Extracting Transcriptional Regulatory Information From DNA Microarray Expression Data

by

William A. Schmitt, Jr.

MSCEP, Massachusetts Institute of Technology, 2000

BSChE, Michigan State University, 1998

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Chemical Engineering

at the
Massachusetts Institute of Technology
[September 2003]
July, 2003

Author _____
Department of Chemical Engineering
July 9, 2003

Certified by _____
Gregory Stephanopoulos
Professor of Chemical Engineering
Thesis Supervisor

Accepted by _____
Daniel Blankschtein
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Extracting Transcriptional Regulatory
Information From DNA Microarray Expression Data

by

William A. Schmitt, Jr.

Submitted to the Department of Chemical Engineering on
June 13th, 2003 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering

ABSTRACT

Recent technological developments allow all the genes of a species to be monitored simultaneously at the transcriptional level. This necessitates a more global approach to biology that includes consideration of complex interactions between many genes and other intracellular species. The metaphor of a cell as a miniature chemical plant with inputs, outputs, and controls gives chemical engineers a foothold in this type of analysis. Networks of interacting genes are fertile ground for the application of the methods developed by engineers for the analysis and monitoring of industrial chemical processes.

The DNA microarray has been established as a tool for efficient collection of mRNA expression data for a large number of genes simultaneously. Although great strides have been made in the methodology and instrumentation of this technique, the development of computational tools needed to interpret the results have received relatively inadequate attention. Existing analyses, such a clustering techniques applied to static data from cells at many different states, provide insight into co-expression of genes and are an important basis for exploration of the cell's genetic programming. We propose that an even greater level of regulatory detail may be gained by dynamically changing experimental conditions (the input signal) and measuring the time-delayed response of the genes (the output signal). The addition of temporal information to DNA microarray experiments should suggest potential cause/effect relationships among genes with significant regulatory responses to the conditions of interest.

This thesis aims to develop computational techniques to maximize the information gained from such dynamic experiments. As a model system, we have chosen the unicellular, photoautotrophic cyanobacteria *Synechocystis* sp. PCC6803 for study, as it is 1) fully sequenced, 2) has an easily manipulated input signal (light for photosynthesis), and 3) fixes carbon dioxide into the commercially interesting, biodegradable polymer *polyhydroxyalkanoate* (PHA). We have created DNA microarrays with ~97% of the *Synechocystis* genome represented in duplicate to monitor the cellular transcriptional profile. These arrays are used in time-series experiments of differing light levels to measure dynamic transcriptional response to changing environmental conditions.

We have developed networks of potential genetic regulatory interactions through time-series analysis based on the data from our studies. An algorithm for combining gene position information, clustering, and time-lagged correlations has been created to generate networks of

hypothetical biological links. Analysis of these networks indicates that good correlation exists between the input signal and certain groups of photosynthesis- and metabolism-related genes. Furthermore, this analysis technique placed these in a temporal context, showing the sequence of potential effects from changes in the experimental conditions.

This data and hypothetical interaction networks have been used to construct AutoRegressive with eXogenous input (ARX) models. These provide dynamic, state-space models for prediction of transcriptional profiles given a dynamically changing set of environmental perturbations. We have shown that these models provide information for the design of additional experiments, and their accuracy has been validated with independent data. The derived networks and the models based on them have therefore been shown to hold not only predictive capabilities for transcriptional-level phenomena but also to provide hypotheses for the nature of the underlying biochemical relationships.

Thesis Supervisor: Greg Stephanopoulos
Title: Bayer Professor of Chemical Engineering

ACKNOWLEDGEMENTS

Professor Gregory Stephanopoulos has been an unparalleled source of guidance and patience. His ability to maintain a broad perspective while demanding attention to the most minute detail have set my standards for quality work both here and in later life. Professor George Stephanopoulos is also owed many thanks, particularly for helping me to structure my problem and get oriented in the right direction from the very start of my work. The additional guidance and insight of my committee members Professor Dane Wittrup, Dr. Isidore Rigoutsos, and Dr. Gaspar Taroncher-Oldenburg are greatly appreciated.

Jatin Misra is owed immeasurable thanks for his intellectual input, constructive criticism, and support throughout qualifying exams, Practice School, and this thesis. My work on this project has been singularly inspired by our discussions and his insights into the key challenges facing our then-nascent Bioinformatics sub-group. The rest of the Bioinformatics group, including Ameya "Cash out" Agge, Faisal "Spastic Twitch" Reza, Vipin "Pumpatude" Gupta, Joel "Checkmated" Moxley, Daehee "Godammit" Hwang, and Kyle "Ahoy!" Jensen, has provided similar assistance and camaraderie. The positive atmosphere they created in the lab far outweighed the numerous distractions they caused, although one of us must now become a chess grandmaster to justify the effort collectively expended on the game. Thanks also to LISPE members Orhan Karsligil, Matthew Dyer, and Ahmed Ismail for their friendship and participation as additional sounding boards for ideas. None of these group members would have accomplished half as much without the far-reaching care and support of Susan Lanza. Of course, all the little chocolates and other sweets she provided did not bother me in the least.

My thanks goes to all of the Metabolic Engineering lab members, but especially to Ryan Gill, Saliya Silva, and Angelo Mondragon for teaching me not only how to extract transcriptional data from *Synechocystis*, but also how not to kill all of my cultures, all of the time. I also appreciate the friendship and research insight of Javier Femenia throughout my work. Thanks also to Brett Roth for keeping the Metabolic Engineering lab running smoothly while sharing anecdotes from his vast experience to enrich my every visit to the office.

Thanks to my friends at both MIT TechLink and MIT Winos for providing an outlet for not only interaction with the rest of MIT's community but also for giving me wonderful excuses to either

go to the Muddy Charles Pub or purchase lovely wines as frequently as necessary. John Lock and Monica Rixman have particularly been key in this endeavor.

All of my apartment mates from 504A in its many permutations have been wonderfully supportive throughout. I'm not sure such living conditions would have been conducive to this work without the help of Deborah Gerson, Anita Stout, Katie Adams, John Choe, Jason Fuller, Mike Johnson, and Tammy Mann in capturing mice and maintaining a positive outlook about our most fascinating living situation. Special thanks go to Christa Beranek for making lots of food to share and listening patiently to my near-continuous rants about my research, my life, or ESPN"s sports center commercials. Her patience and friendship has been a remarkable anchor during the last 5 years.

Thanks go to Sophie Louvel for putting up with me during the rather stressful experience of writing up and defending this work. Her love and support helped me not only through writing this thesis, but also through the nearly impossible tasking of choosing the right career path – and with any luck, we have both made the right choices.

Finally, infinite gratitude is owed to my parents and grandparents for their love, unwavering support, and unshakable faith that whatever my research project was, I was sure to be doing the best possible job on it, no matter what anyone else said. And to my sister and best friend Erica, for helping me to keep perspective while ensuring that my ego was held in check as often as necessary, which turned out to be rather often. Surely I would not have managed any of this work without the inspiration provided by such a wonderful, intelligent, and motivated sibling.

TABLE OF CONTENTS

THESIS ABSTRACT 2
ACKNOWLEDGEMENTS 4
TABLE OF CONTENTS 6
LIST OF TABLES 8
LIST OF FIGURES 8
CHAPTER 1 INTRODUCTION 10
1.1 Modeling biological systems 11
1.2 Thesis outline 16
1.3 References 17
CHAPTER 2 PRIOR DNA MICROARRAY DATA MODELING EFFORTS 18
2.1 Assumption of deterministic model form 18
2.2 Stochastic modeling 20
2.3 Bayesian networks 22
2.4 Information theory approach 26
2.5 References 28
CHAPTER 3 EXAMPLES OF TRANSCRIPTIONAL REGULATION NETWORKS 31
3.1 Lactose/arabinose example 31
3.1.1 Transcription-only analysis 34
3.1.2 The role of experimental conditions 37
3.1.3 Conclusions 39
3.2 T7 regulatory reconstruction 40
3.2.1 The model 40
3.2.2 Model mRNA results 42
3.2.3 Analysis of mRNA data 44
3.2.4 Regulatory reconstruction 46
3.3 References 49
CHAPTER 4 COMPUTATIONAL TOOLS FOR DYNAMIC DATA 50
4.1 Time-lagged correlations 50
4.1.1 Formulation and prior work 50
4.1.2 Lactose/arabinose Example 53
4.1.3 Graphviz for correlation network visualization 56
4.2 ARX models 57
4.2.1 Model building 59
4.2.2 Choosing between models 60
4.3 Conclusions 61
4.4 References 61
CHAPTER 5 BIOLOGICAL SYSTEM AND EXPERIMENTAL PROTOCOLS 63
5.1 Cyanobacteria 63
5.1.1 *Synechocystis* PCC6803 64
5.2 Experimental protocols for *Synechocystis* 67
5.2.1 Growth and maintenance 67
5.2.2 Sample collection for RNA extraction 69
5.2.3 *Synechocystis* RNA extraction 70

5.3 DNA microarray protocols 73
5.3.1 Printing of arrays 73
5.3.2 RNA processing into labeled cDNA 74
5.3.3 Hybridization of samples 77
5.3.4 Cleaning and scanning of arrays 78
5.4 References 79
CHAPTER 6 NETWORK DISCOVERY EXPERIMENTS 81
6.1 Experiment 1: network training data 81
6.1.1 Experimental details 81
6.1.2 Time-lagged correlation implementation methodology 84
6.1.3 Results 90
6.2 Experiment 2: network validation data 97
6.2.1 Design of experiment 97
6.2.2 Experimental details 101
6.2.3 Comparison of testing and validation data 101
6.3 Network validation 106
6.3.1 Network robustness 106
6.3.2 Prediction of new profiles 112
6.4 References 116
CHAPTER 7 OTHER TOOLS FOR ANALYSIS OF MICROARRAY DATA 119
7.1 Data sets 120
7.2 Discovery of interesting variables in known classes 121
7.2.1 *P* tests 121
7.2.2 *t*-tests 122
7.2.3 Wilks' lambda 125
7.2.4 Cross-validation 127
7.3 Classification of samples 130
7.3.1 Likelihood ratio tests as classifiers 130
7.3.2 Multi-dimensional discriminant analysis 136
7.4 Statistical robustness 139
7.4.1 Power analysis 140
7.4.2 Algorithm 146
7.4.3 Implementation and results 147
7.4.4 Reliability discussion and conclusions 150
7.5 Conclusions 152
7.6 References 153
CHAPTER 8 SUMMARY AND SIGNIFICANCE OF WORK 157
8.1 Summary of thesis results 157
8.2 Significance of Results 159

LIST OF TABLES

Table 2-1: Theoretical data, three-gene network case 23
Table 3-1: Simplified model equations for lactose/arabinose regulation 33
Table 3-2: T7 mRNA events 45
Table 4-1: Simplified dynamic model equations for lactose/arabinose regulation 54
Table 4-2: Connections derived from time-lagged correlation 55
Table 5-1: BG-11 medium composition 67
Table 6-1: Correlation between sequential genes 87
Table 6-2: Genes in correlation network at appropriate time-lags 95
Table 6-3: Time-lagged correlation values, experiments 1 and 2 102
Table 6-4: Validation of the 20 best ARX models by the AIC criterion 113
Table 7-1: Discriminatory genes for the distinction of T-ALL from B-ALL samples 124

LIST OF FIGURES

Figure 1-1: Simplified diagram of *lacZYA* and *araBAD* regulation 13
Figure 1-2: T7 genome and regulatory connections 15
Figure 2-1: Stochastic petri net representation of gene activation 21
Figure 2-2: Alternative relationship hypotheses 22
Figure 2-3: Hypothetical networks for galactose example (from Hartemink *et al.*) 25
Figure 3-1: Simplified diagram of *lacZYA* and *araBAD* regulation 33
Figure 3-2: Hypothetical mRNA expression profiles for *lac* and *ara* regulation 35
Figure 3-3: Regulatory reconstruction based only on mRNA data 36
Figure 3-4: Expression and condition information 38
Figure 3-5: Reconstructed regulatory network 39
Figure 3-6: T7 genome 41
Figure 3-7: T7 regulatory structure 41
Figure 3-8: mRNA Levels for T7 Genes 0.7-3.5 43
Figure 3-9: Changes in mRNA transcription, genes 0.7 and 1.0 43
Figure 3-10: Changes in mRNA transcription, genes 1.7 - 3.5 44
Figure 3-11: Reconstruction of the first segment of T7 47
Figure 3-12: Reconstruction of T7 expression regulation 48
Figure 4-1: Glucose metabolism system (from Arkin, Shen, & Ross) 52
Figure 4-2: Glucose metabolism reconstruction (from Arkin, Shen, & Ross) 53
Figure 4-3: Graphical network derived from time lagged correlation 55
Figure 4-4: Graphviz sample input 56
Figure 4-5: Graphviz representation of lactose/arbinose example reconstruction 57
Figure 5-1: Simplified structure of *Synechocystis* PCC6803 central metabolism 64
Figure 5-2: Overview of *Synechocystis* photosynthesis machinery (from J. Barber) 66

