1	ERT2	Ethylene-responsive nuclear protein / ethylene-regulated nuclear protein; similar to ethylene-responsive nuclear protein -related [Arabidopsis thaliana] (TAIR:AT5G44350.1); similar to IMP dehydrogenase/GMP reductase [Medicago truncatula] (GB:ABE90052.1)
At3g23240.1	ERF1 (ETHYLENE RESPONSE FACTOR 1)	DNA binding / transcription factor/ transcriptional activator; encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family (ERF1). The protein contains one AP2 domain. There are 18 members in this subfamily including ATERF-1, ATERF-2, AND ATERF-5. EREBP like protein that binds GCC box of ethylene regulated promoters such as basic chitinases. Constitutive expression of ERF1 phenocopies ethylene over production. Involved in ethylene signaling cascade,downstream of EIN2 and EIN3.

Continued on next page

At3g04580.1 ,At3g04580.2	EIN4 (ETHYLENE INSENSITIVE 4)	Ethylene receptor, subfamily 2. Has serine kinase activity.
At3g20770.1	EIN3 (ETHYLENE INSENSITIVE3)	Transcription factor; ethylene-insensitive3
At5g03280.1	EIN2 (ETHYLENE INSENSITIVE 2)	Transporter; Involved in ethylene signal transduction. Acts downstream of CTR1.
At5g21120.1	EIL2 (ETHYLENE INSENSITIVE3-LIKE 2)	Transcription factor; ethylene-insensitive3-like2 (EIL2)
At2g27050.1	EIL1 (ETHYLENE INSENSITIVE3-LIKE 1)	Transcription factor; ethylene-insensitive3-like1 (EIL1)
At1g73730.1	EIL3 (ETHYLENE INSENSITIVE3-LIKE3)	Transcription factor; Encodes a putative transcription factor involved in ethylene signalling. Isolated DNA binding domain has been shown to bind DNA in vitro.
At5g10120.1	Identical to Putative ETHYLENE-INSENSITIVE3-like 4 protein (EIL4)	Similar to ethylene insensitive 3 family protein [Arabidopsis thaliana] (TAIR:AT5G65100.1); similar to 52O08_27 [Brassica rapa subsp. pekinensis] (GB:AAZ67573.1); contains InterPro domain Ethylene insensitive 3; (InterPro:IPR006957)
At1g04310.1	ERS2 (ETHYLENE RESPONSE SENSOR 2)	receptor; encodes an ethylene receptor related to bacterial two-component histidine kinases.
At2g40940.1	ERS1 (ETHYLENE RESPONSE SENSOR 1)	receptor; Ethylene receptor, subfamily 1. Has histidine kinase activity.

Continued on next page

At3g33520.1	EIN6	structural constituent of cytoskeleton; Encodes ACTIN-RELATED PROTEIN6 (ARP6), a putative component of a chromatin-remodeling complex. Required for both histone acetylation and methylation of the FLC chromatin in Arabidopsis. Located at specific regions of the nuclear periphery. Expression throughout plants shown by in-situ and immunolocalization methods. Mutants show defects in fertility, leaf, flower and inflorescence development and shorter flowering times.
At1g55010.1	PDF1.2	similar to PDF 1.5 ; similar to Cysteine-rich antifungal protein 4 precursor (AFP4) (GB:O24331); contains InterPro domain Gamma thionin; (InterPro:IPR008176)
At5g65100.1	ethylene insensitive 3 family protein; Identical to Putative ETHYLENE-INSENSITIVE3-like 5 protein (EIL5)	similar to ethylene insensitive 3 family protein [Arabidopsis thaliana] (TAIR:AT5G10120.1); similar to 52O08_27 [Brassica rapa subsp. pekinensis] (GB:AAZ67573.1); contains InterPro domain Ethylene insensitive 3; (InterPro:IPR006957)

Table B.2: Gene names and descriptions in ethylene circuit.

