January 2009

Modeling And Identification Of Differentially Regulated Genes Using Transcriptomics And Proteomics Data

Thanura Elvitigala
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Electrical and Systems Engineering

Dissertation Examination Committee:
Bijoy K. Ghosh, Chair
Norman Katz
Jr Shin Li
Hiro Mukai
Himadri B. Pakrasi
Heinz Schaettler

MODELING AND IDENTIFICATION OF DIFFERENTIALLY REGULATED GENES USING TRANSCRIPTOMICS AND PROTEOMICS DATA

by

Thanura Ranmal Elvitigala

A dissertation presented to the Graduate School of Arts and Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009
Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Modeling and Identification of Differentially Regulated Genes using Transcriptomics and Proteomics Data

by

Thanura Ranmal Elvitigala

Doctor of Philosophy in Systems Science and Mathematics

Washington University in St. Louis, 2009

Research Advisor: Professor Bijoy Kumar Ghosh

Photosynthetic organisms are complex dynamical systems, showing a remarkable ability to adapt to different environmental conditions for their survival. Mechanisms underlying the coordination between different cellular processes in these organisms are still poorly understood. In this dissertation we utilize various computational and modeling techniques to analyze transcriptomics and proteomics data sets from several photosynthetic organisms. We try to use changes in expression levels of genes to study responses of these organisms to various environmental conditions such as availability of nutrients, concentrations of chemicals in growth media, and temperature. Three specific problems studied here are transcriptomics modifications in photosynthetic organisms under reduction-oxidation (redox) stress conditions, circadian and diurnal rhythms of cyanobacteria and the effect of incident light patterns on these rhythms, and the coordination between biological processes in cyanobacteria under various growth conditions.

Under redox stresses caused by high light treatments, a strong transcriptomic level response, spread across many biological processes, is discovered in the cyanobacterium *Synechocystis* sp. PCC 6803. Based on statistical tests, expression levels of about 20% of genes in *Synechocystis* 6803 are identified as significantly affected due to influence of high light. Gene clustering methods reveal that these responses can mainly be classified as transient and consistent responses, depending on the duration of modified behaviors. Many genes related to energy production as well as energy utilization are shown to be strongly affected. Analysis of microarray data under two stress conditions, high light and DCMU treatment, combined with data mining and motif finding algorithms led to a discovery of novel transcription factor, $RRTF1$ that responds to redox stresses in *Arabidopsis thaliana*.

Time course transcriptomics data from *Cyanothece* sp. ATCC 51142 have shown strong diurnal rhythms. By combining multiple experimental conditions and using gene classification algorithms based on Fourier scores and angular distances, it is shown that majority of the diurnal genes are in fact light responding. Only about 10% of genes in the genome are categorized as being circadian controlled. A transcription control model based on dynamical systems is employed to identify the interactions between diurnal genes. A phase oscillator network is proposed to model the behavior of different biological processes. Both these models are shown to carry biologically meaningful features.

To study the coordination between different biological processes to various environment and genetic modifications, an interaction model is derived using Bayesian network approach, combining all publicly available microarray data sets for *Synechocystis* sp. PCC 6803. Several novel relationships between biological processes are discovered

from the model. Model is used to simulate several experimental conditions, and the response of the model is shown to agree with the experimentally observed behaviors.

Acknowledgments

I would like to take this opportunity to express my greatest gratitude to my research advisors professor Bijoy K. Ghosh and professor Himadri B. Pakrasi for their invaluable guidance, advices, encouragement and support, extended to me throughout my graduate studies in Washington University in St Louis. Without their academic and financial support, this dissertation would not have been possible.

I thank professor Hiro Mukai of Electrical and Systems Engineering, for accepting to be my academic advisor and the co-chair of my dissertation committee. I also like to thank professors Norman Katz, Heinz Schaettler and Jr-Shin Li, who agreed to participate in my dissertation committee.

I thank professor Ralph Quatrano of Washington University and professor Rajeev Aurora of Saint Louis University for numerous fruitful discussions. I thank all the members of Pakrasi Lab, the Center for Biocybernetics and Intelligent Systems of Washington University and the proteomics group in Pacific Northwest National Lab, for their help throughout this period. Special thanks go to Dr. A.K. Singh, Dr. A. Khandelwal, Dr. J. Stockel, Dr. A. Polpitiya and Dr. E.A. Welsh.

I also thank all the faculty members, fellow students and staff of the department of Electrical and Systems Engineering in Washington university.

I also acknowledge the financial support from Washington University Engineering School, which enabled me completing my graduate studies at Washington University in St Louis.

Last but not least, I express my heartfelt gratitude to my wife, Hiranda, parents, two brothers, relatives and friends for their guidance, support and encouragement.

Thanura Ranmal Elvitigala

Washington University in Saint Louis
December 2009

Dedicated to my parents and wife.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Living cells are complex dynamical systems, showing a remarkable ability to adapt to different environmental conditions for their survival. Unraveling principles governing regulation of different biological processes in cells has been a fundamental challenge to humankind for a long time. Proper understanding of cellular regulation and utilizing that knowledge to control cellular behaviors are vital in many aspects. In the field of medicine, it helps people to find new cures for numerous diseases such as cancers and diabetes, produce more effective drugs and treat patients with different disorders. In the field of agriculture, it helps developing new varieties of crops, which generate higher yields, possess tolerance to harsh environmental conditions such as drought and cold, and produce foods containing additional nutrients. It also provides answers to some of the important problems faced by humans currently; such as discovering ways to reduce global warming and finding alternatives to replace depleting sources of fossil fuels.

1.1 Photosynthesis Organisms

Photosynthetic organisms represent the most important class of organisms on the earth. They created basis for the life on earth. Through oxygenic photosynthesis process, they convert carbon dioxide in environment into organic compounds, especially sugars, utilizing energy from the sunlight. As a by product, these organisms evolve oxygen, thus creating a conducive environment for other species. Photosynthetic organisms are widely accepted as an essential component in answering current global problems including global warming, pollution and energy crisis.

Due to their critical role on life, lot of research efforts have been invested to understand photosynthetic organisms. These organisms represent a wide variety of living forms from simple prokaryotic unicellular organisms, such as cyanobacteria, to complex eukaryotic systems, such as vascular plants. Researchers have focused on understanding general biological principles of these organisms as well as answering specific questions. Examples for these specific topics include studying how these organisms respond to stress conditions and how to improve their stress tolerance, how to improve growth rates and bio-mass production, and how to modify these organisms to introduce novel abilities to produce useful chemical compounds. Some of the organisms studied in detail and discussed in subsequent sections include, *Synechocystis* sp. PCC 6803 , a fresh water cyanobacterium, *Cyanothece* sp. ATCC 51142, a nitrogen fixing cyanobacterium, and *Arabidopsis thaliana*, a vascular plant. More details on these organisms and data sets available from them are given in Appendix A.

1.2 Central Dogma of Molecular Biology

Central to regulation of different biological processes and pathways in living cells are dynamic interactions between deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and protein molecules, described using the central dogma of molecular biology. In Figure 1.1, the main components of central dogma are illustrated. Genetic instructions for cellular responses to various external environmental conditions, such as availability of nutrients and variations in temperature, and internal conditions, such as presence or absence of essential regulatory or structural proteins, are stored in the form of DNA. The DNA composed of four nucleotide bases, namely adenine (A), cytosine (C), guanine (G) and thymine (T), attached to two backbones made of sugars and phosphate groups which are joined by ester bonds. These two strands form a helical structure and consist of millions of bases of A, C, G and T, and commonly known as chromosomes. A typical cell may consists of one or more chromosomes.

Genetic information stored in DNA are decoded through a process known as transcription, where a protein complex called RNA polymerase produces the corresponding RNA molecules. The DNA subsequences that have the capability to generate specific RNA molecules are called genes. A typical cell consists of thousands of such genes. One type of RNA, known as the messenger RNA (mRNA), gives rise to corresponding proteins through translation. Proteins are key players of all the biological processes performed by a living cell. Among these functions, is transcription control itself. Recently various mechanisms, outside the central dogma, which control the functions of living cells have been discovered. One such example is regulatory role played by the non-protein coding RNAs commonly known as micro RNAs (miRNA). However, still most of the cellular responses and behaviors can be attributed to the regulation

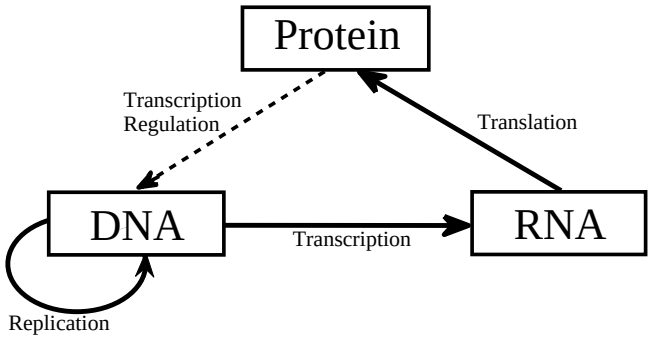


Figure 1.1: Central dogma in molecular biology.
Interactions between different molecules in a living cell are defined by the central dogma in molecular biology. Genetic information contained in DNA is transferred from one generation to the next through *replication*. Depending on requirements of the cells, different RNA molecules are produced in *transcription* and subsequently *translated* into corresponding proteins. Some of the proteins act as regulators to control the transcription process.

occurring through the central dogma. As a result, understanding principles governing the interactions in central dogma is still considered the key to uncover the secrets of life.

1.3 Motivation

During last few decades, a significant improvement in our understanding of the cellular systems has been achieved. The introduction of various advanced high throughput technologies such as genome sequencing [49], microarrays [65], and proteomics [6]; and

collaboration of diverse disciplines including biologists, computer scientists, engineers and mathematicians; have contributed to the progress of this field to a great extent. Several new fields of research including systems biology, synthetic biology, comparative genomics, bioinformatics and computational biology have shown tremendous development during last few years. Difficulties in integrating knowledge from diverse disciplines is seen as one of the main challenges hindering a rapid progress in these fields. Despite many notable achievements, dynamics of many cellular processes are still poorly understood.

Questions related to molecular biology are numerous. Studies related to genome sequencing and comparative genomics focus on comparing and contrasting genome sequences of different organisms with the objective of relating their similarities and differences with specific features of corresponding organisms. Systems biology tries to understand overall behavior of cells using global transcriptomics, metabolomics and proteomics measurements. Main focus of synthetic biology is to use genetic tools to modify genome sequences by deleting existing genes and inducting new genes, with the objective of achieving desired behaviors from those organisms. Bioinformatics involve in deriving novel data analysis tools so that hidden details in biological data could be extracted and interpreted. Computational biology mainly focus on modeling various aspects of biological processes and generating new hypotheses, which can be tested through subsequent experiments. Though various disciplines targets specific areas using different approaches and tools, they all contribute to one global aim namely: understanding regulation of biological processes in a living cell.

1.4 Outline

We focus on applying several computational and systems engineering tools to analyze different high throughput data sets. In Chapter 2, details on two such high throughput techniques, transcriptomics and proteomics, are presented. After introducing biochemical principles, we focus on two specific technologies, namely two-color microarrays based transcriptomics and bottom-up label free liquid chromatography - mass spectrometry(LC-MS) based proteomics. Various aspects involved in experimental design and preliminary data processing including quality assessment and data normalization are discussed. Some challenges specific to proteomics data processing are looked at, before concluding the chapter.

Chapter 3 focuses on one of the important biological questions, relevant to photosynthetic organisms, namely understating mechanisms important to maintain homeostasis inside a cell under various redox stress conditions. Data from several experimental conditions that produce redox stresses in photosynthesis organisms are analyzed. Genes showing differential behaviors under these stresses are identified using statistical tests and clustered together to gain an understanding on cellular responses. A co-expression network is obtained for the differentially expressed genes in *Arabidopsis thaliana* and regulatory region motifs are discovered in possible co-regulated gene groups.

In Chapter 4, the overall response and coordinated behaviors of different biological pathways in *Synechocystis* sp. PCC 6803 are studied using probabilistic approaches. Biological pathway level behaviors are derived using individual gene expressions and a Bayesian network is obtained for these processes. Biological significance of the network is discussed and simulation results for several experiment conditions are

presented. Chapter is concluded with a comparison between Bayesian network and correlation based results.

Chapter 5 introduces the diurnal rhythms in cyanobacterium *Cyanothece* sp. ATCC 51142. Several methods are introduced to separate diurnal behaviors into circadian controlled and light responding groups. In Chapter 6, a transcription control model based on dynamical systems is proposed to infer relationships between different diurnal genes. Parameters are selected to model different gene groups and most probable associations between genes are selected using biological insights.

In Chapter 7, a phase oscillator network is proposed to model the behaviors of main biological processes with diurnal rhythms. Model parameters are tuned to reconstruct the actual expressions. Network is used to simulate several experimental conditions and results are shown to be consistent with actual observations.

Chapter 8 introduces the use of proteomics data to gain further understanding on cellular behaviors. Proteomics data from several growth conditions are used to study the cellular response of *Synechocystis* sp. PCC 6803 at translational level. Proteomics data from *Cyanothece* sp. ATCC 51142 revealed that number of genes with oscillatory behaviors at translational level is much less compared to the those at transcription level.

Chapter 9 summarizes the findings of this dissertations and discusses directions for future research.

Chapter 2

Transcriptomics and Proteomics Data

2.1 Introduction to Transcriptomics

Transcriptomics, also known as genome-wide expression profiling, is one of the tools used to study changes in activities of genes in response to various modifications in internal and external cell environments. DNA microarrays (referred to as microarrays hereafter) [65] is the most common high throughput technique used to generate transcriptomic data sets. Instead of monitoring activities of few selected genes, microarrays facilitate measurement of activities of thousands and often tens of thousands of genes representing the entire or the most part of the genome of an organism in a single experiment. During the last two decades, numerous improvements have been made to microarrays, which are now capable of generating high throughput data with an increased level of accuracy.

Although different types of microarray technologies are currently available, the underlying science is mostly similar. Usually microarray chips are made out of glass plates. DNA sequences corresponding to different genes are printed on to different

locations on the chip using covalent bonds. This can be performed by directly embedding already synthesized DNA sequences on to the chip, as commonly done in custom made microarrays, or by synthesizing relevant sequence nucleotide-by-nucleotide on the chip, as done in oligonucleotide microarrays. During an experiment, mRNA is extracted from a biological sample and tested for quality and quantity using capillary electrophoresis and nanodrop spectrogram respectively. Complementary DNA sequences are obtained from mRNA using reverse transcription, labeled with dyes and hybridized on to the microarray chip. On the chip, DNA sequences bind to corresponding complementary DNA sequences more tightly, so that it is possible to remove the non-specific bindings. The intensity of dyes at each spot is proportional to its relative abundance compared to the total mRNA extraction.

Much of the differences between microarray technologies are related to the length of DNA sequences printed on to the chip, the number of different sequences embedded on a single chip, the number of replicates for a given sequence and the types of dyes used to label the mRNA. The cDNA microarrays use longer DNA sequences, usually in the range of 300–400 nucleotides, while the oligonucleotide microarrays use shorter sequences, usually in the range of 15–75 bases. For example, Affymetrix is an oligonucleotide type microarray and uses 15–18 bases long sequences in their chips. However, in order to achieve gene specificity, several sequences from a given gene are included. The samples are labeled using a single dye, so a global level data scaling is required for the comparison between different microarrays. On the other hand, Agilent microarrays, another oligonucleotide type chips, use relatively longer sequences of around 60-bases, but contain just one or two different sequences for a given gene. The mRNA from control and target experiments are labeled with two different dyes, usually cyanine varieties of green (cy3) and red (cy5) colors. Labeled complementary

DNA strands from two samples are hybridized onto the same chip so that differences of gene activities under two conditions can directly be compared. Chips are excited using two lasers of the same wavelengths as two dyes and fluorescence emission is measured. Two scanned images may later be merged to get a single image for each chip. In the combined images, red and green color spots correspond to those genes having different mRNA concentrations under target and control experimental conditions while yellow color spots corresponds to those genes having a similar level of mRNA concentrations. Two channel microarray technology introduces a variability to the data due to differences in dyes and this needs to be taken into account during experiment design and data processing. Figure 2.1, illustrates different steps involved in conducting a two-color microarray experiment.

Figure 2.2, shows a scanned image of a two channel microarray. Most of the spots in the chip are yellow in color, indicating that corresponding genes are expressed to similar levels under two conditions. Spots with shades of red and green correspond to genes having different mRNA concentrations due to their differential behaviors under two experimental conditions.

Microarray experiments are extensively used to identify interactions between genes. These computational methods view the process of transcription and translation in the central dogma as gene interactions, where the transcribed mRNA from one gene controls the activity levels of the others. The implicit assumption is that the abundance of a regulator protein is proportional to its mRNA level. This assumption is reasonable under many experimental conditions.

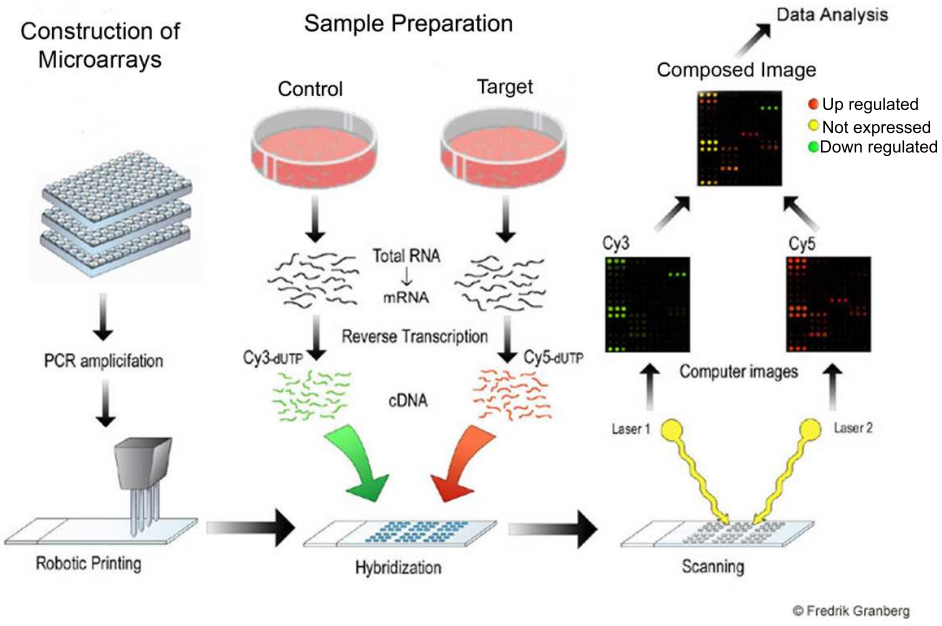


Figure 2.1: Steps involved in performing a two-color DNA microarray experiment. In microarray construction, DNA sequences corresponds to different genes are printed onto glass slides. During the experiment, mRNA extracts from two experimental conditions are converted to corresponding complementary DNA (cDNA) through reverse transcription. These cDNA from the two samples are labeled separately with two dyes and hybridized on to microarray chips. After washing away non-specific bindings, the chips are scanned with lasers of two different colors, and the scanned images are combined to get a composite image. Individual gene expressions are extracted from these images and used for further analysis (Image courtesy: Ashoka Polpitiya).

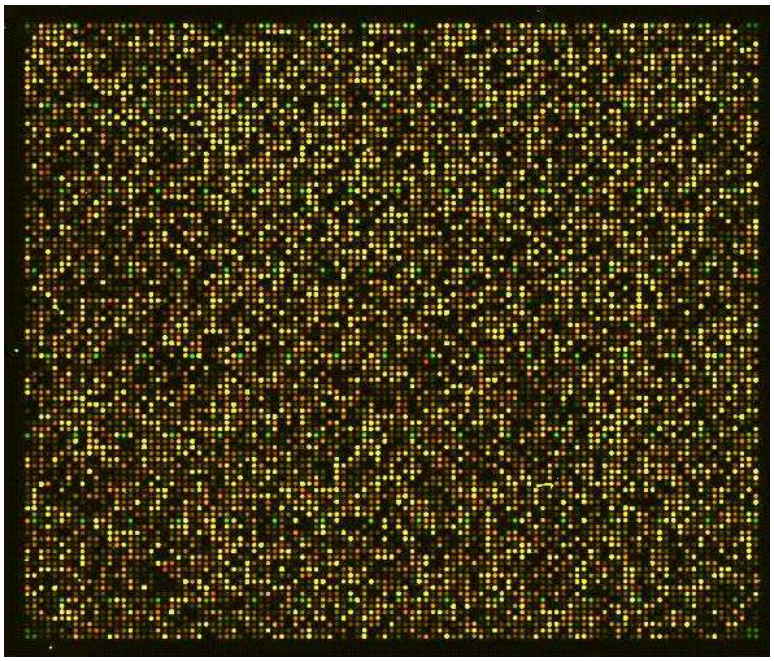


Figure 2.2: Scanned image of a two channel DNA microarray.
Different colors represent the relative abundance of gene expressions under two experimental conditions. Spots with shades of red and green correspond to genes having different mRNA concentrations where as spots in yellow corresponds to genes whose expressions did not get significantly affected.

Analysis of data generated in microarray experiments involves several steps. These steps include quality assessments, preliminary data processing, efficient representation of data to facilitate identification of salient features, and categorizing data into different groups in order to reduce dimensionality. Various techniques including correlation measurements, probabilistic methods such as Bayesian networks and linear and nonlinear dynamical systems, can be utilized to infer mutual interactions between genes. In subsequent sections, we apply some of these techniques to analyze several microarray data sets.

2.2 Introduction to Proteomics

Proteomics [87] is the logical continuation of the widely popular transcriptional profiling methodology, that is the microarrays. Proteomics focus on studying multi-protein systems in organisms, commonly known as the proteome, or the complete protein complement of its genome, with the aim of understanding distinct proteins and their roles as a part of a larger networked system. This is a vital component of the modern systems biology approaches, where the key goal is to characterize the system level behavior rather than behavior of single components. Measuring mRNA levels as in DNA microarrays alone does not necessarily tell us much about the levels of corresponding proteins in a cell and their regulatory behavior since they are subjected to many post-translational modifications and other modifications by environmental agents. The role of the proteins can not be overstated as they are responsible for the structure, energy production, communications, movements and division of all cells.

While genome-wide microarrays are ubiquitous, proteome microarrays are missing due to the fact that proteins do not share the same hybridization properties of nucleic acids. Mass spectrometry methods have effectively been used for the characterization of proteins and has now become the platform of choice for the analysis of complex protein samples. Here we analyze several proteomics data sets generated by bottom-up approach using mass spectrometry. The essential feature of bottom-up proteomics is that it uses small amino acid sequences, known as peptides, obtained via digestion of proteins, to detect original proteins. Usually an approximately six or more amino-acids-long peptide sequence uniquely maps to a protein. This enables identification to be performed by simply searching for the peptide sequence in a database of protein sequences.

Mass spectrometer is central to the current proteomics research [6]. Mass spectrometer measures the mass-to-charge ratio (m/z) of molecules. Recent years have seen a tremendous improvement in the mass spectrometer technology and there are about 20 different commercial versions available for proteomics. All mass spectrometers are designed to carry out the distinct functions of ionization and mass analysis.

A standard bottom-up experiment has the following key steps: (a) extraction of proteins from a sample, (b) fractionation to remove contaminants and proteins that are not of interest, (c) digestion of proteins into peptides using an enzyme such as trypsin, (d) post-digestion separations to obtain a more homogeneous mixture of peptides, and (e) analysis by mass spectrometry. Although many informatics tools can process the resulting data from the mass spectrometer, accurate identification and quantization of the proteins in a sample remain as a fundamental challenge.

When analyzing protein samples from an organism, first a database of peptides in that organism is created. The database is typically constructed using liquid chromatography (LC) based tandem mass spectrometry (LC-MS/MS) approach, where samples, sent through tiny liquid columns, are analyzed using two step mass spectrometry to achieve a higher level of resolution (Please refer Figure 2.3 for more details). The identity of the peptide is obtained by constructing theoretical mass spectra for peptide sequences in a genome and comparing them against the observed peaks to determine the best match. The matching criteria can be either a cross-correlation value [47] or a probability-based method [15]. Each observed peptide is then mapped onto a unique spot in a two-dimensional space, with the mass-to-charge ratio and time of observing the particular peptide (elution time) as corresponding coordinates. These maps are known as accurate mass and time (AMT) tags. Once the AMT database is in place, the subsequent experiments involve a single LC-MS step, where observed

peptides are later matched with corresponding entries in the database. Since LC-MS step is faster than LC-MS/MS step, higher throughput levels are achieved.

Quantitative proteomics techniques primarily evolve under two categories, namely stable isotope labeling and label free methods [6]. The stable isotope labeling techniques are analogous to the two-channel microarrays in transcriptome analysis. Samples from different experiments are analyzed using isotopes of N, O, or C. These isotopes are introduced, metabolically, chemically, or enzymatically, to the sample from one experimental condition. The two samples are then mixed and analyzed in a single cycle. Since the chemical properties of isotopes are same, the isotope-labeled and native peptides differ only by their mass and are separately detected. The relative intensities of a given peptide under two conditions are determined by measuring the abundance of native and isotope-labeled forms.

The label-free quantification method is analogous to single channel microarrays. No labeling of proteins is involved and the two samples are analyzed separately. While these techniques are free of the complexities related to labeling, the measurements are more prone to variations caused by the use of equipment in multiple runs. Peptide abundances are given as intensities of the detected signals or as *spectral counts*. The spectral count refers to the number of times a peptide is detected in various reads of mass spectrometry. Individual peptide measurements are then mapped back to their corresponding proteins. The mapping process can be complicated when it is not one-to-one, which occurs in certain cases where multiple isoforms of a protein are present. Various algorithms are used to infer protein abundance levels by combining corresponding peptides [58].

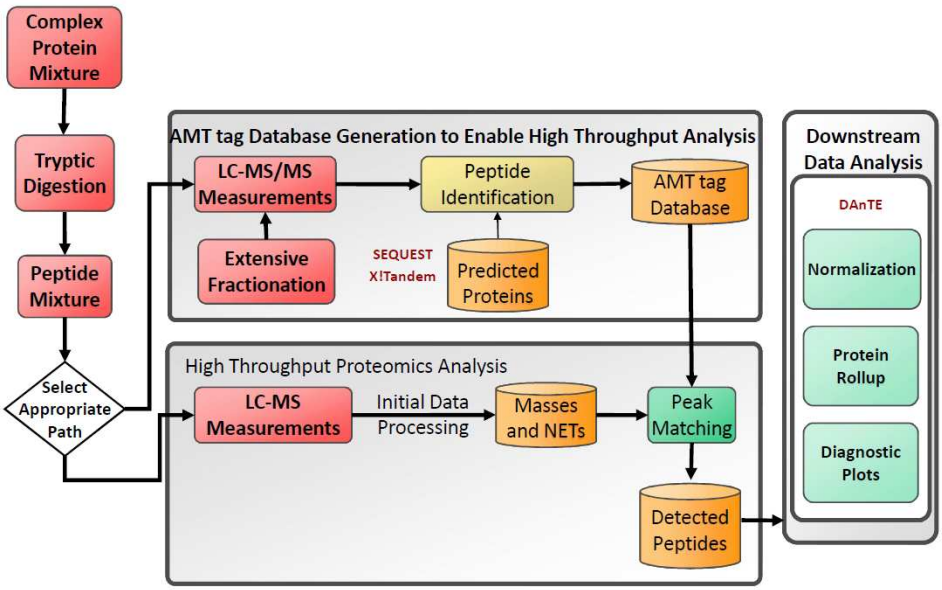
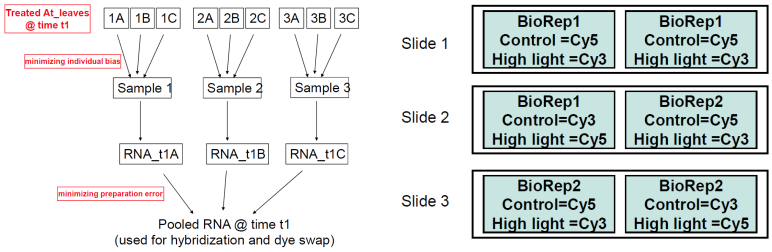


Figure 2.3: Main steps in performing a proteomics experiment using label-free, bottom-up approach.
Protein extracts from biological samples are digested using enzymes such as Trypsin to get corresponding peptides. Peptides are analyzed using either *liquid chromatography based tandem mass spectrometry* LC-MS/MS, (higher resolution) or *liquid chromatography - mass spectrometry* LC-MS, (lower resolution) levels. LC-MS/MS is useful in generating the *accurate mass and time* AMT, database for different organisms. Once AMT database is ready, subsequent samples can be analyzed using high throughput LC-MS. The observed peptides are identified by mapping them to entries in AMT database (Image courtesy: Ashoka Polpitiya).

2.3 Experimental Design

Experiment design is the most important preliminary step in high throughput experiments, since it affects all subsequent steps in the analysis and the interpretation of results. Goal of an experimental design is to reduce the undesirable effects from variations, which are not in focus of the experiment. These variabilities could arise from both biological diversity of the samples and technical factors. Differences in the biological materials are mainly due to changes in growth conditions and cell densities of the cultures. Typically 2–3 biological replicates are included in a single microarray experiment to take these differences into account. The technical variabilities can occur at any step during the experiment, from extraction of mRNA to scanning of the microarrays. These changes occur mainly due to inconsistencies in sample preparation or technical limitations in the instruments, for example inherent variability in microarray printing techniques. One of the main source of variabilities, unique to the two-color microarrays, results from different characteristics of the dyes used, which is commonly known as the dye bias. In order to address the problem of dye bias, microarray experiments include a dye swap where, the two dyes for labeling the samples are switched on replicate arrays. As a result, each experiment typically includes 6–8 microarrays. The data from these replicates are analyzed using statistical methods to isolate the variability.

Figure 2.4 shows different experiment design approaches. In [39], several samples are mixed, in order to reduce the variance introduced by the differences in biological samples, as shown in Figure 2.4(a). Figure 2.4(b) shows use of different biological samples and dye swaps to generate multiple technical replicates. This design is used in [70].



(a) (b)

Figure 2.4: Different microarray experimental designs.
In 2.4(a) utilized in [39], samples from different biological samples are mixed at two levels in order to reduce the variance between samples. 2.4(b), employed in [70], shows the use of dye swaps and biological samples to generate multiple technical replicates. (Images extracted from [39] and [70])

2.4 Quality Assessment

Microarray data analysis starts with quality assessment of raw data to ensure that they are of sufficient quality. One approach is computing coefficient of variation of individual spots on a microarray. When a microarray is scanned, the feature extraction software assigns each pixel, either to the signal (area where mRNA is bound) or the background. The final intensity value given to each spot and used for further analysis is the average intensity value for the pixels determined as the signal for a given spot. Coefficient of variation is used to quantify the intensity distribution of individual pixels categorized as the signal. A lower coefficient of variation value for the signal suggests a lower intensity variation among the pixels included as signal. Typically for a good quality two-color microarray, value of coefficient of variation is less then 10% for more than 90% of the spots. Another statistic that is taken into consideration is the overall signal intensity distribution of the spots. Under a 16-bit resolution scanner, intensities of a pixel can vary between 0 and 65535. A good array should show a wide spread of intensities for different spots, within the allowable range. A dense distribution towards the lower range is an indication of insufficient mRNA

quantity and thus likely to give poor separation between background and signal. On the other hand too many spots in the higher range is an indication of excessive use of mRNA and can cause contaminations of the neighboring spots. In general when a chip contains many spots with saturated pixels, a problem in experimental procedure is likely, since these pixels do not represent the true intensities for that spot.

2.5 Data Normalization

In general there are two objectives in any time course microarray experiment. The first objective is to compare gene expressions under different conditions. The second objective is to study the behaviors of genes over time. In order to do these types of comparisons, the observed data need to be normalized. Another reason for the normalization is to remove the systematic biases present in the data. An important observation, typical in two-color microarrays, is non-uniform behavior of dyes at different intensity levels. This behavior is well observed by plotting the intensity-ratio graph (log values of the product and the ratio of intensities of the two channels for each spot) for each microarray. Since majority of genes are not differentially regulated under a given experimental condition, log ratio values are expected to be spread around the value zero. However, data usually reveal a shift and an intensity based trend due to differences in dye behaviors. Normalization can be used to remove such bias in the raw data. In addition, normalization is used to correct for additional factors such as irregularity of the slides and variations introduced by the printing technology.

The local weighted linear regression (LOWESS) based data normalization procedure is widely used for microarray data normalization, since it is capable of removing

many trends present in the data. A robust version of LOWESS normalization, which is more resistant to outliers compared to the standard LOWESS algorithm, is also available. The robust version of normalization performs data smoothing through a two-step procedure. First, the local weights corresponding to each point within a selected window are calculated using a tri-cubic function given by,

$$w_{i1} = \left(1 - |\frac{x - x_i}{d(x)}|^3\right)^3. \tag{2.1}$$

Subsequently a linear regression computations are done incorporating those weights. The smoothed curve thus obtained is used to find the residuals and a second set of weights using,

$$w_{i2} = \begin{cases} (1 - (r_i/6MAD)^2)^2 & \text{if } |r_i| < 6MAD \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

which reduces the effects of outliers. The final weights used to perform smoothing are the product of the two beforehand computed sets of weights. Usually a window size of 25%–40% is selected to ensure that various assumptions made during normalization remain valid.