Figure 5-3: Model of *Synechocystis* phycobilisomes (from W. Schluchter) 66
Figure 5-4: Setup of sparged-gas reactor vessel 69
Figure 5-5: Example *Synechocystis* microarray 74
Figure 6-1: Example two-channel control DNA microarray experiment 83
Figure 6-2: Example two-channel DNA microarray experiment 83
Figure 6-3: Schematic of *Synechocystis* genome section 86
Figure 6-4: Input light (solid line) and gene expression profiles, *psbEFLJ* operon 91
Figure 6-5: Network construction, first iteration 92
Figure 6-6: Network construction, second iteration, zero-lag clusters 93
Figure 6-7: Network construction, second iteration, all clusters 94
Figure 6-8: ARX model fit to the average expression profile of Group 2 97
Figure 6-9: Prediction errors *vs.* AIC criterion for ARX models 99
Figure 6-10: Group differences predicted by ARX models for a collection of inputs 101
Figure 6-11: PCA projection of all 74 samples 104
Figure 6-12: PCA loadings of all 113 genes used for PCA 105
Figure 6-13: First-iteration correlations found with $|R| \geq 0.75$ 107
Figure 6-14: Effect of reduction of cutoff to $|R| \geq 0.65$ 108
Figure 6-15: Correlations observed with shuffled time points 110
Figure 6-16: Affect of random noise on network identity 111
Figure 6-17: Reduction in cutoff R value required to compensate for random noise 112
Figure 6-18: Prediction errors for both data sets *vs.* AIC criterion 114
Figure 6-19: Predicted and validation data for Group 2 115
Figure 6-20: Comparison of predicted and validation data for Groups 2 and 8 116
Figure 7-1: Theoretical discriminatory and non-discriminatory gene distributions 126
Figure 7-2: Cross-validation scheme for selection of genes 129
Figure 7-3: Idealized, normal gene distribution, 2 classes 132
Figure 7-4: Expression distribution for gene X95735 (Zyxin) 133
Figure 7-5: Discrimination of leukemia samples with log likelihood scoring 134
Figure 7-6: Discrimination of tissue samples with log likelihood scoring 135
Figure 7-7: 2-class discrimination of leukemia samples using FDA 137
Figure 7-8: 3-class discrimination of leukemia samples using FDA 137
Figure 7-9: 3-class discrimination boundaries for leukemia samples using FDA 138
Figure 7-10: H_0 and H_1 for 2-class distinction 143
Figure 7-11: Sample size determination for 2-class distinction (AML/ALL) 143
Figure 7-12: H_0 and H_1 for 3-class distinction 145
Figure 7-13: Sample size determination for 3-class distinction (AML/B-/T-ALL) 145
Figure 7-14: Power analysis scheme for determination of sample size 147
Figure 7-15: Cross-validation results for gene selection, AML/ALL classification 149
Figure 7-16: FDA results for sample-size testing, AML/B-/T-ALL classification 150
Figure 8-1: The relationship between ARX coefficients and underlying biology 160

CHAPTER 1 INTRODUCTION

The DNA microarray has been established as a tool for efficient collection of mRNA expression data for a large number of genes simultaneously. To date, experimental design and analysis of the resulting data from microarrays have received less attention than the experimental methods themselves. Typical experimental design strategies for DNA microarrays, however, have been aimed at observing static binary differences between conditions, such as disease vs. non-disease case comparisons. This approach identifies genes with similar expression patterns, and therefore is a valuable tool for grouping annotated genes with potentially related genes and discovering transcription factor binding motifs. Furthermore, such experiments provide information for diagnostics and drug development targets, but are not well suited to uncover the roles of these genes in the larger context of cellular regulation. It is therefore desirable to expand the range of experimental strategies and tools for use with DNA microarrays to improve the information provided by this technique.

DNA microarray data is specifically suited to measure changes in cellular phenotype at the transcriptional level. Monitoring the expression levels of a large number of genes simultaneously yields insight into transcriptional regulation for a given set of experimental conditions. A relatively complete picture of transcriptional regulatory behavior should be possible by combining carefully designed experiments covering a wide range of conditions with DNA microarray assays. Because current experimental approaches usually monitor a pseudo-steady-state level of transcription, there is little retained information that might distinguish regulatory elements from the genes they affect. Dynamic experiments, on the other hand, with data taken over a series of time points, may offer insights into the way cellular phenotype is a function of changing environmental conditions. In this way the regulatory implications of changes in the system can be observed as they unfold and causal relationships can be hypothesized.

To realize this goal of increased knowledge of gene regulation from high-throughput experimental techniques, a framework is needed for both conducting maximally informative experiments and interpreting the output DNA microarray data. The genes with the most relevant transcriptional information must be culled from the total data set, and these genes must then be

placed into a framework of hypotheses about their interactions. Environmental variables, must also be considered for their impact on cellular regulation. Competing hypothesis must then be evaluated using new data or unrelated analysis techniques to rule out inconsistent hypotheses.

This thesis aims to:

- Explore the use of statistical methods for DNA microarray data to identify a subset of maximally informative genes for a given experiment,
- Develop methodologies for identifying genes that have significant changes in expression pattern which appear to be related to changes in environmental conditions,
- Construct hypothetical regulatory networks from the relationships suggested by correlational analyses,
- Apply these methods to elucidate the transcriptional programming of *Synechocystis* sp. PCC6803,
- Identify what information or experimental conditions are required to distinguish between hypothesized networks or establish their existence,
- Address all of these issues in a manner that is compatible with future high-throughput experimental data, from both DNA microarrays and other sources.

1.1 Modeling biological systems

Regulation of gene expression in prokaryotic and eukaryotic systems involves the complex interaction of a host of genes, their expressed proteins, other metabolites, and other species present in the cell. For a gene's mRNA to be expressed, an active RNA polymerase (RNAP) must bind to DNA upstream of the gene in question and transcribe uninterrupted the gene's nucleic acid sequence. Promoters, or short sequences of DNA upstream of a gene which enhance RNAP binding, are vital to this process because they dictate the affinity of RNAP for the gene in question. If other proteins bind competitively to this region, or to the RNAP itself, the transcription rate will be detrimentally changed. On the other hand, metabolites or other

species may enhance the transcription of a gene by assisting the binding of RNAP to promoters or by deactivating a competing protein (see the lactose/arabinose regulation example below).

From a modeling perspective, "perfect" understanding of transcription regulation would be the completion of a detailed set of equations which describe the transcription rates of each gene as a function of the concentrations of other chemical species present. For example, to describe change in concentration of the mRNA of some chemical species A, one may need to consider the concentration of species A, plus the concentrations of its protein product(s), plus other metabolites which might have an affect upon A's expression, etc.

$$\frac{\partial[A_{mRNA}]}{\partial t} = f([A_{mRNA}],[A_{protein}],[B_{protein}],[C_{metabolite}]\ldots) \qquad 1\text{-}1$$

To write explicitly such an equation for all of the species within a cell, three steps must be undertaken:

1) The structure of the network of interactions must be determined. That is, which chemical species affect which other species?

2) The form of each interaction must be discovered. These interactions, in general, may be non-linear and furthermore may have a significant degree of dependence on stochastic processes within the cell such as protein binding[1,2].

3) The appropriate kinetic parameters should be determined, noting that parameters may only approximate processes with significant stochastic features.

To appreciate the complexity of such an undertaking, consider as an example the regulatory networks for lactose and arabinose metabolism in *E. Coli*. Neither sugar is preferred relative to glucose, so it is valuable for the cell to shut down the use of these less efficient metabolites when not needed. However, it is also important for the cell to be able to detect the presence of lactose or arabinose in conditions of low concentrations of glucose to make the determination of how strongly to upregulate the genes in these pathways. A simplified representation of the genetic network is given in Figure 1-1[3,4]. The genes that affect metabolism directly are the operons (a series of genes with a single set of regulatory controls) *lacZYA* and *araBAD*. Metabolites

glucose, lactose, and arabinose are marked in bold and mRNA species are displayed within boxes. Protein species and complexes (marked with *) are given the capitalized names of the gene from which they are translated.

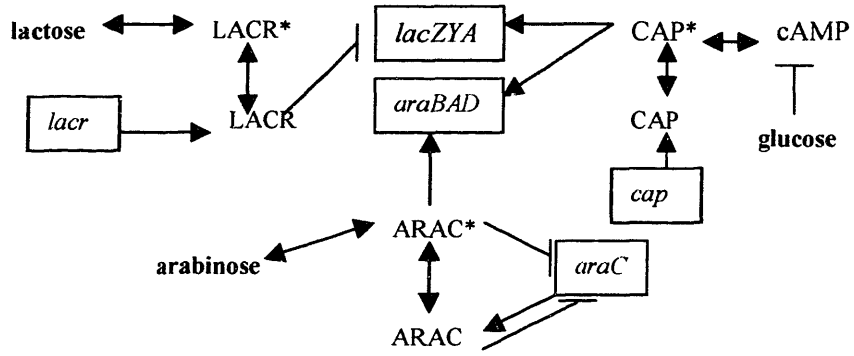

Figure 1-1: Simplified diagram of *lacZYA* and *araBAD* regulation

Figure 1-1 provides a summary representation of 30 years of careful study and countless biochemical experiments probing these interactions. For example, Figure 1-1 shows what has been discovered about the impact of glucose concentration on the regulation of both *lacZYA* and *araBAD*. As glucose concentrations decrease, cAMP levels increase, forming a complex with CAP protein. This complex then assists the transcription of both *lacZYA* and *araBAD*. In turn, both lactose and arabinose metabolism have their own set of regulatory interactions through intermediate complexes to convey information about the concentration of metabolites present.

Inspection of Figure 1-1 illustrates some of the difficulties in creating ideal models for transcriptional regulation as described above. None of the three metabolites shown affect the genes in question through the same interaction model. The wide variety of potential interaction models makes it impossible to distinguish between possibilities without tremendous amounts of data about the concentrations of each species present. DNA microarrays can provide such data; however, only concentrations of some of the members of the network can be sampled with this tool. If only the species that are boxed in Figure 1-1 can be measured, then only 5 measurements are available to characterize 16 chemical interactions. Even if the metabolite concentrations are measured, there is still not nearly enough information to uniquely determine the structure of the system.

This problem of chemical species that cannot be easily measured limits the extent to which reconstruction of the system structure is possible. The existence of co-factors, active and inactive proteins, and other complexes cannot be effectively measured *in vivo* at this time. Thus, the functional form of the deterministic equations outlined above and the parameters associated with them cannot be determined from data unless major assumptions are made about the type and form of interactions that occur. Due to this limitation, the reconstruction effort is limited to the identification of interactions (either direct or indirect) between genes with other genes, and between genes with their environment. Thus, the primary goal of this thesis is to tackle the first of the three modeling stages outlined earlier: the elucidation of the network structure by determining which genes affect which others.

Current DNA microarray analyses, including clustering studies, are often conducted near steady state and help to determine which genes are transcribed in a correlated fashion with other genes. These types of clustering studies are useful in suggesting which genes may have some sort of impact on others. On the other hand, in such studies the action of the regulatory network structure is observed only through the net results, so it is difficult to separate genes responsible for regulation from those that are affected. For example, consider *araC* and *araBAD* in Figure 1-1. If it observed that the transcription of *araC* and *araBAD* are upregulated under the same experimental conditions, clustering analysis suggests correctly that the expression of these genes is related. However, the fact that *araC* regulates *araBAD* and not the opposite cannot be determined from these types of experiments.

A more complete analysis should therefore include the regulation of dynamically evolving species. Including time-series data provides additional information and challenges as regulatory effects may be observed after some time-lag from the original event which triggered the observed event. For the arabinose example, we could theoretically determine that *araC* is up- or down-regulated *before araBAD*, suggesting the cause/effect nature of the interaction and therefore providing more detail about the structure of the network.

In some cases, transcriptional regulation studies are not only enhanced by dynamic data, but in fact cannot be reasonably conducted in any other fashion. Consider the case of T7 viral expression as shown in Figure 1-2[5]. The virus invades a host *Escherichia coli* cell when one tip

of the viral DNA penetrates the cell membrane. *E. coli* RNA polymerase (EcRNAP) recognizes a promoter on this tip and begins to pull the rest of the genome into the cell, transcribing genes as it moves along. In this way, genes are expressed sequentially as they enter the intracellular region: at the first moments of infection, none of the genes located in the rear of the genome are transcribed, but they are eventually transcribed when the infection process is complete.

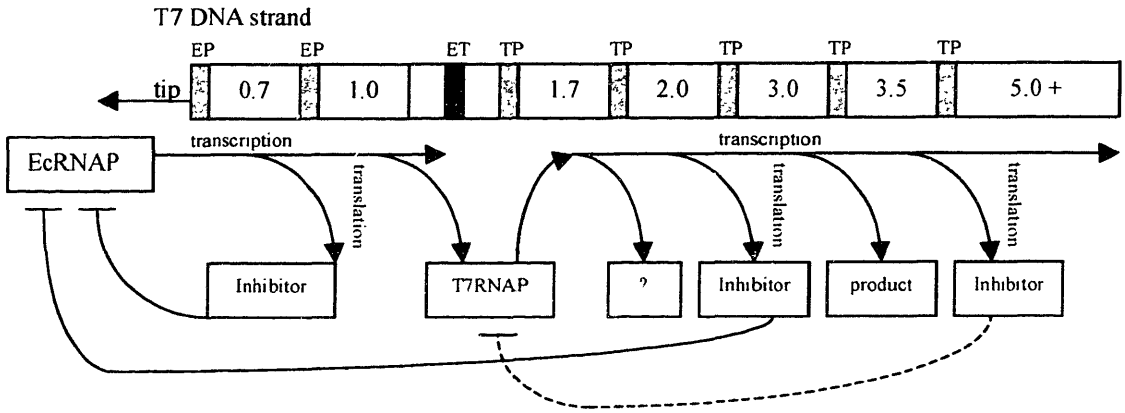

Figure 1-2: T7 genome and regulatory connections

In this figure, genes are named by sequential numbering related to their order in the genome (only a sub-set of both regulatory and structural genes have been shown here). "EP" refers to an EcRNAP promoter, while "ET" refers to an EcRNAP terminator. Gene 1.0 produces the protein for T7's own RNA polymerase (T7RNAP). EcRNAP expresses the first few genes in T7 and the remaining genes are under the influence of T7RNAP promoters marked "TP". Both RNAPs are eventually inhibited by the action of proteins that form complexes with the RNAPs and prevent further transcription. Thus, by the time infection has been completed and a large number of the inhibitory proteins have been produced, no further transcription will be observed. The sequential nature of these events, including the transcription of each gene and the eventual inhibition of these genes, means that static experiments taken out of temporal context will have no relevant information about the underlying regulation.