B.3 Gene Annotations

Gene Name	Description
At2g34960.1	CAT5 (CATIONIC AMINO ACID TRANSPORTER 5);cationic amino acid transporter; Arabidopsis thaliana amino acid permease family protein (At2g34960)
At3g09900.1 AtRABE1e / AtRab8E	Arabidopsis Rab GTPase homolog E1e; GTP binding; similar to AtRABE1d/AtRab8C, GTP binding [Arabidopsis thaliana] (TAIR:AT5G03520.1); similar to ras-related protein RAB8-3 [Nicotiana tabacum] (GB:BAB84324.1); contains InterPro domain Small GTP-binding protein domain; (InterPro:IPR005225); contains InterPro domain Ras small GTPase, Rab type; (InterPro:IPR003579); contains InterPro domain Sigma-54 factor, interaction region; (InterPro:IPR002078); contains InterPro domain Ras GTPase; (InterPro:IPR001806); contains InterPro domain Ras; (InterPro:IPR013753)
At4g38495.1	Unknown protein; similar to conserved hypothetical protein [Aedes aegypti] (GB:EAT47050.1); similar to OSIGBa0138E08-OSIGBa0161L23.9 [Oryza sativa (indica cultivar-group)] (GB:CAH67928.1); similar to Os04g0274400 [Oryza sativa (japonica cultivar-group)] (GB:NP_001052351.1); contains InterPro domain YL1 nuclear, C-terminal; (InterPro:IPR013272)
At1g03550.1	Secretory carrier membrane protein (SCAMP) family protein; similar to secretory carrier membrane protein (SCAMP) family protein [Arabidopsis thaliana] (TAIR:AT2G20840.1); similar to similarity to SCAMP37 [Pisum sativum] (GB:AAC82326.1); similar to Os01g0780500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001044437.1); similar to Os07g0564600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001060004.1); contains InterPro domain SCAMP; (InterPro:IPR007273)
At5g21950.1	Hydrolase, alpha/beta fold family protein; similar to hydrolase, alpha/beta fold family protein [Arabidopsis thaliana] (TAIR:AT4G33180.1); similar to Alpha/beta hydrolase fold [Medicago truncatula] (GB:ABE81749.1); contains InterPro domain Esterase/lipase/thioesterase; (InterPro:IPR000379); contains InterPro domain Alpha/beta hydrolase fold-1; (InterPro:IPR000073); contains InterPro domain Alpha/beta hydrolase; (InterPro:IPR003089)
At1g53170.1	ATERF-8/ATERF8 (ETHYLENE RESPONSE ELEMENT BINDING FACTOR 4); DNA binding / transcription factor/ transcriptional repressor; encodes a member of the ERF (ethylene response factor) subfamily B-1 of ERF/AP2 transcription factor family (ATERF-8). The protein contains one AP2 domain. There are 15 members in this subfamily including ATERF-3, ATERF-4, ATERF-7, and leafy petiole.

Table B.3: Cluster 1: Some gene names and descriptions.

Gene Name	Description
At4g38250.1, At4g38260.1	Amino acid transporter family protein; similar to amino acid transporter family protein [Arabidopsis thaliana] (TAIR:AT2G42005.1); similar to amino acid transport protein (GB:AAB82307.1); similar to OSIGBa0158F05.8 [Oryza sativa (indica cultivar-group)] (GB:CAH66759.1); similar to OSJNBa0017B10.14 [Oryza sativa (japonica cultivar-group)] (GB:CAE03099.2); contains InterPro domain Amino acid/polyamine transporter II; (InterPro:IPR002422); contains InterPro domain Amino acid transporter, transmembrane; (InterPro:IPR013057) @ unknown protein; similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G20740.1); similar to H0409D10.8 [Oryza sativa (indica cultivar-group)] (GB:CAH66750.1); similar to Os09g0323500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001062871.1); contains InterPro domain Protein of unknown function DUF833; (InterPro:IPR008551)
At5g22210.1, At5g22210.2	Unknown protein
At1g78270.1	UDP-glucose glucosyltransferase, putative; similar to UDP-glucoronosyl/UDP-glucosyl transferase family protein [Arabidopsis thaliana] (TAIR:AT1G22360.1); similar to transcription factor/ transferase, transferring glycosyl groups [Arabidopsis thaliana] (TAIR:AT1G22380.1); similar to UGT85A1 (UDP-glucosyl transferase 85A1), UDP-glycosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups [Arabidopsis thaliana] (TAIR:AT1G22400.1); similar to glycosyltransferase NTGT5b [Nicotiana tabacum] (GB:BAD93690.1); contains InterPro domain UDP-glucuronosyl/UDP-glucosyltransferase; (InterPro:IPR002213)
At2g29640.1	Josephin family protein; Identical to Josephin-like protein [Arabidopsis Thaliana] (GB:O82391); similar to josephin protein-related [Arabidopsis thaliana] (TAIR:AT1G07300.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:AAP06835.1); similar to Os03g0265200 [Oryza sativa (japonica cultivar-group)] (GB:NP_001049646.1); contains InterPro domain Machado-Joseph disease protein MJD; (InterPro:IPR006155)
At5g46190.1	KH domain-containing protein; similar to KH domain-containing protein [Arabidopsis thaliana] (TAIR:AT4G18375.2); similar to Os08g0200400 [Oryza sativa (japonica cultivar-group)] (GB:NP_001061211.1); similar to KH, type 1 [Medicago truncatula] (GB:ABE79454.1); contains InterPro domain KH; (InterPro:IPR004087); contains InterPro domain KH, type 1; (InterPro:IPR004088)
At5g41765.1	Unknown protein; similar to unknown protein [Arabidopsis thaliana] (TAIR:AT4G00232.1); contains InterPro domain Protein of unknown function DUF573; (InterPro:IPR007592)
At2g38050.1	DET2 (DE-ETIOLATED 2); Similar to mammalian steroid-5-alpha-reductase. Involved in the brassinolide biosynthetic pathway.