Figure 2.5(a) and Figure 2.5(c) show product-ratio plot for two different microarrays before performing data normalization. The data contain an intensity dependant trend. As shown in Figure 2.5(b) and Figure 2.5(d), these trends are removed by applying the robust LOWESS normalization.

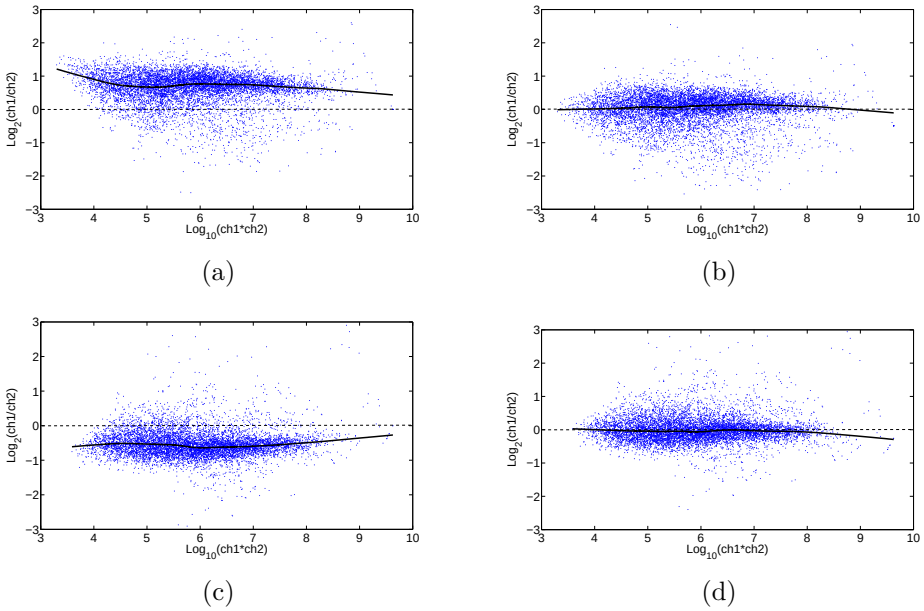


Figure 2.5: Distribution of intensities of spots in microarrays, observed as product-ratio plots.
Data normalization is carried out as the first step in microarray data analysis so that the variations in different microarrays are reduced. Intensity based trend commonly observed in two channel microarray data (a,c), is reduced through the LOWESS normalization (b,d).

2.6 Proteomics Data Processing

The quality assessments and normalization steps discussed above can be applied to the proteomics data as well. However, some additional steps are relevant to processing of the proteomics data. Many of these steps are summarized in [58]. For example, in label-free proteomics, global intensity adjustments, based on *mean absolute deviation* (MAD) or *central tendency* adjustments, might be important to bring the overall intensity values of different samples to comparable levels.

One important challenge unique to proteomics is handling of the missing data points. In contrast to microarray data, where the number of missing data points is negligible, proteomics data contain a lot of missing data points. Inferring from these data points

can be performed using different approaches depending on the specific problem. In [19], the authors discusses about one such method suitable for time course data. Depending on whether the missing data points are at the ends or middle of the time series, imputed values are selected using the closest observed data point or an interpolated value. Additional imputation methods include simple substitutions with mean/median values or pre-chosen values; K-nearest neighbor based approaches where missing point is computed as a weighted average of the observed values of the K-nearest neighbors; and singular value decomposition based approaches, where missing points are determined using a linear combination of eigen vectors correspond to the gene expression matrix [80].

The normalization of individual peptide expressions to get overall protein abundance is another challenge unique to proteomics data. Even though peptides correspond to a single protein should possess a same intensity value; due to differences in their chemical properties; they produce different measurements. There are several algorithms to compute the overall protein intensity values using individual peptide abundances. Such methods are generally referred to as *rollup* techniques. Most of the rollup methods start by removing the outlier peptides from the group. Time course proteomics data can be normalized using *R–rollup*, where peptides are first scaled using a reference peptide. Usually peptide with the highest abundance is selected as the reference. Overall abundance of the corresponding protein at each time point is then computed as the mean/median of the scaled peptide intensities of the corresponding time point. Additional details on rollup techniques are found in [58].

2.7 Conclusions

Transcriptomics and proteomics represent two of the widely used high throughput biological data generation techniques. Transcriptomics data, mainly obtained using DNA microarrays, are used to measure genome-wide gene expression levels under different experimental conditions where as proteomics techniques measure the levels of proteins present in a cell. In order to be able to compare the changes in cell environment at transcriptional level as well as translational level, it is important to handle the data generated from different experiments appropriately.

Some of the preliminary steps common to both transcriptomics as well as proteomics include experiment design, quality assessments, and data normalization. Processing of proteomics data requires additional steps to infer missing data points and derive protein abundance levels from individual peptides. These preliminary steps need careful consideration since all subsequent results as well as interpretations of the data depend on these operations.

In this chapter, we introduced details of both transcriptomics and proteomics high throughput techniques. We presented various steps involved in preliminary data processing related to them. These steps are applied in the analysis of several data sets presented in the subsequent chapters.

Chapter 3

Redox Regulation in Photosynthetic Organisms

3.1 Redox Stress on Photosynthetic Organisms

Redox or reduction-oxidation reactions are chemical reactions that result in changes of oxidation state of the substrates involved. Oxidation refers to a reaction where the oxidation number is increased (or electrons are lost) whereas reduction refers to a reaction resulting in a reduction in the oxidation number (or gain of electrons) of a molecule.

Many of the reactions in biological systems falls into the category of redox [57]. For example, aerobic cellular respiration involves oxidation of glucose to CO_2 and reduction of oxygen to water to generate energy in the form of Adenosine Triphosphate (ATP). Similarly in photosynthetic organisms, in the presence of light energy, reverse reaction of respiration takes place where CO_2 is reduced to glucose and water is oxidized to oxygen. In these reactions, a proton gradient is created by intermediate steps where oxidation and reduction of nicotinamide adenine dinucleotide (NAD^+) and NADH takes place, driving the production of ATP.

In order to these reaction to proceed, it is vital to maintain the proper balance between NAD^+/NADH and $NADP^+$/NADPH, which is referred to as redox state of the cell. Various external and internal changes in cell environment cause alterations to redox status of cells. Due to criticality of maintaining homeostasis, organisms have developed various mechanisms to handle redox stress conditions. However the principles behind these mechanisms are poorly understood.

3.1.1 Aims

We analyze transcriptomics data from several experimental conditions, where photosynthetic organisms are subjected to various redox stress conditions, to study their responses to such stresses. These experiments include *Synechocystis* sp. PCC 6803, subjected to three different stress conditions, namely exposure to high light, treatment with 3-(3,4-dichlorophenyl)-1,1-dimethylurea (DCMU) and preferential excitation of two photosystems, PS-I and PS-II, and *Arabidopsis thaliana*, subjected to two stress conditions, namely high light and DCMU. For each experiment, microarrays are produced using mRNA samples extracted over a time course. Specific aims of the analysis include comparing and contrasting behavior of genes under different stresses and identifying important genes that respond to redox stress to maintain homeostasis in photosynthetic organisms. We apply various statistical tests to select differentially expressed genes from the time course data, use clustering techniques to classify genes to different behavioral groups and combine other biological knowledge, such as pathway level details and DNA sequences, to refine our results.

3.2 Analysis Tools and Techniques

3.2.1 Identification of Differentially Expressed Genes

Analysis of transcriptomics data starts by identifying genes which show different behaviors under two experimental conditions. Using the preliminary data processing steps discussed in Chapter 2, fold changes of gene expression levels under different experimental conditions are computed. Identification of differentially expressed genes is performed either using an absolute fold-change cutoff or using statistical significant tests. When an absolute fold change cutoff is used, it is important to pick an appropriate value as the cutoff. If the selected value is too large, many differentially expressed genes will not be included in the analysis where as a too small value will result in an inclusion of many false positives. Typically determination of the cutoff is performed with the help of additional confirmation experiments such as reverse transcription polymerase chain reactions (RT-PCR) [9]. Few genes with different levels of fold changes are selected and RT-PCR experiments are performed to see whether differential behaviors can be validated independently. The lowest fold change which is verified by RT-PCR is selected as cutoff for the fold change and genes with higher fold changes are selected for further analysis.

An alternative approach is to use statistical significant tests to select differentially expressed genes. The student's t-test [72] is a standard statistical test and widely used to identify differentially expressed genes between two different experimental conditions. It can be applied to most of the situations, where other techniques cannot be used. The test is conducted as one sample test using log ratios of expression values, or as a two sample test using absolute expression values for the two conditions. The

one sample t-test is based on the null hypothesis that observed log-ratio values for a given gene are from a Gaussian population with a zero mean while the two sample t-test is based on the null hypothesis that expression values from two experimental conditions have a same mean and a standard deviation. The acceptance of the alternative hypothesis can be done at different significance levels (p-values) such as 0.1%, 1% or 5%. The t-test is used for both time series as well as non-time series data sets but requires a reasonable number of microarrays for the underlying assumptions (that is the samples come from a Gaussian Distribution) of the test to hold valid.

Extraction of differential gene expressions (EDGE) is mainly designed for time series data [45], but can also be applied to non-time-series data sets. EDGE approximates data using a set of basis functions and fits a model, using either the least squares [89] or expectation maximization [18] algorithms. The null distribution of test statistics is calculated through a bootstrap procedure [17]. Each gene is assessed, using false positive probability or false discovery rate, to determine whether it is differentially expressed or not.

Since EDGE is optimized for detecting genes with an altered behavior over a time course, this method does not pick a gene that is up-regulated or down-regulated throughout the time course. On the other hand, since data are combined and processed as a series, EDGE can be applied to data sets with few replicates per time point.

Statistical tests can also be followed by a threshold cutoff for the log-ratio values. This reduces the number of false positives in the selected gene set and the number of genes needed to be focused on. Furthermore the larger fold changes can easily be verified using RT-PCR. However, filtering genes based on log-ratio values, carries the risk

of missing some important genes, that only show small changes in their abundance levels between different conditions, but play a significant role in the gene regulation.

3.2.2 Clustering of Gene Expressions

The main goals of gene clustering are identifying principal behavioral patterns in the data and grouping genes based on those patterns. Gene clusters make data handling easier and usually carry some biological significance. For example, co-regulated genes, whose activities are controlled by a common promoter, tend to show similar gene expression patterns. Therefore co-regulated genes occur in a single cluster. In addition, some of these clusters are rich in genes of specific biological functions. When a given biological pathway responds to an external or internal cue by changing the expression of its constituent genes, the expression profile of these genes is similar and thus these genes are clustered together. In a scenario where the information about all of the constituents of a cluster is not known, those unknown genes can be predicted to be from the same biological pathway as the most of genes in that cluster. This approach provides a useful means for assigning functions for novel genes whose function had not been previously reported. However, further experiments need to be performed to demonstrate the involvement of the given gene in the particular process.

Any clustering method centers on two key questions, namely, how to measure the similarity between expressions of two genes and how to group similar genes together and separate non-similar genes. Although it is possible to make observations about various clustering techniques, no one method is the best, since no single criterion exists to measure the goodness of the resulting clusters.

Similarity of gene expressions is computed using various distance measures, such as Euclidean distance, Pearson correlation, un-centered correlation, and Hamming distance. Usually when the data are in log-ratio values, as is common in two-channel microarrays, Euclidean distance is used. When the data is expressed in absolute values, as in the case of Affymetrix, correlation or cosine distances are preferred. Finally, the use of Hamming distance is limited to discretized data sets.

The method of measuring the inter-cluster distances and intra-cluster distances, also known as the *linkage function*, is selected next. Commonly used linkage functions include *single linkage* which is the smallest distance between any two members of two clusters, *total linkage* which is the largest distance between any two members of two clusters, *average linkage* which is the distance between centroids of two clusters, and the average distance between any two members of two clusters.

Clustering techniques generate clusters using many approaches. One common clustering method, known as k-means, requires the user to define the number of clusters to be generated. First the initial cluster centroids are selected randomly, uniformly, or from a subset of genes. Then the remaining genes are distributed among the clusters based on the chosen linkage function. Since the starting choice of the centroids are different, the resulting clusters obtained from each run may not agree with the others. As a result, the algorithm is run a large number of times, and the clusters that give the minimum average distance are picked. The k-means clustering algorithm has several limitations. The algorithm tries to distribute all the genes among the selected number of clusters, thus genes with distinct expression patterns frequently end up being grouped together. Also, when a dominant expression pattern is present, which is a common scenario in gene expression data, a large number of seeds might be

obtained from those genes and the resulting clusters from those seeds show a similar pattern.

Another widely used clustering technique is hierarchical clustering algorithm. Here, the algorithm starts by considering all the genes as separate clusters. Then, based on the distance, the closest two genes are joined to build a single cluster. Next the above step is repeated, but now the two genes clustered together are considered as a single node. This procedure is followed until all the genes are put into one cluster. The results are usually viewed as dendrograms. By cutting the tree at different levels, different numbers of clusters are obtained.

The self-organizing map is another clustering technique that utilizes learning algorithms seen in neural networks [69]. Based on a user-defined number, nodes are initialized randomly. An iteration proceeds by picking a gene randomly and moving the nodes toward the selected gene by amounts that depend on the Euclidean distances between expressions of the selected gene and the nodes. The closest node is moved the most, while the furthest node is moved the least. This iteration is repeated for a large number of times (20000–50000), at the end of which the genes are organized as clusters.

Determining the Number of Clusters

Usually, deciding the number of clusters in microarray data is a difficult task. However, as discussed under gene clustering algorithms, the number of clusters is required as an input to many clustering algorithms. Cross validation techniques such as hold-out cross-validation, k-fold cross-validation, or leave-one-out cross-validation [41] are used to determine the number of clusters. All validation techniques use a subset of

the data for cluster identification and use the remainder to evaluate the performance. In the case of gene clustering, the average distance of the remaining genes to the closest cluster is commonly used as a performance measure.

3.2.3 Generating Co-Expression Networks

In a co-expression network, genes are connected to form a network by joining pairs of genes that are close in terms of their expressions. Closeness between genes is measured using one of the distance measures discussed earlier in the Section 3.2.2. Whether to make a link between two genes or not depends on the threshold selected for the distance, and different networks result in accordingly. A visualization software is typically used to view the gene network. One such software, Cytoscape [66], can format the network according to different criteria, in addition to displaying it. As a result, it is possible to identify groups of genes, sometimes referred to as hubs, that are more tightly connected to each other within the group than those outside the group. These hubs are analogous to the clusters obtained from clustering algorithms.

Determining a threshold for the distance measure is a relevant question in co-expression networks. The threshold is sometimes decided in accordance with the power law distribution [14]. Power law distributions are observed in various types of networks arising in fields such as physics, chemistry, biology, computer science, and social sciences. The main idea behind power law distribution is that nodes of these networks contain the relationship

$$\log\left(f(x)\right) = k\log\left(x\right) + \log\left(a\right), \tag{3.1}$$

where x is the number of connected neighbors of a given node, and $f(x)$ is the number of times x is observed. The gradient k is shown to have a value in the range of -1.8 to -2.2. The intercept $\log(a)$ is a constant for a given network. In other words, these networks consist of relatively few nodes with a large number of connected neighbors and many nodes with a small number of neighbors. The threshold for a gene expression network is decided in a way that the power law distribution for the resulting network has a gradient in the same range as the other networks. Although it is still not experimentally proved that gene expression networks do really follow the power law distribution, this criterion can be used as a systematic method for deciding thresholds.

3.2.4 Extracting Probable Interactions among Co-expressed Genes

The main objective of gene clustering and co-expression networks is to identify possible interactions between genes using transcriptomics data. Genes under the control of a single regulator are likely to show similar behaviors and thus being co-expressed. However all the genes that show co-expressions under few experimental conditions are not necessarily be co-regulated. In order to identify actual regulatory relationships between genes, it is required to use additional criteria to process co-expressed genes identified earlier.

Use of existing biological knowledge of relationships between genes and metabolic pathways is one way to identify co-regulatory genes among co-expressed genes. Data mining techniques are increasingly being used to identify the relationships between

genes reported in the literature [20]. Since these techniques typically include interactions from different organisms, it is possible to identify novel relationships for the organism being studied. Also it serves as a validation step for the known genes.

Transcription factors are a group of proteins which play a main role in transcription regulation. Transcription factors regulate the transcription of genes by targeting specific sequences of DNA. These DNA sequences are commonly referred to as binding site motifs. Transcription factors induce or suppress a gene expression by binding to a DNA sequence adjacent to the gene they regulates. The presence of conserved sequences in the upstream region of a group of genes suggests that they might be regulated by a single transcription factor.

Identification of binding site motifs is not a trivial task since these motifs are not necessarily identical between genes. Several methods are available ([55]–[64]) to identify such motifs. Algorithms presented in [55] and [34] take upstream regions of a group of co-expressed genes, discovered using clustering or transcription networks, and search for conserved regions within them. *Consensus* [34], uses a greedy algorithm to search and align conserved sequences in a set of upstream DNA sequences, so that the final alignment matrix maximizes the information content. Since an exhaustive search over all the combinations is usually computationally not practical, these algorithms employ heuristic search techniques. In contrast, [83], an algorithm based on dictionary building models, searches entire genomes and predicts over-represented sequences.

3.3 Results

3.3.1 Both Transient and Consistent Changes in Gene Expressions are Observed in *Synechocystis* sp. PCC 6803 Subjected to High Light Conditions.

Genes in *Synechocystis* 6803, showing differential behaviors under high light exposure, are identified using the t-test and and applying a cutoff for the log-ratio values. 762 genes among 3459 total genes in the microarray had a statistical significance value less than 1% and a fold change of more than 1.3 compared to the control condition, in at least one time point. These genes are selected for further analysis. Fold change cutoff of 1.3 is successfully verified using independent RT-PCR experiments. Expressions of 762 genes are discretized to three levels. A gene is assigned $+1$, if fold change is greater than 1.3, -1, if fold change is less than $1/1.3$, that is more than 1.3 time reduction in the target expression compared to the control, or 0, otherwise. Main behaviors among the genes are identified using discrete expressions and clustered together. In Figure 3.1, we display the largest 11 clusters. Expressions of the remaining genes are observed manually and some of them are associated with the relevant clusters based on their functional categories. Expressions of the remaining genes are shown in the last sub-figure.

We observe genes with both transient as well as consistent modifications of their expression levels when cells are subjected to high light conditions. Genes in clusters 1–3 are down-regulated during the experiment. These genes have different delays till they start responding to the high light, with those in cluster-1 responding immediately and those in cluster-3 responding with about 2-hour delay. They continued to be

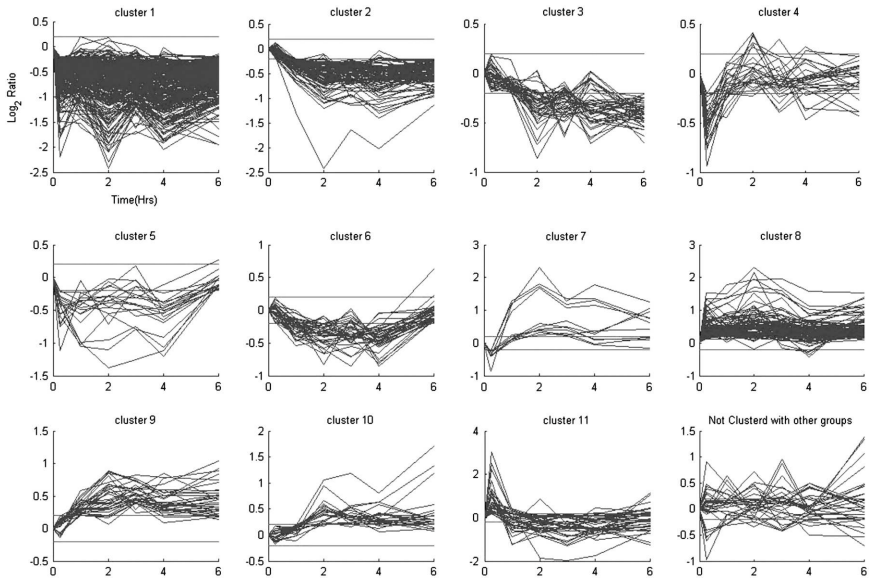


Figure 3.1: Gene clusters for transcriptomics data from *Synechocystis* sp. PCC 6803, subjected to high light stress conditions.
Eleven main behaviors of genes are identified using discretized gene expressions. Among them, gene groups with both transient and consistent changes in their expressions are observed. Genes take different amounts of delays to respond, allowing us to infer the sequence of events occur in the cell.

down-regulated during the remaining time of the experiment. Genes in clusters 8–10 show analogous behaviors except that their expressions levels are increased under high light. Genes in clusters 4–7 and those in cluster-11 show transient behaviors, where gene expressions reaching to normal levels towards the end of the experiment.

Analysis of the genes in each cluster reveals that the genes from different biological functions behave similar manner under the influence of high light. Figure 3.2, extracted from [70], shows the distribution of genes from different gene functions among various clusters. These clusters allow us to derive conclusions on overall response of *Synechocystis* sp. PCC 6803 to high light. Especially we observe down-regulation of photosynthesis and pigment biosynthesis related genes, soon after cells are subjected to high light (sub-figure 1), followed by carbon fixation and nitrogen assimilation

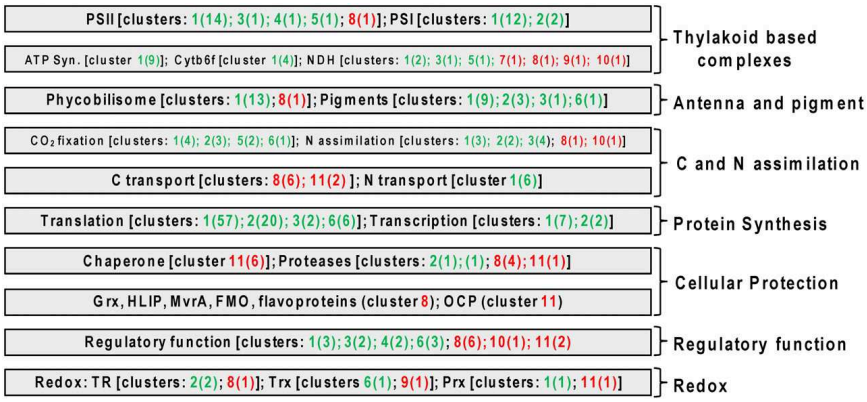


Figure 3.2: Composition of gene clusters for transcriptomics data from *Synechocystis* sp. PCC 6803, subjected to high light. Genes belonging to same biological functions show similar overall behavior.
(Extracted from [70])

related genes after sometime. This allows us to come to conclusions on integrated responses on energy production (photosynthesis and pigment biosynthesis) and energy consumption (carbon and nitrogen fixation) related processes.

3.3.2 Preferential Excitation of Photosystem-I and Photosystem-II Gives Rise to Different Cellular Responses.

In response to preferential excitation of Photosystem-I (PS-I) and Photosystem-II (PS-II), a total of 1202 genes show differences in their expression levels with at least 1.3 fold change between two conditions, measured at 1% significance level. Of these genes, 224 genes with greater transcripts abundance under PS-I excitation and 243 genes with greater transcripts abundance under PS-II excitation, show significant changes in abundances only in one time point. Remaining genes are differentially expressed in multiple time points. Similar to high light treatment we observe both transient and consistent changes in the gene expressions. In Figure 3.3, resulting clusters with distinct behaviors are shown.

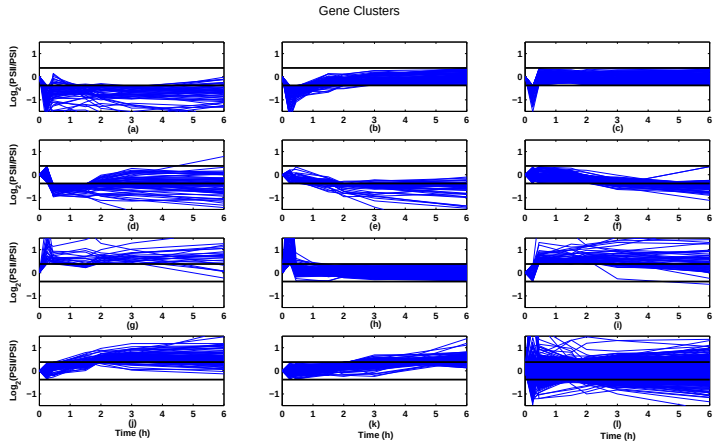


Figure 3.3: Gene clusters for transcriptomics data from *Synechocystis* sp. PCC 6803, subjected to preferential excitation of Photosystems I and II. Eleven distinct behaviors are identified using discretized expressions.

As discussed in [71], distinct transcriptome response is observed in the two treatments, where cyclic photosynthetic electron transport chain becoming active under preferential excitation of photosystem-I and cytochrome-c-oxidase and photosystem-I becoming active during preferential excitation of photosystem-II.

3.3.3 About 10% of the Genes in *Synechocystis* sp. PCC 6803 Respond to All Three Types of Redox Stresses; High Light, DCMU and Preferential Excitation of PS-I and PS-II

We discovered 342 genes, whose expressions are significantly affected by all three redox stress conditions, namely high light, DCMU and preferential excitation of PS-I and PS-II. Figure 3.4 shows the number of genes differentially expressed under different conditions. Three stresses have significant effect on the transcriptome of *Synechocystis*

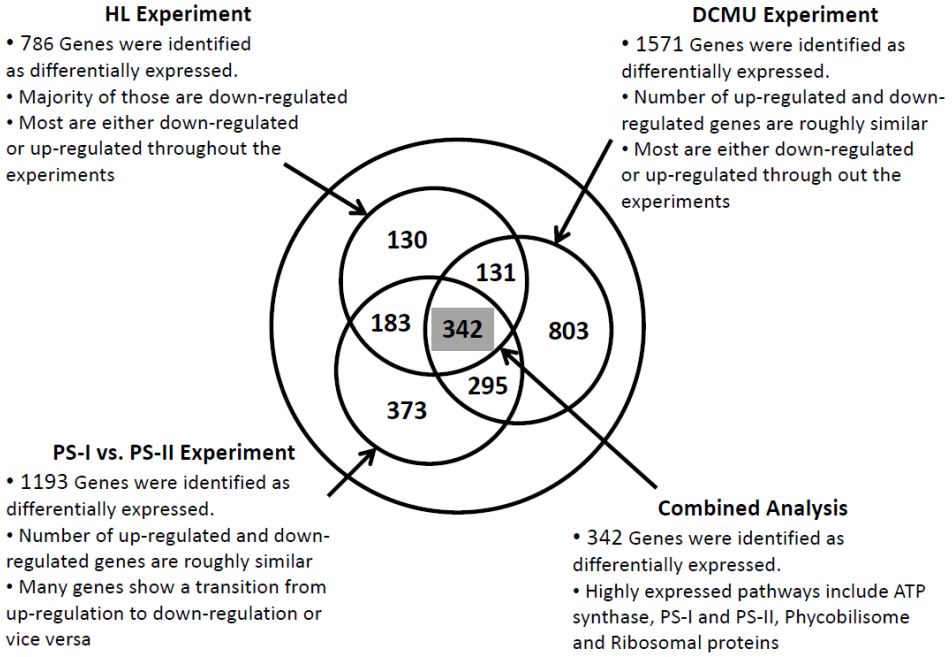


Figure 3.4: Number of differentially expressed genes in three redox experiments performed using *Synechocystis* sp. PCC 6803. Many energy generation related processes get affected under these stress conditions.

sp. PCC 6803. The number differentially expressed genes in each experiment varied between 25% to 50%. As observed from Table 3.1, we see clear differences in cellular responses to these three stresses. However, as a general trend, we see cells reduce their energy generation activities in adapting to all three stresses, as observed by modifications in the processes ATP synthase and photosystems.

Self organizing maps are utilized to identify the main gene behavioral patterns among 342 genes that are affected in all three conditions. Based on k-fold cross validation, we determined that we may identify 12 gene clusters in the data. In Figure 3.5, we show the results from k-fold cross validation and the clusters obtained using self organizing maps, plotted in the first two-principle component space. These 342 genes may be

Table 3.1: Percentages of differentially expressed genes in various biological pathways of *Synechocystis* sp. PCC 6803, under three different Redox stress conditions

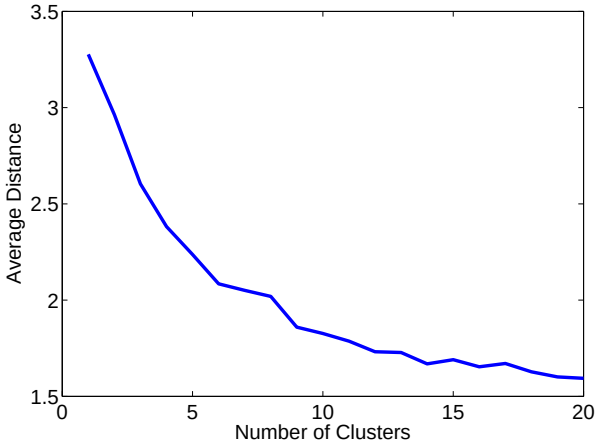
Cyanobase Pathway	Total Genes	All Conditions	In HL	In DCMU	In PS-I vs PS-II
ATP synthase	10	80%	90%	90%	90%
CO2 fixation	15	40%	53%	73%	53%
NADH dehydrogenase	23	30%	43%	70%	78%
Photosystem-I	16	63%	81%	75%	88%
Photosystem-II	27	48%	63%	78%	74%
Phycobilisom	18	67%	78%	72%	100%
Ribosomal proteins	63	60%	73%	81%	76%
RNA synthesis	23	22%	35%	78%	61%

Level of effects from the three stresses: high light, DCMU and preferential excitation of two photosystems; is different for different biological pathways. However, many energy generation related processes and growth related processes get significantly affected in all three conditions.

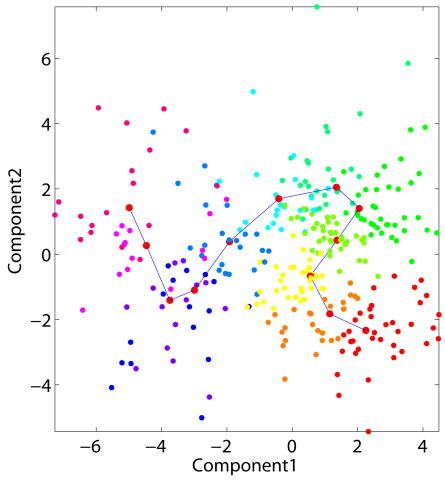
utilized to identify the key genes responsible to redox stress response in cyanobacteria and target genes can be verified using additional experiments.

3.3.4 Transcriptomics Data Analysis Leads to Discovery of a Novel Transcription Factor in *Arabidopsis thaliana*

Microarray data analysis of *Arabidopsis thaliana* revealed that 20% and 8% of expressions of 20436 genes are differentially regulated under high light and DCMU treatments, respectively. Approximately 6% of genes are identified in both perturbations and are classified as potential redox responsive genes (RRGs). Two co-expression networks are generated in an attempt to identify genes whose expressions are correlated during adjustment to homeostasis under high light and DCMU conditions. As shown



(a)



(b)

Figure 3.5: k-fold cross validation provides a guideline to determine the optimal number of clusters. Self organizing maps can be used to classify genes among these clusters.

We selected twelve clusters as the optimum number of clusters for the gene expressions. This is the minimum number of clusters, where no significant reduction in average distance is achieved by increasing the cluster count. Resulting clusters from self organizing maps based gene classification are shown in 3.5(b).

in Figure 3.6, ten subnetworks are identified from the high light network. These clusters are further classified considering the expressions under DCMU experiment.

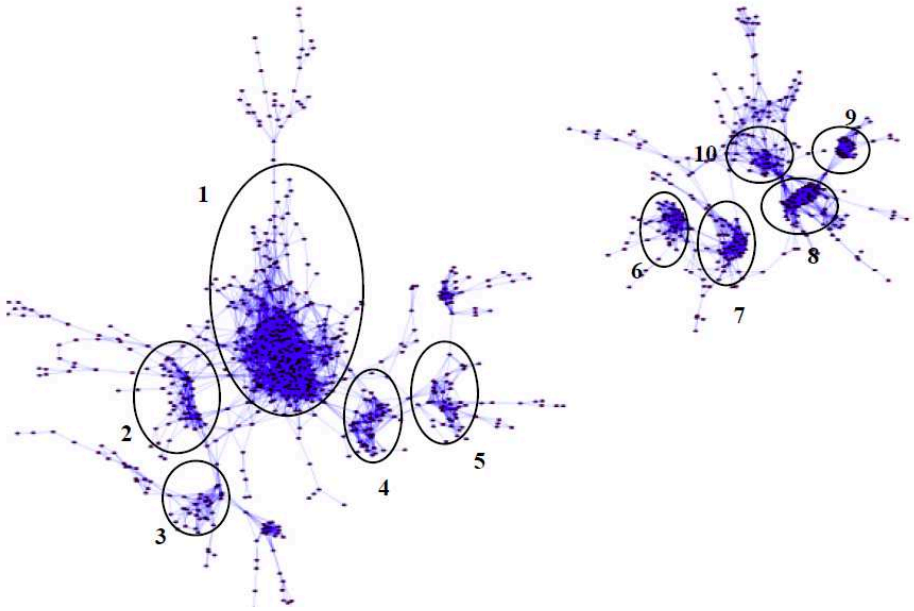


Figure 3.6: Correlation network obtained for *Arabidopsis* microarray data under highlight treatment. Ten gene clusters are identified from the network and used for further analysis [39].