Here, an analysis of the features of time series data is required to deduce which events in the expression profiles mark the onset of other changes. One technique, discussed at length in this thesis, is time-series correlations. This technique can be applied to measurements of each of the mRNAs generated to answer whether or not the expression of one gene may affect the

downstream (later in time) expression of another gene. Some archetypical expression patterns may be obvious through direct observation of the expression profiles: one class of genes may behave in one way while another clearly responds in a different fashion. Additionally, genes (or classes of genes) may be consistently correlated: if the increase in the transcription rate of one gene corresponds to the down-regulation of another gene, then we may hypothesize there may be an inhibitory relationship between the two, even if it occurs through unmeasured intermediates. The case of T7 regulation shows how this may occur, as the transcription of Gene 3.5 ultimately leads to the down-regulation of all other genes due to the inhibition of the T7RNAP. This down-regulation is observable even without knowledge of the specific protein intermediate.

Ultimately, the knowledge gained from gene-expression data alone cannot complete regulatory network reconstruction, as shown by both of the examples. Unmeasured intermediate species will hamper this effort until such a time as a wider spectrum of measurements can be taken. However, development of methodologies to maximize the useful information that can be gained from DNA microarray data can help to frame the network of interactions, and such outlines can be used to probe the system further.

1.2 Thesis outline

This thesis focuses broadly on computational tools relevant to the effect use of DNA microarrays. Methods for effectively handling not only data from current experimental strategies as well as data from the dynamic studies proposed in this work are examined in detail. However, to maintain a single, cohesive line of discussion, other computational tools developed during this work are saved for the final chapter as a related but self-contained body of work. The bulk of the thesis focuses on the tools required to elucidate the light-dependent transcriptional network structure of a model organism (*Synechocystis* PCC6803) through dynamic DNA microarray experiments.

Chapter 2 reviews prior modeling efforts of DNA microarray data. The two illustrative examples introduced in this chapter are then discussed in greater detail in Chapter 3 to demonstrate the possibilities and limitations for understanding transcriptional regulation under a variety of different experimental situations. These examples are used as the basis for an analysis framework based on time-lagged correlations and AutoRegressive with eXogenous input (ARX)

models, explained in detail in Chapter 4. The biology and features of the model system of *Synechocystis* is discussed in Chapter 5, along with protocols for culture maintenance and DNA microarray experiments. Chapter 6 shows the specific application of the proposed computational techniques experiments on the model system in order to elucidate transcriptional regulatory features. This chapter also includes a discussion of the derived network features, robustness, and likelihood of making false conclusions from the experimental data. Finally, Chapter 7 discusses computational tools applicable to a wider range of DNA microarray data, including discriminatory gene selection, sample classification, and sample size required for statistically significant conclusions from microarray data. Chapter 8 summaries the findings of this work as a whole.

1.3 References

1. McAdams, H. H. & Arkin, A. "Stochastic mechanisms in gene expression." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **94**, 814-819 (1997).

2. Goss, P. J. E. & Peccoud, J. "Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 6750-6755 (1998).

3. Voet, D. & Voet, J. G. *Biochemistry* (John Wiley & Sons, Inc., New York, 1995).

4. Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P. & Darnell, J. *Molecular Cell Biology* (Scientific American Books, New York, 1995).

5. Endy, D., Kong, D. & Yin, J. "Intracellular kinetics of a growing virus: A genetically structured simulation for bacteriophage T7." *Biotechnology and Bioengineering* **55**, 375-389 (1997).

CHAPTER 2 PRIOR DNA MICROARRAY DATA MODELING EFFORTS

It has been pointed out that both protein and mRNA expression data is required for the mechanistic analysis of gene expression networks[1,2]. Even without such protein data, several research groups have attempted to lay out frameworks for the consideration of DNA microarray data as a partial solution of the gene regulation problem. For the most part, prior modeling efforts have been hamstrung by a lack of properly designed experimental data with which to derive parameters for the proposed models or test their conclusions. Nevertheless, these attempts explicitly connect transcriptional data to underlying biological phenomenon, and therefore all offer to add insight into cellular regulation at the transcriptional level. Some major schools of thought are presented below.

2.1 Assumption of deterministic model form

Some authors have suggested "reasonable" models of genetic interaction, to which DNA microarray data should be fit. Simple linear relationships and more complicated functions have been proposed to uncover relationships from both static and time-series data. For genes whose transcription can be accurately expressed by such functions, these strategies allow for explicit parameter estimation, at the cost poor data fit for other genes.

The most far-reaching of these attempts was made by Chen *et al.*[1], who suggested a linear transcription model where each mRNA and protein species present is linearly dependent on the concentration of the other mRNA and protein species in the cell:

$$\frac{d\mathbf{r}}{dt} = C\mathbf{p} - V\mathbf{r} \qquad \frac{d\mathbf{p}}{dt} = L\mathbf{r} - U\mathbf{p} \tag{2-1}$$

where $\mathbf{r}$ is the vector of mRNA concentrations and $\mathbf{p}$ is the vector of protein concentrations. C and L are the rates constants for transcription and translation, respectively. Likewise, V and U specify degradation rate constants. Given enough quality, dynamic measurements of both $\mathbf{r}$ and $\mathbf{p}$ and the assumption that both degradation terms U and L are diagonal matrices (*i.e.* degradation of a chemical species is only a function of the concentration of that species), the authors show that Equation 2-1 can be rearranged to determine completely the rate matrices C, V, L, and R.

Furthermore, they present an extension of this model to the case where **p** is not measured dynamically. In this way they seek to address the reconstruction problem from microarray data alone, if functional form is known. Specifically, **p** in Equation 2-1 can be eliminated by substituting $\mathbf{p} = e^{-Ut}\mathbf{p}_1$ they calculate

$$L\mathbf{r} - U\mathbf{p} = \frac{d\mathbf{p}}{dt} = -Ue^{-Ut}\mathbf{p}_1 + e^{-Ut}\frac{d\mathbf{p}_1}{dt} = -U\mathbf{p} + e^{-Ut}\frac{d\mathbf{p}_1}{dt} \qquad 2\text{-}2$$

and therefore

$$\frac{d\mathbf{p}_1}{dt} = e^{Ut}L\mathbf{r} \quad \Rightarrow \quad \mathbf{p} = e^{-Ut}\int e^{Ut}L\mathbf{r}dt \qquad 2\text{-}3$$

Plugging this solution for **p** back into Equation 2-1 gives, after some manipulation, a dynamic expression for transcript concentration **r** independent of other measurements

$$\frac{d^2\mathbf{r}}{dt^2} = \left(-CUC^{-1} - V\right)\frac{d\mathbf{r}}{dt} + \left(-CUC^{-1}V + CL\right)\mathbf{r} \qquad 2\text{-}4$$

where C^{-1} indicates the inverse of C. Inspection of this relationship, however, shows that the transcription matrix C is degenerate, and thus any solution depends on the initial value of **p**. All of this analysis assumes that the relationships between the transcription/translation rates and the RNA/protein concentrations are indeed linear.

Models similar to this one in concept, if not the specific model, have also been presented by Weaver *et al.*[3] and D'Haeseleer *et al.*[4]. From the standpoint of modeling, the choice of "reasonable" assumptions for the form of interactions may yield models which are satisfactorily similar to the actual system. However, attempting to simplify the systems present or fit them to a modeling scheme, real biological interactions may be obscured. As an example, Chen *et al.* assumes that both mRNA and protein degradation rates (V and U) are diagonal and therefore the degradation of each chemical species is dependent only on its own concentration. Although this assumption simplifies the calculations, these degradations may be a function of the concentration of proteins that digest these species or other proteins that stabilize them, which would cause matrices V and U to be non-diagonal. Because there is no room for such interactions in the current solution, the regression of parameters in the simplified case will give unsatisfactory

results. Even with such limitations, these methodologies offer computationally tractable problems with explicit inclusion of known biological information when available.

2.2 Stochastic modeling

Another problem with a purely deterministic approach is that it ignores the stochastic nature of some biological processes[5-7]. Because some species within the cell exist in very small concentrations, all of the biochemical processes involving these species are dependent on Brownian motion and therefore have some characteristics of random processes. Such distinctions may play a very important role in some systems where the state of the system is defined by a stochastic first step. One particularly well-studied example is that of λ-phage infected *Escherichia coli* cells[6,8]. Whether an infected cell enters the lysogenic or lytic mode is dependent on the binding of a few key species (such as Cro protein and cI repressor) to the phage genome immediately after infection. The exact sequence of events (*i.e.* which proteins bind to the DNA first) determines whether the lytic mode is initiated or if a self-sustaining lysogenic cycle is started.

The difficulty with attempting a rigorous formulation in such cases is that new parameters are necessary to describe the stochastic nature of the system. Arkin *et al.*[6] present a framework for modeling such systems using stochastic kinetic models as an extension of differential equations similar to those shown in section 2.1. As a test case, λ-phage infected *E. coli* systems were simulated and compared to experimental results from other sources. The authors were able to predict the percentages of lysogenic and lytic cells for populations of cells at a range of conditions. On the other hand, deterministic models are only capable of predicting whether the average cell in a population would enter the lytic or lysogenic cycle, and thus fail to model the overall behavior of, for example, a cell culture. Although results showed that this stochastic formulation best represented the actual situation, the simulation was effective only because a large body of knowledge on the individual probabilities of each event was known. Such an approach is therefore unrealistic for exploring poorly-understood systems.

Another way to capture the state of the molecular species governed by stochastic processes is that of *Stochastic Petri Nets* (SPNs). SPNs are a collection of *places* (chemical species), tokens

(number of molecules in a *place*), and *transitions* (chemical reactions). Transitions become "active" only when certain criteria are met and then move *tokens* from one *place* to another with an exponentially distributed time delay. When the number of *tokens* becomes small, the stochastic nature of a *transition* is important: as the number of *tokens* approaches infinity, this simplifies to a deterministic system[7]. This formulation can be expanded to include continuous factors if appropriate, in the form of *hybrid Petri Nets* (HPNs)[9].

Consider the model shown in Figure 2-1 (adapted from Goss and Peccound[7]) for a simplified model of gene expression. If the gene is not active, *i.e.* no promoters are bound upstream of it, then a *token* exists in position 1. This *token* may transit to the active *place* (position 2) through the rate decay parameter λ. From position 2 it may either return to the inactive state through inactivation (μ) or it may generate a mRNA *token* (ν) and return itself to the active state. In this way, the active gene may transcribe many mRNA *tokens* at position 3, which themselves may be degraded via transition δ. Thus only species that have many molecules (*tokens*) may be simplified in the limit of infinite copies to purely deterministic representations. Deterministic representations therefore fall short of perfect modeling of any system where there are only a few copies of relevant species present.

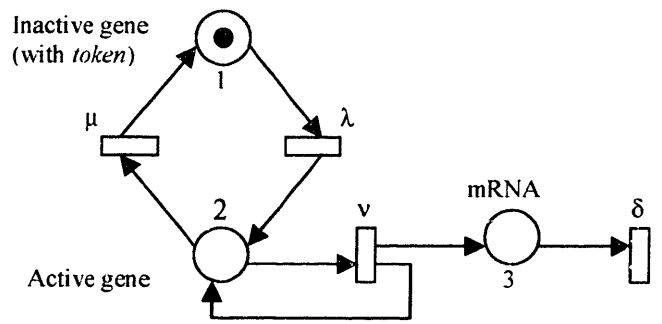

Figure 2-1: Stochastic petri net representation of gene activation

Petri nets have been used successfully to represent the λ-phage infected *E. coli* lysogeny/lysis switch[9] as well as ColE1 plasmid replication[7]. Although techniques exist for evaluating the steady-state or transient behavior of the Markov chains generated by such systems, Petri net methodologies are still based on *modeling* rather than *discovery*. As such, they are very useful for hypothesis testing and representation, but they still do not directly examine the genetic relationships that are not hypothesized *a priori*.

2.3 Bayesian networks

If details about needed for mechanistic modeling of microarray expression is not available, then statistical tools must be applied to hypothesize the existence of connections between chemical species within the regulatory networks. For example, the well-studied statistical tools associated with *Bayesian networks* can be used to test the dependence and conditional independence in data[10]. Given multiple measurements, these tests can be used to evaluate the likelihood that a measurement is conditionally dependent on another. If the expression levels of two genes show strong interdependence, we hypothesize the existence of underlying biochemical relationships connecting the two.

Consider the case of three genes labeled g_1-g_3 (adapted from Jaakkola[11]). These genes might be unrelated (independent), or there may be some inter-dependence. If we hypothesize that g_3 is conditionally dependent on g_1, then this case should be compared to the null hypothesis to determine if this additional relation better describes collected data in a statistically significant way. These two hypotheses can be drawn as graphs, with the null hypothesis (no relation) labeled as H_0 and the alternative as H_1, as shown in Figure 2-2.

Figure 2-2: Alternative relationship hypotheses

For some data set X, the *likelihood ratio statistic* comparing the two models' ability to describe the data is given as

$$T(X) = 2\log_2 \frac{\hat{P}(X|H_1)}{\hat{P}(X|H_0)} \sim \chi_v^2 \qquad 2\text{-}5$$

where ν is the difference in degrees of freedom between model H_1 and H_0 and χ_ν^2 is the chi-square distribution with ν degrees of freedom. If DNA microarray data were to be taken for all

three genes, then we could evaluate the likelihood ratio by using the expression intensities to determine the relative probabilities of the two hypotheses.

Specifically, microarray data can first be quantized into discrete levels of expression to provide finite degrees of freedom. If all data is reduced to 0's and 1's (corresponding to inactive or activated genes, for example, or "high" expression and "low" expression, as determined by the experiment), then in the null case there are 3 degrees of freedom ($P(g_1=0)$, $P(g_2=0)$, $P(g_3=0)$ – all $P(g_i=1)$ probabilities are the complements of these), while in the hypothesized case, there are 4 degrees of freedom ($P(g_1=0)$, $P(g_2=0)$, $P(g_3=0|g_2=0)$, $P(g_3=0|g_2=1)$). These graphs and their underlying probability distributions are called a *Bayesian network.* If, for instance, 100 experiments were conducted, then the probability of finding $g_3 = 0$ can be calculated in addition to the probability that $g_3 = 0$ when $g_2 = 1$. Consider the hypothetical data presented in Table 2-1, corresponding to the example in Figure 2-2.