Table B.4: Cluster 2: Some gene names and descriptions.

Gene Name	Description
At3g52500.1	Aspartyl protease family protein; similar to aspartyl protease family protein [Arabidopsis thaliana] (TAIR:AT4G16563.1); similar to aspartic protease [Fagopyrum esculentum] (GB:AAS48510.2); contains InterPro domain Peptidase A1, pepsin; (InterPro:IPR001461); contains InterPro domain Peptidase aspartic, catalytic; (InterPro:IPR009007)
At4g23820.1	Glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein; similar to glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein [Arabidopsis thaliana] (TAIR:AT5G41870.1); similar to Os05g0587000 [Oryza sativa (japonica cultivar-group)] (GB:NP_001056466.1); similar to Os02g0256100 [Oryza sativa (japonica cultivar-group)] (GB:NP_001046468.1); similar to putative polygalacturonase [Oryza sativa (japonica cultivar-group)] (GB:AAT44156.1); contains InterPro domain Virulence factor, pectin lyase fold; (InterPro:IPR011050); contains InterPro domain Glycoside hydrolase, family 28; (InterPro:IPR000743); contains InterPro domain Pectolytic enzyme, Pectin lyase fold; (InterPro:IPR012334)

Table B.5: Cluster 3: Some gene names and descriptions.

Gene Name	Description
At3g15190.1	Chloroplast 30S ribosomal protein S20, putative; Identical to 30S ribosomal protein S20, chloroplast precursor (RPS20) [Arabidopsis Thaliana] (GB:Q9ASV6;GB:Q9LIL6); similar to Os01g0678600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001043859.1); similar to ribosomal protein rpS20 [Bigelowiella natans] (GB:AAP79183.1); contains InterPro domain Ribosomal protein S20; (InterPro:IPR002583); contains InterPro domain Ribosomal protein S20p; (InterPro:IPR010013)
At1g15290.1	Binding; similar to binding [Arabidopsis thaliana] (TAIR:AT4G28080.1); similar to tetratricopeptide repeat (TPR)-containing protein [Arabidopsis thaliana] (TAIR:AT1G01320.1); similar to putative tetratricopeptide repeat(TPR)-containing protein [Oryza sativa (japonica cultivar-group)] (GB:BAC84544.1); similar to TPR repeat [Medicago truncatula] (GB:ABE77904.1); similar to H0811D08.1 [Oryza sativa (indica cultivar-group)] (GB:CAJ86110.1); contains InterPro domain Tetratricopeptide region; (InterPro:IPR013026); contains InterPro domain Tetratricopeptide TPR_1; (InterPro:IPR001440); contains InterPro domain Tetratricopeptide TPR_2; (InterPro:IPR013105); contains InterPro domain Tetratricopeptide-like helical; (InterPro:IPR011990)

Table B.6: Cluster 4: Some gene names and descriptions.

Gene Name	Description
At5g36170.1, At5g36170.2, At5g36170.3	HCF109 (HIGH CHLOROPHYLL FLUORESCENT 109); translation release factor; Required for normal processing of polycistronic plastidial transcripts
At1g72310.1	ATL3 (Arabidopsis Tx́icos en Levadura 3); protein binding / zinc ion binding; Encodes a putative RING-H2 zinc finger protein ATL3 (ATL3).

Table B.7: Cluster 5: Some gene names and descriptions.

Gene Name	Description
At1g47900.1	Unknown protein; similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G19835.1); similar to Putative myosin-like protein [Oryza sativa (japonica cultivar-group)] (GB:AAL77142.1); similar to Os03g0246500 [Oryza sativa (japonica cultivar-group)] (GB:NP_001049544.1); contains InterPro domain Protein of unknown function DUF869, plant; (InterPro:IPR008587)

Table B.8: Cluster 6: Some gene names and descriptions.

Gene Name	Description
At2g39725.1, At2g39725.2	Complex 1 family protein / LVR family protein; similar to Os08g0278600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001061438.1); similar to unknown protein [Oryza sativa (japonica cultivar-group)] (GB:BAC99750.1); contains InterPro domain Complex 1 LYR protein; (InterPro:IPR008011)
At5g07410.1	Pectinesterase family protein; similar to ATPPME1, pectinesterase [Arabidopsis thaliana] (TAIR:AT1G69940.1); similar to pectin methylesterase allergenic protein [Salsola kali] (GB:AAX11262.1); contains InterPro domain Virulence factor, pectin lyase fold; (InterPro:IPR011050); contains InterPro domain Pectinesterase; (InterPro:IPR000070)

Table B.9: Cluster 7: Some gene names and descriptions.