In order to examine the biological significance of these gene clusters, the upstream regions of the relevant genes are analyzed using the Consensus algorithm [34]. Up to 500 base pairs of DNA sequences from the upstream of the co-expressed genes are searched for conserved sequences of eight bases. Discovered regulatory region motifs together with their significance values are given in Figure 3.7. Several motifs that are previously identified related to other stresses, such as light, dehydration, and abscisic acid, are among them [39].

In order to further investigate the significance of these gene clusters, individual expressions of genes belonging to the largest sub-cluster are examined. Among these genes, 30 genes are consistently down regulated by more than two fold in all time points under both experimental condition. Figure 3.8 shows the relevant 30 genes and the connections between them. Several well characterized stress responsive genes

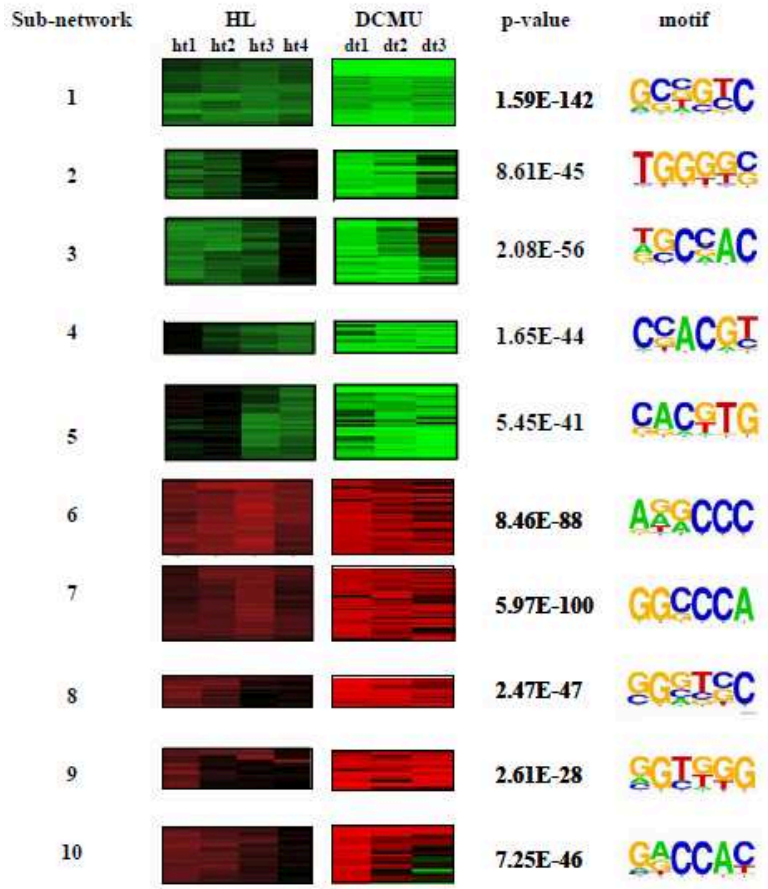


Figure 3.7: Regulatory region motif analysis for the gene subnetworks, identified using transcriptomics data for *Arabidopsis thaliana*.
Consensus algorithm [34] is used to search for conserved regulatory region motifs in the upstream regions in the co-expressed gene groups. Expressions of genes belonging to each cluster under two experimental conditions; highlight and DCMU treatments, conserved regulatory region motifs and their significance values (p-values) discovered using *Consensus* are shown.

are identified among these genes. A novel regulatory gene, redox-responsive transcription factor 1 ($RRTF1$), is connected to many of these stress response genes. Literature search reveals that $RRTF1$ gene is differentially expressed in the majority of previously reported transcriptomics experiments. With additional biological experiments, this genes is later shown to have an important role in redox stress response of *Arabidopsis* [39].

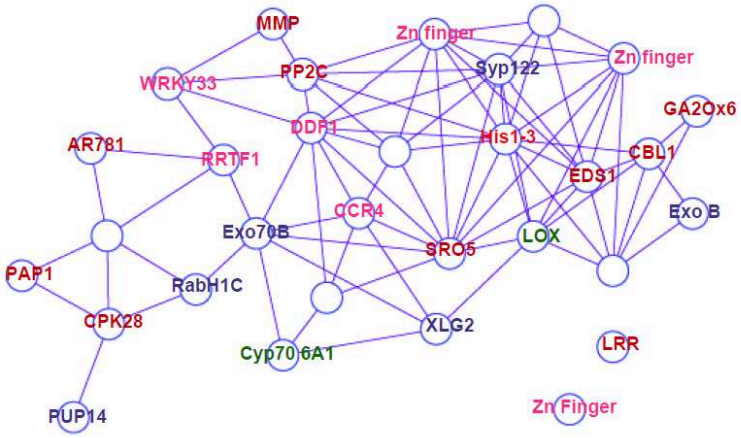


Figure 3.8: Subnetwork of thirty genes consisting of many stress responsive genes. We focused on a thirty-genes subnetwork comprises of many stress responsive genes, selected from the correlation network in Figure 3.6. Searching through the related publications, a novel transcription factor, $RRTF1$, related to several of these stress responsive genes is identified in the network. Later biological experiments verified that it plays an important role in redox stress response in *Arabidopsis thaliana* [39].

3.4 Discussion and Conclusions

In this work we presented several techniques used to identify differential behaviors in gene expressions under various stress conditions, that gives rise to redox perturbations in photosynthetic organisms. We used these techniques to analyze five transcriptomics data sets from two photosynthetic organisms.

We discovered that cyanobacteria show significant transcriptional level response to redox stress conditions, with about 10% genes getting affected by all stresses. We observed both stress specific responses as well as general responses by the cells. These genes can be targeted in additional experiments to identify those that are central to redox stress response in cyanobacteria.

Transcriptomics data analysis in *Arabidopsis thaliana* led to the discovery of novel transcription factor that was later shown to be a key player in maintaining redox

homeostasis in plants. We utilized both regulatory region motif finding algorithms and literature search to identify this gene among many possible targets.

Chapter 4

Coordination between Biological Pathways in Response to Different Environmental and Genetic Modifications

4.1 Motivation

Living organisms modify activity level of their biological processes depending environmental conditions. Though response of individual genes might be different between conditions, some general behavior patterns in biological processes can be observed. For example in [70], it is shown that overall expression level of genes in energy generating photosynthesis process is lower when bacterial cells are subjected to high light conditions. It is also shown that level of activity in energy consuming processors such as carbon and nitrogen fixation becomes lower subsequently. Similar behaviors were observed in DCMU and preferential excitation of PS-I and PS-II systems. As discussed in Chapter 5, genes in many biological processes in *Cyanothece* sp. ATCC 51142 peak during specific time of the day. This suggests the exitances of highly

coordinated regulatory relationship between genes in main biological processes in cyanobacteria.

With increasing number of publicly available microarrays, ability to derive reliable gene regulatory networks from transcriptomics data has been shown. In [8], 266 microarray data sets from *Halobacterium salinarum NRC-1* under different environmental and genetical perturbations are used to get a gene regulatory network with prediction capabilities. In [25], existing gene regulatory network for *Escherichia coli* was extended using expression data from 445 microarrays. Many of the predicted relationships were validated using experimental procedures. However a comprehensive regulatory network for cyanobacteria still does not exist.

In this chapter, we propose the use of Bayesian network approach to study cellular response of cyanobacteria. We discuss how to combine individual gene expressions, obtained using microarrays from different platforms, to get biological process level behaviors. Biological process level information carry more information towards understanding the overall cellular behavior. We then discuss several approaches available for identifying the structure of a Bayesian network and derive corresponding system level regulatory network for cyanobacteria, *Synechocystis* sp. PCC 6803. We discuss a method to quantify the strengths of the associations between different biological processes. The resultant network is used to simulate some of the experiment conditions and the responses of the network to those conditions are inferred. We show that these inferences agree with the observations made in the original experiments. Finally, we discuss how these type of networks could be helpful, in making decisions on controlling the cellular activities, so that the desired behaviors are achieved.

4.2 Probabilistic Approaches: Bayesian Networks

There are several approaches to derive regulatory networks using transcriptomics data including dynamical system modeling based on continues time and discrete time models and correlation networks. These models try to identify regulatory relationships between different genes. However, due to under-determined nature of the problem where number of variables (genes) are significantly higher compared to number of observations (experiments), these models are unreliable and typically need extensive verifications using additional methods. The problem becomes more significant when the data are obtained from different microarray platforms and experimental procedures. In such situations, probabilistic networks are shown to perform better [68].

Bayesian networks have been very popular in number of fields, including artificial intelligence, decision theory, data fusion and medicine [59]. This approach is shown to be very powerful, when one has to work with imperfect data. This feature makes Bayesian networks an important tool in the field of biology. The data generated in biological experiments are, most of the times, noisy and contained lot of missing values. Bayesian networks can analyze such data sets very effectively. Bayesian approach for biological systems has several desirable properties including, the solid probabilistic background behind the algorithms, the ability to combine data from different conditions and platforms and the ability to make inferences on the network responses under different perturbations. However, the use of Bayesian networks in biology had been constrained, for a long time, due to limited availability of data. Previous applications of Bayesian networks, for studying gene regulation, has been limited to a selected set of genes. For example, [27] focused on the cell-cycle related genes in *Saccharomyces cerevisiae*.

A Bayesian network is a graphical model representing the probabilistic relationships between random variables. The network is usually presented as a directed acyclic graph (DAG), which encodes the conditional independence for the joint probability distribution of the variables. Once the network is restricted to a DAG, given the values of its parent nodes, $parents(X_i)$, the probability of a child node X_i becomes independent of all other non-parent nodes. For example, using Bayes' rule, the joint probability distribution for a four node Bayesian network can be written as

$$P(X_1, X_2, X_3, X_4) = P(X_1) \times P(X_2/X_1) \times P(X_3/X_1, X_2) \times P(X_4/X_1, X_2, X_3), \quad (4.1)$$

and based on the conditional independence represented by the structure of the network shown in Figure 4.1, this expression is simplified to

$$P(X_1, X_2, X_3, X_4) = P(X_1) \times P(X_2/X_1) \times P(X_3/X_1) \times P(X_4/X_2). \quad (4.2)$$

In general, the joint probability distribution of a graph can be given as

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i/parents(X_i)). \quad (4.3)$$

Conditional independence makes the computation of joint probability distribution of a Bayesian network much simpler. We would like to refer to [36] for more details on Bayesian networks.

In biology, Bayesian networks are useful for inferring relationships between genes or gene functions from microarray data. One of the hurdles in this approach is, again, the small number of observed values compared to the variables in the system. Also,

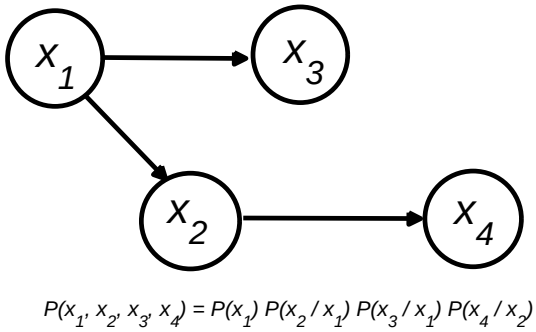


Figure 4.1: A Bayesian network with four nodes, presented as a directed acyclic graph.
Arrows are drawn from parent nodes to child nodes. Given values of the parent nodes, the probability distribution for a child node becomes independent of all other non-parent nodes in the network.

learning the structure of a network with more than 20 variables is a computationally challenging problem.

4.3 Learning the Structure of the Network

Since searching entire domain of structures is super-exponential [63], in most of the practical applications, structure learning is done using heuristic methods. Structure learning is performed either as a constraint-based procedure, where links are removed from the network using conditional independence criteria, or as a score-based procedure, where links are added or removed to minimize/maximize a particular score function.

In [12] greed equivalence search (GES) algorithm using Markov equivalence classes is introduced. The GES algorithm starts with an empty network and derives the

optimal network following a two-step procedure. In the first step network is extended by adding links to the network, one at each cycle. The second step involves removing links from the resultant network. Algorithm stops when score cannot be improved using any of these two steps. Structure learning algorithms based on Maximum Weight Spanning Tree (MWST) [13] restrict the search space only to tree structures and thereby improves the execution time. K2 algorithm requires the user to specify the hierarchy of the nodes and algorithm searches for the best structure only among the networks that satisfy the given hierarchy. When node hierarchy is unknown, K2 algorithm can be initiated with MWST [33] or using mutual information approach (K2-MI) [11].

One of the commonly used score functions, Bayesian information criteria (BIC), involves maximizing

$$BIC(S/D) = \log_2 P(D/\hat{\theta}_s, S) - \frac{\text{size}(S)}{2} \log_2(N), \tag{4.4}$$

where S is the structure of the network defining the nodes and the links between the nodes, $\hat{\theta}_s$ is the set of estimated parameters. D is observed data given as an $M \times N$ matrix, where N is the number of nodes and M is the number of observations. Since the entire domain of structures is super-exponential, searching for the correct structure in large networks with more than 20 nodes is done using heuristic methods.

4.4 Quantifying Influence between Nodes: Links Strengths in the Network

All the links in a network do not have the same level of influence from the parent nodes to the child node. In order to quantify the link strengths, the true link strength percentage [81], given by

$$LS_{true}(X \rightarrow Y) = \frac{U(Y/\mathbf{Z}) - \mathbf{U}(\mathbf{Y}/\mathbf{X}, \mathbf{Z})}{U(Y)} \times 100\%, \tag{4.5}$$

where

$$U(Y/\mathbf{Z}) = -\sum_{\mathbf{z}} p(\mathbf{z}) \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y}/\mathbf{z}) \log_{\mathbf{2}} \mathbf{p}(\mathbf{y}/\mathbf{z}),$$
$$U(Y/X, \mathbf{Z}) = -\sum_{x,\mathbf{z}} p(x, \mathbf{z}) \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y}/\mathbf{x}, \mathbf{z}) \log_{\mathbf{2}} \mathbf{p}(\mathbf{y}/\mathbf{x}, \mathbf{z}),$$

is used. Here $LS_{true}(X \rightarrow Y)$ is true link strength of the arrow from X to Y. $\mathbf{Z}$ corresponds to the parents of Y, except for X. The corresponding probability densities are represented by $p()$, and the summations are taken over all combinations.

True link strength quantifies the percentage reduction of uncertainty on the state of a child node given the state of a parent node. It is computed as the ratio between reduction of entropy of child node given the parent node and the original entropy of the child node.

4.5 Inferring Behavior of the Network under Different Conditions

One of the powerful features of Bayesian networks is making inferences on expected changes in the networks under different perturbations. This allows one to make predictions on optimal changes to be made so that a desired behavior could be obtained from the system. There are several existing algorithms to perform the inferences on Bayesian networks. The junction tree algorithm; an exact algorithm is one of the popular technique to get marginal probabilities of a bayesian network given an evidence(s) [36].

4.6 Bayesian Network for Biological Processes in *Synechocystis* sp. PCC 6803

4.6.1 Data Processing

Transcriptomics data from 164 published and unpublished microarray experiments are combined to derive a regulatory network for *Synechocystis* sp. PCC 6803. Some of these data sets are from time course data on a single perturbation, while the others are single time point data with different perturbations. Published data sets were collected from the both NCBI-GEO [21] and KEGG expression [37] databases.

Since data is obtained from different sources, the differences in the experimental conditions and microarray platforms give rise variations in the data and make the combined analysis difficult. Data need to be processed and combined carefully, so

that the variations between the platforms had minimal effect on the final conclusions. Raw data sets are processed using the robust version of LOWESS normalization [60], to remove the bias in the data. For other data sets, the normalized data from the corresponding databases are used.

In order to avoid the effects of local changes in different microarray chips, differential behaviors of genes are identified using statistical significance tests only. Further the data is discretized into three levels; up, down and not expressed, so that the individual experiments have the same contribution towards the final conclusions. This step was performed for each experiment separately to ensure, that it is independent of the microarray platform variations. Figure 4.2 shows the histogram for distribution of genes and the differentially expressed experimental conditions. Majority of the genes are differentially expressed in about 20%-35% of experiments while in about 10% of genes, expressions are modified in more than 90% of conditions. Biological significant of these highly expressed genes are explored in detail in [71].

4.6.2 Obtaining Process Level Behavior using Gene Expressions

Since, *Synechocystis* sp. PCC 6803 consists of in excess of 3000 genes, deriving a global regulatory network at gene level, using Bayesian network approach, is computationally infeasible. Further gene level networks, for most of the times, are difficult to interpret and do not provide complete picture of the cellular response. This problem becomes further complicated, since the role of many of the genes are currently unknown.

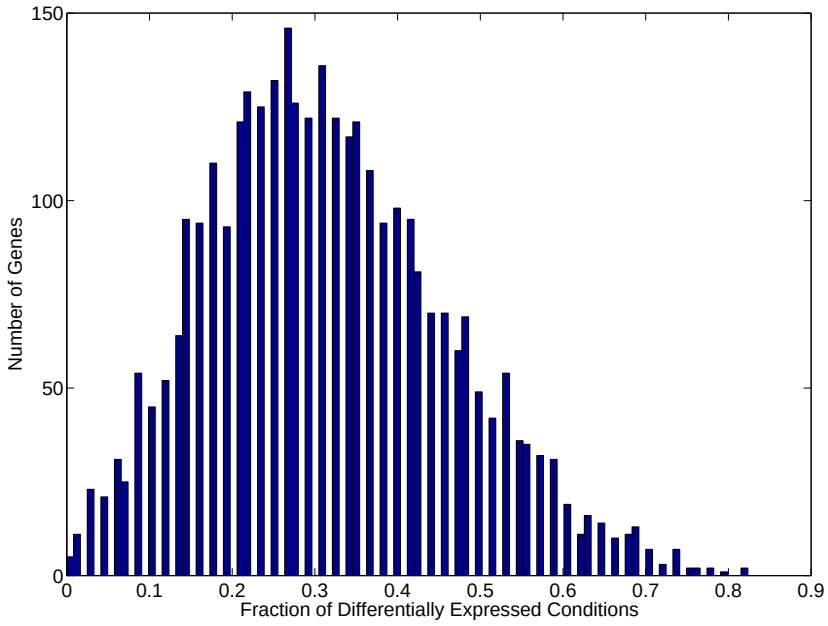


Figure 4.2: Histogram for number of genes vs. fractions of differentially expressed conditions.
The most of the genes are differentially expressed in about 20%-35% of the conditions.

In contrast, a network at biological processes level provides more useful information for biologists. Behaviors of biological processes provide direct interpretations on the nature of the overall response of cells to different experiment conditions. Further, the process level behaviors are determined using a group of genes. As a result the missing values of individual genes have a minimal effect on the computations. In this analysis, the KEGG metabolic pathway [38] classifications are used to group related genes into biological processes.

The distributions of individual gene expressions from different pathways showed a shift of their sample means to different levels depending on stress conditions. The level of the shift in sample means of the distributions was quantified using one sample 'Kolmogorov-Smirnov (KS)' test [50]. KS-test is utilized to determine whether the observed log-ratio values of genes in each pathway are significantly different from a distribution with a zero mean. If the null hypothesis is rejected at a significance level of 5%, the particular pathway is assigned $+1$ or -1, depending on whether the mean value is > 0 or < 0, representing an up and down regulation respectively. If null hypothesis could not be rejected, we assign that pathway a value 0, indicating that the particular pathway is not differentially expressed under the given condition. In Figure 4.3, the distribution of individual gene expressions of genes belonging to the process ribosome is shown. Based on KS-test at 5% significance level, ribosome pathway is assigned values -1, 0 and $+1$ in Figure 4.3 (a), (b) and (c) respectively.

The *Synechocystis* sp. PCC 6803 genes represent 100 different KEGG metabolic pathways. After considering the percentage of experiments each pathway is differentially regulated and the number of genes included under each pathway, we selected 51 pathways as informative and used for further analysis.

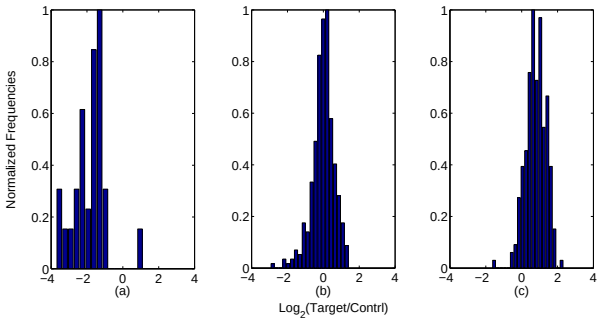


Figure 4.3: Distribution of $log_2(Target/Control)$ values of individual genes in ribosome pathway under three different conditions.
Using KS-test, this pathway was assigned three states namely DOWN (-1) , NOT CHANGED (0) and UP (+1) in distributions shown in (a),(b) and (c), respectively.

4.6.3 Identification of Network Structure

Assuming observations are independent of each other, formulae for BIC in 4.4 can be simplified to

$$BIC(S/D) = \sum_{i=1}^{M} log_2 P(D_i/\hat{\theta}_s, S) - log_2(M) * size(S)/2. \tag{4.6}$$

where D_i is expression values of processes in i^{th} experiment condition, given as an $N \times 1$ vector.

Since the process level expressions are discretized, computation is performed as a frequency counting step given by

$$BIC(S/D) = \sum_{i=1}^{N} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} C_{ijk} log_2 \left(\frac{C_{ijk}}{C_{ij}} \right) - \frac{log_2(M)}{2} \sum_{i=1}^{N} q_i(r_i - 1), \tag{4.7}$$

where N corresponds to the number of nodes in the network, which is 51, corresponding to the number of biological processes selected for the analysis. r_i corresponds to the number of states of process X_i, which is 3 for all the nodes in the network.

$q_i = \prod_{X_l \in parents(X_i)} r_l$ denote the number of possible configurations for the parent nodes of X_i which reduces to $q_i = 3^{p_i}$ where $p_i = \#parents\, of\, X_i$. C_{ijk} corresponds to the number observations for particular combination of i, j and k. C_{ij} is computed as $C_{ij} = \sum_{k=1}^{r_i} C_{ijk}$.

With discretized data, conditional entropy, U(.), calculations to quantify the link strengths also simplified to

$$\begin{aligned} U(Y/\mathbf{Z}) &= -\frac{1}{M} \sum_{\mathbf{z}} \left(\sum_{y} N_{\mathbf{zy}} \log_2 \frac{N_{\mathbf{zy}}}{N_{\mathbf{z}}} \right) \\ U(Y/X, \mathbf{Z}) &= -\frac{1}{M} \sum_{\mathbf{z},\mathbf{x}} \left(\sum_{y} N_{\mathbf{zxy}} \log_2 \frac{N_{\mathbf{zxy}}}{N_{\mathbf{zx}}} \right) \end{aligned} \tag{4.8}$$

where total observations, $M = 164$. $N_{\mathbf{zy}}$, $N_{\mathbf{z}}$, $N_{\mathbf{zxy}}$ and $N_{\mathbf{zx}}$ correspond to the relevant counts for different configurations of $\mathbf{z}$, y and x.

4.6.4 Software Implementation

Matlab (www.mathworks.com) versions of the related algorithms have been implemented in different toolboxes by Kevin Murphy [54], Olivier Francois [26] and Imme Ebert-Uphoff [81]. However some of these implementations scaled poorly with the networks having large number of nodes. As a result, modifications were needed to improve the speed. We re-implemented routings for cache management used to save scores for already computed sub-graphs, algorithm for conversions from partially directed acyclic graphs to directed acyclic graphs and algorithm for calculation of local scores using BIC, in C++, which improved the total execution time by orders of magnitudes. This enabled us to derive the relevant network using a regular personal computer within short period of time.

Table 4.1: Bayesian information criterion (BIC) scores for networks of biological pathways, obtained using different structure learning algorithms.

Method	BIC Score ($\times 10^3$)
GES	-7.4709
MWST	-7.4922
K2-MWST	-7.4775
K2-MI	-7.6647

Greed equivalence search (GES) algorithm resulted in a network having the highest score. However it takes more computational time compared to other algorithms. All algorithms other than GES limit their search domain to certain classes of DAGs.

4.7 Results and Discussion

4.7.1 Network Structure

Final network for biological processes in *Synechocystis* sp. PCC 6803 is derived using GES algorithm with BIC as the score function. This algorithm resulted in a network with the highest score compared to other algorithms available for structure learning, including MWST, K2-MWST and K2-MI. Table 4.1 gives the highest BIC scores obtained using different algorithms.

Figure 4.4, shows the resulting network for the selected 51 pathways. Colors of the links represent the influence of corresponding parent node on the child. The link strength percentages for the network varied between 15.8% - 45.8%. The strongest links are observed between Carbon Fixation and Glycolysis / Gluconeogenesis metabolites; Purine and Pyrimidine metabolism; and Citrate cycle (TCA cycle) and Reductive carboxylate cycle (CO2 fixation). Also strong connections are observed connecting many central metabolic pathways in the cell, including energy generation related pathways such as oxidative phosphorylation and pentose phosphate pathway; energy storing pathways such as carbon fixation and; energy consumption and growth related

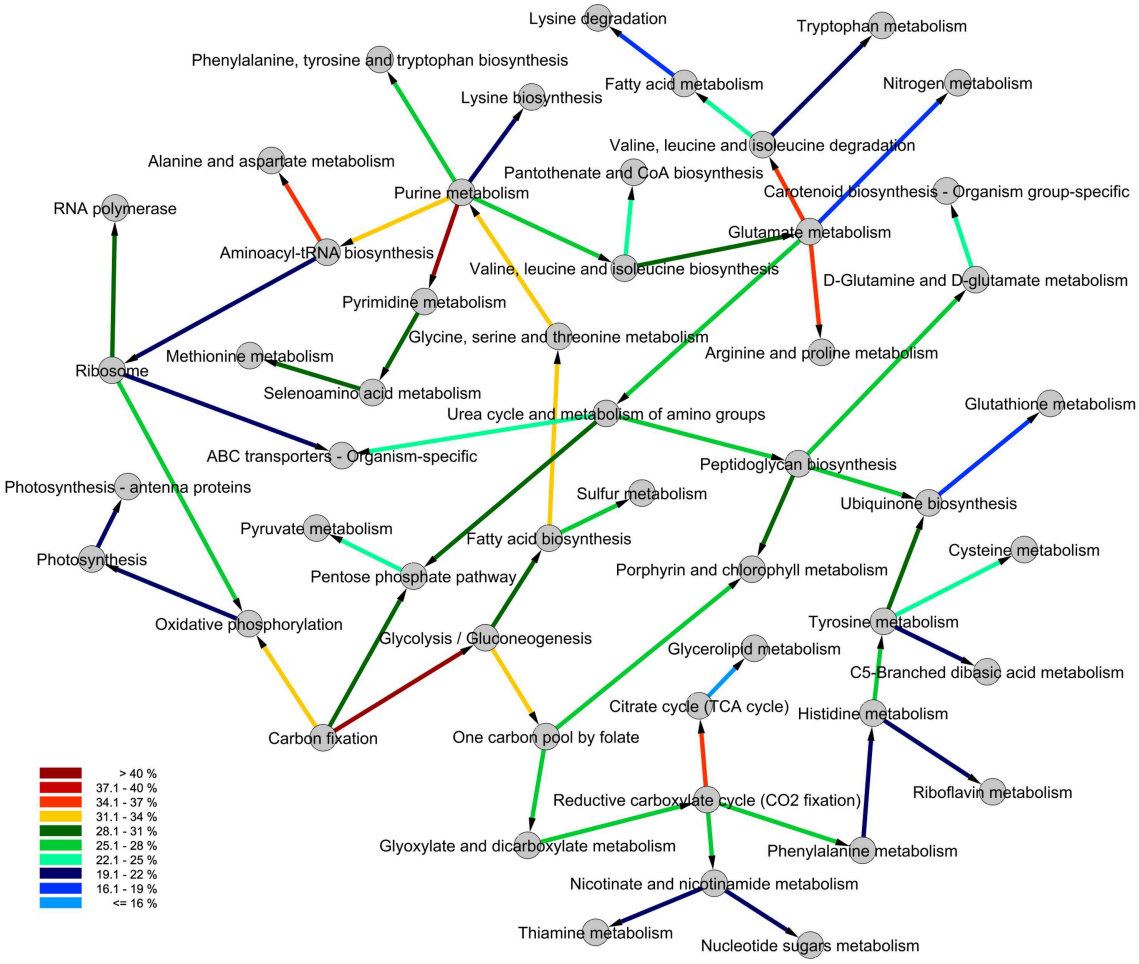


Figure 4.4: Bayesian network for KEGG pathways derived using GES algorithm and BIC scoring criteria.
The colors of the arrows represents the strength of the links, quantified using the true link strength percent (4.5).

pathways such as ribosome, glutamate and purine metabolism etc. This suggests a well coordinated behaviors of the processes critical to the survival of the cells. Table 4.2 lists strengths of the links in the final network.

Table 4.2: Association between different biological pathways in *Synechocystis* sp. PCC 6803 computed using true link strength percentage.

Pathway-1	Pathway-2	Strength
Carbon fixation	Glycolysis/Gluconeogenesis	45.8
Purine metabolism	Pyrimidine metabolism	42.6
Aminoacyl-tRNA biosynthesis	Alanine and aspartate metabolism	38.6
CO2 fixation	Citrate cycle (TCA cycle)	38.3
Glutamate metabolism	Arginine/proline metabolism	36.5
Glutamate metabolism	Valine, leucine & isoleucine degradation	35.9
Purine metabolism	Aminoacyl-tRNA biosynthesis	35.5
Glycine, serine & threonine metabolism	Purine metabolism	34.2
Glycolysis / Gluconeogenesis	One carbon pool by folate	33.0
Carbon fixation	Oxidative phosphorylation	32.8
Fatty acid biosynthesis	Glycine, serine & threonine metabolism	32.8
Carbon fixation	Pentose phosphate pathway	31.8
Valine, leucine and isoleucine biosynthesis	Glutamate metabolism	31.2
Tyrosine metabolism	Ubiquinone biosynthesis	30.9
Urea cycle and metabolism of amino groups	Pentose phosphate pathway	30.7
Pyrimidine metabolism	Selenoamino acid metabolism	29.8
Selenoamino acid metabolism	Methionine metabolism	29.6
Ribosome	RNA polymerase	29.5
Glycolysis / Gluconeogenesis	Fatty acid biosynthesis	29.4
Peptidoglycan biosynthesis	Porphyrin and chlorophyll metabolism	29.4
Reductive carboxylate cycle (CO2 fixation)	Phenylalanine metabolism	28.6
Purine metabolism	Phenylalanine & tryptophan biosynthesis	27.1
Glutamate metabolism	Urea cycle and metabolism of amino groups	27.1
Reductive carboxylate cycle (CO2 fixation)	Nicotinate and nicotinamide metabolism	27.0
Purine metabolism	Valine, leucine and isoleucine biosynthesis	26.9
Glyoxylate and dicarboxylate metabolism	Reductive carboxylate cycle (CO2 fixation)	26.9
Histidine metabolism	Tyrosine metabolism	26.9
Fatty acid biosynthesis	Sulfur metabolism	26.8
Ribosome	Oxidative phosphorylation	26.8
Peptidoglycan biosynthesis	D-Glutamine and D-glutamate metabolism	26.6
Urea cycle and metabolism of amino groups	Peptidoglycan biosynthesis	26.0
One carbon pool by folate	Glyoxylate and dicarboxylate metabolism	26.0
One carbon pool by folate	Porphyrin and chlorophyll metabolism	26.0
Peptidoglycan biosynthesis	Ubiquinone biosynthesis	25.9
Urea cycle and metabolism of amino groups	ABC transporters	25.5
D-Glutamine and D-glutamate metabolism	Carotenoid biosynthesis	24.0
Tyrosine metabolism	Cysteine metabolism	23.6
Pentose phosphate pathway	Pyruvate metabolism	23.1
Valine, leucine and isoleucine biosynthesis	Pantothenate and CoA biosynthesis	22.8
Valine, leucine and isoleucine degradation	Fatty acid metabolism	22.6
Purine metabolism	Lysine biosynthesis	21.7
Tyrosine metabolism	C5-Branched dibasic acid metabolism	21.3
Valine, leucine and isoleucine degradation	Tryptophan metabolism	21.1
Ribosome	ABC transporters - Organism-specific	21.0

Continued on next page

Table 4.2 – continued from previous page

Pathway-1	Pathway-2	Strength
Aminoacyl-tRNA biosynthesis	Ribosome	20.8
Nicotinate and nicotinamide metabolism	Thiamine metabolism	20.1
Nicotinate and nicotinamide metabolism	Nucleotide sugars metabolism	20.1
Phenylalanine metabolism	Histidine metabolism	20.1
Oxidative phosphorylation	Photosynthesis	20.0
Photosynthesis	Photosynthesis - antenna proteins	20.0
Histidine metabolism	Riboflavin metabolism	19.8
Ubiquinone biosynthesis	Glutathione metabolism	19.0
Glutamate metabolism	Nitrogen metabolism	17.8
Fatty acid metabolism	Lysine degradation	16.8
Citrate cycle (TCA cycle)	Glycerolipid metabolism	15.8

Higher links strengths suggest stronger connection between corresponding pathways. In general stronger connections are observed between central metabolic pathways, suggesting a higher level of coordination between vital biological processes.