Table 2-1: Theoretical data, three-gene network case

n = 100	Observed values				Expected counts	
	g1	g2	g3	counts	H_0	H_1
Experimental profiles	0	0	0	19	11.25	20.00
	0	0	1	6	13.75	5.00
	0	1	0	2	11.25	2.50
	0	1	1	23	13.75	22.50
	1	0	0	21	11.25	20.00
	1	0	1	4	13.75	5.00
	1	1	0	3	11.25	2.50
	1	1	1	22	13.75	22.50
				100	100	100
counts, gi = 1	50	50	55			
counts, g3 = 1|g2=1			45			
counts, g3 = 1|g2=0			10			

The expected counts for each of the eight experimental profiles ($[g_1,g_2,g_3]$ as [0,0,0], [0,0,1], *etc.*) give the probabilities for each gene's expression: $P(g_1 = 1) = 0.5$, $P(g_3 = 1) = 0.55$, *etc.* For the H_0 network, the three genes are assumed to be independent, so the expected probability of measuring [0,0,0] = $P(g_1 = 0)$ x $P(g_2 = 0)$ x $P(g_3 = 0)$ = (0.5) x (0.5) x (.45) = 0.1125, corresponding to 11.25 expected observations. On the other hand, for H_1, g_2 and g_3 are not independent, so the probability of measuring [0,0,0] = $P(g_1 = 0)$ x $P(g_2 = 0)$ x $P(g_3 = 0|g_2 = 0)$ = (0.5) x (0.5) x (40/50) = 0.20, or 20 expected observations. Simple inspection of Figure 2-1 the

second model fits the observations much better, but the likelihood ratio test allows us to quantify how much better the hypothesized network fits the data – in this case, confidence in the H_1 hypothesis relative to the null is much greater than 99%, allowing us to reject the null hypothesis.

This framework gives a rigorous methodology for comparing hypothesized networks statistically. This methodology is particularly attractive because it prevents overfitting of the data by demanding statistical significance to every relationship added to the system (*ie.* each increase in the degrees of freedom). Recall from section 2.1 that in the differential modeling case presented by Chen *et al.* all parameters would be fit if enough data existed, with no explicit means of determining whether those parameters had any intrinsic value or merely modeled experimental noise.

This technique has been used to compare possible networks of transcriptional regulation of galactose metabolism genes in *Saccharomyces cerevisiae*[12]. Two hypothetical diagrams of the regulatory structure affecting Gal2, Gal5, and Gal7 transcription are shown in Figure 2-3. Both suggest that the gal4 gene is transcribed into mRNA (Gal4m) that is then translated into the corresponding protein (Gal4p), which in turn regulations the transcription of gal2, gal5, and gal7. The two models differ, however, by the action of Gal80 protein (Gal80p). M1 assumes that Gal80 protein affects the transcription of Gal4, while M2 assumes that Gal80 protein interacts directly with the Gal4 protein. Therefore, if the probability of observing a given level of transcription for Gal4 is calculated to be dependent on the transcriptional level of Gal80 in a statistically significant way, M1 should be accepted over M2.

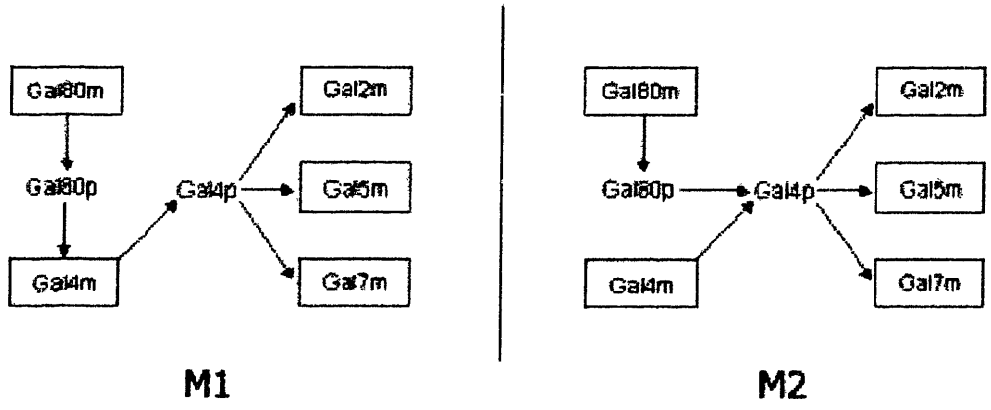

Figure 2-3: Hypothetical networks for galactose example (from Hartemink *et al.*)

The authors discretized 52 experiments worth of Affymetrix DNA microarray data from experiments on *S. cerevisiae* into 0's and 1's. Specifically, the distribution of data for each gene individually was used to create a cutoff above which all values were simplified to 1's, and below which all values were set to 0's. The conditional probabilities were then calculated for the expression level of each gene, assuming the interactions suggested by M1 and M2. The likelihood ratio test was then used to compare the two models, and it was discovered that that the M2 model is approximately 13,000 times more likely than M1. This model currently enjoys more widespread acceptance than the model originally proposed by biologists (M1)[12].

Note that in this example, potential networks were enumerated beforehand, but in many cases this may not be reasonable if prior knowledge does not exist. If used for discovery, a major difficulty with this method is that there are many possible networks to consider. For example, in the case of three genes shown in Figure 2-2 above, there are 12 different *equivalence classes*, or models that are indistinguishable because they make the same independence assumptions. For example, the case where "g_1 affects g_2" is in an equivalence class with the case where "g_2 affects g_1" because they both make the same set of interdependence assumptions and therefore have both equal degrees of freedom and related parameters. This equivalence class is distinct from the classes "g_1 affects g_3", "g_1 and g_3 affect g_2", *etc.* Testing each of these classes against the others to find the best-fit class becomes computationally challenging as the number of measured variables becomes large. Therefore heuristics are commonly employed: the number of possible

inputs to a node may be limited, for example, or a local maximum based on searching a sub-set of candidates may be found[13].

Another challenge is that of distinguishing *causal* relationships from co-dependence. In a Bayesian network, $X \rightarrow Y$ is in an equivalence class with $Y \rightarrow X$: based on the measurements alone we may not always be able to tell which event is the cause and which is the effect[10]. Because the dynamic nature of the system is not taken into account in this modeling framework as discussed here, genes that cause an effect in other genes dynamically are not identified. Thus determining all members of an equivalence class is the most information that can be gained from this methodology, in the absence of other experiments (such as knockouts) that can distinguish between the two possibilities. Adjustments may be made for dynamic data, but the number of potential hypotheses to compare increases proportionally to the number of time-offsets to be considered (*i.e.* comparison of gene 1 with gene 2 at the same time point, at one point later, at two points later, *etc.*). Given the already large number of potential networks to consider, the addition of time to each interaction may create an unrealistically complex problem.

Finally, attempts to use this framework to date have involved the discretization of data into 2 or more levels. Extensions of the methodology to include nearly continuous representation (*i.e.* many levels) would require extremely large amounts of data to accurately determine the conditional probabilities. Alternatively, a distribution could be assumed to describe each gene's expression. However, DNA microarray data to date has not suggested simple distributions for most genes[14-19], so conclusions drawn from assumed distributions will be suspect.

Even with these limitations, Bayesian networks provide a statistically sound method of comparing hypothetical networks with experimental data and rigorously determining which networks match the data more clearly than others.

2.4 Information theory approach

Another way of looking at the dependence of two genes is to calculate the amount of unique information gained by observing both. If knowing only one of the gene expression profiles gives equivalent information to knowing both, then the transcription of these genes may be connected, and we can hypothesize that some biochemical relationship or series of interactions exist.

Information theory makes use of *Shannon entropy* (H) as a quantitative measure of mutual information content. It is defined in terms of the probability (p) of observing a particular state (i) from a set of total states (n) for a collection of (k) variables (in this case, genes) as[20]

$$H(g_1, g_2, \ldots g_k) = -\sum_i^n p_i \log_2 p_i \qquad 2\text{-}6$$

This function is maximized when all (n) states are equiprobable, $H_{MAX} = \log_2(n)$. For example, if there are two unrelated genes (k = 2) which can each take the expression values "0" or "1" then n = 4 for this set of genes: [0,0],[0,1],[1,0],[1,1]. Since each state is equiprobable, $H(g_1,g_2) = \log_2(4) = 2 = H_{MAX}$. However, if the genes have some co-dependency, then all states are not equiprobable. For example, if one gene is only expressed under the same conditions as the other is expressed (*i.e.*, g_1 expression = g_2 expression) then only two of the four possible states are expressed: [0,0],[1,1], giving $H(g_1,g_2) = \log_2(2) = 1 < H_{MAX}$.

To establish relationships between genes, the *mutual information* (M) can be calculated as

$$M(g_1, g_2) = H(g_1) + H(g_2) - H(g_1, g_2) \qquad 2\text{-}7$$

In other words, mutual information is an expression of the information contained in each gene individually minus the information contained in their intersection. For the unrelated gene example, M = 0. On the other hand, when g_1 expression = g_2 expression then M = 1, so mutual information exists. In fact, since $M = H(g_1) = H(g_2)$ then the mutual information is *maximal*: knowledge of one is sufficient to know the value of the other[20].

Butte and Kohane[21] used this methodology to examine all pair-wise relationships in *S. cerevisiae* data from Stanford[22]. The *S. cerevisiae* study included 79 separate DNA microarrays were performed covering cells in a variety of conditions, such as different points in the cell cycle. In order to discretize the gene expression data, a histogram was drawn for each gene individually and divided into 10 evenly-spaced bins, with $p(g_i, bin_n)$ = # of measurements in bin_n/total number of measurements for g_i. It was then hypothesized that genes with higher mutual information values are more likely to be biologically correlated. The authors then selected cutoff value (by comparison of randomized data to the original data set) of M > 1.3 to distinguish only "highly correlated" genes. Note that in this example M > 1 is possible due to the 10 bins chosen, in

contrast to the example above where only 2 bins were selected. These highly correlated genes were then clustered into 22 "Relevance Networks" of genes with no inter-network connections. The relationships between genes within each network were then examined.

Liang *et al.*[23] extended this formulation to a methodology that includes connections beyond pairwise interactions. Using Boolean characterizations of all genes in a test set, the mutual information between each input (I) and output (O) is compared to the information content of that output (O). If the mutual information of the pair M(O, I) is equal to the output information H(O), then the input completely determines the output. After describing as many outputs as possible with single inputs, their algorithm, called REVEAL[23], attempts a pairwise analysis for the remaining outputs. If all pairs of inputs also fail to describe an output, the algorithm moves to triplets, *etc.* In this way, the program avoids calculating all possible combinations of inputs and outputs at each iteration while still attempting to explore all possible input explanations of each output.

Although information theory outlines a rigorous framework for generating gene interactions, it has some of the many of the same difficulties as the Bayesian approach. Both systems require discretized data for reasonable determination of the probabilities required in their calculations. Furthermore, neither considers the impact of time in their current formulations, and time-based adjustments will involve further increases in computational complexity. Nonetheless, both Bayesian networks and information theory provide some insight into the genetic regulatory problem.

2.5 References

1. Chen, T., He, H. L. & Church, G. M. "Modeling gene expression with differential equations." *Pacific Symposium of Biocomputing* Hawaii, (1999).

2. Hatzimanikatis, V. & Lee, K. H. "Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information." *Metabolic Engineering* **1**, 275-281 (1999).

3. Weaver, D. C., Workman, C. T. & Stormo, G. D. "Modeling regulatory networks with weight matrices." *Pacific Symposium of Biocomputing* Hawaii, (1999).

4. D'Haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. "Linear modeling of mRNA expression levels during CNS development and injury." *Pacific Symposium of Biocomputing* Hawaii, (1999).

5. McAdams, H. H. & Arkin, A. "Stochastic mechanisms in gene expression." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **94**, 814-819 (1997).

6. Arkin, A., Ross, J. & McAdams, H. H. "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." *Genetics* **149**, 1633-1648 (1998).

7. Goss, P. J. E. & Peccoud, J. "Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 6750-6755 (1998).

8. Voet, D. & Voet, J. G. *Biochemistry* (John Wiley & Sons, Inc., New York, 1995).

9. Matsuno, H., Doi, A., Nagasaki, M. & Miyano, S. "Hybrid Petri net representation of gene regulatory network." *Pacific Symposium of Biocomputing* Hawaii, (2000).

10. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. "Using Bayesian Networks to Analyze Expression Data." *Fourth Annual Inter. Conf. on Computational Molecular Biology* Tokyo, (2000).

11. Jaakkola, T. "Beyond Clustering - Graph Models." *Course 6.892 Computational and Functional Genomics, Lecture 12* MIT, (Spring 2000).

12. Hartemink, A. J., D.K. Gifford, T.S. Jaakkola, R.A. Young. "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks." *Pacific Symposium on Biocomputing* Hawaii, (2001).

13. Friedman, N., Goldsmidt, M. & Wyner, A. "Data Analysis with Bayesian Networks: A Bootstrap Approach." *Fifteenth Conf. on Uncertainty in Artificial Intelligence* Stockholm, (1999).

14. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. "Quantitative Monitoring Of Gene-Expression Patterns With a Complementary-Dna Microarray." *Science* **270**, 467-470 (1995).

15. DeRisi, J. L., Iyer, V. R. & Brown, P. O. "Exploring the metabolic and genetic control of gene expression on a genomic scale." *Science* **278**, 680-686 (1997).

16. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herskowitz, I. "The transcriptional program of sporulation in budding yeast." *Science* **282**, 699-705 (1998).

17. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." *Molecular Biology Of the Cell* **9**, 3273-3297 (1998).

18. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring." *Science* **286**, 531-537 (1999).

19. Schmitt, W. A. & Stephanopoulos, G. "Prediction of transcriptional profiles of *Synechosystis* PCC6803 by dynamic autoregressive modeling of DNA microarray data." *Biotechnol Bioeng* **submitted** (2003).