Gene Name	Description
At1g59670.1	ATGSTU15 (Arabidopsis thaliana Glutathione S-transferase (class tau) 15); glutathione transferase; Encodes glutathione transferase belonging to the tau class of GSTs. Naming convention according to Wagner et al. (2002).
At1g32960.1	Subtilase family protein; similar to subtilase family protein [Arabidopsis thaliana] (TAIR:AT1G32950.1); similar to subtilase family protein [Arabidopsis thaliana] (TAIR:AT4G10540.1); similar to subtilase family protein [Arabidopsis thaliana] (TAIR:AT1G32940.1); similar to Os09g0530800 [Oryza sativa (japonica cultivar-group)] (GB:NP_001063751.1); similar to Protease-associated PA; Proteinase inhibitor I9, subtilisin propeptide [Medicago truncatula] (GB:ABE90461.1); contains InterPro domain Protease-associated PA; (InterPro:IPR003137); contains InterPro domain Peptidase S8 and S53, subtilisin, kexin, sedolisin; (InterPro:IPR000209); contains InterPro domain Proteinase inhibitor I9, subtilisin propeptide; (InterPro:IPR010259); contains InterPro domain Proteinase inhibitor, propeptide; (InterPro:IPR009020)
At2g39440.1	Unknown protein; similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G61280.1); similar to hypothetical protein MtrDRAFT_AC126784g11v2 [Medicago truncatula] (GB:ABE94681.1); contains InterPro domain PIG-P; (InterPro:IPR013717)
At3g54140.1	Proton-dependent oligopeptide transport (POT) family protein; similar to ATPTR2-B (NITRATE TRANSPORTER 1), transporter [Arabidopsis thaliana] (TAIR:AT2G02040.1); similar to proton-dependent oligopeptide transport (POT) family protein [Arabidopsis thaliana] (TAIR:AT1G62200.1); similar to proton-dependent oligopeptide transport (POT) family protein [Arabidopsis thaliana] (TAIR:AT5G01180.1); similar to LeOPT1 [Lycopersicon esculentum] (GB:AAD01600.1); similar to putative peptide transport protein [Oryza sativa (japonica cultivar-group)] (GB:BAD31819.1); similar to peptide transporter [Hordeum vulgare] (GB:AAC32034.1); contains InterPro domain TGF-beta receptor, type I/II extracellular region; (InterPro:IPR000109)

Table B.10: Cluster 8: Some gene names and descriptions.

Gene Name	Description
At3g55470.1, At3g55470.2	C2 domain-containing protein; similar to C2 domain-containing protein [Arabidopsis thaliana] (TAIR:AT1G63220.1); similar to Os-FIERG2 gene product [Oryza sativa] (GB:AAC04628.1); contains InterPro domain C2; (InterPro:IPR000008); contains InterPro domain C2 calcium/lipid-binding region, CaLB; (InterPro:IPR008973)
At2g35070.1	Unknown protein; similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G35090.1); similar to conserved hypothetical protein [Medicago truncatula] (GB:ABE89621.1); contains domain UNCHARACTERIZED (PTHR14360)
At1g27300.1	Unknown protein; similar to Os02g0509600 [Oryza sativa (japonica cultivar-group)] (GB:NP_001046928.1)

Table B.11: Cluster 9: Some gene names and descriptions.

Appendix C

Publications

- R. Krishna and S. Guo, "A Partial Granger Causality approach to explore causal networks derived from multi-parameter data," Lecture notes in Computer Science, Springer, vol. 5307, pp. 9–27, 2008.

- J. Feng, D. Yi, R. Krishna, S. Guo, V. Buchanan-Wollaston, "Listen to Genes: Dealing with Microarray Data in the Frequency Domain," PLoS ONE 4(4): e5098. doi:10.1371/journal.pone.0005098, 2009.

- R. Krishna, C-T. Li and V. Buchanan-Wollaston,"Interaction Based Functional Clustering of Genomic Data," IEEE International Conference on Bioinformatics and Bioengineering(BIBE),Taichung, Taiwan, 22-24 June, 2009.

Bibliography

[ACDC+08] C. Angelini, L. Cutillo, D. De Canditiis, M. Mutarelli, and M. Pensky. Bats: a bayesian user-friendly software for analyzing time series microarray experiments. *BMC Bioinformatics*, 9(1):415, 2008.