4.7.2 Network Inference: Using Network to Make Predictions on Cell Behavior Under Different Treatments

In this section, we try to simulate some of experimental conditions, observe the responses of the network and compare them with results obtained in actual biological experiments. In Table 4.3, we list some of the experimental conditions considered, main process(s) affected by the treatment and the evidence entered into network to simulate these conditions. We select the pathways that are expected to get affected directly by the corresponding growth condition as inputs and observe changes in the remaining pathways due to changes in the status of the inputs.

In Figure 4.5 we show the changes in probabilities of being up-regulated for some of the processes under different CO_2 conditions under elevated light inputs. Level of light is entered by changing status of photosynthesis to +1 while the CO_2 level is

Table 4.3: Simulating response of the Bayesian network under different experimental conditions.

Experiment Condition	Process	Evidence
Low light growth	Photosynthesis	-1
Low light growth with glucose	Photosynthesis	-1
	Glycolysis	+1
High light growth with ambient CO_2	Photosynthesis	+1
	CO_2 Fixation	0
High light growth with limited CO_2	Photosynthesis	+1
	CO_2 Fixation	-1
High light growth with high CO_2	Photosynthesis	+1
	CO_2 Fixation	+1
High light growth with limited N	Photosynthesis	+1
	Glutamate	-1

Evidences are entered into the network by setting appropriate status for the relevant processes. Pathways which are expected to get affected directly by the corresponding growth conditions are selected as inputs to the network.

represented by changing the status of CO_2 fixation to appropriate levels. With these evidences, inferences made from the network reveal a slight increase in probabilities of being up-regulated for many processes. However these probabilities reduce if the CO_2 supply is limited and increase significantly under higher level of CO_2 supply. These results agree with the observations made in the original experiments, where it is shown that under elevated light conditions, higher growth rates are achieved with high level of CO_2 inputs but limited CO_2 levels hinder the growth ([35] and [70]).

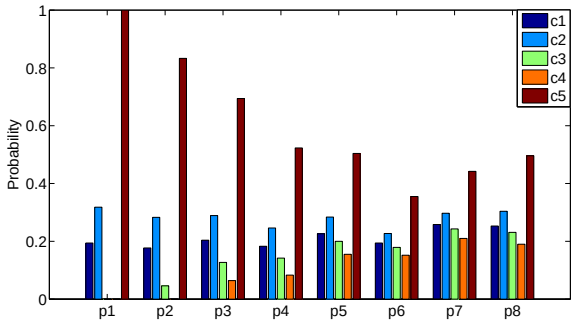


Figure 4.5: Inference from the network simulating some of the experimental conditions.
Probabilities of some of the biological processes being up-regulated under different growth conditions with the presence of high light are presented. Probabilities of being up-regulated are significantly increased for many processes when a higher level of CO_2 is present. These simulation results agree with the observation made in the original experiments, where higher growth rates were achieved under such conditions. c1: original, c2: photosynthesis UP, c3: photosynthesis UP and carbon fixation NOT CHANGED, c4: photosynthesis UP and carbon Fixation DOWN, c5: photosynthesis UP and carbon fixation UP, p1: carbon fixation, p2: glycolysis/gluconeogenesis, p3: fatty acid biosynthesis, p4: glycine, serine and threonine metabolism, p5: purine metabolism, p6: valine, leucine and isoleucine biosynthesis, p7: lysine biosynthesis and p8: pyrimidine metabolism

4.7.3 Comparison between Bayesian Networks and Correlation Measurements

Correlation measurement is an alternating approach for determining relationships between variables. In order to see how Bayesian and correlation approaches compare with each other we generated a correlation network as follows.

Since the expressions are discretized, we use Hamming distance to measure the similarity between different pathways. The Hamming distance measures the fraction of times two expressions differ from each other and is defined as,

$$D(X,Y) = \frac{\sum_{M_{xy}} I(X_j \neq Y_j)}{M_{xy}} \tag{4.9}$$

where X and Y are two pathways, M_{xy} is number of experiment conditions considered for the distance measurement between X and Y, and $I()$ is the indicator function which takes values $I(true) = 1$ and $I(false) = 0$.

Two pathways, which are not differentially expressed in a large number of conditions, can give rise to a smaller Hamming distance and thus can be misleading. In order to avoid this, we included only those conditions, where at least one of the two expressions was differentially expressed.

In order to compare with the Bayesian network, a correlation network is generate by connecting those nodes where $Hammingdistance \leq 0.3$. If any node is unconnected to the network based on this criterion, it is linked to its nearest neighbor in terms of the distance. This resulted in a network having the same number of nodes and connections to that of the Bayesian network. There are 26 links among the total of 55, which are common to both networks.

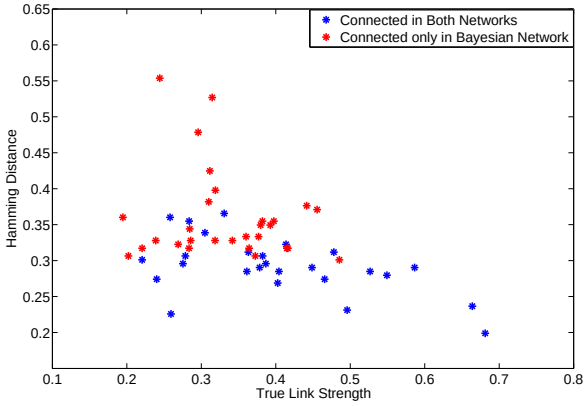


Figure 4.6: Hamming distance and true link strength measurements for connections in the Bayesian network.
Connections common to both Bayesian and Correlation networks consist of small Hamming distances. However Bayesian network captured some non-linear associations which are not identified by correlation based measurements.

In Figure 4.6, Hamming distance and true link Strength measurements for links in the Bayesian network are shown. It should be noted that the links common to both Bayesian and correlation networks consist of small Hamming distances (thus more similar in their expressions). In additions Bayesian network consisted of several links with larger Hamming distances, indicating that it captured some non-linear relationships, which were not observed by linear measurements such as Hamming distance.

4.8 Conclusions

In this section we presented the possibility of using probabilistic approaches to integrate the transcriptomics data from numerous sources. We proposed a statistical approach to derive the biological pathway level behaviors using expressions of individual genes. A probabilistic network based on Bayesian approach is derived for

Synechocystis sp. PCC 6803 considering 164 transcriptomics data sets. We quantified the association between different pathways using true link strength percentage, a measure of reduction in entropy of a given child node due to inclusion of different parent nodes.

The resultant network is used to simulate various experimental conditions. We show the inferences made from the network to agree with the observations made in the original biological experiments. We also compare the networks obtained Bayesian-based network with a correlation-based network and show that it captures some associations not picked by correlation.

Chapter 5

Elucidating Diurnal Rhythms in Cyanobacteria

5.1 Diurnal Rhythms in Cyanobacteria

Diurnal rhythms or day-night cycles have been observed in wide range of organisms from bacteria to mammals [56]. Filamentous fungus *Neurospora crassa* shows a daily rhythm in production of asexual spores [48]. The common fruit-fly, *Drosophila melanogaster*, shows different activity levels depending on the time of the day; higher activity levels at the sunrise and sunset and lower activity level during other times of the day [62]. Photosynthetic plant *Arabidopsis thaliana* shows diurnal movement in its leaves [40]. In mammals wake-sleep cycle synchronizes with the day-night cycle of the earth. Activities of many organs including liver is shown to be diurnally regulated in mice [2]. Relatively recent times, daily cycles have been observed in single cell organisms including many cyanobacteria [30].

Diurnal rhythms can be driven by two main causes, namely the external environment cues, particularly light and the temperature and internal time keeping mechanisms.

Many organisms including species of cyanobacteria, fungus, incests, plants and mammals have developed specialized genes and/or cells for keeping the time and these mechanisms are commonly known as circadian clocks.

Cyanobacterium *Cyanothece* sp. ATCC 51142 shows strong diurnal behavior. It has been observed many of the biological processes in *Cyanothece* are diurnally regulated [74]. This behavior is critical for the survival of the organism, as it needs to well coordinate two essential but incompatible processes, photosynthesis and nitrogen fixation within a single cellular environment [67].

5.1.1 Aims

Two transcriptomics experiments have been conducted to identify the diurnal behaviors in *Cyanothece*. In [74], *Cyanothece* sp. ATCC 51142 is grown under alternating 12 hour light and dark cycles, while in [79], cells are grown under a 12 hour light and 12 hour dark period followed by a constant light period of 24 hours. In both experiments *Cyanothece* sp. ATCC 51142 cells are in nitrogen-fixing conditions. Global transcriptomics measurements are made for two consecutive diurnal periods with a sampling rate of every four hours and a shift in sampling time of one hour between the experiments. The studies were conducted using Agilent (www.agilent.com) custom made two-channel microarrays.

Aims of this analysis include identification of genes showing oscillatory behavior at transcription level, classification of oscillatory genes, and characterizing altered behaviors due to changes in light input patterns.

5.2 Identifying Rhythmic Behaviors in Gene Expressions: Fourier Score and False Discovery Rates

Fourier score and false discovery rates (FDR) based approach is originally proposed for the detection of cell cycle related genes [10]. The Fourier score of any signal $x(t)$ given as a finite dimension vector is defined by

$$F = \sqrt{\left(\sum_t \sin \omega t \cdot x(t)\right)^2 + \left(\sum_t \cos \omega t \cdot x(t)\right)^2}, \tag{5.1}$$

where $\omega = 2\pi f$ is the angular frequency of the expected oscillations. In order to identify the main frequency components of the gene expressions, fast Fourier transform (FFT) can be performed on the mean deducted data.

When a given signal is oscillatory and of the same frequency as the reference signal, it gives rise to a larger Fourier score. In order to quantify the significance of the Fourier score, we compare the value for the original signal with the Fourier scores for large collection of random signals. These random signals are obtained by using different permutations of the original signal. The significance of the Fourier scores can be quantified using p-value measurements or using the false discovery rate.

The p-value is a significance measurement which is computed for each gene separately. P-value of a given Fourier score is defined as

$$pval = \frac{\sum_{j=1}^{M} I(FS_j \geq FS_0)}{M}, \tag{5.2}$$

where FS_j is Fourier score for the j^{th} random signal obtained using a permutation of the original expression for a given gene, FS_0 is Fourier score for the original gene expression, and M is number of permutations which is selected to be a large number such as 10000. $I(x)$ is an indicator function taking values,

$$I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

False discovery rate (FDR) is a global measurement computed using all the gene expressions. An empirical FDR for a chosen threshold t for the Fourier score is define as

$$FDR(t) = \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} I(F_{j,k} \geq t)/M}{\sum_{k=1}^{N} I(F_k^o \geq t)}, \tag{5.3}$$

where M is number of permutations used for the null hypothesis, N is total number of genes, $F_{j,k}$ Fourier score for the j^{th} random signal obtained using the k^{th} gene and F_j^o is Fourier score for the original expression of k^{th} gene. The original signals are scaled to have a unit standard deviation, so that Fourier scores for different genes are comparable.

Under alternating light input conditions, all the diurnal genes, irrespective of whether they are under the regulation of the circadian clock or the external light input, show oscillatory expression levels. Therefore Fourier score based approach can be utilized to identify the diurnal genes using the expressions levels measured in [74].

5.3 Angular Distance based Classification for Identification of Transient Behaviors

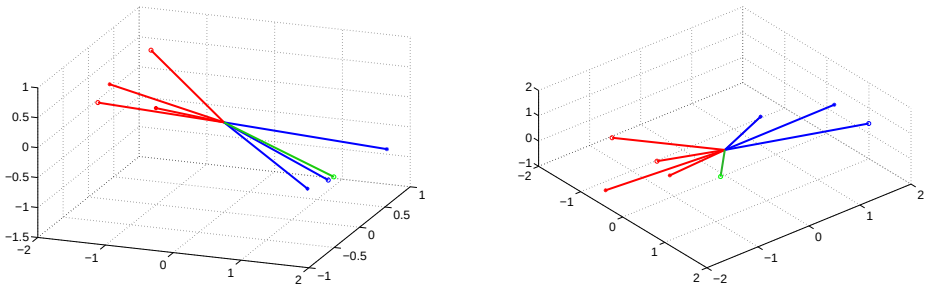
The Fourier score and false discovery rate based method is primarily derived to separate cyclic behaviors from non-cyclic ones. However it cannot be reliably used to detect the altered gene behaviors with the changing light conditions. For example, in [79], genes were under alternating light conditions for first 36 hours and then switched to a different input condition in the last 12 hours. Since there is an oscillatory behavior for the first three thirds of the measurements, even for a gene which altered its behavior during the last 12 hours, Fourier score method is likely to give a significant value and fail to detect the change in expression. Here we propose a classification method based on angular distance to correctly classify the transient behaviors under altered light conditions.

The data is separated into four 12 hour data sets, which correspond to the different light and dark periods in each experiment. Accordingly we obtain four 3-dimensional vectors for each gene for each experiment. The pair wise angular distances between different vectors for a given gene is calculated as,

$$d_{1,2} = \left(1 - x_1 {x_2}^T / (x_1 {x_1}^T)^{1/2} (x_2 {x_2}^T)^{1/2}\right), \tag{5.4}$$

where x_1 and x_2 are the vectors correspond to two different 12 hour periods. The distances $d_{1,2}$ can have any value between 0 and 2, with 0 representing the vectors in the same direction and 2 representing vectors with the opposite direction.

With this approach, for oscillating genes, smaller distance measures are obtained for expression vectors coming from similar light regimes and larger distance measures



(a) Hydrogenase gene cce_2318 (b) Hydrogenase gene cce_1063

Figure 5.1: Distribution of vectors corresponding to different light regimes for two Hydrogenase genes.
A gene which does not change its behavior significantly under subjective dark is shown in Figure 5.1(a) while Figure 5.1(b) shows a gene which changes its behavior significantly. Red: under light, Blue: under dark and Green: under last 12h in [79]

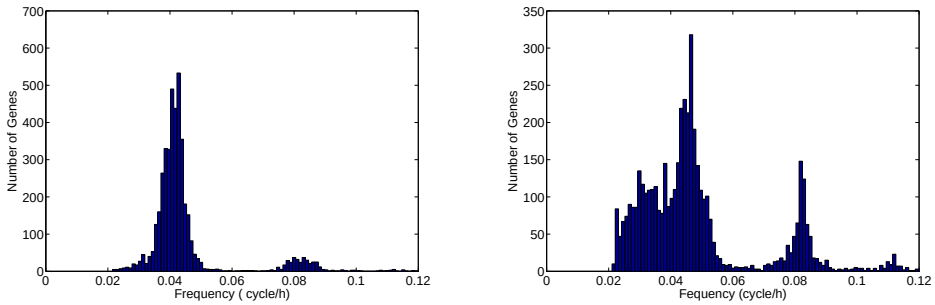
are obtained for expression vectors coming from different light regimes. If a gene did not change its behavior during last 12 hours in [79], that gene is expected to give a small distance measurement for two vectors corresponds to the second and the fourth 12 hour regimes. Accordingly angular distance based approach can be utilized to characterize the altered behaviors due to changes in light input pattern.

The idea of using angular distance for characterizing gene behavior under different light regimes is graphically shown in Figure 5.1. It shows the distribution of vectors corresponding to different light regimes for two selected genes. First gene shows oscillations under both conditions while second gene ceases to oscillate under constant light conditions. Clearly for the gene which show change in its behavior under constant light conditions, the vector corresponds to last 12 hour in LDLL is located away from vectors corresponds to regular light and dark regimes.

5.4 Combining Fourier Score and Angular Distance based Approaches

Two methods discussed above can be combined to get an accurate classification of gene behaviors. Diurnal genes identified using Fourier score based method can be classified into two groups, namely circadian controlled genes (CCGs) and light responding genes (LRGs). CCGs are primarily under control of the circadian clock and do not get their behavior altered by changes in light input while LRGs readily respond to incident light patterns.

When combining Fourier score and angular distance based methods, it is required to determine a threshold for angular distance measurements in order to decide whether two expression vectors from different light regimes are similar or not. One logical approach is to pick a threshold value that results in the highest agreement between two methods for diurnal genes under alternating light conditions. Genes that are identified as cyclic, using gene expressions under [74], by Fourier scores are analyzed using angular distance. With different thresholds for distance measure, number of genes classified as cyclic by angular distance criterion is computed. The value which produce the maximum agreement between two methods is picked as the threshold for gene classification. By looking at the expressions of these genes under [79], they are classified as CCGs or LRGs.



(a) Distribution of main frequencies in [74] (b) Distribution of main frequencies in [79]

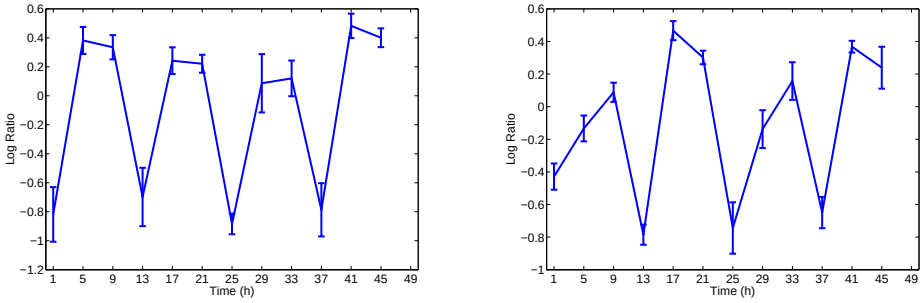
Figure 5.2: Main frequencies present in the gene expressions are found using fast Fourier transform.
Distribution of frequencies in two experiments shows clear differences, suggesting significant influence by the incident light pattern.

5.5 Diurnal Genes in *Cyanothece* sp. ATCC 51142

5.5.1 Fast Fourier Transform to Identify Main Oscillatory Frequencies in Gene Expressions

Fast Fourier transform analysis of gene expressions in [74] reveals the existence of two main frequencies correspond to 24 hour and 12 hour under alternating light conditions. This is clear from the distribution of main frequencies shown in Figure 5.2(a). These frequencies remain even after switching to continues light conditions [79], as observed in Figure 5.2(b). However number of genes showing same oscillatory frequencies are much lesser compared to the alternating input condition indicating that many genes altered their behavior with the changes in the light input.

One of the novel finding of this analysis is identification of ultradian genes; genes that oscillate with shorter periods (in this case 12 hour) compared to regular 24 hour period. This is a novel discovery for any cyanobacteria. In Figure 5.3, two genes with 12 hour oscillations are shown.



(a) cce_1889 (b) cce_3225

Figure 5.3: Two genes showing 12h oscillations. Identification of ultradian genes is a novel finding for any cyanobacteria.

5.5.2 Fourier Score and False Discovery Rate based Method Identified more Diurnal Genes than Previously Reported

Since two main oscillatory frequencies are identified in the gene expressions, Fourier score for each gene is calculated using two reference signals, one having 12 hour period and the other having 24 hour period. Threshold for Fourier score is selected at 2% FDR level. Compared to 1445 genes reported in the original analysis [74], 2138 genes representing 43% of the genome of *Cyanothece* sp. ATCC 51142 are determined as diurnally regulated by the Fourier score based approach. This suggest that the diurnal behavior of genes at transcriptomics level is much wide-spread than previously reported.

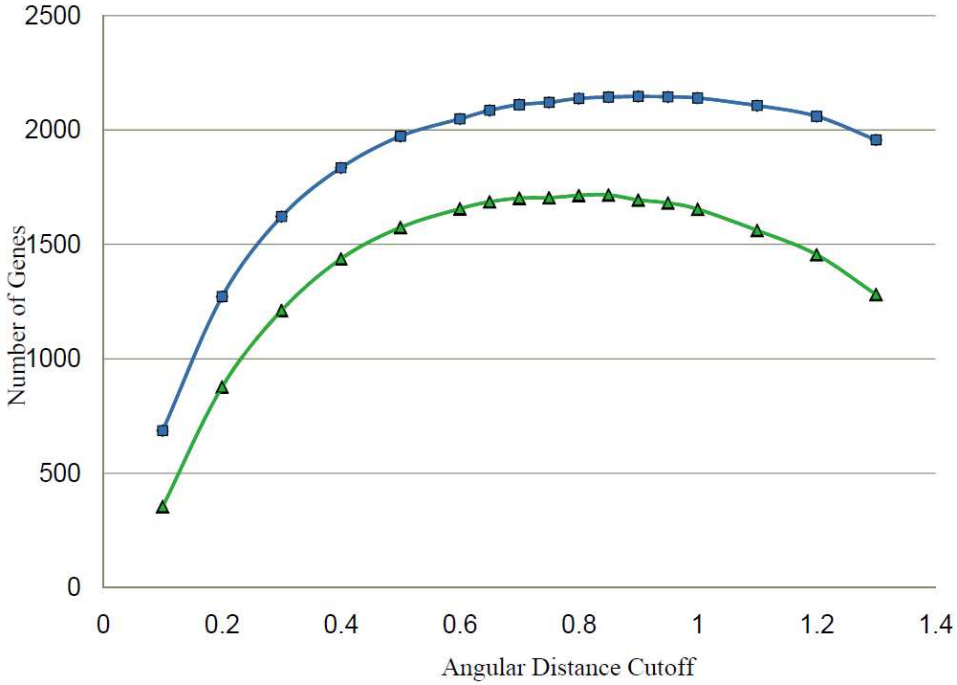


Figure 5.4: Threshold for angular distance is selected so that the agreement with the Fourier score based method is maximum.
For different thresholds, we compute the number of genes classified as diurnal by the angular distance based method. The blue color curve represents the number of genes in agreement if we consider data from only [74]. Similar results for considering first 36 hours in [79] also is given by the green curve. We picked 0.8 as suitable cutoff, since it corresponded to the maximum agreement between two methods.

5.5.3 Majority of the Diurnal Genes Respond to External Input Patterns

The 2138 diurnal genes identified from the analysis of [74] are used to determine the threshold for angular distance measurements. For different thresholds, number of genes classified as oscillatory by having similar expression vectors for similar light regimes that is in two light regimes or in two dark regimes and different expression vectors for opposite light regimes, namely light vs dark, is computed. The results are shown in Figure 5.4.

Based on these calculations, a cutoff of 0.8 for the angular distance is selected. This cutoff resulted in 97% agreement between two methods for the classification of genes using expressions in [74]. After including data from first 36 hours in [79], 78% agreement between the two methods is obtained. Accordingly the expressions of a gene in two different 12h periods are considered to be similar if the corresponding vectors are within a distance of 0.8 to each other, and disparate if the distance is higher. The vectors of genes transcribed with an ultradian period of 12 hour is assumed to be similar to each other.

After combining two criteria, six main groups of gene behaviors are identified within expression data. Gene counts corresponds to these groups are given in Table 5.1.

Table 5.1: Classification of diurnal genes in *Cyanothece* sp. ATCC 51142, based on their behavior in two experimental conditions.

	Stöckel et. al [74]	
Toepel et. al[79]	24h	12h
24h	448	3
12h	49	5
N.C.	722	45

Periods $24h$ and $12h$ correspond to the periods of the primary oscillations. N.C: Not Cyclic

Accordingly 448 genes that show 24 hour oscillations under both conditions are identified as being under circadian clock (CCGs). 722 genes that oscillated with 24 hour period only under alternating light conditions are classified as light responding genes (LRGs). Additionally 50 genes with ultradian oscillations were detected. Among these genes, 5 genes shows consistent oscillations irrespective of changes in incident light patterns.

Table 5.2: Pairwise angular distance measurements for different light regimes.

	Stöckel et. al [74]						Toepel et. al[79]					
Group	D1L1	D1D2	D1L2	L1D2	L1L2	D2L2	L1D1	L1L2	L1S2	D1L2	D1S2	L2S2
1	1.78	0.11	1.79	1.85	0.08	1.79	1.79	0.12	1.68	1.70	0.33	1.69
2	1.78	0.17	1.72	1.88	0.14	1.76	1.73	0.18	1.57	0.68	1.42	0.78
3	0.23	0.44	0.29	0.37	0.05	0.44	0.58	0.17	0.27	0.46	0.68	0.27
4	0.33	0.17	0.29	0.43	0.20	0.30	1.44	0.40	0.63	1.26	1.04	0.92

Smaller distances represent vectors in same direction (and thus have similar expression pattern) while large distances represent vectors in opposite directions. L:Light, D:Dark and S:Subjective Dark

We observe clear difference for the angular distances for these gene categories. In Table 5.2, the average distance for gene groups under various light regimes are given.

5.6 Analysis of Diurnal Genes

Behaviors of diurnal genes provide vital information on coordination of different biological processes within *Cyanothece* cells. In order to gain more details on gene behaviors, we focused on additional features within diurnal gene expressions.

5.6.1 Clustering Based on Phase of Oscillatory Genes

In order to identify the co-expressed genes, diurnal genes are clustered based on the phase of their expression profiles. Phase of the oscillation is determined using the first term of the Fourier approximation of each gene expression. Peak time is derived from the phase of the oscillation. In Figure 5.5 we present the various gene groups using a graphical representation. Two main gene groups; circadian controlled and light responding genes with 24h oscillations are shown as two rings. Genes belonging

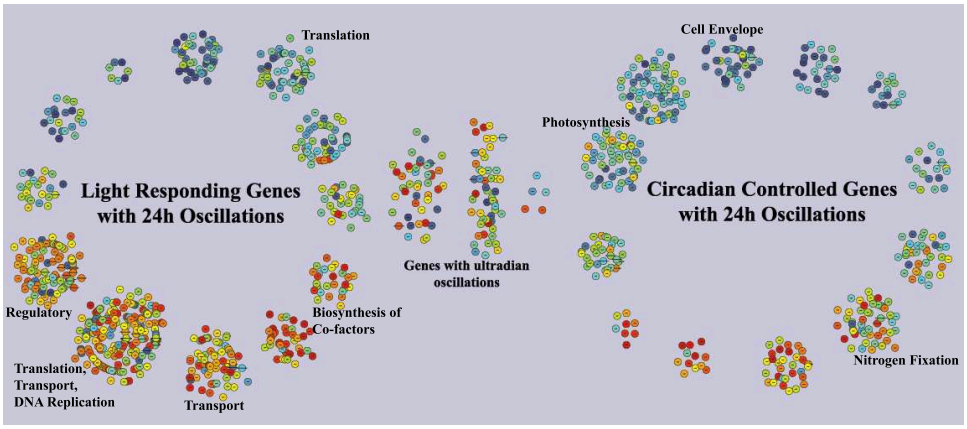
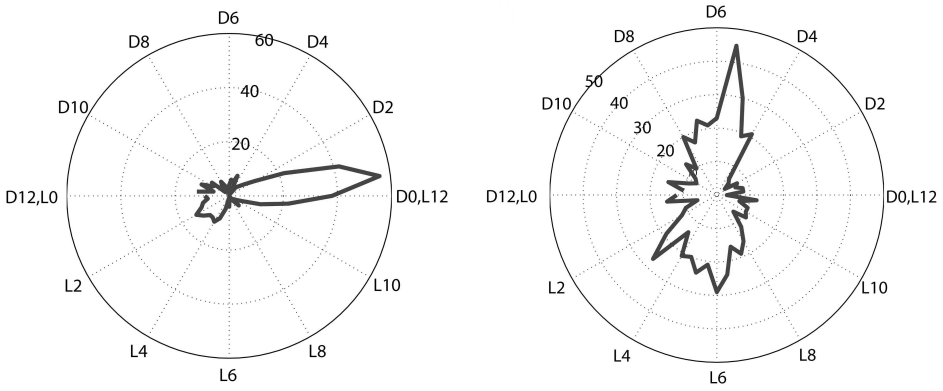


Figure 5.5: Main Gene categories identified using gene classification methods. The CCGs and LRGs are further clustered based on their phases of oscillations and are colored based on their activity levels; red representing high and blue representing low, at a given point of time. Several genes with ultradian oscillations; those with less than 24h periods, are also observed.

to each group are separated into 12 different sub-clusters each, based on the peak times of their activities so that genes which peak during a 2-hour period are grouped together. Number of genes having ultradian oscillations, that is oscillations with less than 24h periods, are also shown separately. Genes are colored based on their activity levels; red representing high activity level while blue indicating lower activity level. Genes belonging to some of the gene functions are over represented in these clusters, thus allowing to identify the sequence of activation of different functions over a course of 24h cycle. Certain biological processes such as ribosomal genes, photosynthesis and nitrogen fixation are highly coordinated with majority of genes belonging to each process peaking at specific time of the day. These gene clusters are used in Chapter 7 to derive process level model for diurnal genes.



(a) *circadian controlled* genes (b) *light responding* genes

Figure 5.6: Distribution of peak times for *circadian controlled* and *light responding* genes.
Majority of the circadian controlled genes peak at the onset of dark period where as light responding genes peak mainly during the mid of light period.

5.6.2 Peak Time Distribution for CCGs and LRGs

Analysis of gene clusters revealed that genes belonging to different biological functions peak at different times of the day, thus clustered together. In order to examine whether there is any difference between behavior patterns of CCGs and LRGs, we compute the distribution of peak times for these two categories of genes. Clear differences in the distribution of peak times are observed from this calculation. Majority of the circadian controlled genes peak at the onset of dark period where as light responding genes peak mainly during the mid of light period, as shown in Figure 5.6.

5.6.3 Localization of Genes in the Genome

Genes physically located close to each other sometimes share a common promoter region and are transcribed as a group. These genes are referred to as 'operons'. Genes belonging to a single operon show similar expression patterns, though there

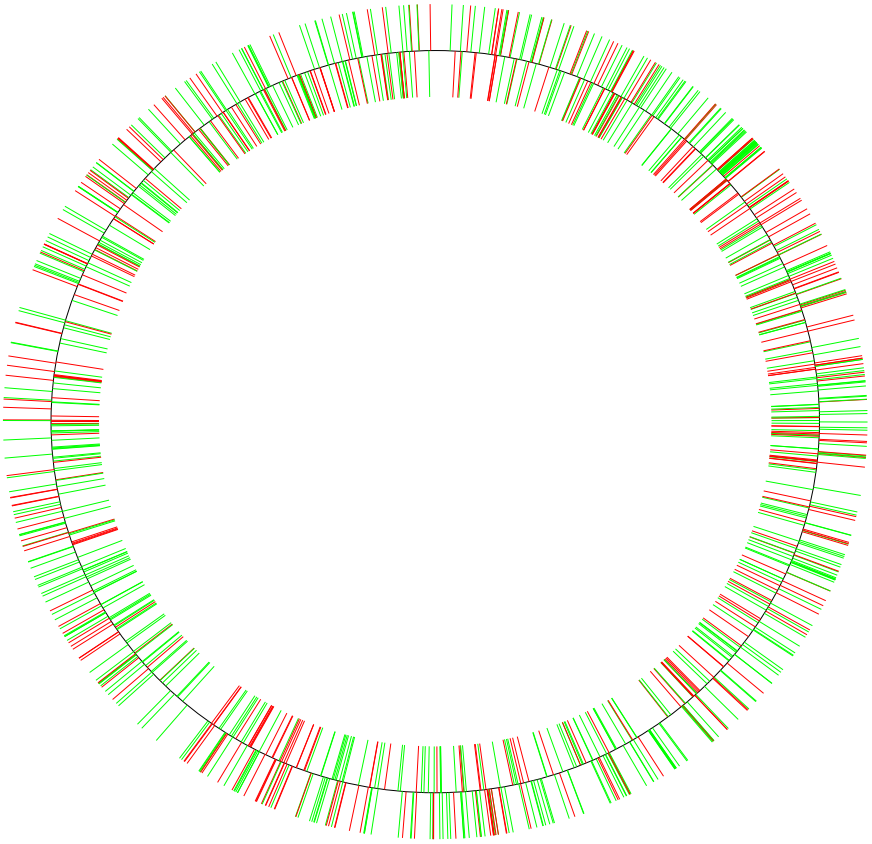


Figure 5.7: Locations of diurnal genes in the circular chromosome of the *Cyanothece* sp. ATCC 51142.
The circadian controlled genes are shown in red and light responding genes are in green. We observe many consecutive genes classified as either CCGs or LRGs. These genes might represent different operons.

can be some variations due to binding efficiency of the RNA polymerase. We examine the positions of CCGs and LRGs in the genome to see whether any localization of these genes is observed. Figure 5.7 shows the locations of the diurnal genes in the genome. Many genes occur in the groups of three or more genes and many such groups consist of CCGs or LRGs only.