20. Somogyi, R. & Fuhrman, S. "Distributivity, a general information theoretic network measurement, or why the whole is more than the sum of its parts." *The International Workshop on Information Proces.. ng in Cells and Tissues 1997* Sheffield, (1997).

21. Butte, A. J. & Kohane, I. S. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pacific Symposium on Biocomputing* Hawaii, (2000).

22. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. "Cluster analysis and display of genome-wide expression patterns." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 14863-14868 (1998).

23. Liang, S., Fuhrman, S. & Somogyi, R. "Reveal, a general reverse engineering algorithm for inference of genetic network architectures." *Pacific Symposium on Biocomputing* Hawaii, (1998).

CHAPTER 3 EXAMPLES OF TRANSCRIPTIONAL REGULATION NETWORKS

To appreciate the level of complexity inherent to dynamic regulatory reconstruction problems, the two examples presented in the Introduction (Chapter 1) are presented in greater detail. In both case, models of transcription were created to approximate the response of each system to certain environmental situations. This data was then analyzed independently of the underlying modelsto determine how much of each underlying model's structure could be uncovered from data alone. Each example highlights some of the opportunities and challenges with reconstructing gene regulatory networks.

In the first example, we consider the regulatory networks of lactose and arabinose studied under different extracellular conditions. This study shows how incomplete understanding or consideration of relevant environmental variables hamstrings the reconstruction effort, and demonstrates that transcriptional data in a vacuum of other information is insufficient to gain more than cursory insight into cellular function. For this example, time is mostly ignored, and its only role is to distinguish experimental conditions.

For the second case, a model of the expression pattern for T7 infected *E. Coli* cells was studied. Because all of the transcriptional events are specifically sequential for T7 infection, static experiments hold almost no information about the regulation of the infection process. This study shows how very fine time-series data with many points lends itself to a host of time-series analysis methods, but still requires that some external knowledge be applied for more complete reconstruction.

3.1 Lactose/arabinose example

Attempts to infer information from a variety of experimental conditions are at the heart of many types of gene expression analysis, especially information theory approaches[1,2] (as discussed in section 2.4). The implicit assumption is that if enough unique combinations of environmental conditions are studied, then correlated (or anti-correlated) genes can be identified from among other genes. As a model for this type of analysis, the regulation of metabolism genes for both lactose and arabinose were used to create a simplified, deterministic model of transcriptional

expression. The gene expression levels predicted by the system were adopted to show the merits and limitations of an essentially time-independent rule-based model for uncovering regulatory information.

The preferred energy source for *E. Coli* is glucose. However, other sugars such as lactose and L-arabinose are utilized when present during conditions of low glucose concentration. In this example, the concentrations of glucose, L-arabinose, and lactose were manipulated as inputs to the model, representing in the external environment of a hypothetical *E. Coli* cell. The model was built to take into account only the *lacZYA* operon as described in Lodish *et al.*[3] and the *araBAD* operon as described in Voet and Voet[4]. It shows how the genes of these operons would be expressed if changes in the concentrations of glucose, arabinose, and lactose were altered assuming no other part of cell physiology was altered. The model was based on Figure 3-1 (a copy of Figure 1-1 from Chapter 1), with a few key assumptions made for modeling purposes:

- The time scale considered was assumed to be large compared to transcription and translation processes. Thus, cellular information is transferred instantly: time lag is assumed to be negligible for transcription and translation to occur. During an "experiment" the extracellular conditions were changed to cause changes in the expression patterns.

- The level of *cap* mRNA and *lac repressor* mRNA (*lacr)* were assumed not to be regulated: that is, these genes were assumed to be always transcribed at some basal level.

- Complex formation for both lac repressor and CAP/cAMP was assumed to be very favorable and dependent only on which of the two species in question was limiting.

- The concentration of AraC was assumed to be constant but partitioned between the free and complexed states, dependent only on the level of arbinose present. In reality, the concentration of AraC is self-repressing and should lead to an oscillatory expression profile. For the time scale considered, it was assumed that these oscillations would be minor, so that only the amount of arabinose present would determine how much araC is transcribed.

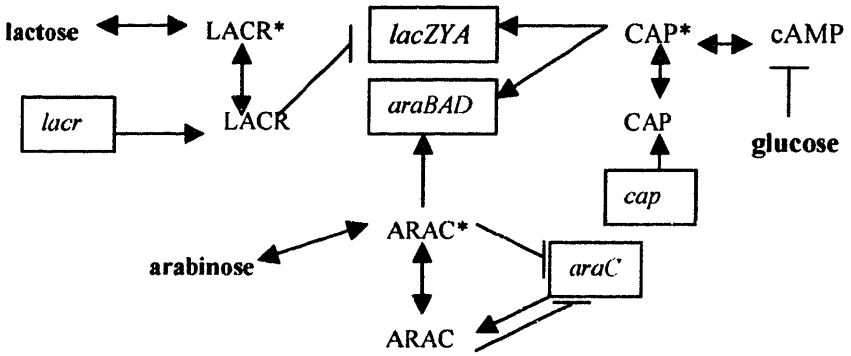

Figure 3-1: Simplified diagram of *lacZYA* and *araBAD* regulation

The equations used are listed below – note that the equations are not written to exactly match interactions listed in Figure 3-1, but rather are written to best approximate the relationships in as simple a form as possible. Note also that a square-relationship has been proposed for the concentrations of *lacZYA* to add more a complicated relationships to the proportionalities otherwise proposed below.

Table 3-1: Simplified model equations for lactose/arabinose regulation

(all concentration scales and rate constants arbitrary)

Lactose, arabinose, and glucose: manually varied in the range (0, 10)

cap, CAP, *lacr*, and LACR fixed at (5)

cAMP = 10 * 1/glucose

CAP* = min(CAP, cAMP)

LACR* = min(LACR, lactose)

$lacZ = lacY = lacA = 10\,(CAP^*/5)\,(LACR^*/5)^2$

ARAC* = 10*arabinose

ARAC = 10 – ARAC*

araC = 1/ARAC*

$araB = araA = araD = 10\,(CAP^*/5)\,(ARAC^*/10)$

3.1.1 Transcription-only analysis

The model was used to predict the transcriptional profile of *E. Coli* for a variety of experimental conditions that represent combinations of the three input sugars. These specific combinations of sugars were ignored for the transcription-only analysis, in order to determine how much of the regulatory network can be reconstructed without knowledge of the relevant environmental variables. This is analogous to the approach used by prior authors where a compendium of transcriptional data is analyzed without consideration of the specific experimental conditions, as discussed in section 2.3 and section 2.4. For the specific sugar concentrations used to create these profiles, see section 3.1.2.

The first step in the analysis of data generated from this computational experiment of varying sugar concentrations was to determine which outputs' effects were indistinguishable from others. This was achieved by clustering the gene expression data. In this experiment *lacZ*, *lacY*, and *lacA* have exactly the same profile and were therefore indistinguishable, since they occur in an operon. Likewise *araB*, *araA*, and *araD* have the same expression profile. In a real system, these profiles might not be exact matches but presumably would be too similar to distinguish given the noise expected in DNA microarray experiments. Note that *cap* and *lacr* can also be clustered according to this method in our model expression data as neither is regulated by outside factors.

Using any of a host of well-studied clustering methods[5,6], the genes *lacZ*, *lacY*, and *lacA* can be easily collected into a single group (*lacZYA* from here on) as well as some of the *ara* genes (*i.e. araBAD*). This grouping simplifies the picture to a smaller sub-set of genes to be considered. Next, these mRNA profiles were compared to see if the gene expression patterns revealed any potential cause/effect relationships. Figure 3-2 shows these profiles, with the profile for *araC* shown on the right axis because it differed significantly in scale from the other genes.

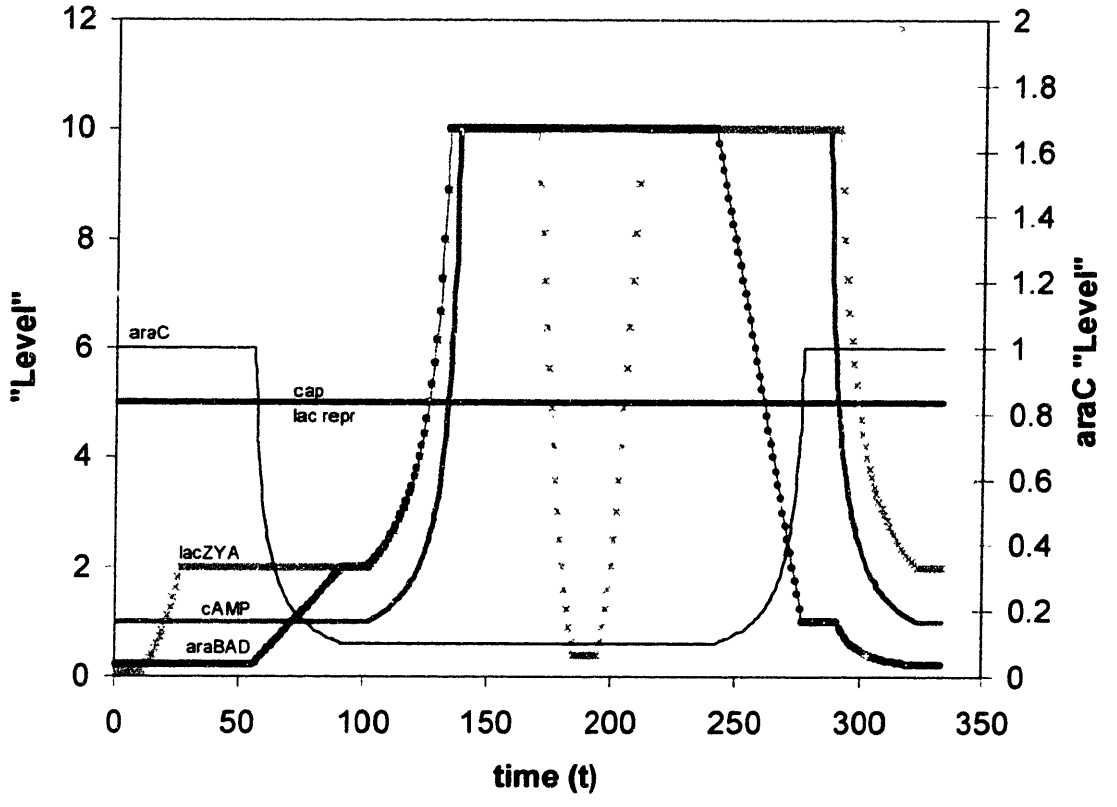

Figure 3-2: Hypothetical mRNA expression profiles for *lac* and *ara* regulation

Simple inspection revealed some key features of Figure 3-2. First, since their expression profiles were completely flat, it is apparent that neither *cap* nor *lac* repressor were affected at the experimental time-scale by any of the other genes present under these experimental conditions. Second, if *cap* or *lac* repressor had any affect on the other genes present, there is no event in the expression profiles that would uncover this relationship. These two genes were therefore put into a group with questionable impact on the rest of the system, as depicted in Figure 3-3. The remaining genes were then considered separately.

Because the time scale in this example was only used to distinguish experimental conditions (that is, rates of changes are without true physiological meaning), analysis of each event that occurred for a given gene can be limited to comparison with events that occurred at the same time for other genes. For example, analysis of the *lacZYA* expression profile change beginning at "time" t = 10, where no other genes showed a correlated change, revealed that *lacZYA* seems to have no

effect on the other species present. Thus the correlation of *lacZYA* with other genes in this experimental window would be very low, using any of the techniques described in Chapter 2.

The next event started at about t = 55, where *araBAD* was upregulated and *araC* was downregulated. Here either a) *araBAD* might repress *araC* or b) *araC* might repress *araBAD* or c) some third factor might affect both. How can these cases distinguished? A clue exists at point t = 100, when *araBAD* expression was upregulated while *araC* remained unchanged. Therefore the flow of information appears to be one way: that is, *araBAD* expression does not impact *araC* expression, which ruled out option "b" (see Figure 3-3). This shows the importance of looking at a variety of experimental conditions to distinguish relationships.

The final gene to consider was *cAMP*. This gene was upregulated at t = 100 with both *araBAD* and *lacZYA*. Similarly, all three were downregulated at t = 280. Since it has already been established that *cAMP* expression does not necessarily change when *araBAD* and *lacZYA* expression levels do, it seems that *cAMP* may have participated in the upregulation of both. This information, as well as all other conclusions reached up to this point, is summarized in the graph shown in Figure 3-3.

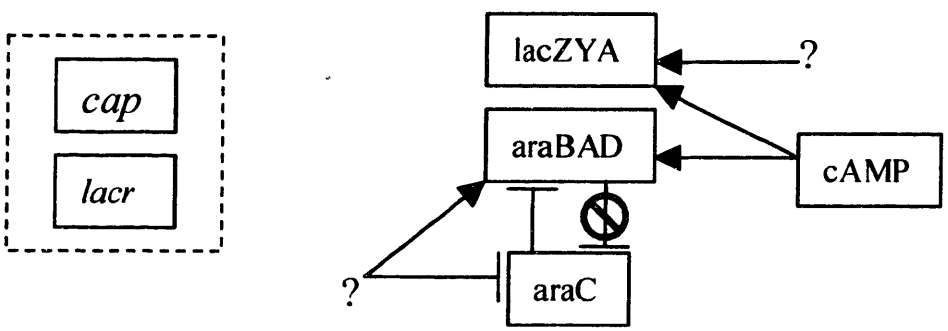

Figure 3-3: Regulatory reconstruction based only on mRNA data

The graph in Figure 3-3 has some features in common with Figure 3-1, but is missing some of the relationships. On the other hand, Figure 3-3 also includes some spurious relationships that need to be eliminated from the graph. Therefore other information will be required to refine the network.

3.1.2 The role of experimental conditions

The preceding example shows the limitations of the use of transcription data as the sole source of information in the regulatory reconstruction effort. In the model experiment, the relative levels of the sugars glucose, arabinose, and lactose were manipulated to drive the change in cellular physiology and gene expression. By keeping these experimental conditions in mind, a greater level of clarity can be achieved. Figure 3-4 shows the expression data (minus *cap* and *lacr*) overlaid with information on the transitions between different experimental conditions (presence or absence of each of the three sugars under consideration marked by + or –). Since "time" in this experiment has been arbitrarily assigned without physiological meaning, the shape of the curves during these transitions is unimportant, only the resulting conditions that are evidenced upon achieving steady-state.