[Aka69] H. Akaike. Fitting autoregressive models for regression. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.

[Alo07] U. Alon. *An introduction to Systems Biology : Design Principles of Biological Circuits*. Chapman & HallCRC, 2007.

[AMS04] N. Ancona, D. Marinazzo, and S. Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70:056221, 2004.

[AYA07] I. P. Androulakis, E. Yang, and R. R. Almon. Analysis of time-series gene expression data: Methods, challenges and opportunities. *Annual Review of Biomedical Engineering*, 9:205–228, 2007.

[BA99] A. L. Barabsi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[Bar02] A. Barabsi. *Linked: The New Science of Networks*. Basic Books, 2002.

[Ber01] N. Berkum. Dna microarrays: raising the profile. *Current Opinion in Biotechnology*, 12(1):48–52, 2001.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R.Stat. Soc.*, B 57:289–300, 1995.

[BH03] G. Bader and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[BHWK05] R. Balasubramaniyan, E. Hullermeier, N. Weskamp, and J. Kamper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–77, 2005.

[BIAS03] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. Variance and bias to compare normalization methods for high density oligonucleotide array data. *Bioinformatics*, 19:185–193, 2003.

[BJGJ+03] Z. Bar-Joseph, G. Gerber, T.S. Jaakkola, D.K. Gifford, and I. Simon. Continuous representations of time series gene expression data. *J. Comput. Biol.*, 3(4):341–356, 2003.

[BJGS+03] Z. Bar-Joseph, G. Gerber, I. Simon, D.K. Gifford, and T.S. Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *PNAS*, 100(18):10146–10151, 2003.

[BS01] L. Baccala and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84:463–474, 2001.

[CCWN+01] A. Claridge-Chang, H. Wijnen, F. Naef, C. Boothroyd, and N. et al. Rajewsky. Circadian regulation of gene expression systems in the drosophila head. *Neuron*, 32:657–671, 2001.

[CDB97] Y. Chen, E.R. Dougherty, and M.L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *J. Biomed. Optics*, 2:364–374, 1997.

[CFS99] T. Chen, V. Filkov, and S. Skiena, editors. *Identifying gene regulatory networks from experimental data*, 1999.

[CKS+03] Y.J. Chen, R. Kodell, F. Sistare, K.L. Thompson, S. Morris, and J.J. Chen. Normalization methods for analysis of microarray gene-expression data. *J Biopharm Stat.*, 13(1):57–74, 2003.

[Cle79] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc*, 74:829–836, 1979.

[Com09] Wikimedia Commons. Main page — wikimedia commons, 2009.

[CONG02] L. Cheng, C. Ohlen, B. Nelson, and P. Greenberg. Enhanced signaling through the il-2 receptor in cd8+ t cells regulated by antigen recognition results in preferential proliferation and expansion of responding cd8+ t cells rather than promotion of cell death. *PNAS*, 99(5):3001–3006, 2002.

[CP91] S. Chatterjee and B. Price. *Regression Analysis by Example*. John Wiley & Sons, 1991.

[CSC+92] C. Cambiaggi, M. Scupoli, F. Cestari, T.and Gerosa, G. Carra, G. Tridente, and R. Accolla. Constitutive expression of cd69 in in-

terspecies t-cell hybrids and locus assignment to human chromosome 12. *Immunogenetics*, 36:117–120, 1992.

[DDB+02] M.R. Doyle, S.J. Davis, R.M. Bastow, H.G. McWatters, L.A. Kozma-Bogn, F. Nagy, A.J. Millar, and R.M. Amasino. The elf4 gene controls circadian rhythms and flowering time in arabidopsis thaliana. *Nature*, 419:74–77, 2002).

[DES08] M. Dehmer and F. Emmert-Streib, editors. *Analysis of Microarray Data: A Network-Based Approach*. Wiley-VCH, 2008.

[DGM+06] N. Dojer, A. Gambin, A. Mizera, B. Wilczynski, and J. Tiuryn. Applying dynamic bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, 7:249, 2006.

[DH05] P. DHaeseleer. How does gene expression clustering work? *Nat. Biotechnol.*, 23(12):1499–1501, 2005.

[DIB97] J.L. DeRisi, V.R. Iyer, and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6, 1997.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J Royal Stat. Soc.*, B-39:1–38, 1977.

[DSS03] K. Dobbin, J. H. Shih, and R. Simon. Statistical design of reverse dye microarrays. *Bioinformatics*, 19(7):803–810, 2003.

[EBJ06] J. Ernst and Z. Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(1):191, 2006.

[Edw03] D Edwards. Non-linear normalization and background correction in one-channel cdna microarray studies. *Bioinformatics*, 19:825–833, 2003.