5.7 Discussion and Conclusions

After the application of two criteria, Fourier score and angular distance, current analysis of the [74] data set identified 43% of genes (2138 genes) in the *Cyanothece* sp. ATCC 51142 genome with oscillating expression patterns under alternating light and dark conditions. Compared to previously reported 1445 genes, this represents a significant increase in the number of diurnal genes detected in *Cyanothece* sp. ATCC 51142. This observation suggests that diurnal regulation of gene expressions in *Cyanothece* sp. ATCC 51142 might be greater than previously thought. However, after combining and analyzing both data sets using the two different methods, only 722 (14.8%) of genes in the genome were found to be diurnally regulated or light inducible, and 448 genes (9.2%) could be classified as circadian controlled. This relatively small number of diurnally regulated genes common in both data sets results from the stringent criteria used for the gene classification. Use of strict criteria ensures that we pick genes that are insensitive to differences in growth and culture conditions and therefore comprises of robust cyclic behaviors. Interestingly, five circadian controlled genes and 45 genes with transient expression patterns oscillate with an ultradian frequency of 12h.

Taken together, the combination of the angular distance and Fourier Score based methods results in higher level of confidence on identification of cyclic expressed genes in *Cyanothece* sp. ATCC 51142. These analysis uncovered that most of the previously identified diurnal genes are indeed light responsive.

Chapter 6

Modeling Interactions between Diurnal Genes

6.1 Modeling and Identification of Interactions between Genes

Understanding dependence between different genes in transcription and translation activities is very important aspect in studying behavior of cells. However, identifying transcription regulatory links between genes has always been a challenging task. This is primarily due to limited availability of gene expression data, compared to large number of variables (genes) involved in the system. Despite these limitations, numerous methods have been developed to identify possible relationships between genes.

Different approaches have been proposed to model interactions between genes. Whether gene interaction should be modeled as a deterministic or a stochastic process has been debated for a long time. Arguments in favor of the stochastic modeling, is based on the randomness observed during the molecular interactions. However stochastic modeling requires considerably large number of data points and is therefore difficult to use. Though there is inherent randomness in interactions at a molecular level, in

order to understand overall response of genes, it is usually sufficient to study the average behavior of gene products. Furthermore some of the environmental changes, such as day/night cycle, take place at a much slower time scale compared to molecular interactions. As discussed by [78], gene behaviors under such input conditions can be considered as purely deterministic.

Several probabilistic methods to model gene interactions are available. Some methods determine gene interactions based on entropy and mutual information [5]. One of the limitations of these methods is their inability to detect causal relationships between genes; namely separating regulators from the targets. In [25], the authors overcome this hurdle by limiting the regulators to the already known transcription factors. In [46], conditional mutual information was used to establish causal relationships. Various methods based on Boolean networks [3], Probabilistic Boolean networks [68], and Bayesian networks [27] have been applied successfully, to model relatively small number of genes.

Deterministic systems can successfully be modeled using differential equations. Many such models have been proposed based on the interaction patterns observed in the actual system. In [82], feed-forward loop (FFL) has been identified as a dominant motif in gene interaction networks. Coherent FFL based models are used in [10] to study the dynamic interactions between genes and three FFLs are successfully identified in yeast.

6.1.1 Aims

In this section we try to identify possible interactions between diurnal genes in *Cyanothece* sp. ATCC 51142 based on a biological realistic model. The model needs to be

able to explain diverse behavioral patterns observed among diurnal genes including the existence of diverse frequencies and modifications of behaviors under changing input conditions. We also try to bring existing biological insight on gene interactions to refine the resultant interaction model. Finally the resulting network is analyzed to identify its biological relevance.

6.2 Dynamical System Model to Explain Interactions between Diurnal Genes

In order to explain the existence of different behavioral patterns and to study possible interactions between genes we propose a dynamical systems model, given by

$$\dot{Y}(t) = -\alpha_y Y(t) + \beta_y f(X(t), K_{xy}), \tag{6.1}$$

$$\dot{Z}(t) = -\alpha_z Z(t) + \beta_z g(X(t), Y(t), K_{xz}, K_{yz}), \tag{6.2}$$

where $X(t)$, $Y(t)$ and $Z(t)$ represent expression levels of genes X, Y and Z respectively. The activation function $f(X(t), K) = (X(t)/K)^H/(1 + (X(t)/K)^H)$ has two parameters H and K. The parameter H controls the steepness of $f(u, K)$. Its value is shown to be in the range of 1–4 in many biological applications [82]. As discussed later, we select both $H = 1$ and $H = 2$ depending on the gene groups we model. The parameter K defines the expression of Gene X required to significantly activate the expression of the other genes. We assume that the regulators operate away from the saturated regions and pick $K >> X(t)$. The regulator genes X and Y of (6.2), are assumed to be acting independently or additively so that $g(t)$ is selected to have

the form $g(t) = f_x(X(t), K_{xz})f_y(Y(t), K_{yz})$ or $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$ respectively.

The models 6.1 and 6.2 are linear time invariant dynamical systems with $f(t)$ and $g(t)$ being inputs. These models can be solved analytically and the solutions are given by

$$Y(t) = e^{-\alpha_y t}Y(0) + \beta_y \int_0^t e^{-\alpha_y (t-\sigma)}u(\sigma)d\sigma, \tag{6.3}$$

with $u(t)$ as input to the system.

Since the system is asymptotically stable, for large values of time t, the first term in the solution can be ignored. Moreover when $u(t)$ is a periodic function, the expression of the target gene $Y(t)$ would also be oscillating with the same frequency but possibly with some phase shift.

6.3 Explaining Different Gene Groups using the Model

Based on the model, oscillations of the target genes are determined by the oscillations of their regulators. Different types of regulatory relationships give rise to different patterns of behaviors. We assume that some of the higher level regulators get input from two global factors, namely circadian oscillator and/or external light input and subsequently propagate those signals to their target genes.

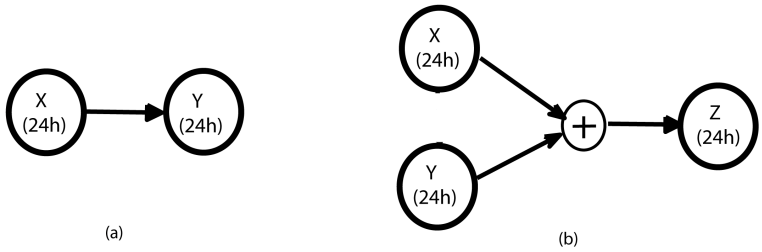


Figure 6.1: Possible regulatory relationships for genes with 24h oscillations. In (a) the target gene X is controlled by a single regulator Y where as in (b), the target gene Z is controlled by two regulators X and Y, acting additively on the target.

Genes with a Main Period of 24h

We select $H = 1$ and assume that most of the genes are regulated by a single regulator, which also has a main period of 24h. These regulatory relationships are first modeled using 6.1. For those genes which could not be explained using a single regulator, we assume the regulation relation to be of 6.2, where two regulators act additively. In this case we try to fit the data using $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$. Figure 6.1 shows possible regulatory mechanisms for genes having 24h oscillations.

Based on the model 6.3, target gene would also be oscillating with a period of 24h. If a gene is under circadian control directly or indirectly then it continues to show the same behavior when the light pattern changes to constant conditions as well. However if it has a significant direct influence from the incident light pattern, then it ceases to oscillate under such conditions. This explains the possible mechanism to observe two different groups of genes, first having 24h oscillations under both experiments and second having 24h oscillations only under alternating inputs.

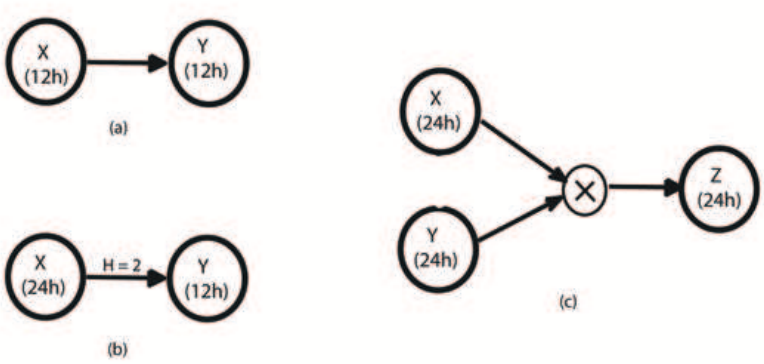


Figure 6.2: Possible regulatory relationships for genes with 12h oscillations. In (a) the target gene X is controlled by a single regulator Y with 12h oscillations where as in (b), the target gene Y is controlled by a single regulators X with 24h oscillations. In the second case we use $H = 2$ in Hill function. In (c), the target gene Z is regulated by two regulators X and Y with 24h oscillations, acting independently on the target.

Genes with a Main Period of 12h

Similar to the explanation given for the 24h genes, if the regulator itself has a 12h oscillation, then the target would also have the same period. This is just one of the possible scenarios. However it is still not clear how 12h oscillations are originated at the first place, since the natural oscillations are of 24h period irrespective of whether they are coming from the circadian clock or the oscillatory diurnal cycle of light input.

We propose two possible scenarios where a regulator with 24h oscillations can give rise to 12h oscillations in the target. First, it can be according to the model 6.1 with $H = 2$. In this case there is a single regulator gene. Second, it might be based on 6.2 with two independent regulators targeting a single target. In this case $g(t)$ takes the form $g(t) = f_x(X(t), K_{xz})f_y(Y(t), K_{yz})$. Both these models can generate 12h oscillations with an input having 24h period. Figure 6.2 shows possible regulatory mechanisms for genes having 12h oscillations.

Genes Oscillate with Different Periods in the Two Experiments

These type of behaviors can easily be modeled using 6.2 with two regulators working additively, thus $g(t)$ taking the form $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$. Two regulators oscillate with two different frequencies and depending on the external conditions, their influence on the target would be different, giving rise to different frequencies in the target under two conditions.

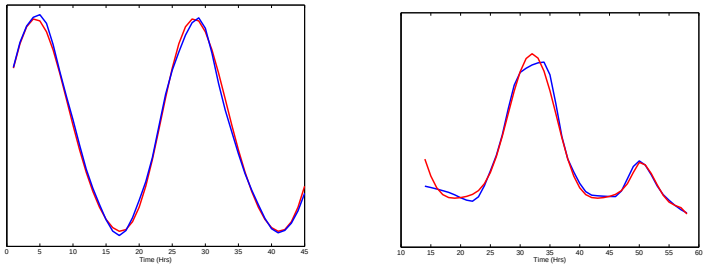
6.3.1 Approximation of Gene Expressions

In [10], ways to simplify the process of finding numerical solutions to ordinary differential equations are discussed in detail. They propose to expand the original expression data using differentiable basis functions so that the derivatives can be computed directly. The approximation problem can be efficiently solved using least square techniques.

Since the main behavioral patterns of the data analyzed here are oscillations, we use sinusoidal functions as basis functions for this problem. Original expressions are approximated as a linear combination of sinusoidal functions along with a linear trend, as given by

$$X(t) = a + bt + \sum_{j=1}^{N} \alpha_j \sin(j\omega t + \phi_j), \tag{6.4}$$

where $\omega = 2\pi/24$ is the angular frequency corresponding to 24h and ϕ_i is the phase angles of the approximated signals. Parameters a, b, α_j and ϕ_j are estimated using least square optimization method. Since the original data is sampled at 4h, N is limited to 2.



(a) Expressions under DLDL (b) Expressions under LDLL

Figure 6.3: Good approximation of a gene expression under two experimental conditions.
Model in 6.4 was sufficient to get good approximations for expressions of more than 75% of diurnal genes in *Cyanothece* sp. ATCC 51142 under both alternating and constant light input conditions.

In order to better capture the transient behavior of light responding genes under constant light conditions, this approximation is done separately for oscillatory region and the transient region. With the selected parameters, model captures at least 75% of total energy in the original signal for more than 99% and 80% of genes in [74] and [79] respectively. In Figure 6.3, we show approximation of an expression of a light responding gene using 6.4 under two experiments. Genes that are not approximated accurately with this model are excluded from further analysis.

Once the original gene expression is approximated, its derivative can be calculated easily as

$$\dot{X}(t) = b + \sum_{j=1}^{N} j\alpha_j\omega_t \cos\left(j\omega t + \phi_j\right). \tag{6.5}$$

6.3.2 Model Fitting

Model fitting is done in several steps. For all possible gene pairs, approximated expressions and their derivatives are fitted using model 6.1. Optimal parameter

values α and β are obtained using nonlinear least square method, minimizing

$$F(\alpha_y, \beta_y) = \| \dot{Y}(t) + \alpha_y Y(t) - \beta_y f(X(t)) \| . \tag{6.6}$$

For the optimal parameter values, the error is calculated using

$$\text{Error} = \mathrm{F}(\mathrm{k_x}, \mathrm{k_u})^2_{\mathrm{opt}} / \| \dot{\mathrm{x}}(\mathrm{t}) \|^2. \tag{6.7}$$

Gene pairs giving rise to a normalized error $\leq 10\%$ are considered as possible regulator-target pairs.

If a gene cannot be approximated using a single regulator, we try to fit the data using 6.2. If a particular target is approximated well using a single regulator in the other experimental condition, that regulator is picked as one of the candidates. This is based on the assumption that most regulatory relationships are preserved under changing conditions but additional regulators can be recruited, specific to the different conditions. If the selected gene does not produce a good model fitting in conjunction with any another gene, acting as the second regulator, we try the possibility of additional gene pairs as regulators, starting with those that gives rise to smaller errors.

6.4 Finalizing the Network Connections

6.4.1 Robustness of the Regulatory Links

Robustness is an essential feature in gene regulations. Biological systems are required to be able to maintain the proper target-regulator relationship in the presence of various disturbances arising from external and internal causes. In order to evaluate the robustness of regulatory links identified using the model, we changed values of the parameters α and β by $\pm 5\%$ from their optimal values and error in model fitting is calculated using 6.7 for the modified parameters. Only those links which give rise to a normalized error $\leq 10\%$ for modified parameters are considered in the final gene regulatory network.

6.4.2 Selecting Most Probable Regulators Among Few Candidates

One of the challenges in deriving a gene regulatory network (GRN) is identifying valid links between genes from many possible candidates. Since the number of time points in the data is significantly less than the variables in the system, these problems are mostly under-determined. As a result identification of most likely relationships needs to be performed using known biological insight about the system. Following are some of the assumptions generally made about gene interactions in bacteria.

1. Genes having same phase are likely to regulated by a single regulator.

2. Biological networks tend to follow power law; few hubs with many genes and many hubs with few genes.

3. Regulatory links between genes are likely to be preserved under changing conditions. Level of influence of regulators might change under different treatment/condition and may become visible only under a specific condition.

4. Genes located in close proximity in the genome may belong to a single operon and are regulated by a single regulator.

5. Regulatory relationships between genes are resilient to external noise.

The assumptions described above can be used to filter out some of the possible links between genes. On the other hand a realistic model should be capable of preserving some of these basic assumptions. So we can use them as a criteria to measure the acceptability of the model for the purpose of explaining observed data.

6.5 Results and Discussion

6.5.1 Gene Interaction Network for *Cyanothece* sp. ATCC 51142 Diurnal Genes

A total of 1251 genes identified as light responding or circadian controlled are used in the analysis. Using 6.4, a total of 1012 genes are well approximated with the approximated signal capturing $\geq 75\%$ of energy of the original signal in both experiments. The network design is limited to these well approximated genes. We found that, for [74], expressions of 968 genes, representing 95% of those included in the network,

could be explained using the single regulator-target model given by 6.1. Remaining genes required at least two regulators and are fitted with the model 6.2. In the case of [79], only 476 (47%) genes are approximated using 6.1. They consist of 334 circadian controlled genes and 137 light responding genes. Furthermore 24 circadian controlled genes, 307 light responding genes and 44 other genes are approximated using 6.2. Behavior of 166 genes are not captured using either of the models. This clearly shows the existence of a more complex level of gene interactions under transient light patterns.

It is observed that the majority of possible regulator-target links are resilient to parameter variations. With 5% deviation from the optimal values, more than 75% of those links remain valid, with the model fitting producing an error $< 10\%$. Final GRN is derived while preserving the properties described in Section 6.4.2. Only those links which were resilient to parameter fluctuations are considered in the final network. The network for [74] consists of 167 unique regulator genes while the network for [79] consists of 250 unique regulators. This represent about 3.5-5% of the total genome. It should be noted that in other well studied bacterial systems such as E.*coli*, percentage of transcription factors is around 3.7%.

Number of targets for a given regulator varied from 1–65, following a power-law distribution with an exponent of -1.9. Using a robust least squares fit we note that the correlation coefficient is 97% for the log-log plot between the distribution of the number of targets and their frequencies, indicating a good approximation for a power-law distribution.

In Figure 6.4, the resulting gene regulatory network is presented under [74] is shown. It was noted that additional links occurs in the network under [79], which are needed

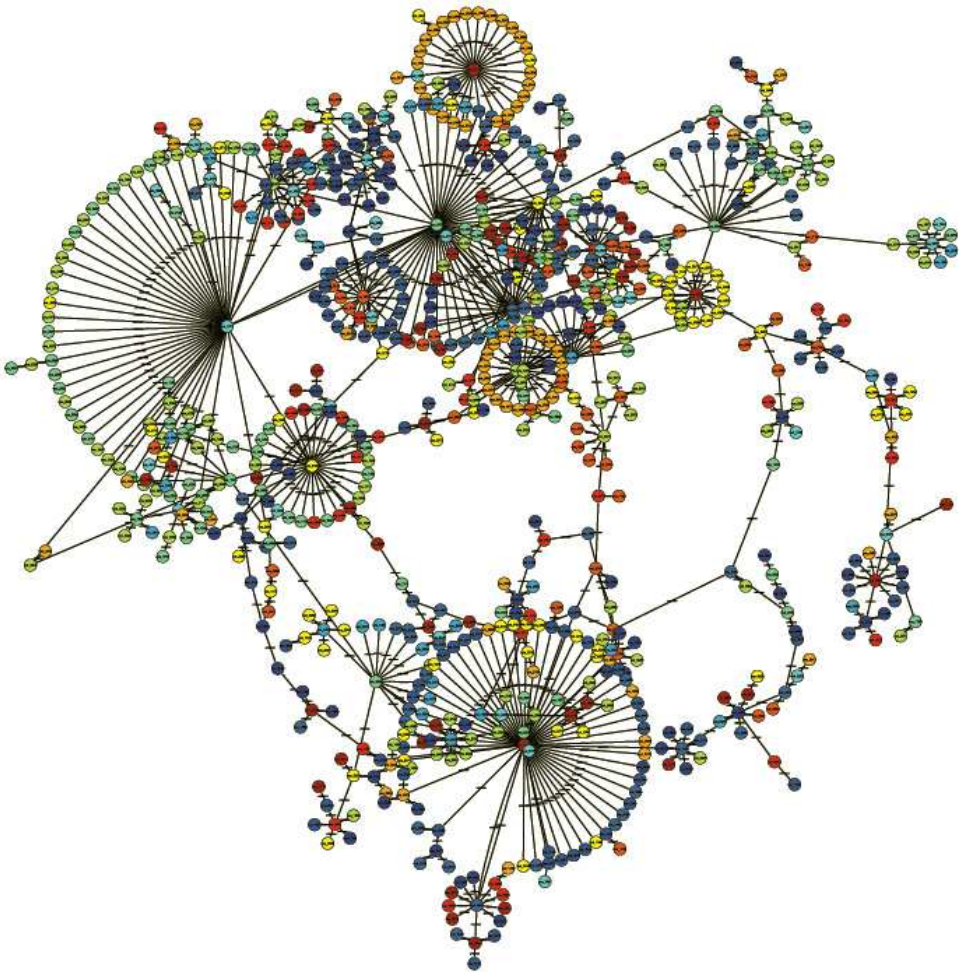


Figure 6.4: Gene regulatory network showing the possible links between diurnal genes.
Genes are colored based on their expression level at a given point of time. Various regulatory relationships structures, already characterized in other biological systems, are identified in the network.

to capture the transient behaviors in gene expressions resulted from changes in light input patterns.

6.5.2 Direct Regulation Vs Indirect Regulation

Cyanothece genome consists of 194 annotated regulatory-function genes representing about 4% of total genes. Out of them 28 genes were included in the network, representing 2.7% of genes included in the network. We find that 304 genes in the network can be associated with these 28 genes using either 6.1 or 6.2. We identify these links as likely direct regulation between genes. Other links might represent either indirect regulations or unclassified regulatory functions.

6.5.3 Core Network and Extended Network

Minimum network for [74] consists of 607 regulatory links while minimum network for [79] consists of 822 links. We see that some of the interactions have more influence in one condition compared to the other. As observed in [52], it suggests the existence of superimposed circadian signaling and diurnal signaling, where one type becomes significant under specific conditions.

There are 130 essential links in regulatory networks under two conditions. This number represents close to 10% of the combined network. We identify that these genes belong to a core gene network. The remaining links are possibly condition specific, indicating that they have a significant influence only during one experimental condition. These genes give rise to an extended network. In the core network, genes with known regulator-functions is present as regulators only in 5% of the times. In contrast, among extended network, this percentage rises to 26%. We believe that this is an indication of the dynamic role played by the regulatory genes helping cells to adapt to changing environmental conditions.

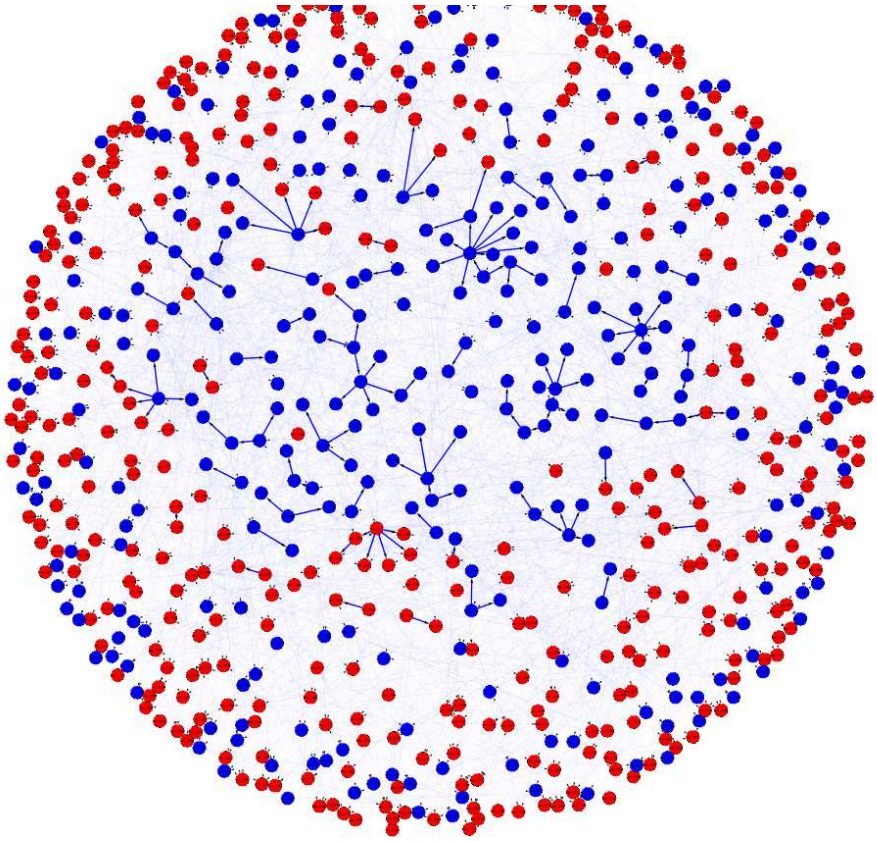


Figure 6.5: Consistent links between diurnal genes under different input conditions. Genes belonging to the core network are connected in minimum networks for both [74] and [79]. These genes are located in the center of network and are rich in circadian controlled genes. Blue: circadian controlled, Red: light responding

Furthermore in the core-network, 70% of the target genes belong to the circadian controlled group. Here 80% of the regulators came from the same group. However in the extended network, circadian controlled genes represent only 35% of the targets and regulators. The remaining genes are from light responding group. Figure 6.5 show the distribution of genes in the core network and the extended network.

We observe clear correspondence between number of links in the network and gene categories identified in the previous work [28]. In the combined network, 33% and 61% of the circadian controlled genes had just 1 and 2 regulators respectively. In contrast only 4% of light responding genes had a single regulator. Further 28% and

67% of light responding genes contained 2 and 3 regulators respectively. Among genes identified as having two dominant frequencies in the two conditions, 92% had 3 regulators in the final network.

6.5.4 Regulation of Possible Operons

Genes belonging to a single operon consist of a single regulatory region and are transcribed as a group. However depending on the respective positions in the operon their transcription levels show differences. A transcription control model should be flexible enough to assign genes belonging to possible operons to a single regulator, despite the changes in the transcription levels. As explained in Section 6.5.8, we treated those genes, located in the same DNA strand and have a separation of less than 100 base pairs between their Open Reading Frames (ORFs), as members of an operon. Among the genes in the network there were 275 such genes giving rise to 110 operons. We observe that genes in 43 operons can be associated with the same regulator. Expressions of genes from different groups are significantly different so that they are not associated with the same regulators.

6.5.5 Regulators of Different Biological Processes

Some of the regulators in the network are associated with specific biological processes. The significance of the dependence between the regulator and the biological process is measured using Fisher‘s exact test,[1]. In Figure 6.6, distributions of target genes for top regulators are shown. It is clear that many regulators are associated with only a few pathways. Except for the first 10 regulators, others are associated only with less

than five different pathways. Similarly most of the important biological processes are associated only with few regulators.

6.5.6 Phase Difference between Regulator-Target Pairs

One of the important features of the transcription control model proposed in this analysis is its ability to associate genes with possible phase differences. Moreover, using phase difference between a regulator and a target, it is possible to identify if the particular interaction is positive (inductive) or negative (repressive). Based on the final gene regulatory network, majority of the phase differences between regulator-target pairs are observed to be between 4–5h. Based on the value of the parameter β in 6.1 and 6.2, regulation relationships are identified as positive or negative. We observed that close to 45% of genes show negative regulation. This suggests that in a bacterial system, both inductive as well as repressive regulation takes place with similar proportions. This has previously being observed in E.*coli* also, where activator and repressor percentages are 48% and 52% respectively.

A close examination of expressions of genes classified as light responding shows that majority of them alter their regular oscillatory behavior only after some delay when they are switched to constant light conditions. This fact supports the time delay observed between the regulator and target genes in the model.

6.5.7 Network Motifs

Various regulatory relationships structures, already characterized in other biological systems are identified in the network. These structures include connections that

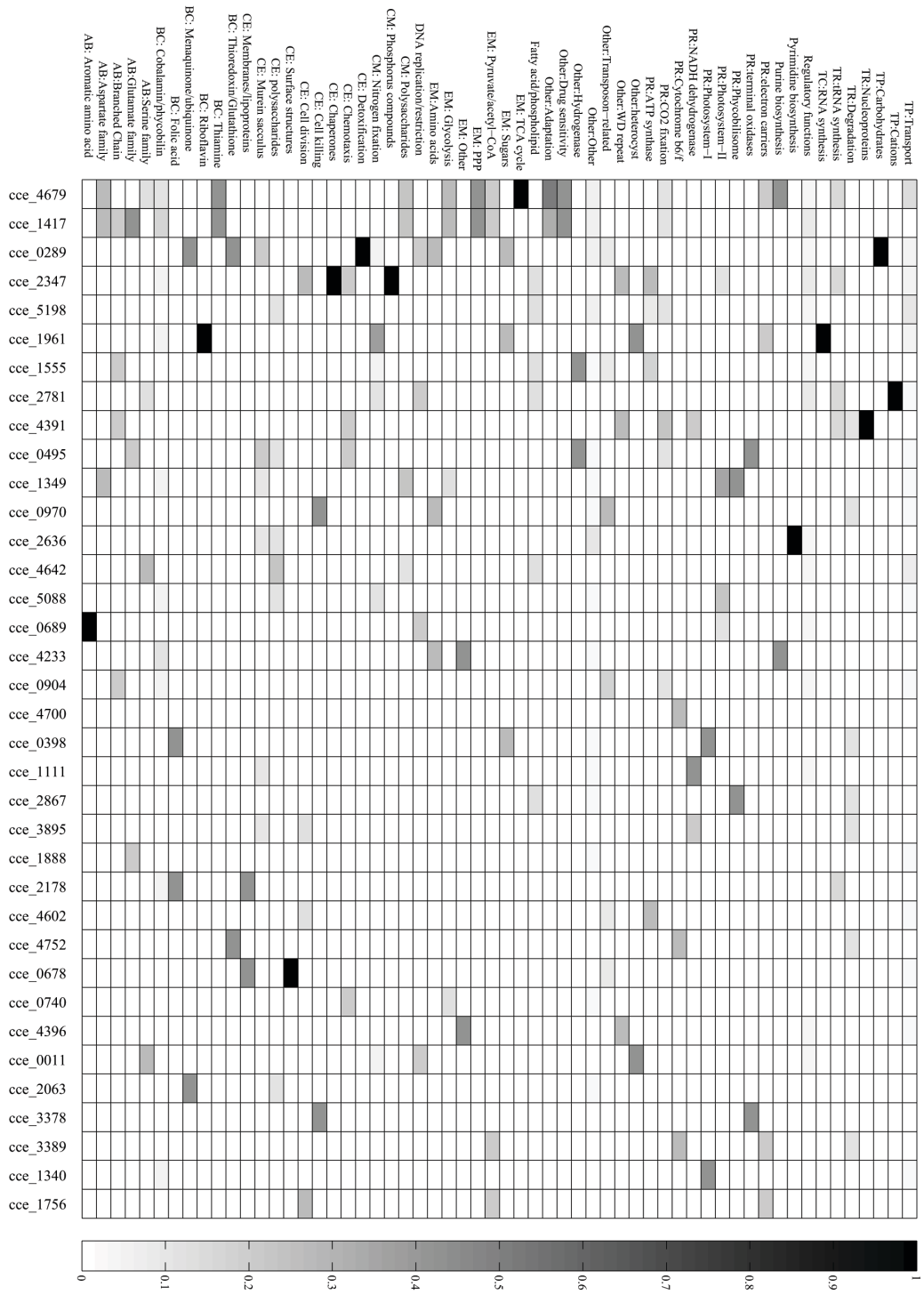
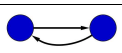
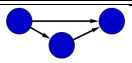
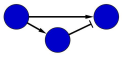
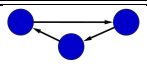
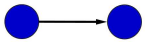
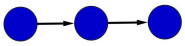


Figure 6.6: Top regulators and the fractions of genes from different processes associated with them. Except for the first 10 regulators, others are associated only with less than five different pathways. Similarly most of the important biological processes are associated with only a few regulators.

represent auto-regulation, coherent and incoherent feed forward loops, and single and multi input regulations. These connections can be verified by conducting follow-up experiments.

Table 6.1: Some of the regulatory modules identified within the gene regulatory network.

Type of Motif	Structure	Number of Occurrences
Auto Regulation		4
Coherent FFL		10
Incoherent FFL		9
Cyclic		1
Single Input		70
Multiple Input		> 300
Chain		70

Many regulatory structures already characterized in other biological networks could be found within the interaction network obtained for diurnal genes in *Cyanothece* sp. ATCC 51142.

6.5.8 Regulatory Region Motifs

If target genes associated with a given regulator are truly interacting, we expect them to share a common regulatory region motif. This idea can be used as a method of measuring the accuracy of the GRN. If we can find over-represented motif among the possible targets of a given regulator, it increases the chances of those regulatory relationships to be actual and direct.

In order to identify conserved regions in the upstream regions of the genes we use multiple sequence alignment program *Consensus* [34]. In bacterial systems, the upstream regions of genes are not well characterized. As a result we use following criteria to extract the relevant regions.

1. If two genes in the same strand are separated by less than 100 nucleotides, we consider them to be a part of an operon. Then we move forward in the strand until we have a wider separation between genes and consider the upstream regions corresponding to relevant gene, so obtained. We make sure that the upstream region of an operon is included only once in the calculation.

2. Criteria for minimum separation is applied only for genes in the same strand. If consecutive genes are on opposite strand we do not treat them as co-regulated.

3. Upstream region is limited to 500 base pairs forward or sequence up to end of the gene located ahead, which ever the shorter.

We search for the consensus sequence of the length 8 in the upstream regions of the relevant genes. Significance of the selected motifs are evaluated by comparing the proportion of genes containing the given motif among the possible targets and among the rest of the genes in the network.

Analysis of the upstream regions using Consensus results in several conserved regions. The significance of the obtained motifs to be non-random is calculated as p-values. Additionally, we calculate the ratio of observing the motifs among the target genes and compare that to all the remaining genes. There are many motifs for which this ratio exceeded 20. Using these two criteria we are able to identify several highly specific probable binding site motifs.

Table 6.2 lists some of the highest ranked motifs. Figures of conserved motifs are generated using *WebLogo3* [16].