When sugar data was compared to the mRNA expression levels, it became clear that the upregulation of the *lacZYA* genes near time t = 10 was correlated to the increase in lactose concentration (see Figure 3-4). This was verified by the removal of lactose near time t = 170, which again affected only *lacZYA*. On the other hand, glucose concentration levels affected *lacZYA*, *araBAD*, and *cAMP* levels as is seen by comparison of the glucose+ (t < 100) and glucose- (t > 140) systems. Finally, the concentration of arabinose explains changes in both *araC* and *araBAD* expression levels while affecting none of the other genes.

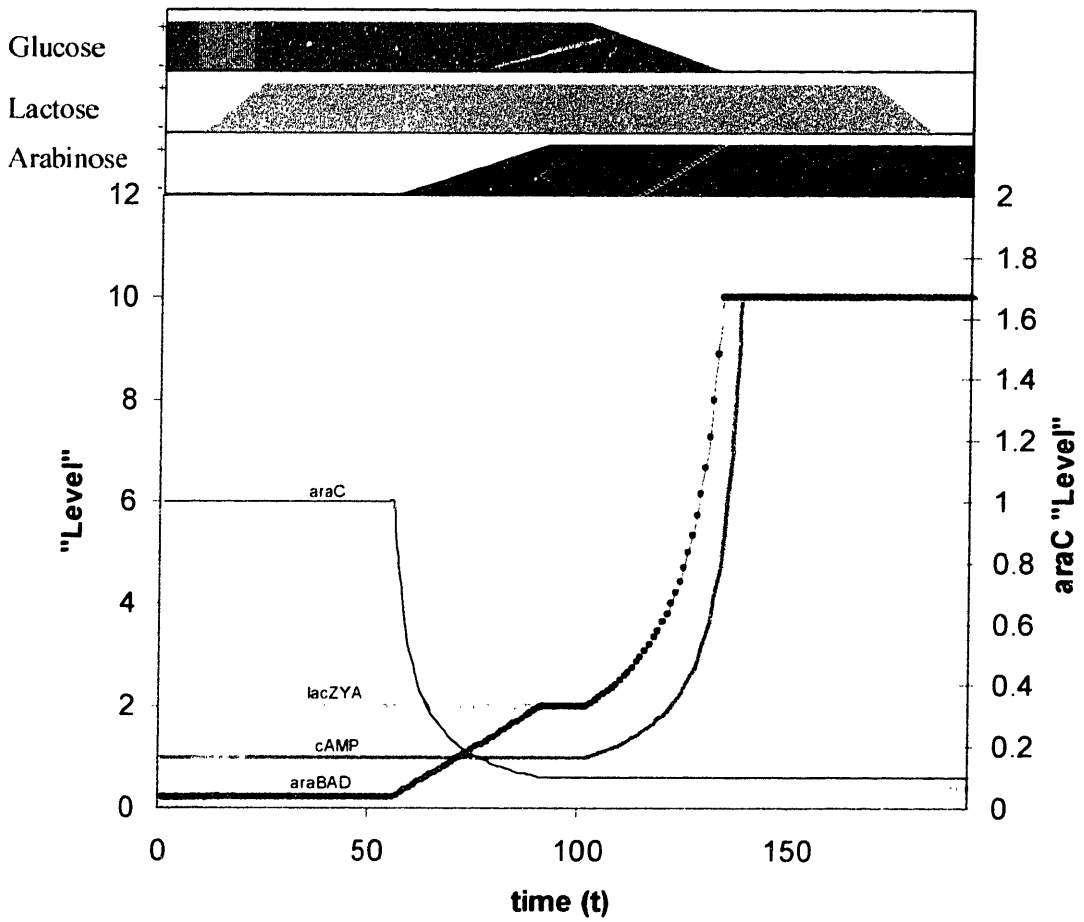

Figure 3-4: Expression and condition information

The updated regulatory network is shown in Figure 3-5. Note that although more details have been filled in, alternative hypotheses have been formed as well. For example, since *cAMP* is exactly inversely correlated to glucose concentration, it is impossible to tell which of the two chemical species directly affected *lacZYA* and *araBAD*.

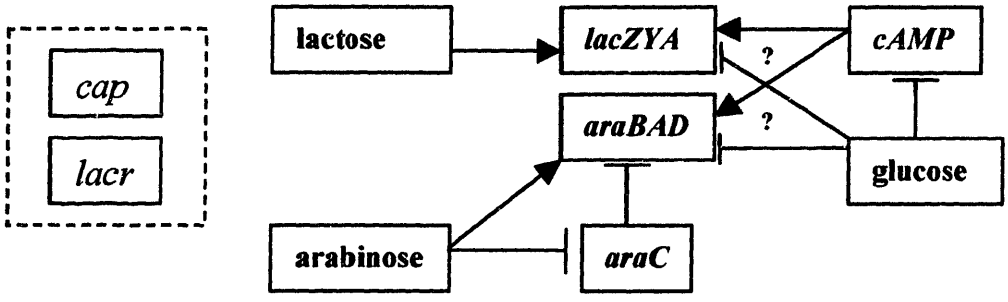

Figure 3-5: Reconstructed regulatory network

3.1.3 Conclusions

Compare the reconstruction show in Figure 3-5 to that of the original biological network shown in Figure 3-1. Note that in some ways the simple representation given in Figure 3-5 is a function of the simplified model system. For example, the relationship between ARAC protein and *araC* mRNA has been ignored in the model, and thus is not evidenced in the reconstruction process. Other differences, however, are due to issues of observability. In this system, lactose actually represses the *LacR* protein, which represses *lacZYA*. While the overall effect can be seen (lactose activates *lacZYA*) the effect of the intermediate species *lacr* is missed due to the difficulties of observability.

Note, however, that Figure 3-5 overlaps the network in Figure 1-1. Some of the relationships are in a more compact representation, while others are tentative in nature. These two differences are key to the understanding of regulatory reconstruction centered on microarray data: a) intermediate species such as protein complexes are ignored and their effects summed to simple relationships between genes and b) more candidate potential interactions than are actually required are hypothesized, and the reconciliation of these relationships may not be possible without additional information.

This example has also shown the importance of having a variety of experimental conditions to best attempt network reconstruction. If one set of experimental conditions fails to distinguish between possible relationships, another may reveal the structure. However, analysis of the transcriptional data without clear understanding of the environmental conditions limits the usefulness of such experiments. Furthermore, some greater level of experimental detail will be required to capture a greater level of detail about the network structure. Dynamic studies offer

one source of additional experimental information, and an example of such a study is shown in section 3.2.

3.2 T7 regulatory reconstruction

The T7 bacteriophage infection process is another well-studied system, but in this case the dynamic rates of the biochemical reactions involved a better understood, allowing for a more detailed and physiologically relevant model. The phage injects its DNA into an *E. coli* host, setting off a sequential series of steps that ultimately leads to cell lysis. This injection does not occur instantaneously, but rather requires about 9-12 minutes[7]. Because of this, genes are expressed sequentially as their DNA appears in the host. This type of system provides a model for the interpretation of sequential gene expression events for the discovery of genetic networks.

3.2.1 The model

A model of the lytic cycle, beginning with infection of a single cell with a single virus, has been proposed by other researchers[8]. This simulation draws from a body of literature available on *E. coli* and T7 biology to estimate parameters such as T7 DNA injection rate, T7 and *E. coli* transcription rates, T7 packaging rates, and equilibrium constants between chemically active species. This simulation allows the user to track the concentration of all T7-related species including mRNAs and proteins, as well as the concentration of a few *E. coli* species key to the replication of T7 such as *E. coli* mRNA polymerase (EcRNAP).

A diagram of the T7 genome is shown in Figure 3-6[8]. The genome enters the cell at the left tip and is transcribed from left to right. Vertical bars with circles at the top represent EcRNAP binding sites, while lighter vertical bars represent T7 RNA polymerase (T7RNAP) binding positions. Also labeled above the geneome are both EcRNAP and T7RNAP terminators (TE and Tϕ). The heights of these bars represent relative efficiency of that site: for promoters, this is the affinity of the polymerase complexes to bind to these sites, while for terminators the height indicates the how frequently the polymerases are forced from the genome when the terminator site is reached. Each gene product is labeled with respect to its ordering in the genome and therefore order of expression[9]. Some genes are labeled with non-integer values for historical reasons: these are genes that were discovered after the integer-valued genes were already named.

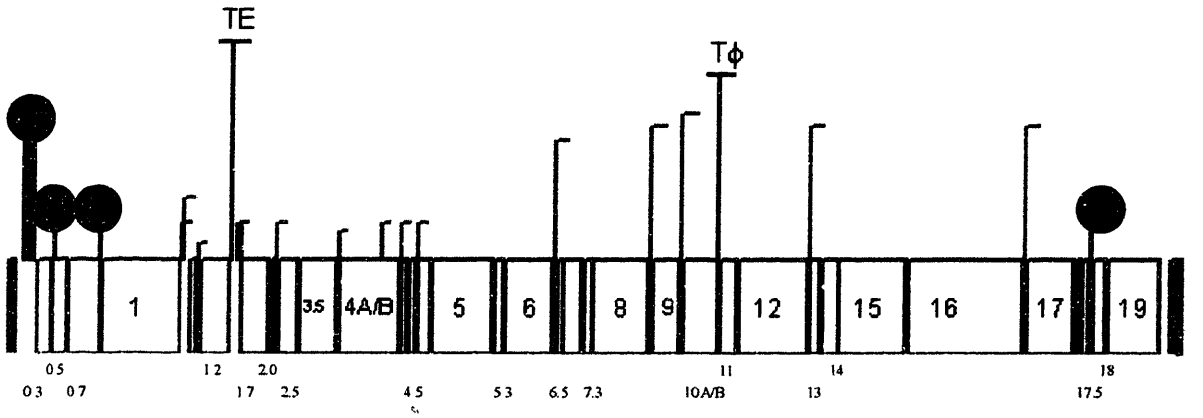

Figure 3-6: T7 genome

Not all of the genes in Figure 3-6 were used for this example: many only have structural roles important for the creation of new T7 viruses, and therefore do not have regulatory roles. The genes considered here and their functions (if known), have been shown earlier in Figure 1-2 (from Chapter 1). This figure is reproduced here as Figure 3-7 for the reader's convenience.

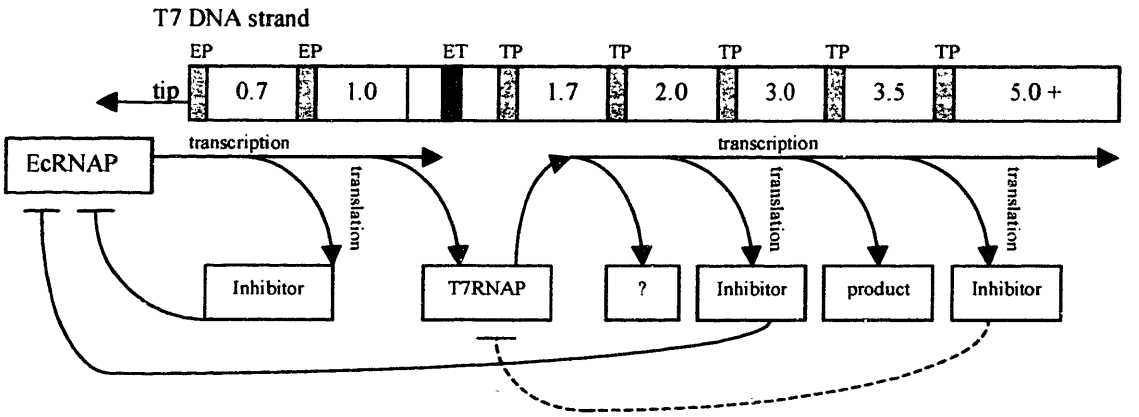

Figure 3-7: T7 regulatory structure

Both genes known to be important for regulation and some purely structural genes which have no effect on regulation have been included. This was done to evaluate if the reconstruction of gene regulation could be undertaken from mRNA data (as in a gene-chip experiment) with excess information that might confound the analysis.

3.2.2 Model mRNA results

As an example of the output from this model, consider the expression profiles for the subset of genes shown in Figure 3-8. Note that genes 1.7 and 2.0, as well as genes 3.0 and 3.5, are expressed in such rapid succession and with such similar expression profiles so as to be nearly indistinguishable. The only obvious difference between these species is the time at which each is first expressed. Thus these genes should be clustered to simplify the analysis, with knowledge of the small time lag between them retained.

Although this graph shows the sequential nature of the genes expressed, it does not give insight into the events that cause changes in the rates of expression. For these regulatory effects, a more directly apparent metric is the rate of transcription, or the derivative of the mRNA expression profile. Even more instructive are the points where the transcription rate changes, *i.e.* the points where the second derivative changes sharply from near zero. An event that suddenly slows the rate of transcription of a gene indicates an inhibitory effect, while an event that speeds the rate of transcription indicates a stimulatory response.

The plots of the 2^{nd} derivative for each of the gene products are given in Figure 3-9 and Figure 3-10. Note genes 1.7 – 3.5 (and debatably 5.0) all have roughly the same 2^{nd} derivative profile.