[EHI03] G.S. Eichler, S. Huang, and D.E. Ingber. Gene expression dynamics inspector (gedi): for integrative analysis of expression profiles. *Bioinformatics*, 19(17):2321–22, 2003.

[ESBB98] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–68, 1998.

[FHP05] J. Fan, T. Huang, and H. Peng. Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *Journal of American Statistical Association*, 100:781–813, 2005.

[FJ09] A.M. Fitch and M.B. Jones. Shortest path analysis using partial correlations for classifying gene functions from gene expression data. *Bioinformatics*, 25(1):42–47, 2009.

[FLNP00] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J. Computational Biology*, 7:601–620, 2000.

[FTVWY04] J. Fan, P. Tam, G. Vande Woude, and Ren Y. Normalization and analysis of cdna micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *PNAS*, 101:1135–1140, 2004.

[FYK+09] J. F. Feng, D. Yi, R. Krishna, S. Guo, and V. Buchanan-Wollaston.

Listen to genes: Dealing with microarray data in the frequency domain. *PLos ONE*, 4(4):e5098, 2009.

[gen00] Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.

[gen09] Genespring gx software version 10, 2009.

[Gew82] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77:304–313, 1982.

[GN86] C. Granger and P. Newbold. *Forecasting Economic Time Series.* Academic Press, 1986.

[Gol84] A.V. Goldberg. Finding a maximum density subgraph. Technical report, EECS Department, University of California, Berkeley, 1984.

[GR69] J. C. Gower and G. J.S. Ross. Minimum spanning trees and single linkage analysis. *Appl. Stat.*, 18:54–64, 1969.

[GR03] D. S. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100:4372–4376, 2003.

[Gra69] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[GSK+00] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, and M.B. et al. Eisen. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–57, 2000.

[GWDF08] S. Guo, J. H. Wu, M. Z. Ding, and J. F. Feng. Uncovering interactions in the frequency domain. *PLoS Comp. Biology*, 4(5):e1000087, 2008.

[HE04] Guo H. and J.R. Ecker. The ethylene signaling pathway: New insights. *Curr. Opin. Plant Biol.*, 7:40–49, 2004.

[HHJ06] R.S. Hacker and A. Hatemi-J. Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application. *Applied Economics*, 38(13):1489–1500, 2006.

[HHS+00] S. L. Harmer, J. B. Hogenesch, M. Straume, H. S. Chang, and Han B. et al. Orchestrated transcription of key pathways in arabidopsis by the circadian clock. *Science*, 290:2110–2113, 2000.

[HRAA+00] R. Hidi, V. Riches, M. Al-Ali, W. W. Cruikshank, D. M. Center, S. T. Holgate, and R. Djukanovic. Role of b7-cd28/ctla-4 costimulation and nf-kappa b in allergen-induced t cell chemotaxis by il-16 and rantes. *J. Immunol.*, 164(1):412–8, 2000.

[HVV+04] K. Himanen, M. Vuylsteke, S. Vanneste, S. Vercruysse, E. Boucheron, P. Alard, D. Chriqui, M. Van Montagu, D. Inze, and T. Beeckman. Transcript profling of early lateral root initiation. *PNAS*, 101(14):5146–5151, 2004.

[ima09] Imagene software, 2009.

[JMBO01] H. Jeong, S. P. Mason, A.-L. Barabsi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[JT05] L. Ji and K-L. Tan. Identifying time-lagged gene clusters using gene expression data. *Bioinformatics*, 21(4):509–516, 2005.

[JW88] R. Johnson and D. Wichern. *Applied multivariate statistical analysis.* Prentice-Hall, 1988.

[Kau93] S. A. Kauffman. *The Origins of Order.* Oxford University Press, 1993.

[KC01] M. K. Kerr and G. A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 77:123–128, 2001.

[KCM02] T.B. Kepler, L. Crosby, and K.T. Morgan. Normalization and analysis of dna microarray data by self-consistency and local regression. *Genome Biology*, 3:RESEARCH0037, 2002.

[KG08] R. Krishna and S. Guo. A partial granger causality approach to explore causal networks derived from multi-parameter data. *Lecture notes in Computer Science*, 5307:9–27, 2008.

[KHN03] A.T. Kwon, H.H. Hoos, and R. Ng. Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, 19:905–912, 2003.

[KIM03] S. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Bioinformatics*, 4(3):228–235, 2003.

[Kit02] H. Kitano. Computational system biology. *Nature*, 420:206–210, 2002.

[KKC00] M. K. Kerr, M. Kartin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

[KLBW09] R. Krishna, C-T. Li, and V. Buchanan-Wollaston. Interaction based functional clustering of genomic data. *In Proceedings of the IEEE International Conference on Bioinformatics and Bioengineering(BIBE'09)*, 2009.