Table 6.2: Selected regulator genes and over-represented upstream region motifs identified within their targets.

Gene	Function	Motifs	P-Value	Ratio
cce_1349	Other categories		6.10E-09	96.8
cce_3378	Regulatory		5.70E-15	54.3
cce_2124	Branched chain		4.09E-16	51.1
cce_2540	Regulatory		1.23E-15	51.1
cce_0206	Other categories		3.54E-13	41.4
cce_0398	not available		5.50E-23	40.3
cce_3206	not available		4.40E-17	35.9
cce_0970	Regulatory		2.57E-19	29.1
cce_1083	not available		5.31E-18	27.8
cce_4602	not available		1.86E-23	26.4
cce_1555	not available		8.04E-23	23.0
cce_1978	not available		8.57E-13	20.7

Consensus algorithm detected several highly conserved regulatory sequences in the upstream of the target genes, associated with different regulators. Significant values for the $ratio = \frac{\%\ of\ times\ motif\ was\ present\ in\ target\ genes}{\%\ of\ times\ motif\ was\ present\ in\ rest\ of\ the\ genes}$ as well as p-values computed by *Consensus* suggest that probability of having these sequences by chance is highly unlikely.

As observed in many experimentally verified transcription factor binding sites, we see some conserved nucleotides in the vicinity of the predicted motifs. Figure 6.7 shows alignments of upstream regions of few selected target genes. Presence of conserved bases in the vicinity of the main motif increases the chances of these motifs being true binding sites for transcription factors.

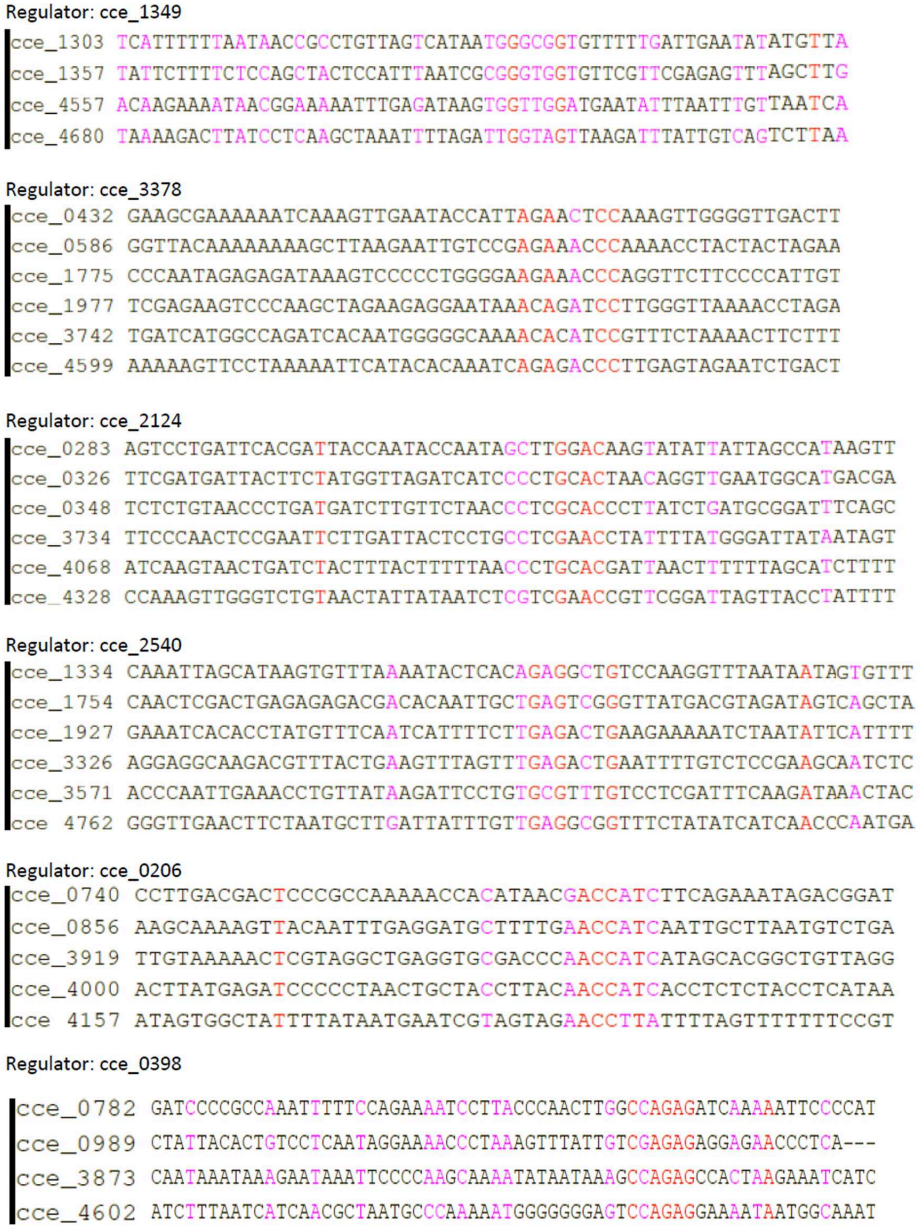


Figure 6.7: Upstream regions of the co-regulated genes aligned using *Consensus*. Several conserved nucleotides in the vicinity of the main regulatory motif are observed for these genes. These types of conserved nucleotides in the vicinity have been observed in many experimentally verified binding sites also.

6.6 Conclusions

In this work we propose using a biologically realistic dynamical systems based on differential equations to model interactions between diurnal genes; both circadian controlled and light responding. We describe the specific simplification made to a general model to suit the data set being analyzed. We discuss how to select the appropriate parameters and function formats based on the types of genes being modeled and show that this model is sufficient to explain interactions between diurnal genes under regular light/dark cycles and transient light conditions. We discuss how one can obtain a global gene interaction network based on the proposed model and how it can be improved by utilizing the already existing biological insight. Various features in the resultant network are discussed in details. We study the changes in the network under different light conditions. The resultant network is shown to be rich, with various interaction patterns already identified in other biological systems. Within the target gene groups picked by the model, we identify many regulatory region motifs that are highly significant, which suggest that many interactions predicted by the model are likely to be actually present.

The model proposed here is clearly stable, which is an essential feature of any biological system. Interactions identified using the model are directional and the targets and the regulators are clearly defined. Since model allows phase shifts between inputs and outputs, it can accommodate the delay between transcription of a regulator and action of the corresponding protein on its target, occur after translation and post translational modifications. These types of relationships are not modeled by traditional correlation based methods.

Majority of the transcriptional relationships inferred by the model are shown to be consistent under parameter modifications. It implies that the relationships we detect are resilient to small variations in the signals and parameters. This is an important feature, any realistic biological model should posses.

The model is able to infer interactions for more than 80% of the genes considered to be diurnal and used for the analysis. We have shown that the network for [74], where cells were under regular dark/light cycles, has considerably less number of interactions compared to that for the [79]. This is due to various alterations of gene expressions occurred under constant light conditions. We make hypothesis that this added complexity of the network indicate additional regulatory relationships that become visible under altered environmental conditions. We have identified consistent links between two conditions and found that majority of the genes involved in those links are previously categorized as circadian controlled.

Using the model we are able to associate about 30% of the genes that are already known to be involved in regulatory roles. We have also identified about 100 possible operons based on gene locations in the genome and we show that genes in 43 of them could be associated with single regulators. Model also suggested that many of the important biological processes are primarily controlled by a relatively small number of regulators. We see that there is about 4–5h time lag between regulator and target genes. This is in good agreement with the delay observed in gene expression data, once the incident light is switched from oscillatory to constant condition.

The final network is rich of many known network motifs. In addition to feed forward loops, a variety of other network structures such as auto-regulations, cyclic regulations, single and multi input genes and chain type of regulations are observed. We have identified many hierarchical regulatory relationships as well.

From the upstream regions of target gene groups, we are able to detect many conserved binding site motifs. We have shown that many of these motifs are very specific to selected groups of genes. Also we are able to detect several conserved nucleotides in the vicinity of the identified motifs. These observations increase the possibility of these regions being actual transcription factor binding sites.

Finally we would like to acknowledge that the proposed network is not complete. In this work our focus has been limited only to the diurnally regulated genes. We show that the network under regular day/night conditions identified as the core network requires extensions to capture the gene expressions under modified light conditions. It is quite possible that more interactions would become visible if system is perturbed by other conditions. With the availability of such data, the model might need to be refined. Also there was no explicit input corresponds to the external light. To capture the effect of light, it might required to incorporate these input channels to the model.

Chapter 7

Modeling Diurnal Behaviors using Phase Oscillators

7.1 Phase modeling : Modeling Biological Processes as an Oscillatory Network

Phase oscillators were originally used for modeling oscillatory systems having large number of weakly interacting oscillators ([88],[77]). Phase oscillator models are appropriate for modeling circadian rhythms, as they directly model the phase dynamics of the system. Phase dynamics is the most important feature in understanding circadian rhythms. A phase model is used in [4] to represent the circadian clock of cyanobacteria and to establish that the interaction between cyanobacteria cells are negligible. In [84], a coupled phase oscillator network was proposed for modeling circadian-controlled genes in cyanobacteria.

7.1.1 Aims

As discussed in Chapter 5 as well as in Chapter 6, many of the diurnally regulated genes from a single biological processes tend to peak their activities at a specific time of the day. This suggests that we can focus on group behaviors of genes instead of looking at individual genes. In Chapter 4, we discussed several advantages of moving from individual genes to groups of genes that are co-expressed and belong to a same biological process.

In this chapter we develop a simple phase oscillatory network to capture the salient features in the diurnally regulated genes. The proposed oscillator network requires to reproduce the actual gene behaviors observed under different light input patterns. It also needs to be resilient to noise, which is an essential feature in biological systems. We use the proposed model to understand synchronization between different processes, modulation of internal clock by external light inputs, and changes expected in circadian clock and other peripheral processes under different light patterns. We relate some of the simulation results with already available biological knowledge.

7.2 Oscillator Network

The coupled oscillator model proposed here consists of a structure shown in Figure 7.1. In Chapter 5, diurnally regulated genes are classified as circadian controlled genes (CCGs) and light responding genes (LRGs). Accordingly network is modeled to have two subnetworks representing these two categories. Each subnetwork consists of a center oscillator and six peripheral oscillators. Two center oscillators correspond to the circadian oscillator and the light sensor. Coupling between the light sensor and

the circadian oscillator represents entrainment of circadian clock by external light input.

Main gene-behaviors in each sub category, CCGs and LRGs, are represented using six peripheral oscillators. The six-oscillator networks are selected due to two observations made in previous analysis in [28], namely:

1. Distribution of phases of genes belonging to well clustered biological processes are mostly localized within a 4h period;

2. The gene regulatory network, generated using a linear dynamical model, indicated that, for the majority of the genes in the network, the phase difference between the target and regulator was 4h. We can capture this relationship using 6 oscillators with approximately $\pi/3$ phase difference.

Six ring oscillators corresponding to LRGs are connected to the light sensor while those corresponding to CCGs are connected to the circadian oscillator. These connections represent the reference phase provided by respective central oscillators to their peripheral oscillators. Between peripheral oscillators, unidirectional interactions are assumed, representing a regulator-target relationship between genes from different processes.

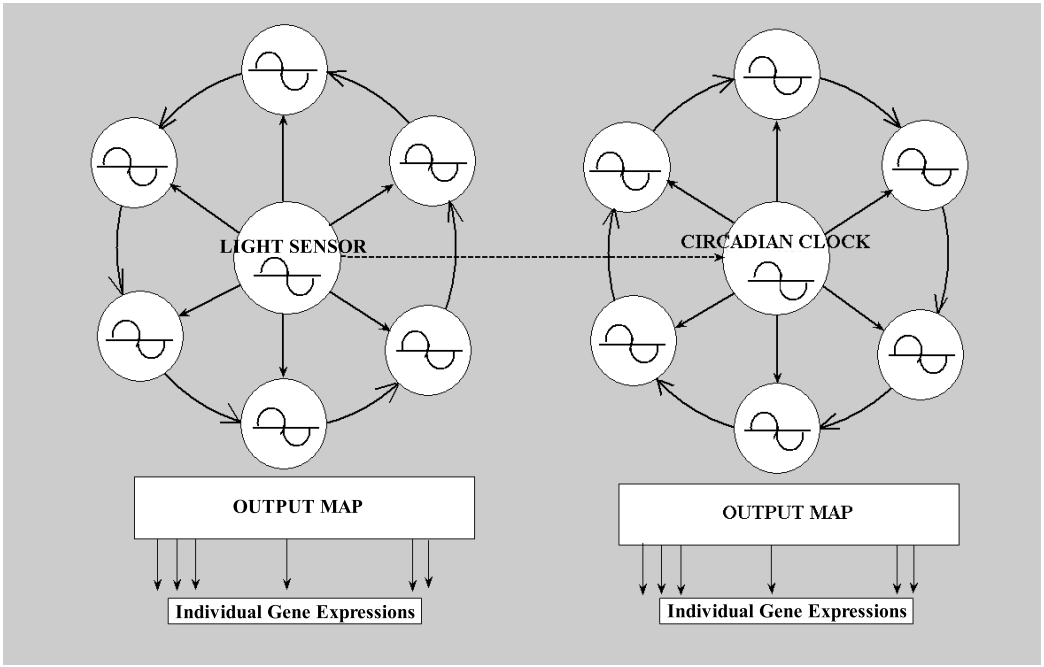


Figure 7.1: Coupled oscillator model representing 24h LRGs and CCGs. Central oscillators, correspond to light sensor and the circadian oscillator, provide reference phases for their ring oscillators representing 24h LRGs and CCGs respectively. Individual gene expressions are obtained as a linear map of the oscillator outputs.

7.3 Phase Oscillator Model

Each oscillator in the network is modeled as a phase oscillator. Due to the lack of knowledge on the light sensor and the output channel of circadian clock in cyanobacteria, central oscillators are assumed to be harmonic oscillators and modeled as,

$$\dot{\phi_{lc}} = \omega_{lc_0}, \tag{7.1}$$

$$\dot{\phi_{cc}} = \omega_{cc_0} + \varepsilon_1 \sin(\phi_{lc} - \phi_{cc}), \tag{7.2}$$

where ϕ_{lc} and ϕ_{cc} are phases of light sensor and the circadian clock respectively. The ω_{lc_0} and ω_{cc_0} are their Eigen frequencies and set to $2\pi/24$, corresponding to a 24h oscillatory period.

Oscillators in the rings are non-harmonic oscillators and are modeled to reproduce actual gene expressions they represent. Their behaviors are modeled as,

$$\begin{aligned}\dot{\phi_{l_i}} = \omega_{l_i} &+ \sum_{k=1}^{N} \varepsilon_{l_i k} \sin(k\phi_{l_i} + \delta_{l_i k}) \\ &+ \varepsilon_2 \sin(\phi_{lc} - \phi_{l_i} - \xi_{l_i}) + \varepsilon_3 \sin(\phi_{ri} - \phi_{l_i} - \upsilon_{l_i}),\end{aligned} \tag{7.3}$$

$$\begin{aligned}\dot{\phi_{c_j}} = \omega_{c_j} &+ \sum_{k=1}^{N} \varepsilon_{c_j k} \sin\left(k\phi_{c_j} + \delta_{c_j k}\right) \\ &+ \varepsilon_4 \sin\left(\phi_{cc} - \phi_{c_j} - \xi_{c_j}\right) + \varepsilon_5 \sin(\phi_{rj} - \phi_{c_k} - \upsilon_{c_k}),\end{aligned} \tag{7.4}$$

where ϕ_{li} and ϕ_{ri} are phases of the i^{th} oscillator for LRGs and the oscillator preceding i^{th} oscillator respectively. Analogously ϕ_{cj}, ϕ_{cc} and ϕ_{rj} correspond to phases of the j^{th} oscillator for CCGs, the circadian clock and the ring oscillator preceding j^{th} oscillator respectively.

7.3.1 Determining Coupling Strengths

The network consists of four types of coupling between oscillators, namely the light sensor–and–circadian clock (ε_1), the light sensor–and–ring oscillator (ε_2), the circadian clock–and–ring oscillator (ε_4), and the ring oscillator–and–ring oscillator ($\varepsilon_3, \varepsilon_5$). Values of these coupling coefficients were determined considering several features that the model needs to produce, including:

1. Faster Entrainment: The cyanobacterium circadian clock is capable of being rapidly entrained/phase reset by the external light ([31]). In order to obtain a faster entrainment, we would like to have a strong coupling strength between the light sensor and the circadian clock. However, since the circadian clock should be able to maintain its oscillations under changing light inputs, we need to ensure $\dot{\phi_{cc}} > 0$ for any phase differences between the light sensor and the circadian clock. Considering these two factors we picked $\varepsilon_1 = 0.1$

2. Cessation of process oscillations: Diurnal biological processes, responding to light pattern, stop their oscillations under constant light conditions. These changes in behavior are noticeable soon after the change in light input pattern, within the first few hours, as observed in [28]. In addition, the circadian clock mutants show changes in oscillation periods and arrhythmic behaviors their biological processes ([42]). In order to achieve these behaviors we pick $\varepsilon_2 = 0.3$ and $\varepsilon_4 = 0.3$.

3. Phase relationship between biological processes: Though clock plays an important role in coordinating other biological processes, regulator-target interactions between genes are also a key determinant on transcriptome levels of a cell. These interactions are taken into account by the coupling between ring oscillators. We picked a relatively weak coupling strengths for these connections and set $\varepsilon_3 = 0.05$ and $\varepsilon_5 = 0.05$.

7.3.2 Parameter Identification

Each of the oscillators in the rings is modeled to capture the average expression of genes it represents. For this purpose we group together genes having a close phase relationship and their mean expressions are calculated. In order to have the same contribution from each gene towards the mean, the original expressions are scaled and shifted. Figure 7.2 shows the normalized expressions for one groups of genes and their mean expression.

Once the mean curve is obtained, it is concatenated several times to get an expression for multiple cycles. The resulting curve is smoothed using cubic interpolation to remove discontinuities. The phase is defined as the angle of a rotating vector, whose projection on the real axis would give the actual mean expression. The phase curve is also smoothed using zeroth order Savitzky-Golay FIR filter [73] with a frame size of 41, since any sudden changes in the slope would produce jumps in the phase derivative. The phase derivative is calculated using two point approximation. For all oscillations, these calculations are done using the gene expressions obtained from the first experiment.

Optimal values for parameters $\omega_{l_i}, \omega_{c_j}, \varepsilon_{l_i k}, \varepsilon_{c_j k}, \delta_{l_i k}$ and $\delta_{c_j k}$ in (7.3) and (7.4) are found by the least square optimization method minimizing the errors, given by

$$\begin{aligned} E_{li} = \| \dot{\phi_{l_i}} - \varepsilon_2 \sin\left(\phi_{lc} - \phi_{l_i} - \xi_{l_i}\right) - \varepsilon_3 \sin\left(\phi_{ri} - \phi_{l_i} - \upsilon_{l_i}\right) \\ -\omega_{l_i} - \sum_{k=1}^{N} \varepsilon_{l_i k} \sin\left(k\phi_{l_i} + \delta_{l_i k}\right) \| \end{aligned} \tag{7.5}$$

and

$$\begin{aligned} E_{cj} = \| \dot{\phi_{c_j}} - \varepsilon_4 \sin\left(\phi_{cc} - \phi_{c_j} - \xi_{c_j}\right) - \varepsilon_5 \sin\left(\phi_{rj} - \phi_{c_j} - \upsilon_{c_j}\right) \\ -\omega_{c_j} - \sum_{k=1}^{N} \varepsilon_{c_j k} \sin\left(k\phi_{c_j} + \delta_{c_j k}\right) \| . \end{aligned} \tag{7.6}$$

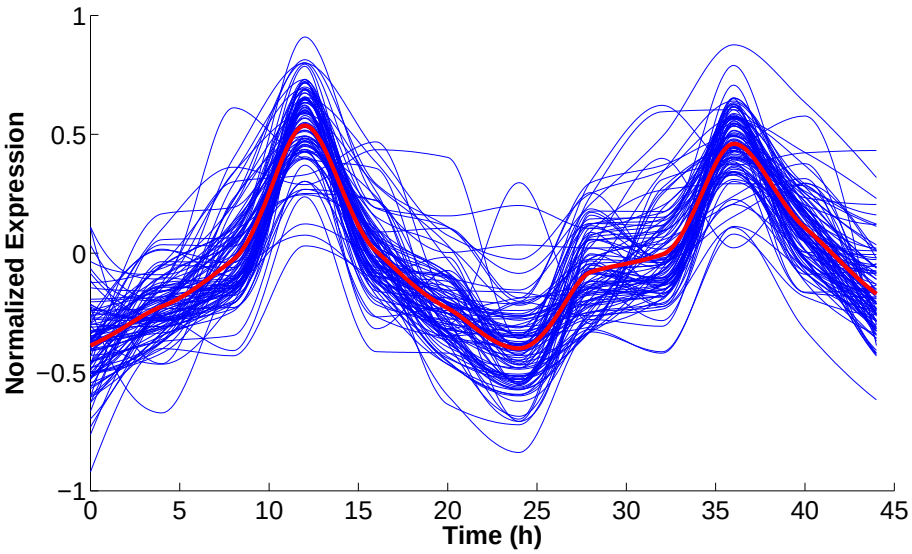


Figure 7.2: Normalized expressions of genes with close phase relationship and their mean expression. Individual oscillators were designed to reproduce these mean expressions.

We picked $N = 5$ to get a good reconstruction. Figure 7.3 shows the approximation of the phase derivative for one of the oscillators. It is clear that with $N = 5$, phase model can approximate the phase derivatives with a good accuracy. With this choice, the error of reconstructing the phase derivative is $\leq 8\%$ for all the oscillations in the system.

Parameters ξ_{xi} and υ_{xi} correspond to average phase differences between the i^{th} oscillator compared to the center oscillator and the $i - 1^{th}$ oscillator respectively.

In order to get oscillator outputs under constant light conditions, we set the Eigen frequency of light sensor to zero during the subjective dark regime (last 12h period in the second experiment). This makes phase of the light sensor a constant during this period and the other oscillators show a transient behavior due to this change. Figure 7.4 shows the outputs of oscillators corresponding to light responding genes, under transient light input pattern in the second experiment.

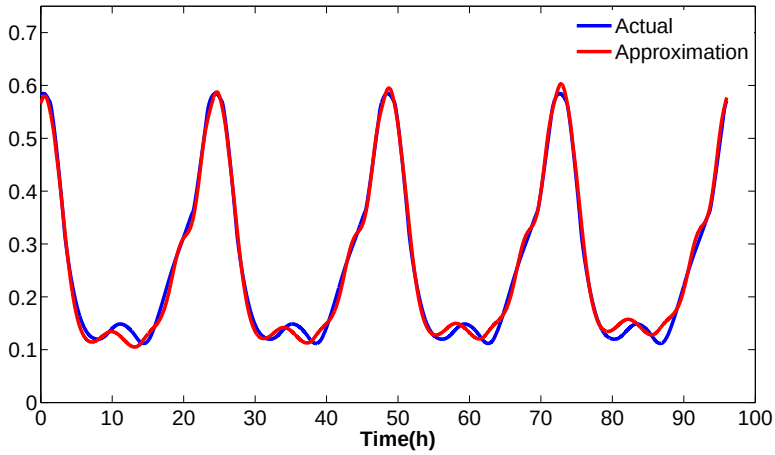


Figure 7.3: Approximation of a phase derivative using the phase model. The proposed oscillator model is sufficient to get a good reconstruction of the actual phase dynamics.

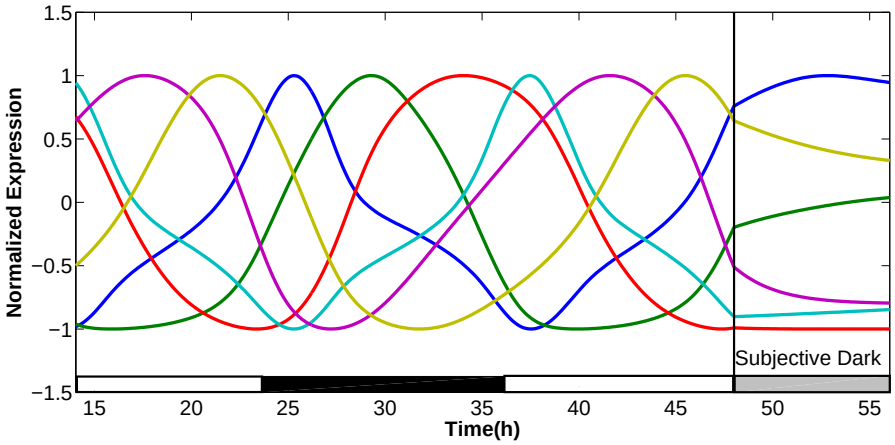


Figure 7.4: Output of six ring oscillators corresponding to LRGs, simulated under transient light conditions. During last 12h, the light sensor is kept at constant phase. Under this condition, phases of ring oscillators reached steady states within few hours.

7.4 Use of Oscillator Model to Study Gene Behaviors

The oscillator model presented here can be used for various purposes. It can be used as a method of filtering and categorizing genes into groups. Oscillator outputs can be treated

as a set of basis functions for this data set, which are better representatives of the actual gene expressions than sine/cosine functions. In addition, the model can be used to simulate gene behavior under various light conditions. It is also possible to study the effect of the oscillator output with changes in parameter values. Predictions from these simulations can be verified using experiments.

7.4.1 Categorization of Genes using Oscillator Model

The actual gene expressions are projected onto oscillator outputs in order to filter those genes, which can be explained using the model. Each gene expression is explained using two closest oscillator outputs in terms of their phases. Goodness of fit was measured using correlation between the approximation and the original expression.

A gene is picked only if it is well approximated using two oscillator outputs. We selected a correlation threshold of 0.8. In addition to a good approximation, we also require that the gene is explained by the same oscillators in both experiments. This ensures the extraction of genes with consistent behavior in two experiments. Figure 7.5 shows the approximation of an actual gene expression using outputs of two closest oscillators.

Based on the reconstruction, 501 and 651 genes are approximated well using oscillator outputs corresponding to circadian controlled and light responding processes respectively. Among these, there were 345 genes which could be classified as both CCG and LRG. We assign them to the group, which results in lower error in the approximation.

Among 501 genes which were associated with circadian controlled oscillators, there are 387 genes which were categorized as CCGs in the previous chapter. However among 651 genes associated with light responding oscillators, only 218 are categorized as LRGs previously.

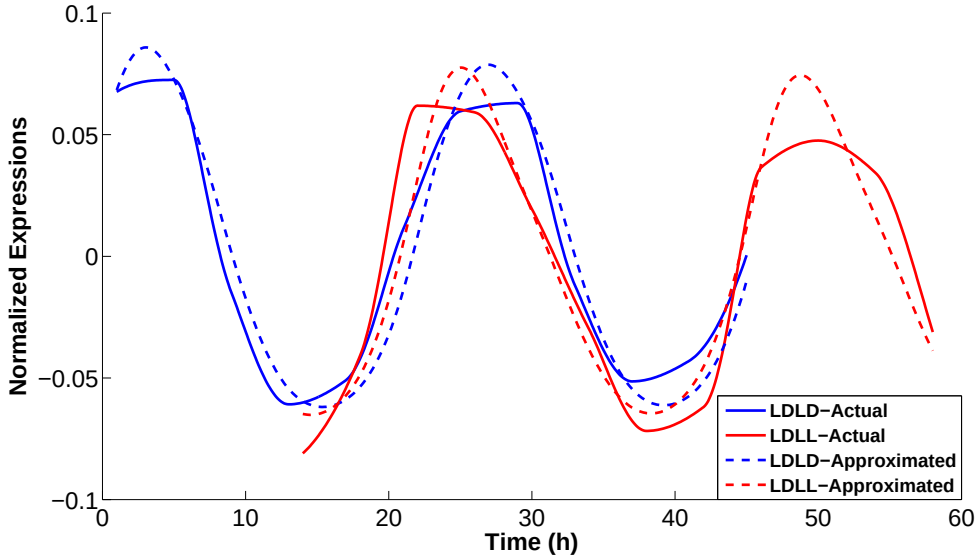


Figure 7.5: Reconstruction of an gene expression using two oscillator outputs. Many diurnal gene expressions could be reconstructed as a linear map of two neighboring oscillators.

7.4.2 Clustering Genes based on the Projections

Those well-approximated genes are clustered based on the oscillators used to represent them. Figure 7.6 contains the distribution of genes for some of the well clustered biological processes. One of the important observations made here is tight co-regulation of genes that belong to processes, which become active at the onset of light or dark phases. Also, compared to the middle of the night or day, more number of genes become active during these periods. This clearly shows the preparation of cells to adapt to changing light conditions.

7.5 Simulation Results

The oscillator network can be simulated under different conditions to make predictions on behaviors of genes. These predictions can be verified by further experiments. Here we discuss some of the simulation results. We specifically focus on the effects on circadian clock

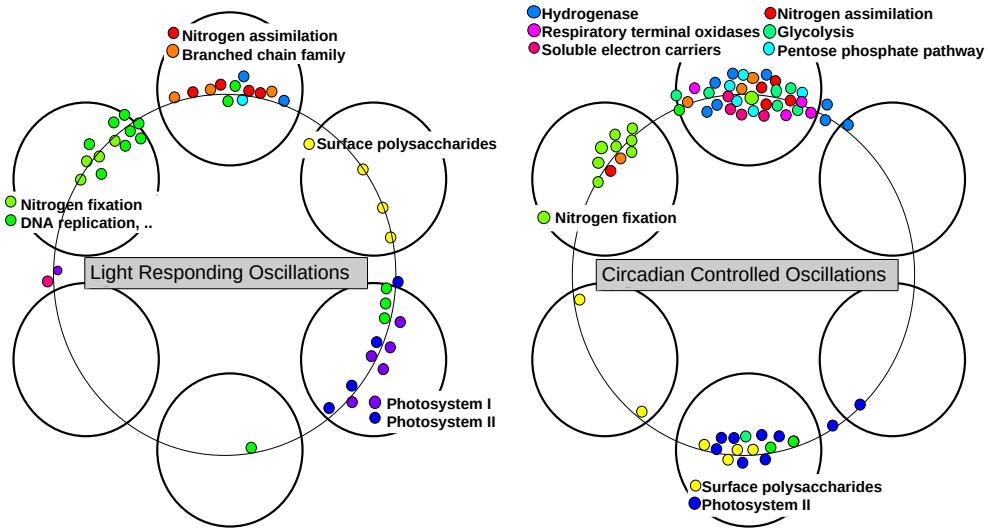


Figure 7.6: Some of the processes which can be directly associated with the individual oscillators in the network.
These processes include many vital processes such as nitrogen fixation, photosynthesis, glycolysis and DNA replication needed for the survival of the cells.

and its associated processes by changes in light inputs. We relate some of the simulation results with actual observations in the literature.

7.5.1 Different Network Topologies

The oscillator network is simulated after removing the clock–and–process coupling, the process–and–process coupling and both these couplings, to study the effect of these changes on phases of oscillators. For this part of simulation, we kept the strengths of both types of coupling at 0.05, so that the role of each type of coupling can directly be compared. The phase differences between two of the oscillator-outputs under different coupling configurations are shown in Figure 7.7.

Based on the simulations, removal of the coupling between the center oscillator and peripheral oscillators gives rise to a larger shift in the phase relations, compared to removal of the coupling between peripheral oscillators. This agrees with the common notion that the

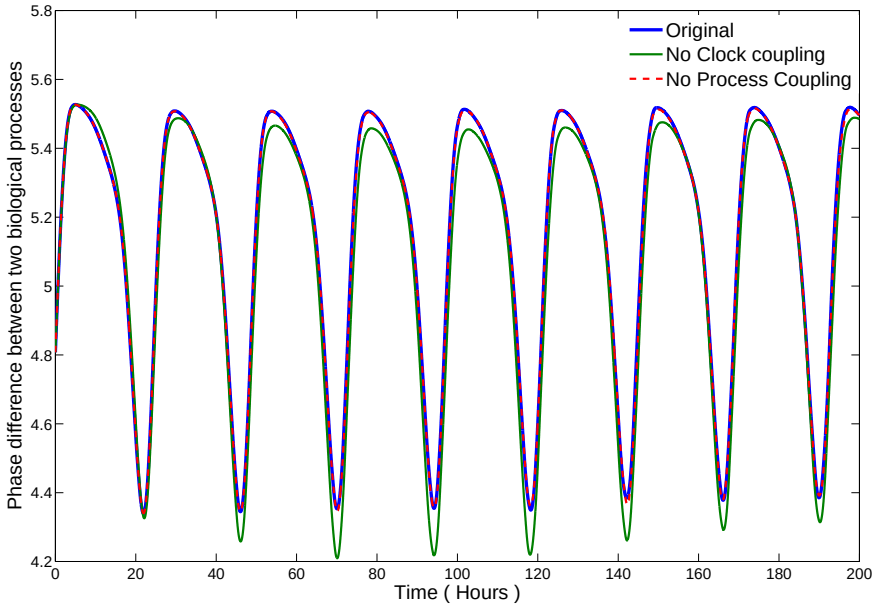


Figure 7.7: Effects on phases of circadian controlled processes under different coupling topologies, measured as phase difference between two process. The effect of removing the coupling between the processes is negligible, compared to the effect of removing couplings between the clock and the processes. This simulation result agree with the experimental observations that show the vital role of circadian clock in maintaining accurate phase relationships between different biological processes.

circadian clock might have more significant role in maintaining the exact phase relationships between biological processes.