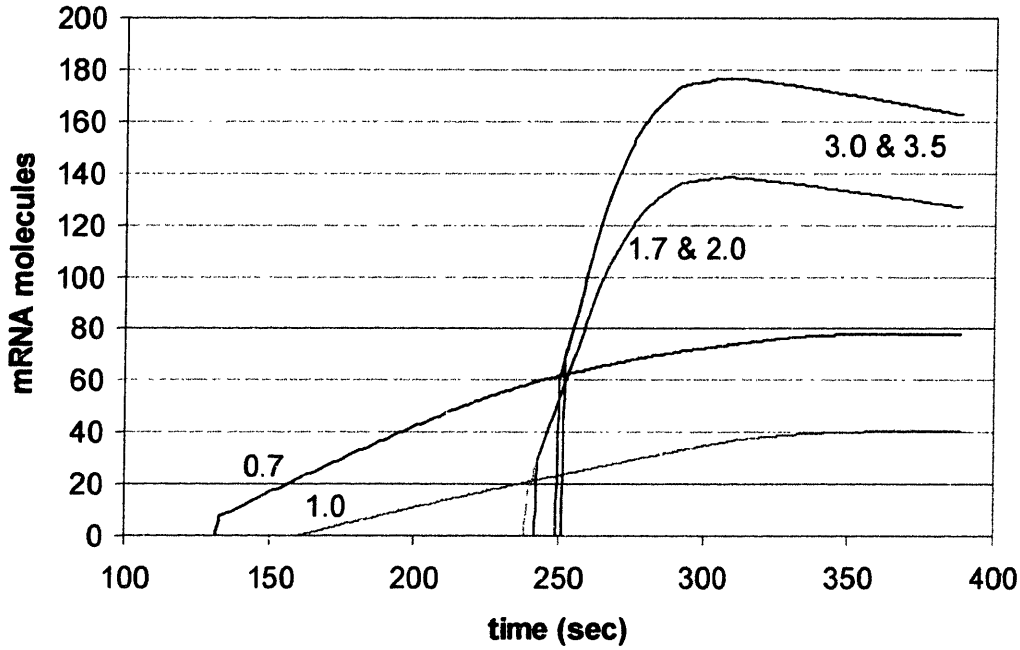

Figure 3-8: mRNA Levels for T7 Genes 0.7-3.5

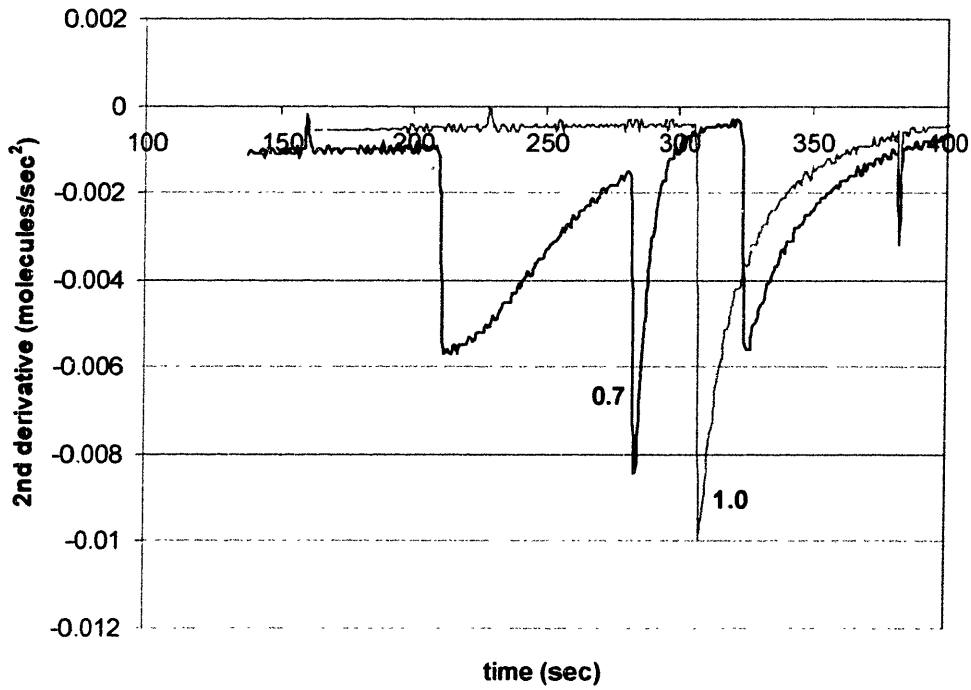

Figure 3-9: Changes in mRNA transcription, genes 0.7 and 1.0

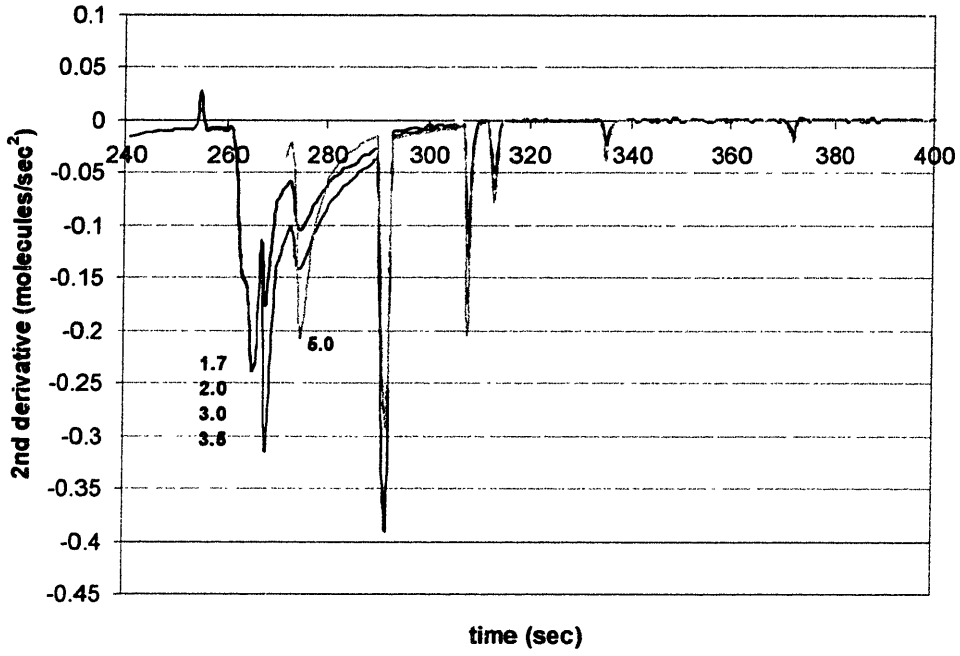

Figure 3-10: Changes in mRNA transcription, genes 1.7 - 3.5

When interpreting these plots, it is important to bear in mind that since this is a numerical simulation, there are some artifacts that are exaggerated by the 2nd derivative calculations. Peaks in the 2nd derivative that do not change the overall trend of the plot but are merely spikes are ignored for this reason (*i.e.* the peak near 380 in Figure 3-9 and the peaks at 290 and beyond in Figure 3-10). In practice, the noise of actual data will obscure interpretation, so the direct application of derivatives to raw data is unlikely to be informative. Use of frequency domain information or smoothed signals could provide a more direct method of analysis. However, for the purposes of the current model example, derivatives demonstrate data features clearly.

3.2.3 Analysis of mRNA data

As in the lactose/arabinose example, clustering of the gene expression levels can be used to combine genes with high correlation into a single cohesive group. Clustering of this data places genes 1.7 and 2.0 into one group and genes 3.0 and 3.5 into another. Effects of these genes on other genes, as well as effects of other genes upon the group, cannot be distinguished in the current experiment.

The behavior of the second derivative can also be used to perform clustering, but with a different interpretation. Genes clustered by this measure would indicate that they may have been subject to the same transcription mechanism and that the same regulatory forces may have been applied to them. Analysis of Figure 3-10 reinforces the clustering implied by the raw expression levels, as well as suggesting grouping genes 1.7 through 3.5 together. Even gene 5.0 seems to follow the same pattern as these, although the case is somewhat weaker because of its late transcription start. On the other hand, there is no obvious transcriptional relationship between genes 0.7 and 1.0 (Figure 3-9). These observations provide significantly more information than could be gained by solely static data as discussed in the lactose/arabinose gene regulation example of section 3.1, but a finer degree of transcriptional data (that is, with very small time intervals) is required than might generally be possible for real-world application at this time. Nevertheless, such detailed data adds significant information, indicating some of the future potential promised by enhanced DNA microarray technologies.

Examination of the mRNA data also leads to the identification of *events*, which will be defined either as a point where a gene is first expressed or when its second derivative reveals a change in the transcription rate. The points discovered are summarized in Table 3-2. Note that the inhibition events for genes 1.7 through 5.0 at 270 and 275 seconds are marked with "?" because it seems likely that only one inhibitory event occurred to produce a broad, unclear peak which includes both time points.

Table 3-2: T7 mRNA events

Approximate time (sec)	Gene	Event
130	0.7	Expression begins
160	1.0	Expression begins
210	0.7	Inhibited
235	1.7	Expression begins
240	2.0	Expression begins
248	3.0	Expression begins
252	3.5	Expression begins
270	1.7,2.0,3.0,3.5	Inhibited (?)
270	5.0	Expression begins
275	1.7,2.0,3.0,3.5,5.0	Inhibited (?)
280	0.7	Inhibited
310	1.0	Inhibited
325	0.7	Inhibited

3.2.4 Regulatory reconstruction

The interpretation of the events uncovered in section 3.2.3 is dependent upon how much biological knowledge is assumed beforehand. The present analysis begins with only the knowledge of the clusters and the events shown in Table 3-2.

The first regulatory event, the inhibition of gene 0.7, is seen is at t ~ 210 sec. At this time only two genes (0.7 and 1.0) are expressed, which limits the possible causes to: a) inhibition at the promoter site for gene 0.7; or b) inhibition of the factor responsible for gene 0.7 transcription. However, there is no reasonable way to distinguish between which gene actually causes the effect and the mechanism involved. The reason is because the mRNA microarray scientist may not have access to the number of active EcRNAP molecules present. If such information were available, it would be obvious that the reduced number of active molecules must be due to interaction with gene 0.7 or its protein product. Here again the major difficulty in reconstruction of regulatory networks, observability, limits the useful knowledge that can be extracted.

However, genome information from sequence data could be used to help solve this problem. Knowledge of Figure 3-6, which identifies EcRNAP promoters upstream of genes 0.7 and 1.0, shows that one of the two gene products must code for a gene which inhibits the action of EcRNAP to transcribe the T7 genome. Finally, if it is known that gene 1.0 produces T7RNAP (which would only increase transcription of the genome and is unlikely to simultaneously inhibit EcRNAP) then all ambiguities are removed from the system and gene 0.7 must be responsible for inhibition of EcRNAP activity.

This progression of understanding by applying outside biological knowledge is shown in Figure 3-11. Note again only a partial picture of the network is developed by gene chip data alone, but that protein data or functionality knowledge from other sources combined with gene-chip data can eliminate ambiguities in the regulatory network.

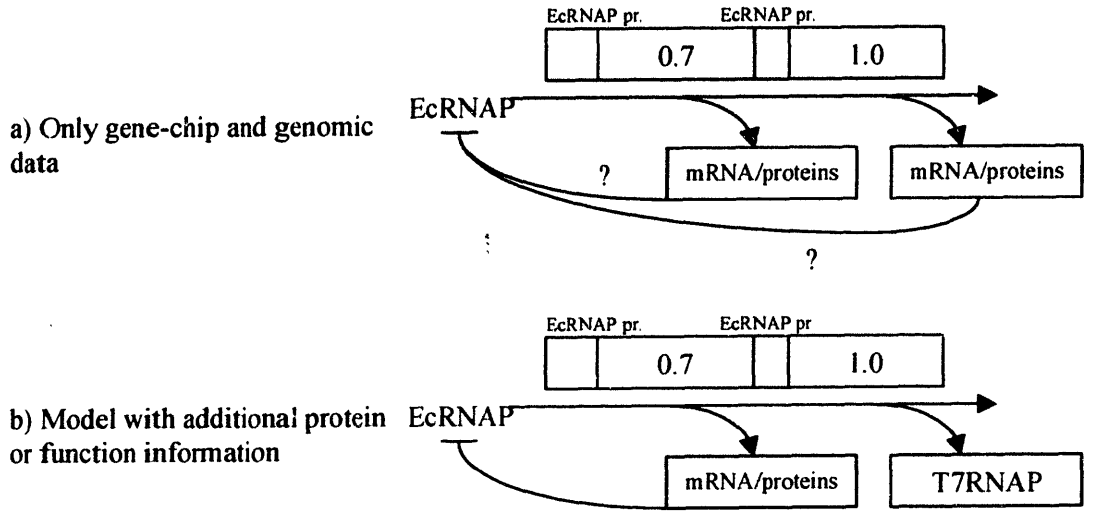

Figure 3-11: Reconstruction of the first segment of T7

The next regulatory item of interest is the inhibition of the gene group 1.7 through 5.0 (270~275 sec), all of which have T7 promoters instead of *E. coli* promoters upstream of their start sequences. If the function of genes 0.7 and 1.0 have been satisfactorily explained as drawn in Figure 3-11, then some member of this new group inhibits the transcription of the whole group. (Note: the assumption that genes 0.7 and 1.0 can be removed from further consideration is not a trivial matter, as these genes could presumably be inhibiting the latter genes. However, if gene 1.0 is identified as the factor that transcribes T7RNAP, and therefore gene 0.7 as the inhibitor of EcRNAP, then the following analysis applies.) Again, some biological insight is required to narrow the field of possible interactions. Gene 5.0 can be eliminated from consideration as the inhibitor as its transcription begins only a few seconds before inhibition is observed: sufficient time has not passed for accumulation of gene 0.5's mRNA to have an impact on the system (*i.e.* translation is unlikely to have occurred, *etc.*). This leaves two groups of undistinguishable genes to consider: the group of genes 1.7 and 2.0 as well as the group of genes 3.0 and 3.5.

Figure 3-8 shows that both of these groups have similar transcriptional profiles and only a small time lag between them. Without further information about the affinity of these genes for T7RNAP or for its promoters, the most concrete conclusion that can be drawn is that either of the two groups could be responsible. This analysis can be repeated for the inhibitory events at times 280 sec (gene 0.7) and 310 sec (gene 1.0), where the impact of genes 1.7 through 3.5 cannot be

distinguished uniquely from the other genes in the set. Note that the time lag of 30 seconds between inhibition of gene 0.7 and gene 1.0 in this case is roughly equal to the time lag in their expression start, which may lead to the conclusion that the same mechanism is being affected (in this case, EcRNAP transcription).