[KLW06] B.-R. Kim, R. C. Littell, and R. L. Wu. Clustering the periodic pattern of gene expression using fourier series approximations. *Curr. Genomics*, 7:197–203, 2006.

[Kol33] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin, 1933.

[LKB+07] P. O. Lim, Y. Kim, E. Breeze, J. C. Koo, H. R. Woo, J. S. Ryu, D. H. Park, J. Beynon, A. Tabrett, V. Buchanan-Wollaston, and H. G. Nam. Overexpression of a chromatin architecture-controlling at-hook protein extends leaf longevity and increases the post-harvest storage life of plants. *The Plant Journal*, 52:1140–1153, 2007.

[LKBG+06] J. C. W. Locke, L. Kozma-Bognar, P. D. Gould, B. Feher, E. Kevei, F. Nagy, M. S. Turner, A. Hall, and A. J. Millar. Experimental validation of a predicted feedback loop in the multi-oscillator clock of arabidopsis thaliana. *Molecular Systems Biology*, 2:59, 2006.

[LL04] Y. Luan and H. Li. Model-based methods for identifying periodically regulated genes based on the time course microarray gene expression data. *Bioinformatics*, 20:332–339, 2004.

[LSGH02] P.D. Lee, R. Sladek, C.M. Greenwood, and T.J. Hudson. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res.*, 12(2):292–7, 2002.

[LZQ+04] X. Lu, W. Zhang, Z.S. Qin, K. Kwast, and J.S. Liu. Statistical resynchronization and bayesian detection of periodically expressed genes. *Nucleic Acids Res.*, 32:447–455, 2004.

[Mar87] S.L. Marple. *Digital Spectral Analysis with Applications.* Prentice-Hall Signal Processing Series, 1987.

[mat] Matlab version 6.5.1. natick, massachusetts: The mathworks inc., 2003.

[MC07] N. Mukhopadhyay and S. Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23:442–449, 2007.

[MHK05] S. Maere, K. Heymans, and M. Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21:3448–3449, 2005.

[MKH+07] H.G. McWatters, E. Kolmos, A. Hall, M.R. Doyle, R.M. Amasino, P. Gyula, F. Nagy, A.J. Millar, and S.J. Davis. Elf4 is required for oscillatory properties of the circadian clock. *Plant Physiology*, 144:391–401, 2007.

[MLW03] K.H. Moller-Levet, C.S.and Chu and O. Wolkenhauer. Dna microarray data clustering based on temporal variation: Fcv with tsd preclustering. *Appl. Bioinformatics*, 2:35–45, 2003.

[MPS06] D. Marinazzo, M. Pellicoro, and S. Stramaglia. Nonlinear parametric model for granger causality of time series. *Physical Review E*, 73:066216, 2006.

[NJ00] Stepanova A. N. and Ecker J.R. Ethylene signaling: from mutants to molecules. *Current Opinion in Plant Biology*, 3(5):353–360, 2000.

[NJW02] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14, 2002.

[NU08] R. Nagarajan and M. Upreti. Comment on causality and pathway search in microarray time series experiment. *Bioinformatics*, 24(7):1029–1032, 2008.

[OS89] A.V. Oppenheim and R.W. Schafer. *Discrete-time signal processing.* Prentice-Hall, 1989.

[Pes05] M. Pessarakli, editor. *Handbook of photosynthesis.* Taylor & Francis, 2005.

[PGQ+99] J. M. Pasque, B. Gross, L. Quek, N. Asazuma, W. Zhang, C. L. Sommers, E. Schweighoffer, V. Tybulewicz, B. Judd, J. R. Lee, G. Koretzky, P. E. Love, L. E. Samelson, and S. P. Watson. Lat is required for tyrosine phosphorylation of phospholipase cgamma2 and platelet activation by the collagen receptor gpvi. *Mol. Cell Biol.*, 19:8326–34, 1999.

[PLL02] W. Pan, J. Lin, and C. T. Le. Model-based cluster analysis of microarray geneexpression data. *Genome Biol.*, 3(2):RESEARCH0009, 2002.

[PREF01] D. Pe'er, A. Regev, E. Elidan, and N. Friedman. Inferring subnetworks from preturbed expression profiles. *Bioinformatics*, 17:S215–S224, 2001.

[PYK+03] T. Park, S-G Yi, S-H Kang, S.Y. Lee, Y-S Lee, and R. Simon. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4:33, 2003.

[Qua01] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet.*, 2(6):418–427, 2001.

[Qua02] J. Quackenbush. Microarray data normalization and transformation. *Natuture Genetics*, Suppl 32:496–501, 2002.

[Qua07] J. Quackenbush. Extracting biology from high-dimensional biological data. *The Journal of Experimental Biology*, 210:1507–1517, 2007.