We also studied the transient behavior of different network topologies, once they were perturbed by shifting the phase of one of the processes (oscillator) by π compared to its original phase. The perturbed oscillator returned to its original phase very quickly, when the coupling with the circadian clock was present and the other processes had little effect from the disturbance. However when the clock links were not present, the perturbed oscillator settled to a different phase, compared to its original. All the other processes were shifted in their phases as a result of the perturbation. Also under this configuration, a much longer period was required to regain the stationary phase behavior. Figure 7.8 shows the

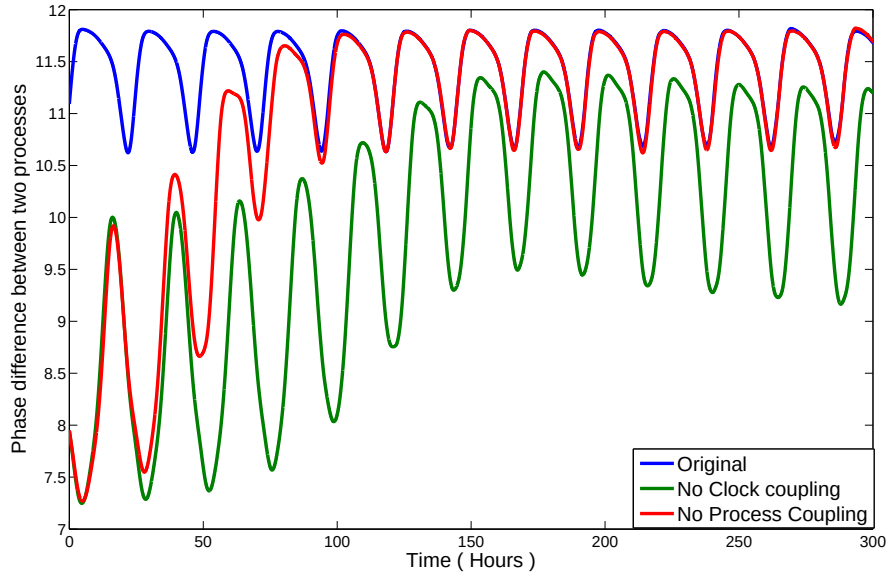


Figure 7.8: Phase differed between two processes, resulting due to a phase shift of one, under different network topologies.
With the connections to the clock, system recovers from the perturbation very quickly, with no significant effect on the other processes.

simulation results. This again suggests the vital role, the circadian clock has in maintaining robust dynamics of the other biological processes.

These behaviors support the observation that the circadian clock is not essential for the survival of the cells but increases the competence of the cells by improving the coordination between different biological processes [43]. This has well been established for other organisms also, which include plants and humans.

7.5.2 Effects of Providing Constant Light Input

In circadian control literature, it is known that the free running period of the clock is not exactly equal to 24h. Usually it can be slightly shorter or longer. Based on [32], *S. elongatus* has a free running period of around 25h. In order to see whether the model is

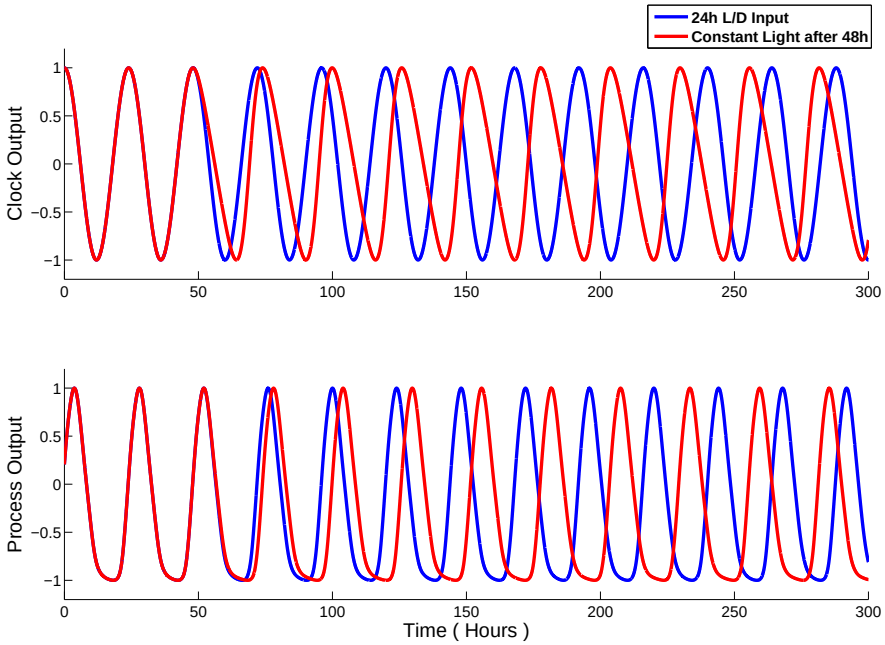


Figure 7.9: Circadian clock and one of the ring oscillator outputs under periodic and constant light input conditions.
Effect of constant light is reflected in clock output immediately, but only observed in the processes outputs with some delay.

capable of generating such a behavior, a simulation is run under constant light conditions. This is achieved by keeping the phase of light oscillator constant. The natural period of the circadian clock is kept at 24h. Figure 7.9 shows the output of the circadian clock and one of the ring oscillators for periodic and constant light inputs. Figure 7.10 shows the corresponding periods of oscillations. As a result of the coupling with the light oscillator, the circadian clock and the ring oscillator show a oscillatory period of around 26h. The free running period varied with the coupling strength. One other observation is that, while the circadian clock oscillations are immediately affected by the changes in light input, the processes under circadian-control are affected with some time delay. This is clear from Figure 7.9.

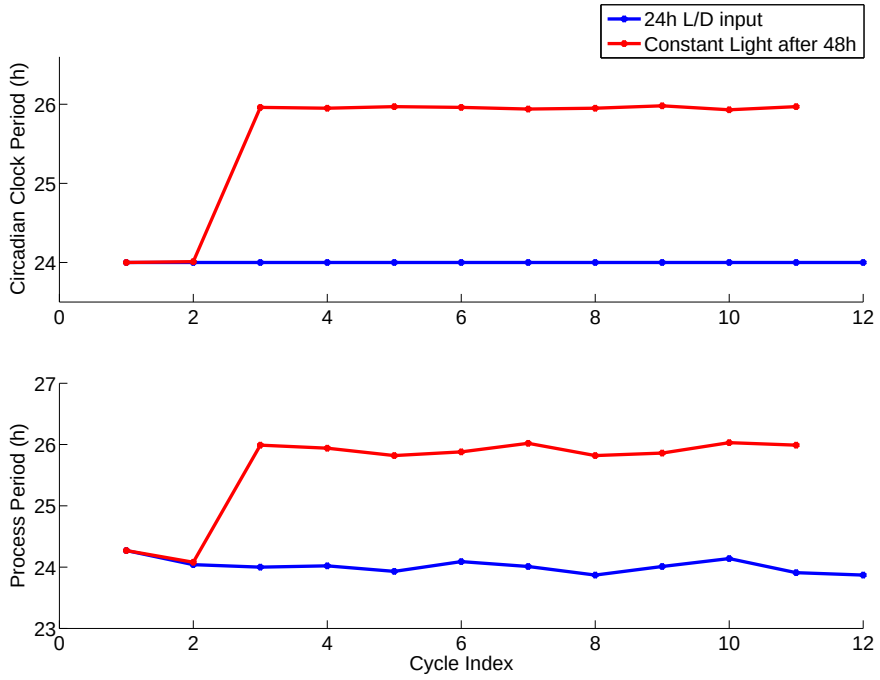


Figure 7.10: Periods of Oscillators under 24h periodic and constant light input conditions.
Free running period of the oscillators shifted to 26h under constant input conditions. This is in agreement with the experimental observations.

7.5.3 Adaptation to Light Patterns with Different Periods

The ability of the circadian clock to follow the different periods in the light input depends on the strength of the coupling between the circadian clock and the light sensor. Figure 7.11 shows the period of oscillations of the circadian oscillator, under light cycles with different periods, for two different coupling strengths. Clearly the circadian clock follows the light period in a wider range with an increased coupling strength between two oscillators. This observation can be used to determine the actual strength of coupling between the light sensor and the circadian clock.

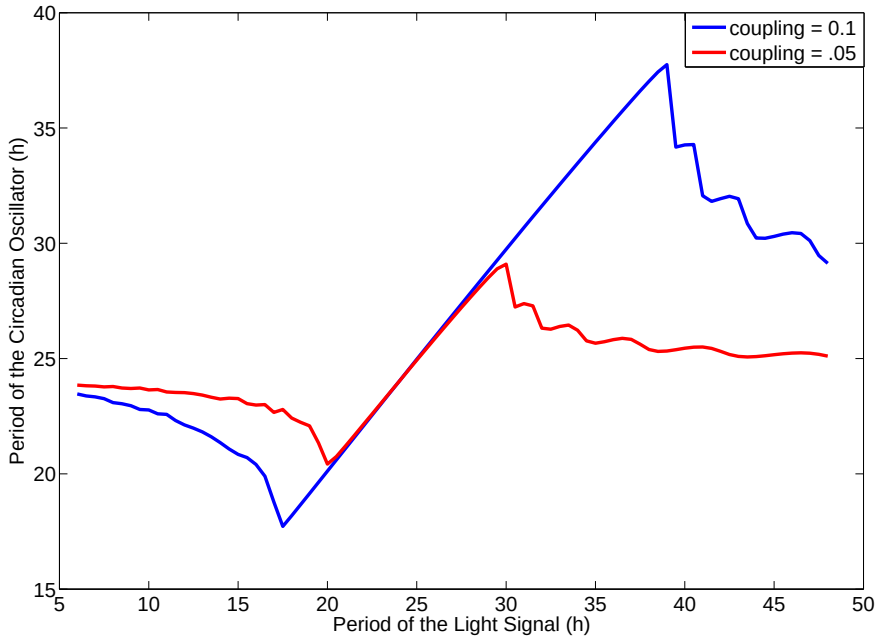


Figure 7.11: Adaptation of circadian clock to different periods of light input. Period of the circadian clock oscillations can be entrained by the external input. The range of entrainment depends on the coupling strength between the light sensor and circadian clock mechanism.

7.5.4 Effect of the Noise

Most of the biological systems are robust to the noise inherent to them. As a result, any realistic model should be robust to fluctuations caused by noise. In order to test the resilience of the current model to the external noise, we add a noise component to the original model. We assume, that the effect of noise changes the Eigen frequency of the oscillators. Therefore we replaced the ω terms with,

$$\omega_x = \omega_{x0}(1 + N_x), \tag{7.7}$$

where N_x represents the White Gaussian noise. We limited the noise signal to be between -0.1 and 0.1 representing 10% deviation of oscillator frequencies from their normal values. This is sufficient to capture the range of frequencies usually observed in the cyanobacteria

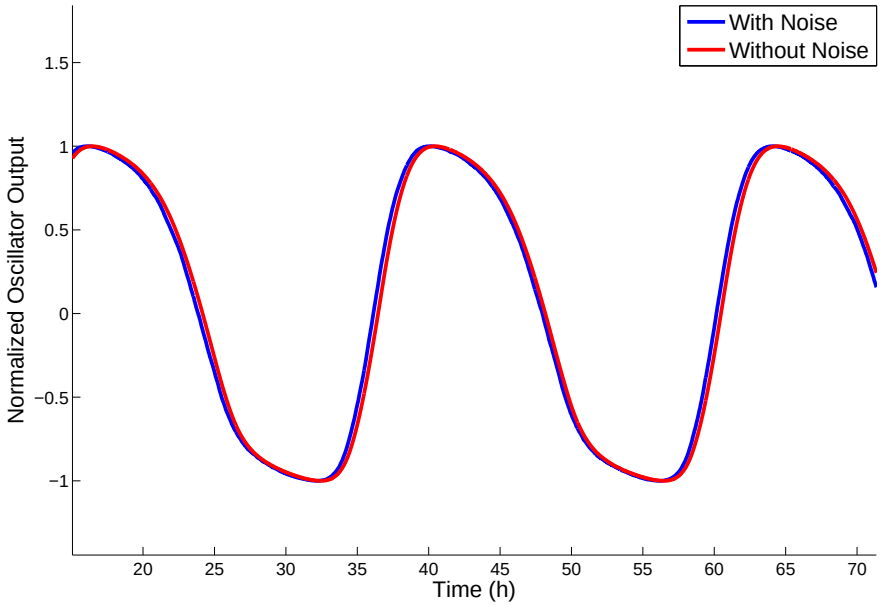


Figure 7.12: Output of a ring oscillator with and without external noise. Effect of noise was negligible on the output of the oscillator. The robustness to the noise is an essential feature of the most of the biological systems.

circadian clock. Noise was added to all oscillators except the light sensor. Equations were solved using Eular method. We observed that the system is extremely robust and the effect of noise on the ring oscillators is negligible. Figure 7.12 shows the simulation results for one of the ring oscillators with and without noise.

7.6 Conclusions and Discussion

In this chapter, we propose a simple coupled oscillator network to model the gene behaviors under different light input patterns. We show that the model proposed here is capable of capturing important dynamics of the gene behaviors. The oscillator outputs are used to classify genes into different groups based on the phases of their expressions. We show that some of the biological processes could directly be mapped to the relevant oscillators. Based

on the simulation results, we argue that the circadian clock is more important for maintaining proper phase relationships between biological processes, compared to the interactions between individual processes. We also discover that there is a noticeable time delay involved in the propagation of changes in light patterns to the circadian-controlled processes. Our model is able to reproduce some of the experimentally observed gene behaviors under altered light conditions. These included the changes in the natural period of circadian clock under constant light. In addition the model was shown to be resilient to noise, an essential feature in most of the biological systems.

It is shown that some behaviors of the network are mainly determined by the coupling strengths between oscillators. The current oscillator model can be improved by determining these coupling strengths using biological experiments.

Chapter 8

Differences and Similarities of Cell Behaviors Observed from Transcriptomics and Proteomics Measurements

Transcriptomics studies only measure steady state expression levels of mRNA concentrations inside a cell. Though transcriptomics data provides vital information on responses of cells to different experimental conditions, these measurements are insufficient to achieve complete understanding on complex regulatory mechanisms in a living cell. It is well known that mRNAs undergo several regulatory controls before corresponding proteins are synthesized [29]. Also steady state protein levels are dependant on the corresponding degradation rates.

Proteomics measure steady state protein levels in a cell. Combination of transcriptomics and proteomics studies reveals differences in mRNA and protein levels and allows identification of possible control steps in determining their levels. In additions, such data sets are useful in improving accuracy of the gene regulatory networks derived using transcription data only.

8.0.1 Aims

We analyze two different proteomics data sets on *Synechocystis* sp. PCC 6803 and *Cyanothece* sp. ATCC 51142. We compare these data with analogous transcriptomics data sets to identify similarities and differences between transcriptional and translational levels.

8.1 Identification of Differentially Regulated Genes using Proteomics Data

Although the statistical methods discussed in Chapter 3 are applicable to proteomics data also, due to limited number of replicates available, the assumptions made in those methods do not hold for the available proteomics data. For example proteomics data for *Synechocystis* sp. PCC 6803 consisted of only two biological replicates. As a result different criteria is used to identify differentially expressed genes using proteomics data sets. This criteria can be given as

1. $mean_1/mean_2 \geq 1.5$
2. $mean_1 - mean_2 > 1$
3. $(mean_1 - 2 \times stddev_1) - (mean_2 + 2 \times stddev_2) > 0$

where $mean_1$ and $stddev_1$ are mean and standard deviation values of either treatment or control, which ever has a larger mean. $Mean_2$ and $stddev_2$ correspond to the other condition. Missing data points are replaced by zeros. If a protein satisfies all three

conditions it is considered as being differentially expressed. Differentially expressed proteins are again categorized as up-regulated or down-regulated, based on whether a particular protein is high or low in the treatment compared to the control experiment respectively.

In the case of *Cyanothece* sp. ATCC 51142 data set, where focus is on diurnally regulated genes, we pick time points with maximum and minimum values as the $mean_1$ and $mean_2$ respectively. In addition to above criteria, we imposed additional condition that the mean values should be more than one in at least four time-points during a period of two days.

8.2 Differentially Expressed Proteins in *Synechocystis* sp. PCC 6803 in Different Growth Conditions

These proteomics data sets are generated using *Synechocystis* cultures, grown under different treatments; namely high CO2, Cold, heat, recovery from NH_4, four nutrient starvation conditions (Fe, N, S and P) and four revery conditions after starvation of Fe,N,S and P. Wild type cells, grown under BG11 growth media, is used as the control experiment. Proteomics data set consists of 17684 different peptides, which are mapped onto possible 2060 different proteins. Protein level expressions are obtained by summing up spectral counts for all peptides correspond to each protein. This approach of getting protein intensities is valid, since all calculations are performed for each protein separately, so that differences in the number of peptides in different proteins do not cause a problem. Number of differentially expressed proteins under

Table 8.1: Number of differentially expressed proteins in *Synechocystis* sp. PCC 6803 under different treatments.

Treatment	Up Regulated	Down Regulated	Total
CO2	192	120	312
Cold	102	165	267
Heat	214	168	382
Fe Starvation	235	166	401
N Starvation	77	476	553
P Starvation	268	88	356
S Starvation	247	148	395
Fe Recovery	244	131	375
N Recovery	231	150	381
NH4 Recovery	257	141	398
P Recovery	316	99	415
S Recovery	275	101	376

All important nutrient starvation and recovery conditions cause significant changes in the protein concentrations of *Synechocystis* sp. PCC 6803. Highest number of genes are affected under nitrogen starvation conditions where more than 85% of the affected proteins are down regulated. Fe: iron, N: nitrogen, P: phosphorus, S: sulfur

different growth conditions varied between 267 and 553. Most number of proteins got differentially expressed under nitrogen starvation condition. In Table 8.1 we list the number of proteins affected under different conditions.

8.2.1 Comparison with mRNA

To study the relationship between protein level and mRNA level changes, proteomics data is compared with similar microarray data sets. We identified five treatments, namely Cold stress, Fe, P, S and N starvation, for which microarray data sets are available under similar conditions. Similarities are measured across different genes as well as across different conditions.

Table 8.2: Correlation measurements between mRNA and proteomics expressions.

Using Log-ratio values for Protein and mRNAs					
	Correlation Measurements				
	Cold Stress	Fe-Starve	P-starve	S-Starve	N-starve
All Genes	-0.057	0.074	0.134	0.061	0.212
Differentially Expressed in mRNA	-0.198	0.162	0.334	0.38	0.37
Differentially Expressed in Proteomics	-0.07	0.1	0.175	0.09	0.299
Using Discretized Expressions					
	Percentage of times values agree				
	Cold Stress	Fe-Starve	P-Starve	S-Starve	N-Starve
All Genes	0.58	0.56	0.66	0.61	0.43
Differentially Expressed in mRNA	0.1	0.12	0.32	0.15	0.34
Differentially Expressed in Proteomics	0.11	0.07	0.17	0.07	0.52

Correlation measurements between mRNA and proteins under comparable experimental conditions are performed. Calculations are done using both log ratio and discretized expression values. Overall correlation is poor under all the treatments. This may be due to experimental variations or different levels of regulations at transcriptome and translational activities.

Comparisons across different treatments did not yield good correlation value. This calculation is done using both actual fold change values between Treatment and Control as well as discretized expressions of these fold change values. In Table 8.2, the relevant results are summarized. Overall correlation between mRNA and Protein level behavior is very low. The correlation values are slightly improved if we perform the calculations using only those genes, which are differentially expressed at mRNA level. For discretized expressions high level of agreement between mRNA and proteins is resulted in due to large number of genes which are not differentially expressed under these conditions.

Correlation measurements for individual genes across different conditions also did not show strong relationship except for few genes in ribosomal 50S complex. Table 8.3 lists some of the genes which are differentially expressed in most of the conditions and have good correlation value.

Table 8.3: Genes with good correlation between mRNA and protein expressions

Gene	Annotation	Correlation	Expressed in Protein	Expressed in mRNA
sll0656	unknown protein	0.955278	5	3
sll1742	transcription antitermination protein NusG nusG	0.724994	5	3
sll1184	heme oxygenase ho1	0.666965	4	4
sll1552	unknown protein	0.989388	5	2
sll0381	hypothetical protein	0.912997	4	3
sll1800	50S ribosomal protein L4 rpl4	0.760709	4	3
sll1799	50S ribosomal protein L3 rpl3	0.757748	4	3
slr1129	ribonuclease E rne	0.937753	5	1
sll1810	50S ribosomal protein L6 rpl6	0.921258	4	2
sll1813	50S ribosomal protein L15 rpl15	0.82778	3	3

Only a handful of genes showed a strong correlation between the expression levels of their mRNA and proteins. These include several ribosome proteins from 50S subunit.

However it is noted that genes belonging to some of the processes including photosystem-II, moved in the same direction, up or down in their expressions, in both mRNA and protein levels under similar experimental conditions. This is not revealed by the correlation measurements. Techniques such as Fisher's exact test, used to identify the association between two variables also could not highlight these observations due to imbalance nature of the contingency tables. In order to capture such behaviors we computed the fraction of genes moving in same or opposite direction for each pathway. Some of the pathways, where majority of the genes move in one direction were highlighted in Table 8.4.

8.3 Diurnal Rhythms in Steady State Protein Levels in *Cyanothece* sp. ATCC 51142

As discussed in Chapter 5, more than 40% of genes in *Cyanothece* sp. ATCC 51142 are shown to be diurnally regulated at the transcription level. To investigate whether

Table 8.4: Fractions of genes that move in the same direction in both mRNA and protein levels

Gene Function	Total Genes	Cold Stress			Fe-Starvation			P-Starvation			S-Starvation			N-Starvation		
		Differentially Expressed Genes	Relationship	Fraction of Similar Behaving Genes	Differentially Expressed Genes	Relationship	Fraction of Similar Behaving Genes	Differentially Expressed Genes	Relationship	Fraction of Similar Behaving Genes	Differentially Expressed Genes	Relationship	Fraction of Similar Behaving Genes	Differentially Expressed Genes	Relationship	Fraction of Similar Behaving Genes
AB:AAAF	21	13	n	.77	14	n	.57	15	n	.67	15	n	.53	14	p	.64
AB:AF	11	10	p	.70	11	n	.55	11	n	.64	10	n	.60	10	p	.60
CP:C	14	9	n	.78	10	p	.50	10	p	.90	8	p	.50	9	p	.56
EM:PPP	8	7	n	.57	7	n	.57	8	p	.88	8	n	.63	6	p	.67
EM:PAM	7	6	p	.50	5	p	.80	6	p	.83	4	p	.50	5	p	.80
EM:TC	8	8	n	.75	8	p	.50	7	p	.71	6	p	.83	7	n	.57
FAM	25	16	p	.50	19	p	.58	19	n	.63	19	n	.53	19	p	.74
PR:AS	9	7	n	.86	8	n	1.0	7	n	.57	9	n	.89	8	p	.63
PR:CF	15	14	n	.64	15	n	.73	15	n	.67	14	p	.57	14	p	.79
PR:PS-I	13	10	p	.50	11	p	.82	10	p	.70	11	p	.73	11	p	.64
PR:PS-II	20	16	p	.63	16	p	.63	17	p	.59	16	p	.88	16	p	1.0
PR:PB	15	14	n	.64	13	n	.69	12	p	.75	14	p	.79	14	p	1.0
PP:PR	19	9	p	.67	10	n	.70	12	p	.50	12	p	.83	10	p	.60
TR:RP	55	43	n	.60	47	n	.64	49	p	.90	47	n	.85	49	p	.96

Even though linear correlation measurements yield poor agreement, we observe genes in many pathways show similar type of response (reduction or increase in expressions) at both mRNA and protein levels. This is clear from the high fractions of genes that move in same direction under a given treatment. Interestingly we observe changes in expressions of mRNA and proteins in some pathways have a negative relationship.

AB:AAAF-Amino acid biosynthesis:Aromatic amino acid family, AB:AF-Amino acid biosynthesis:Aspartate family, CP:C-Cellular processes:Chemotaxis, EM:PPP-Energy metabolism:Pentose phosphate pathway, EM:PAM-Energy metabolism:Pyruvate and acetyl-CoA metabolism, EM-TC-Energy metabolism:TCA cycle, FAM-Fatty acid, phospholipid, and sterol metabolism, PR:AS-Photosynthesis and respiration:ATP synthase, PR:CF-Photosynthesis and respiration:CO2 fixation, PR:PSI-Photosynthesis and respiration:Photosystem I, PR:PSII-Photosynthesis and respiration:Photosystem II, PR:PB-Photosynthesis and respiration:Phycobilisome, PP:PR-Purines and pyrimidines:Purine ribonucleotides, TR:RP-Translation:Ribosomal proteins

Relationship : p-positive, n-negative

these rhythms are present at translational level also we analyzed proteomics measurements from the same experiment. This allowed us to perform direct comparison and identify the similarities and differences between steady state behaviors of mRNA and protein levels.

Proteomics data is generated using *Cyanothece* sp. ATCC 51142 cultures grown under 12h/12h Light/Dark conditions. Samples are extracted every 2h for 48h period. Original data set consisted of 6740 peptides which are mapped onto 1232 different proteins. Oscillatory proteins are identified using combined methods of Fourier scores [23], auto-correlation and trigonometric curve fitting [86]. Total of 166 genes are identified as having strong diurnal rhythms with a main period of 24h. Additional 33 genes are shown to be oscillation with a period of 12h. Compared with the results from transcriptomics analysis, we discovered that 141 genes among these 166 have strong diurnal behavior at mRNA levels. Additional 7 genes also shown to be cyclic but were not detected in transcriptomics analysis. One of the genes with 24h oscillations in protein level is shown to be having 12h oscillations in mRNA.

8.3.1 Time Difference between Transcript and Protein Peak Times

In order to compare the time difference between the peak times of mRNA expressions and the protein expressions, each expression is approximated using the first oscillatory term of the Fourier series expansion. Figure 8.1 gives the number of genes peaked during different times of the day. Time difference between two oscillations is computed as the phase difference of the approximated signals. One notable observation from this comparison of mRNA and protein peak times is significant time differences between

these two times for many genes. Figure 8.2 shows expressions at mRNA and protein levels for two oscillatory genes. There is no time delay between mRNA and protein peak times for the gene in Figure 8.2(a). In contrast, a significant time delay for these two times is observed for the gene in Figure 8.2(b). Figure 8.3 summarizes the distribution of time delays between mRNA peaks and the Protein peaks for various genes. Positive time delays represent genes where mRNA expression leads the Protein expression while negative delays represent genes with leading Protein expressions. In contrast to the observations made in transcriptomics analysis, where genes in many biological processes peak as groups during the same time of the day, wide range of peak times are observed at protein levels for genes within a single biological processes. Only exception is nitrogen fixation where we observe many genes peak at the same time in protein expressions also. With the current transcriptomics and proteomics techniques we are unable to determine the reasons behind these delays. These delays can be due to lag between transcriptional and translational activities or due to variations in synthesis and degradation levels of mRNA and proteins.

8.4 Conclusions and Discussion

Integration of transcriptomics and proteomics data sets revealed many differences between mRNA and protein expressions. Comparison of different growth conditions of *Synechocystis* sp. PCC 6803 showed only a weak correlation between mRNA and proteins. This weak correlation could be due changes in experimental conditions itself. However, by looking at the direction of change in mRNA and protein expressions, we showed that genes from different pathways change as a group with high level of agreement. One important observation made here is that for some pathways behavior

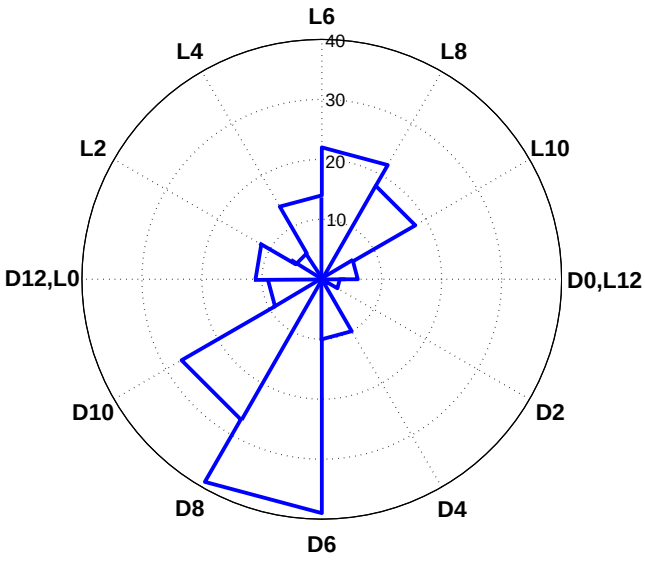
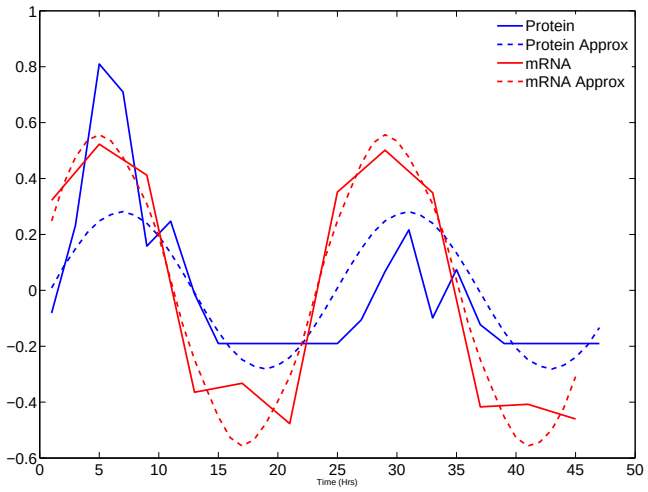


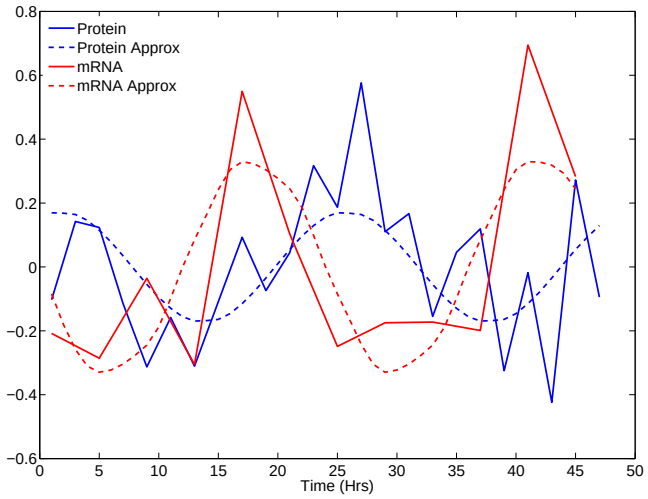
Figure 8.1: Distribution of peak times of protein expression across a single day. Majority of the oscillatory protein expressions reach their peak concentrations after the middle of the dark period. This can be due to higher translation or lower degradation rates during these periods.

of genes at mRNA levels and protein levels show a negative relationship. This might be due to time delays between different events related to transcription and translation as well as rates of degradation of mRNA and proteins.

Since same samples are used to generate both transcriptomics and proteomics data, *Cyanothece* sp. ATCC 51142 data sets provided more direct comparison between two. Out of 1232 detected proteins only 166 are shown to be cyclic. This is in contrast to more than 40% cyclic mRNA detected at transcriptomics level [23]. This suggests that the cells might maintain the protein level changes in lower dynamic range compared to that of mRNA.



(a)



(b)

Figure 8.2: Two genes that show oscillatory behaviors at both mRNA and protein abundance levels. The peak times of mRNA and protein concentrations can vary in a wide range of periods for different genes.

Significant time delays between peak mRNA and protein levels are detected. These time delays might be due to various post transcriptional regulation mechanisms or balance between synthesis and degradation rates of corresponding molecules.