[RAG+04] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. Wild, and F. Falciani. Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

[RK02] P. Ramoni, M.F.and Sebastiani and I.S. Kohane. Cluster analysis of gene expression dynamics. *PNAS*, 99:9121–9126, 2002.

[RSM+02] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002.

[SA05] A. N. Stepanova and J. M. Alonso. Arabidopsis ethylene signaling pathway. *Science*, 276:1872–1874, 2005.

[SAR+09] B. Smith, M. Ashburner, C. Rosse, C. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, and C.J. Mungall. The open biomedical ontologies. `http://www.obofoundry.org/`, April 2009.

[Sch79] G. W. Schwert. Tests of causality: The message in the innovations. *Carnegie-Rochester Conference Series on Public Policy*, 10, (1):55–96, 1979.

[Seb84] G. A. F. Seber. *Multivariate Observations*. John Wiley & Sons Inc., 1984.

[Sma94] C.M. Smart. Gene expresssion during leaf senescence. *New. Phytol*, 126:419448, 1994.

[SMO+03] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504, 2003.

[Spe03] T. Speed. *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC, 2003.

[SS03a] A. Schliep, A.and Schonhuth and C. Steinhoff. Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, 19:264–272, 2003.

[SS03b] G.K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–73, 2003.

[SSS05] I. Schliep, A.and Costa, C. Steinhoff, and A. Schonhuth. Analyzing gene expression time-courses. *IEEE/ACM Transactions on computational biology and bioinformatics*, 2(3):179–193, 2005.

[SSZ+98] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces

cerevisiae by microarray hybridization ,. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[TBW+02] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. Celniker, and G. Rubin. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3(12):research0088, 2002.

[THC+99] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3):281–85, 1999.

[THU+02] S. Takeno, K. Hirakawa, T. Ueda, K. Furukido, R. Osada, and K. Yajin. Nuclear factor-kappa b activation in the nasal polypepithelium: relationship to local cytokine gene expression. *Laryngoscope*, 112(1):53–58, 2002.

[TRLW01] M.K. Tseng, G.C.and Oh, L. Rohlin, J.C. Liao, and W.H Wong. Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, 29:2549–2557, 2001.

[TSKS04] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101:2981–2986, 2004.

[TZL+99] O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen. Housekeeping genes

as internal standards: use and limits. *J Biotechnol.*, 75(2-3):291–5, 1999.

[Ued06] H. Ueda. Systems biology flowering in the plant clock field. *Molecular Systems Biology*, 2:60, 2006.

[Vie99] Andy Vierstraete. The central dogma of molecular biology, 1999.

[VSWBR02] E. P. Van Someren, L. F. Wessels, E. Backer, and M. J. Reinders. Genetic network modeling. *Pharmacogenomics*, 4:507–25, 2002.

[web09] Mged open soruce projects. `http://mged.sourceforge.net/index.php`, April 2009.

[WFS04] S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20:5–20, 2004.

[WGH06] A. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.

[WGW+01] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. Assessing gene significance from cdna microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

[Wie56] N. Wiener. *Modern Mathermatics for Engineers*. McGraw-Hill, New York, 1956.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[XOX02] Y. Xu, V. Olman, and D. Xu. Clustering gene expression data using a graphtheoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–45, 2002.

[XSD+02] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30:303–305, 2002.

[YDL+02] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[YDLS01] Y.H. Yang, S.D. Dudoit, P. Luu, and T.P. Speed. Normalization for cdna microarray data. *SPIE BioE*, 2001.

[YK03] M. J. Yanovsky and S. A. Kay. Living by the calendar: how plants know when to flower. *Nature Reviews Molecular Cell Biology*, 4:265–276, 2003.

[YLW08] Y. Yuan, C-T Li, and R. Wilson. Partial mixture model for tight clustering of gene expression time-course. *BMC Bioinformatics,*, 9:287, 2008.

[YSLY04] L.K. Yeung, L.K. Szeto, A.W. Liew, and H. Yan. Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics*, 20:742–749, 2004.

[YTC02] M. Yeung, J. Tegnrdagger, and J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *PNAS*, 99-9:6163–6168, 2002.

[ZIT+99] W. Zhang, B. Irvin, R. Trible, R. Abraham, and L. Samelson. Functional analysis of lat in tcr-mediated signaling pathways using a lat-deficient jurkat cell line. *International Immunology*, 11(6):943–950, 1999.

[ZPB01] L.P. Zhao, R. Prentice, and L. Breeden. Statistical modeling of large microarray data sets to identify stimulus response profiles. *PNAS*, 98:5631–5636, 2001.

[ZSD06] W. Zhao, E. Serpedin, and E.R. Dougherty. Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, 22(17):2129–2135, 2006.