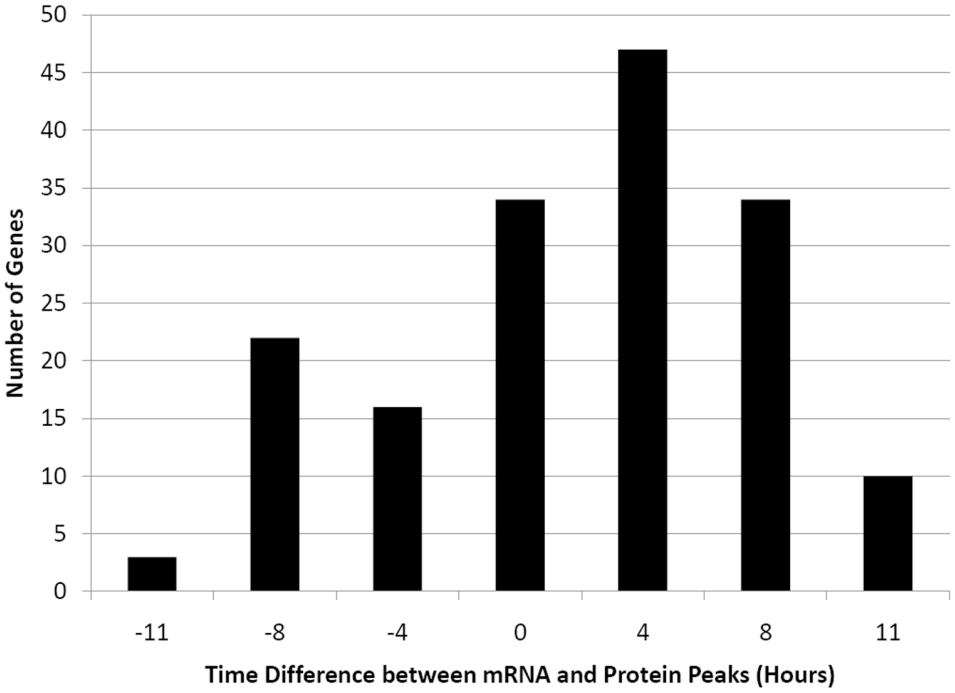


Figure 8.3: Time delays observed between peak times of protein and mRNA expressions.
Positive time delays represent genes where mRNA expression leads the Protein expression while negative delays represent genes with leading Protein expressions. Current techniques are insufficient to explain the exact reasons behind observed delays.

Chapter 9

Conclusions

In this dissertation, we analyzed several high throughput data sets from different photosynthetic organisms to understand their response to changes in their environments. We developed various computational and modeling techniques to analyze these data so that salient features in cellular responses can be extracted. Three specific problems studied here are transcriptomics modifications in photosynthetic organisms to reduction-oxidation (redox) stress conditions, circadian and diurnal rhythms of cyanobacteria and effects of incident light patterns on these rhythms, and coordination between biological processes in cyanobacteria under various growth conditions.

We discussed two commonly used high throughput techniques in transcriptomics and proteomics, namely two-color microarrays and label free bottom-up proteomics. We utilized several computational and statistical algorithms including LOWESS normalization and statistical significance tests to perform preliminary data processing and quality assessments of the data sets. Depending on the objective of the biological experiment, we selected suitable criteria to identify informative genes. These approaches include several statistical tests such as Student's t-test, KS-test, Fourier scores, angular distances and their combinations. Various standard and non-standard classification methods are utilized to group genes to main behavioral categories. We

proposed several deterministic and probabilistic models to explain expressions of these gene groups. We also showed how existing insight on gene interactions and relevant computational algorithms can improve the initial results.

With our analysis we were able to discover system wide transcriptional modifications in the cyanobacterium *Synechocystis* sp. PCC 6803, under various redox stresses caused by high light treatment, DCMU and preferential excitation of photosystem I and II. Gene clustering methods revealed that these responses can mainly be classified as transient responses and consistent responses, depending on durations of modified behaviors. We showed many central pathways related to energy production as well as energy utilization are strongly affected by these stresses. Combined analysis of two stress conditions, high light and DCMU treatment, combined with data mining and motif finding algorithms led to the discovery of a novel transcription factor in *Arabidopsis thaliana*, $RRTF1$, which responds to redox stresses.

Using multiple experimental conditions we were able to show that majority of the diurnal genes in *Cyanothece* sp. ATCC 51142 are in fact light responding. Only about 10% of genes in the genome are categorized as being circadian controlled. We derived two transcription control model based on feed-forward loops and phase oscillators to model and identify interactions between diurnal genes. Both these models are shown to carry biologically meaningful features.

We were able to integrate all transcriptomics data sets available for *Synechocystis* sp. PCC 6803 and utilize probabilistic modeling to obtain a Bayesian network for main biological processes in the cell. Several novel relationships between biological processes are discovered from the model. Model is used to simulate several experimental

conditions, and the response of the model is shown to agree with the experimentally observed behaviors.

Finally we combined the analysis of related proteomics and transcriptomics data sets to study the similarities and differences in cellular responses at these two levels.

Current analysis helps us extending our knowledge on cellular responses to different environment conditions at global level. How ever in order to gain better understanding on these complex dynamical systems, many additional experimental and computational effort is needed. We are hoping move towards this goal by combining newer technologies including metabolomics and genome sequencing.

Appendix A

Experimental Organisms and data sets

A.1 *Synechocystis* sp. PCC 6803

Synechocystis sp. PCC 6803 is the first photosynthesis organism to have a completely sequenced genome. It is capable of growing in numerous environment conditions, ranging from fully autotrophical (growth by fixing environment CO_2 using light energy) to heterotrophic (growth under dark, utilizing sugar through *glycolysis* and *oxidative phosphorylation* to generate required energy). Since its spontaneously transformable, *Synechocystis* is widely used as a model organism in photosynthesis research.

Following data sets from *Synechocystis* sp. PCC 6803 are analyzed:

- High Light Treatment : Microarray data set

  This time course microarrays consist of six time points namely 15min, 1h, 2h, 3h, 4h and 6h. *Synechocystis* cells are grown under high light with an intensity of $300\mu Em^{-2}s^{-1}$ and compared with the cells grown under regular light of intensity $30\mu Em^{-2}s^{-1}$. Each time point consists of 6 microarrays, which include a dye swap and two biological replicates.

- DCMU Treatment : Microarray data set

Figure A.1: *Synechocystis* sp. PCC 6803.
Synechocystis sp. PCC 6803 is the mostly studied photosynthetic cyanobacterium. It is the first cyanobacterium and third prokaryote to have a completely sequenced genome. (Image courtesy: Michelle Liberton)

This data set consists of five time points namely 15min, 45min, 1.5h, 3h, and 6h. *Synechocystis* cells are treated with DCMU (3-(3,4-dichlorophenyl)-1,1-dimethylurea), a very specific and sensitive inhibitor of photosynthesis II system, to reduce the electron flow between photosystem II and plastoquinone, by 20%. Each time point consists of 6 microarrays, which include a dye swap and two biological replicates.

- Preferential Excitation of Photosystem I and Photosystem II : Microarray data set

  Photosystem I and Photosystem II in *Synechocystis* cells are preferentially excited using blue and red light of intensity $10\mu Em^{-2}s^{-1}$, respectively. Samples are obtained at six time points namely 15min, 45min, 1.5h, 2h, 3h and 6h, and 6 microarrays are generated at each time point.

- Comparison of different growth conditions: Proteomics data set

  Proteomics data from twelve different growth conditions where presence of important nutrients are controlled are compared with normal growth conditions under *BG11* growth media. Total of 17920 peptides were detected in different conditions which were later mapped into 2061 unique proteins.

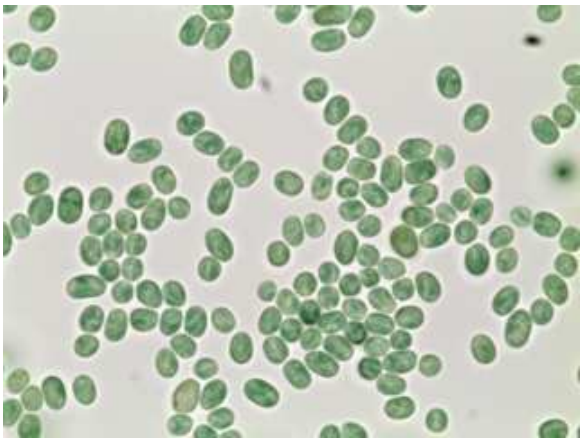


Figure A.2: *Cyanothece* sp. ATCC 51142.
Its ability to fix environmental N_2 (Diazotrophic) as well as performing photosynthesis within a single cell has drawn large research interest during last few years. (Image courtesy: Michelle Liberton)

A.2 *Cyanothece* sp. ATCC 51142

Cyanothece sp. ATCC 51142 is a marine cyanobacteria. Its ability to fix environmental N_2 (Diazotrophic) as well as performing photosynthesis within a single cell, has drawn large research interest during last few years. Because the enzyme which fixes atmospheric N_2 (nitrogenase) is highly sensitive to oxygen, *Cyanothece* sp. ATCC 51142 uses a temporal separation between two processes; namely performing N_2 fixation during dark and photosynthesis during day time. These two processes as well as other metabolic processes are shown to be under strong diurnal regulation [22]. *Cyanothece* sp. ATCC 51142 also consists of a robust *circadian rhythms*; an internal 24h oscillatory mechanism which persists under changing light inputs.

In order to study the cellular behavior under diurnal regulation with regular light and dark inputs and the effects of changing light patterns on different processes, two microarray experiments are conducted ([74] and [79]). In addition proteomics analysis done using the samples from [74]. Following data sets from *Cyanothece* sp. ATCC 51142 are analyzed here.

- Cellular behavior under regular diurnal light inputs : Microarray data set

Cells are grown with regular 12h/12h light-dark input under nitrogen fixing conditions. The time course microarray data set consists of transcriptomics measurements from 4888 genes over a period of 48 hours. Samples are extracted every four hours with the first sample taken after one hour into the dark period.

- Cellular behavior due to changing light input from diurnal to constant light: Microarray data set

  Similar to above experiment except that the cells are kept under constant light input during the second half of the experiment. First sample is extracted after 2 hours into the light period.

- Cellular behavior under regular diurnal light inputs : Proteomics data set

  During the cultures from the first experiment described above, proteomics analysis was done using bottom-up label free approach. In this case samples are taken every 2 hours in contrast to every 4 hours in the case of transcriptomics.

A.3 *Arabidopsis thaliana*

Arabidopsis thaliana is the model organism for plant biology. This vascular plant has been shown to be consisted of more than 29000 genes, which is more than the number of genes in humans. *Arabidopsis* is extensively used in research related to photosynthesis, flowering mechanisms, circadian rhythms, environment stresses etc.

Two time course transcriptomics data sets from *Arabidopsis thaliana* are analyzed:

- High Light Treatment This time course microarray data set consist of four time points namely 45min, 1.5h, 3h, and 6h. For target and control experiments, light intensities of $750\mu Em^{-2}s^{-1}$ and $75\mu Em^{-2}s^{-1}$ respectively, are used.

- DCMU Treatment This data set consists of three time points namely 1.5h, 3h, and 6h.

Figure A.3: *Arabidopsis thaliana*.
Arabidopsis thaliana is the model organism for vascular plants. It is extensively used in research related to photosynthesis, flowering mechanisms, circadian rhythms and environment stresses. (Image courtesy: Abha Khandelwal)

References

[1] A. Agresti, "Categorical Data Analysis, 2nd Edition", *John Wiley & Sons, Inc*, 2002.

[2] R. Akhtar, A. Reddy, E. Maywood, J. Clayton, V. King, A. Smith, T. Gant, M. Hastings, and C. Kyriacou, "Circadian Cycling of the Mouse Liver Transcriptome, as Revealed by cDNA Microarray, Is Driven by the Suprachiasmatic Nucleus", *Current Biology*, vol. 12, pp. 540–550, 2002.

[3] T. Akutsu, S. Miyano, S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model", *Pac Symp Biocomput.*, pp. 17–28, 1999.

[4] M. Amdaoud, M. Vallade, C. Weiss-Schaber, and I. Mihalcescu, "Cyanobacterial clock, a stable phase oscillator with negligible intercellular coupling", *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7051–7056, 2007.

[5] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di-Bernardo, "How to infer gene networks from expression profile", *Molecular Systems Biology*, vol. 3, issue. 78, 2007.

[6] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review", *Anal Bioanal Chem*, vol. 389, pp. 1017–1031, 2007.

[7] A. Blais, and B. D. Dynlacht, "Constructing transcriptional regulatory networks", *Genes Dev.*, vol. 19, pp. 1499–1511, 2005.

[8] R. Bonneau, M. T. Facciotti, D. J. Reiss, A. K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M. H. Johnson, J. C. Bare, W. Longabaugh, M. Vuthoori, K. Whitehead, A. Madar, L. Suzuki, T. Mori, D. Chang, J. DiRuggiero, C. H. Johnson, L. Hood, and N. S. Baliga,"A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell", *Cell*, vol. 131, Issue 7, pp. 1354–1365, 2007

[9] S. A. Bustin, "Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays", *J Mol Endocrinol*, vol. 25, pp. 169–193, 2000.

[10] J. Cao, and H. Zhao, "Estimating dynamic models for gene regulation networks", *Bioinformatics*, vol. 24, pp. 1619–1624, 2008.

[11] X. Chen, G. Anantha, and X. Lin, "Improving Bayesian Network Structure Learning with Mutual Information-Based Node Ordering in the K2 Algorithm", *IEEE Trans. Knowl. Data Eng.*, vol. 20, pp. 628–640, 2008.

[12] D. M. Chickering, "Optimal structure identification with greedy search", *Journal of Machine Learning Research*, vol. 3, pp. 507–554, 2002.

[13] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees", *IEEE Transactions on Information Theory*, vol. 14, pp. 462-467, 1968.

[14] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data", 2007. Available online: `http://arxiv.org/abs/0706.1062`.

[15] R. Craig, R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra", *Bioinformatics*, vol. 12, pp. 1466-1467, 2004.

[16] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: A Sequence Logo Generator", *Genome Res.*, vol. 14, pp. 1188–1190, 2004.

[17] A. C. Davison and D. Hinkley, "Bootstrap Methods and their Applications", *Cambridge Series in Statistical and Probabilistic Mathematics*, 1997.

[18] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 1–3, 1977.

[19] X. Du, S. J. Callister, N. P. Manes, J. N. Adkins, R. A. Alexandridis, X. Zeng, J. H. Roh, W. E. Smith, T. J. Donohue, S. Kaplan, R. D. Smith, and M. S. Lipton, "A Computational Strategy to Analyze Label-Free Temporal Bottom-Up Proteomics Data", *Journal of Proteome Research*, vol. 7, pp. 2595–2604, 2008.

[20] W. Dubitzky, M. Granzow, and D. P. Berrar, "Fundamentals of Data Mining in Genomics and Proteomics", *Springer*, 2007.

[21] R. Edgar, M. Domrachev and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res.*, vol. 30, pp. 207–210, 2002.

[22] T. R. Elvitigala, H. B. Pakrasi, and B. K. Ghosh, "Dynamic Network Modeling of Diurnal Genes in Cyanobacteria", *Springer special edition*, In press.

[23] T. R. Elvitigala, J. Stöckel, B. K. Ghosh and H. B. Pakrasi, "Effect of continuous light on diurnal rhythms in *Cyanothece* sp. ATCC 51142", *BMC Genomics*, vol. 10:226, 2009.

[24] T. R. Elvitigala, H. B. Pakrasi, and B. K. Ghosh, "Modeling and Simulation of Diurnal Biological Processes in Cyanobacteria", *Proceedings of American Control Conference*, St. Louis, MO, 2009.

[25] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles", *PLoS Biol*, vol. 5(1), doi:10.1371/journal.pbio.0050008

[26] P. Leray and O. Francois, "BNT Structure Learning Package, Technical Report", *Laboratoire PSI - INSA Rouen- FRE CNRS 2645*.

[27] N. Friedman, M. Linial, I. Nachman, and D. Peer, "Using Bayesian network to analyze expression data", *J. Comput. Biol*, vol. 7, pp. 601-620, 2000.

[28] B. K. Ghosh, H. B. Pakrasi, and T. R. Elvitigala, "Effect on Diurnal Rhythms by changes in light input", *10th International Conference on Control Automation Robotics & Vision*, 2008.

[29] L. Gold, "Posttranscriptional regulatory mechanisms in Escherichia coli", *Annu Rev Biochem*, vol. 57, pp. 199–233, 1988.

[30] S. S. Golden, M. Ishiura, C. H. Johnson, and T. Kondo, "Cyanobacterial Circadian Rhythms", *Annual Review of Plant Physiology and Plant Molecular Biology*, vol. 48, pp. 327-354, 1997.

[31] S. S. Golden and S. R. Canales, "Cyanobacterial circadian clocks timing is everything", *Nature Reviews Microbiology*, vol. 1, pp. 191–199, 2003.

[32] S. S. Golden, "Timekeeping in bacteria: the cyanobacterial circadian clock", *Current Opinion in Microbiology*, vol. 6, pp. 535–540, 2003.

[33] D. Heckerman, "A Tutorial on Learning with Bayesian Networks", *MIT Press*, 1999.

[34] G. Z. Hertz, and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences", *Bioinformatics*, vol. 15, pp. 563–577, 1999.

[35] Y. Hihara, A. Kameib, M. Kanehisac, A. Kapland, and M. Ikeuchib, "DNA Microarray Analysis of Cyanobacterial Gene Expression during Acclimation to High Light", *Plant Cell*, vol. 13, pp. 793–806, 2001.

[36] F. V. Jensen and T. D. Nielsen, "Bayesian Networks and Decision Graphs", *Springer-Verlag New York, Inc.*, 2007.

[37] M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, "The KEGG databases at GenomeNet", *Nucl. Acids Res.*, vol. 30, pp. 42–46, 2002.

[38] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, "KEGG for linking genomes to life and the environment", *Nucl. Acids Res.*, vol. 36, pp. 480–484, 2008.

[39] A. Khandelwal, T. R. Elvitigala, B. K. Ghosh, and R. S. Quatrano, "Arabidopsis Transcriptome Reveals Control Circuits Regulating Redox Homeostasis and the Role of an AP2 Transcription Factor", *Plant Physiol.*, vol. 148, pp. 2050–2058, 2008.

[40] J. Kim and H. G. Nam, "Instrumentation and Software for Analysis of Arabidopsis Circadian Leaf Movement", *Interdisciplinary Bio Content*, vol. 1(1), pp. 1–4, 2009.

[41] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1137-1143, 1995.

[42] T. Kondo, N. F. Tsinoremas, S. S. Golden, C. H. Johnson, S. Kutsuna, and M. Ishiura, "Circadian clock mutants of cyanobacteria", *Science*, vol. 266, no. 5188, pp. 1233–1236, 1994.

[43] T. Kondo and M. Ishiura, "The circadian clocks of plants and cyanobacteria", *Trends in Plant Science*, vol. 4, pp. 171-176, 1999.

[44] Y. Kuramoto, "Chemical Oscillations, Waves, and Turbulence", *Springer-Verlag, New York*, 1984.

[45] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: extraction and analysis of differential gene expression", *Bioinformatics*, vol. 22, pp. 507–508, 2006.

[46] K. Liang, and X. Wang, "Gene Regulatory Network Reconstruction Using Conditional Mutual Information", *EURASIP Journal on Bioinformatics and Systems Biology*, 2008 .

[47] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates III, " Direct analysis of protein complexes using mass spectrometry", *Nature Biotechnol.*, vol. 17, pp. 676–682, 1999.

[48] Y. Liu and D. Bell-Pedersen, "Circadian Rhythms in Neurospora crassa and Other Filamentous Fungi", *Eukaryotic Cell*, vol. 5, No. 8, pp. 1184–1193, 2006.

[49] D. MacLean, J. D. G. Jones, and D. J. Studholme, "Application of 'next-generation' sequencing technologies to microbial genetics", *Nat Rev Micro*, vol. advanced online publication, 2009.

[50] F. J. Massey, "The Kolmogorov-Smimov test for goodness-of-fit", *J. Amer Statist, Assoc.*, vol. 46, pp. 68–78, 1951.

[51] A. Mehra, C. I. Hong, M. Shi, J. J. Loros, J. C. Dunlap, and P. Ruoff, "Circadian rhythmicity by autocatalysis", *PLos Computational Biology*, vol. 2, pp. 816–823, 2006.

[52] T. P. Michael, T. C. Mockler, G. Breton, C. McEntee, A. Byer, J. D. Trout, S. P. Hazen, R. Shen, H. D. Priest, C. M. Sullivan, S. A. Givan, M. Yanovsky, F. Hong, S. A. Kay, and J. Chory, "Network Discovery Pipeline Elucidates Conserved Time-of-Day Specific cis-Regulatory Modules", *PLoS Genetics*, **4**, e14, 2008.

[53] F. Miyoshi, Y. Nakayama, K. Kaizu, H. Iwasaki, and M. Tomita, "A Mathematical Model for the Kai-ProteinBased Chemical Oscillator and Clock Gene Expression Rhythms in Cyanobacteria", *J Biol Rhythms*, pp. 22–69, 2007.

[54] K. P. Murphy, "The Bayes Net Toolbox for Matlab", *Computing Science and Statistics*, vol. 33, 2001.

[55] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats", *Protein Sci.*, vol. 4, pp. 1618-1632, 1995.

[56] D. Bell-Pedersen, V. M. Cassone, D. J. Earnest, S. S. Golden, P. E. Hardin, T. L. Thomas, and M. J. Zoran, "Circadian rhythms from multiple oscillators: lessons from diverse organisms", *Nat Rev Genet*, vol. 6, pp. 544–556, 2005.

[57] T. Pfannschmidt, K. Brutigam , R. Wagner , L. Dietzel , Y. Schrter , S. Steiner , and A. Nykytenko, "Potential regulation of gene expression in photosynthetic cells by redox and energy state: approaches towards better understanding", *Ann Bot*, vol. 103, pp. 599–607, 2009.

[58] A. D. Polpitiya, W. J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. J. Camp,II, G. A. Anderson, and R. D. Smith, "DAnTE: a statistical tool for quantitative analysis of -omics data", *Bioinformatics*, vol. 24, pp. 1556–1558, 2008.

[59] O. Pourret, P. Nam, and B. Marcot, "Bayesian Networks: A Practical Guide to Applications", *Wiley*, 2008.

[60] J. Quackenbush, "Microarray data normalization and transformation", *Nature Genetics*, vol. 32, pp. 496–501, 2002.

[61] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series", *Pac. Symp. Biocomput*, pp. 455-466, 2000.

[62] D. Rieger, C. Fraunholz, J. Popp, D. Bichler, R. Dittmann, and C. Helfrich-Frster, "The fruit fly Drosophila melanogaster favors dim light and times its activity peaks to early dawn and late dusk", *J Biol Rhythms*, vol. 22, pp. 387–99, 2007.

[63] R. W. Robinson, "Counting Unlabeled Acyclic Digraphs", *Proceedings of the Fifth Australian Conference*, pp. 24-26, 1976.

[64] C. Sabatti and K. Lange, "Genomewide motif identification using a dictionary model", *Proceedings of the IEEE*, vol. 90, no. 11, pp. 1803–1810, 2002.

[65] A. Schulze, and J. Downward, "Navigating gene expression using microarrays - a technology review", *Nat Cell Biol*, vol. 3, pp. E190–E195, 2001.

[66] 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*, vol. 13, pp. 2498–2504, 2003.

[67] L. A. Sherman, P. Meunier and M. S. Coln-Lpez, "Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium", *Photosynthesis Research*, vol. 58, pp. 25–42, 2004.

[68] I. Shmulevich, E. R. Dougherty, S. Kim and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks", *Bioinformatics*, vol. 18, pp. 261–274, 2002.

[69] S. Haykin, "Neural networks - A comprehensive foundation (2nd edition ed.)", *Prentice-Hall*, 1999.

[70] A. K. Singh, T. R. Elvitigala, M. Bhattacharyya-Pakrasi, R. Aurora, B. K. Ghosh, and H. B. Pakrasi, "Integration of Carbon and Nitrogen Metabolism with Energy Production Is Crucial to Light Acclimation in the Cyanobacterium *Synechocystis*", *Plant Physiol.*, vol. 148, pp. 467–478, 2008.

[71] A. K. Singh, T. R. Elvitigala, J. Cameron, B. K. Ghosh, and H. B. Pakrasi, "Strategy of Cellular Adaptations under Perturbations in an Oxygenic Photosynthetic Cyanobacterium", In preparation.

[72] R. R. Sokal, and F.J. Rohlf, "Biometry: The principles and practice of statistics in biological research. 3rd edition", *W.H. Freeman, New York*, 1995.

[73] J. O. Sophocles, "Introduction to signal processing", *Prentice-Hall, Inc.*, 1995.

[74] J. Stöckel, E. A. Welsh, M. Liberton, R. Kunnvakkam, R. Aurora, and H. B. Pakrasi, "Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes", *PNAS*, vol. 105, pp. 6156–6161, 2008.

[75] J. Stöckel, T. R. Elvitigala, M. Liberton, J. Jacobs, E. Welsh, B. K. Ghosh and H. B. Pakrasi, "Cyclic proteins in *Cyanothece* sp. ATCC 51142", In preparation.

[76] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments", *PNAS*, vol. 102, pp. 12837–12842, 2005.

[77] S. H. Strogatz and I. Stewart, "Coupled oscillators and biological synchronization", *Scientific American*, vol. 269, pp. 102–109, 1993.

[78] I. Tagkopoulos, Y. Liu, and S. Tavazoie, "Predictive Behavior Within Microbial Genetic Networks", *Science*, vol. 320. no. 5881, pp. 1313–1317, 2008.

[79] J. Toepel, E. Welsh, T. C. Summerfield, H. B. Pakrasi, L. A. Sherman, "Differential transcriptional analysis of the cyanobacterium *Cyanothece* sp. strain ATCC 51142 during light-dark and continuous-light growth", *J Bacteriol*, vol. 190, pp. 3904–3913, 2008.

[80] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, Missing value estimation methods for DNA microarrays", *Bioinformatics*, vol. 17, pp. 520–525, 2001.

[81] I. Ebert-Uphoff, "Measuring Connection Strengths and Link Strengths in Discrete Bayesian Networks", *Research Report GIT-IIC-07-01, Georgia Institute of Technology, Interactive & Inelligent Computing Division*, 2007. Available online:

[82] A. Uri, "An Introduction to Systems Biology: Design Principles of Biological Circuits". *Chapman & Hall/CRC.*, 2006.

[83] G. Wang, T. Yu, and W. Zhang, "WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar", *Nucl. Acids Res.*, vol. 33, pp. 412–416, 2005.

[84] W. Wang, B. K. Ghosh, and H. B. Pakrasi, "Identification and Modeling of Genes with Diurnal Oscillations from Microarray Time Series Data", *IEEE transactions on Computational Biology and Bioinformatics*, In press.

[85] W. Weckwerth, Metabolomics in Systems Biology, *Annu. Rev. Plant Biol*, vol. 54, pp. 669-689, 2003.

[86] E. A. Welsh, J. S. Stckel and H. B. Pakrasi, "Cycle Detection in Biological Data Sets in Computational and Systems Biology: Applications and Methods", *R. A. Mazzarella and R. D. Head, Editors, Research Signpost: Trivandrum, India*, pp. 51–62, 2009.

[87] M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams, "Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it", *Biotechnology and Genetic Engineering Reviews*, vol. 13, pp. 19-50, 1996.

[88] A. T. Winfree, "The geometry of biological time", *New York: Springer*, 1980.

[89] J. Wolberg, "Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments", *Springer*, 2005.

Vita

Thanura Ranmal Elvitigala

Date of Birth November 19, 1976

Place of Birth Colombo, Sri Lanka

Degrees B.Sc. (Eng), Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka, March 2002
M.S. Electrical Engineering, Washington University in St Louis, December 2006
Ph.D. System Science and Mathematics, Washington University in St Louis, December 2009

Journal Publications Singh, A. K., Bhattacharya-Pakrasi, M., Elvitigala, T. R., Aurora, R., Ghosh, B. K. and Pakrasi, H. B, A systems level analysis of the effects of light quality on the metabolism of a cyanobacterium, *Plant Physiol.*, Accepted for publication.

Elvitigala, T. R., Stöckel, J., Ghosh B. K., and Pakrasi, H. B., Effect of continuous light on diurnal rhythms in *Cyanothece* sp. ATCC 51142, *BMC Genomics*, **10**(1):226, 2009.

Welsh, E. A., Liberton, M., St0̈ckel, J., Loh, T., Elvitigala, T. R., Wang, C., Wollam, A., Fulton, R. S., Clifton, S. W., Jacobs, J. M., Aurora, R., Ghosh, B. K., Sherman, L. A., Smith, R. D., Wilson, R. K., and Pakrasi, H. B., The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle, *PNAS*, **105**(39):15094–9, 2008.

Khandelwal, A., Elvitigala, T. R., Ghosh, B. K., and Quatrano, R. S., Arabidopsis Transcriptome Reveals Control Circuits Regulating Redox Homeostasis and the Role of an AP2 Transcription Factor, *Plant Physiol.*, **148**:2050–2058, 2008.

Singh, A. K., Elvitigala, T. R., Bhattacharya-Pakrasi, M., Aurora, R., Ghosh, B. K. and Pakrasi, H. B, Integration of carbon and nitrogen metabolism with energy production is crucial to light acclimation in the cyanobacterium Synechocystis, *Plant Physiol.*, **148**(1):467–78, 2008.

Book Chapters

Elvitigala, T. R., Pakrasi, H. B., and Ghosh B. K., Dynamic Network Modeling of Diurnal Genes in Cyanobacteria, *Springer Special Editions*, In press.

Conference Publications and Poster Presentations

Elvitigala, T. R., Singh, A. K., Pakrasi, H. B., and Ghosh, B. K, Bayesian Network Approach to understand Regulation of Biological Processes in Cyanobacteria, *Joint* 48^{th} *IEEE Conference on Decision and Control and* 28^{th} *Chinese Control Conference*, Shanghai, China, 2009, Accepted for publication.

Elvitigala, T. R., Pakrasi, H. B., and Ghosh, B. K. Dynamical Systems Modeling of Interactions in Cyanobacteria Diurnal Genes, *Foundations of Systems Biology in Engineering*, Denver, Colorado, 2009.

Elvitigala, T. R., Pakrasi, H. B., and Ghosh, B. K., Modeling and Simulation of Diurnal Biological Processes in Cyanobacteria, *Proceedings of American Control Conference*, St. Louis, MO, 2009.

Liberton, M., Stckel, J., Dohnalkova, A. C., Orr, G., Jacobs, M., Elvitigala, T. R., Welsh, E. A., Min, H., Toepel, J., Metz, T. O., Scholten, H., Kennedy, M. A., Buchko, G. W., Koropatkin, N. M., Aurora, R., Ghosh, B. K., Ogawa, R., McDermott, J. E., Waters, K. M., Oehmen, C., Anderson, G. A., Smith, T. J., Smith, R. D., Sherman, L. A., Koppenaal, D. W., and Pakrasi, H. B., Grand Challenge in Membrane Biology: A Systems Biology Study of the Unicellular Diazotrophic Cyanobacterium *Cyanothece* sp. ATCC 51142, *Genomics:GTL Awardee Workshop VII and USDA-DOE Plant Feedstock Genomics for Bioenergy Awardee Workshop*, Bethesda, Maryland, 2009.

Ghosh, B. K., Pakrasi, H. B., and Elvitigala, T. R., Controlling diurnal rhythms by light, 10^{th} *International Conference on Control, Automation, Robotics and Vision*, Hanoi, Vietnam, 2008.

Elvitigala, T. R., Singh, A. K., Bhattacharya-Pakrasi, M., Aurora, R., Pakrasi, H. B., and Ghosh, B. K., Response to Redox Stresses in *Synechocystis* 6803: Information Revealed from Microarray Experiments, 9^{th} *Cyanobacterial Workshop, Wisconsin*, USA, 2007.

Elvitigala, T. R., Singh, A. K., Khandelwal, A., Bhattacharya-Pakrasi, M., Aurora, R., Pakrasi, H. B., Quatrano, R., and Ghosh, B. K., Computational Analysis of the Redox Stress Response in Organisms, *International Conference on Systems Biology*, Long Beach, California, 2007.

Khandelwal, A., Elvitigala, T. R., Ghosh, B. K., Quatrano, R., Systems Approach Divulges Redox Regulation of Arabidopsis Transcriptome, *International Conference on Systems Biology*, Long Beach, California, 2007.

Wang, W., Elvitigala, T. R., Stockel, J., Pakrasi, H. B., and Ghosh, B. K., Identification and Modeling of Co-Rhythmic Genes from Micro-array Time Series Data, *International Conference on Systems Biology*, Long Beach, California, 2007.

Khandelwal, A., Elvitigala, T. R., Ghosh, B. K., Quatrano, R., Network analysis of the Arabidopsis Transcriptome Reveals Control Circuits Regulating Redox Homeostasis, *International Fall Symposium, Donald Danforth Plant Science Center*, Missouri, 2007.

Khandelwal, A., Elvitigala, T. R., Yiyong, Z., Ben, I., Ghosh, B. K., Quatrano, R., Redox Homeostasis in Plants: The Arabidopsis Transcriptome in Response to Photosynthetic Stress using High Light and DCMU. *Botany and plant biology*, Chicago, Illinois, 2007.

Khandelwal, A., Yiyong, Z., Ben, I., Elvitigala, T. R., Ghosh, B. K., Quatrano, R., Redox Regulation of Arabidopsis Transcriptome, 7^{th} *Annual Fall symposium, Merging and Emerging Disciplines in Biology*, 2005.

December 2009

Biological Data Analysis, Elvitigala, Ph.D. 2009