The regulatory network thus uncovered is presented in Figure 3-12, with "EP" marking EcRNAP promoters, "TP" marking T7RNAP promoters, and "ET" marking the EcRNAP termination site as discovered in the genome sequence. Some feedback loops are marked with "?" because one mechanism cannot be distinguished from another without further experiments into the activity of the proteins. Note, however, that by narrowing each feedback loop to four possibilities (2 groups of 2 genes) the number of potential interactions has been greatly reduced, providing direction for future experiments.

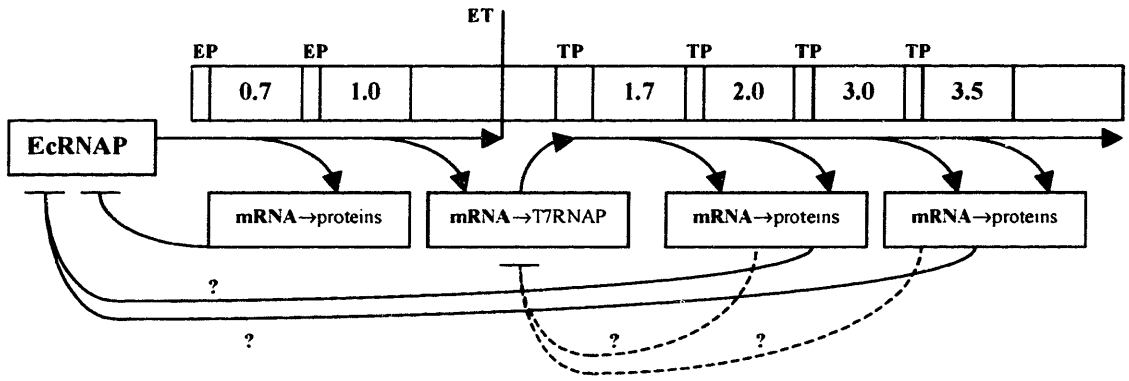

Figure 3-12: Reconstruction of T7 expression regulation

Compare Figure 3-5 to Figure 1-2 in Chapter 1. Even with information from a variety of sources, the reconstructed network could still be only narrowed to a superset of the original system. However, this example has shown that greater information content may be present in time series data if enough time points exist and if a reasonable method of dealing with the lag between events can be generated. Time-lagged correlations seem to best fit this requirement, and are discussed below in section 4.1.

3.3 References

1. Butte, A. J. & Kohane, I. S. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pacific Symposium on Biocomputing* Hawaii, (2000).

2. Somogyi, R. & Fuhrman, S. "Distributivity, a general information theoretic network measurement, or why the whole is more than the sum of its parts." *The International Workshop on Information Processing in Cells and Tissues 1997* Sheffield, (1997).

3. Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P. & Darnell, J. *Molecular Cell Biology* (Scientific American Books, New York, 1995).

4. Voet, D. & Voet, J. G. *Biochemistry* (John Wiley & Sons, Inc., New York, 1995).

5. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. "Cluster analysis and display of genome-wide expression patterns." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 14863-14868 (1998).

6. Dillon, W. R. & Goldstein, M. *Multivariate Analysis* (Wiley, New York, 1984).

7. Garcia, L. R. & Molineux, I. J. "Rate of Translocation of Bacteriophage-T7 DNA across the Membranes of Escherichia-Coli." *Journal of Bacteriology* **177**, 4066-4076 (1995).

8. Endy, D., Kong, D. & Yin, J. "Intracellular kinetics of a growing virus: A genetically structured simulation for bacteriophage T7." *Biotechnology and Bioengineering* **55**, 375-389 (1997).

9. Dunn, J. J. & Studier, F. W. "Complete Nucleotide-Sequence of Bacteriophage-T7 DNA and the Locations of T7 Genetic Elements." *Journal of Molecular Biology* **166**, 477-535 (1983).

CHAPTER 4 COMPUTATIONAL TOOLS FOR DYNAMIC DATA

In this chapter, we introduce tools specifically designed to capture and model systems with dynamic features given time-course experimental data. For the purposes recreating the basic interaction structure, the method of time-lagged correlations is discussed in section 4.1 with typical applications of this method in related fields. As an example, the lactose/arabinose gene regulatory network discussed in section 3.1 is revisited with this tool to determine how much of the network structure could be recaptured with the appropriate dynamic experiments.

Given a set of hypothetical dynamic relationships, it is also desirable to create predictive models of their behavior that can be used to forecast transcriptional profiles for future experiments. For this purpose, we discuss tools associated with AutoRegressive with eXogeneous input (ARX) models (section 4.2), including Akaike's Information Criterion (AIC) for model complexity determination.

Other computational tools for static DNA microarray data analysis have also been studied in this work, but for clarity are not discussed in this chapter. Problems such as clinical diagnosis, discriminatory gene selection, and microarray power analysis are discussed in Chapter 7.

4.1 Time-lagged correlations

Both Bayesian networks and information theory based approaches to network discovery, as presented in Chapter 2, do not make use of the sequential nature of time-series data in their current applications. If enough time points are available to observe gene correlation with some time-lag, a discovery method for uncovering *causal* relationships between genes may be attempted, as presented in the T7 case study of section 3.2. This goal is further distinguished from the modeling approaches discussed in Chapter 2, such as stochastic petri nets or deterministic models, because relationships with few *a priori* assumptions are desired.

4.1.1 Formulation and prior work

Linear Pearson correlations have been used to identify genes which are co-expressed or anti-expressed for clustering purposes[1]. As an extension of this technique, Arkin and Ross[2] make use

of *time lagged correlations* to best correlate species that follow the pattern of others with some time delay. For a series of n time points, for all $\tau < n$ the time lagged correlation $\mathbf{R}(\tau) = (r_{ij}(\tau))$ is defined by

$$S_{ij}(\tau) = \left\langle (x_i(t) - \bar{x}_i)(x_j(t+\tau) - \bar{x}_j) \right\rangle \qquad 4\text{-}8$$

$$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii} S_{jj}}} \qquad 4\text{-}9$$

where $x_i(t)$ denotes the expression of variable i at time t and the angled brackets denote the time average of the product inside. The matrix of lagged correlations $\mathbf{R}(\tau)$ can be used to find the maximum correlation between each species at some lag τ through conversion to a Euclidean distance metric:

$$d_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{1/2} = \sqrt{2}(1.0 - c_{ij})^{1/2} \qquad 4\text{-}10$$

$$c_{ij} = \max_{\tau} |r_{ij}(\tau)| \qquad 4\text{-}11$$

Thus, c_{ij} is a measure of the maximum correlation between two species and the time-lag yielding the maximum correlation: if the value of τ which gives maximum correlation is 0, then the two species are best correlated with no translation in time. $\mathbf{D} = (d_{ij})$ describes the correlation between two species in terms of "distance" by making those species that are least correlated (for any τ) the "farthest" apart[2]. By finding species that are highly correlated and the value of τ that led to this correlation the underlying network of cause and effect relationships may be uncovered. However, note that data points at the extremes of a time-series (*i.e.* the first and last points) are lost when correlation is calculated at values of $\tau > 0$, and more points are lost as τ is increased. Therefore, calculation c_{ij} must be limited to values of τ small enough to ensure enough data points have been included to give statistical significance.

Arkin, Shen, and Ross[3] use this technique to "reconstruct" central carbon metabolism by measuring dynamic concentration profiles of 14 metabolites interconverted by 8 enzymes in a continuous flow reactor system. See Figure 4-1. Boxed chemical species were measured

dynamically, while citrate and adenosine monophosphate (AMP) input concentrations (marked by ovals) were adjusted dynamically to keep the system away from steady-state. Using time-lagged correlations on the output data, these authors were able to recreate most of the features of the original pathway, as shown in Figure 4-2. Note that not all known reactions are present in the reconstructed diagram. For example the inhibitory impact of citrate on the conversion of fructose-6-phosphate (F6P) to fructose 1,6-bisphosphate (F16BP) is not included in Figure 4-2. Furthermore, species that are not measured or adjusted, such as glucose or glyceraldehyde-3-phosphate (GAP), obviously cannot be placed in the network at all.

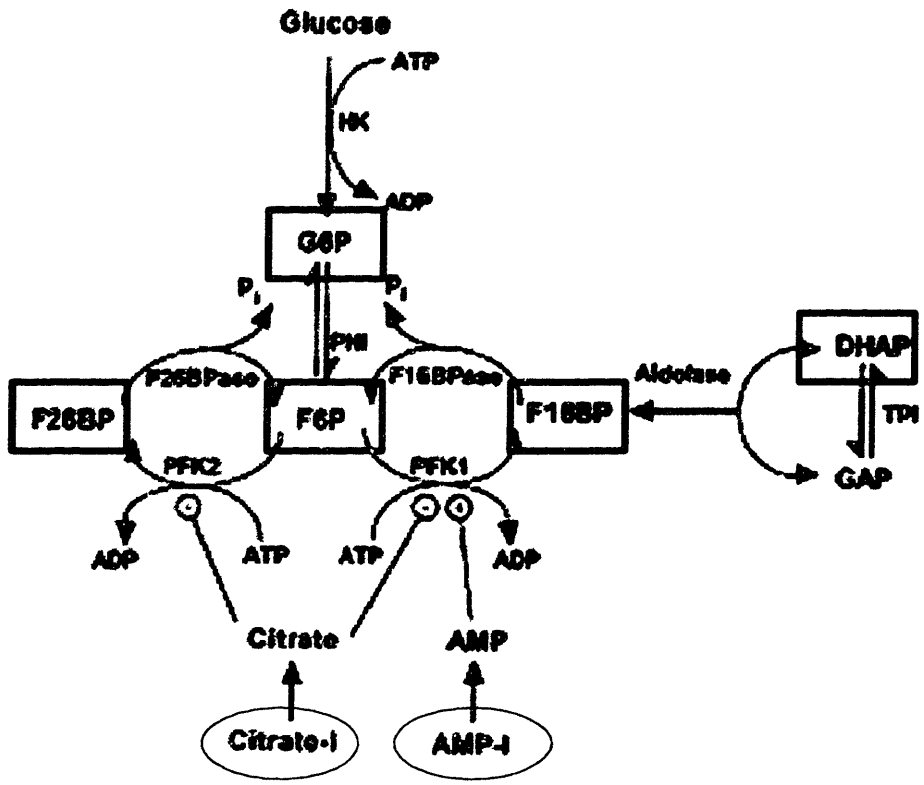

Figure 4-1: Glucose metabolism system (from Arkin, Shen, & Ross)

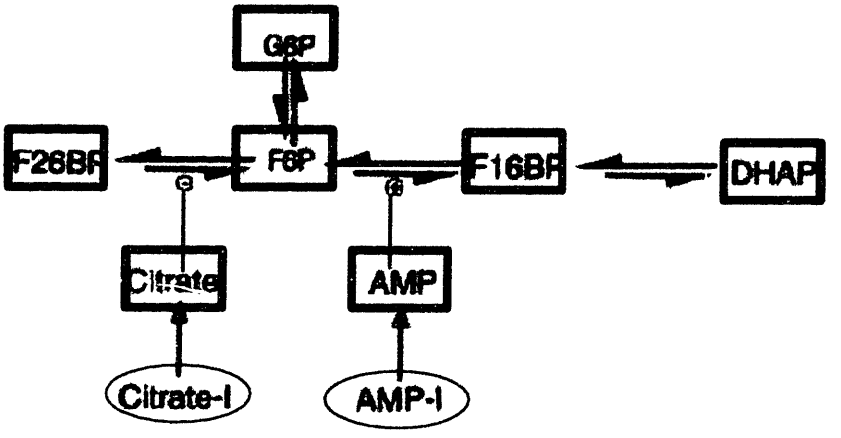

Figure 4-2: Glucose metabolism reconstruction (from Arkin, Shen, & Ross)

These drawbacks aside, this example showed that even when the specific method of interaction is unknown or unmeasured (in this case the enzymes themselves) useful information could be gathered about the overall structure of a given network of interacting chemical species.

An alternative but related method for uncovering activators and inhibitors from time-series data is discussed by Chen *et al.*[4]. Here all peaks, or consecutive time points of greater than average expression, are first identified in data. The "leading" and "trailing" edges of each of these peaks are then compared to other genes' peaks. The temporal relationships between edges of each peak are then scored via an exponentially decaying function of the time lag between the features. In other words, if a peak's leading or trailing edge occurs soon after such a feature in a different gene, then the relationship is scored depending on how close in time the two events are. For the studies shown in Chapter 3, both standard lagged correlations and these peak correlations give largely the same results, and only slight adaptations would be required to use this alternative strategy instead.

4.1.2 Lactose/arabinose Example

Consider an extension to the glucose/arabinose/lactose model shown in section 3.1. The model can be expanded to include time lags for some of the interactions which are assumed to be kinetically controlled (*i.e.* reactions which are not assumed to be at equilibrium). Specifically,

the equations in Table 3-1 were adjusted as shown in Table 4-1, and the simulation of changing sugar concentrations was rerun.

Table 4-1: Simplified dynamic model equations for lactose/arabinose regulation

(all concentration scales and rate constants arbitrary)

cap, CAP, *lacr*, and LACR fixed at (5)

cAMP(t) = 10 * 1/glucose(t-15)

CAP*(t) = min(CAP(t), cAMP(t))

LACR*(t) = min(LACR(t), lactose(t))

$lacZ(t) = lacY(t) = lacA(t) = 10\,(CAP^*(t-3)/5)\,(LACR(t-10)^*/5)^2$

ARAC*(t) = 10*arabinose(t)

ARAC(t) = 10 – ARAC*(t)

araC(t) = 1/ARAC*(t-5)

$araB(t) = araA(t) = araD(t) = 10\,(CAP^*(t-5)/5)\,(ARAC^*(t-10)/10)$

Time-lagged correlations were then calculated for all of the chemical species present. The time lagged correlations give three pieces of information: a) a best time-lag τ, b) a correlation value $|r_{ij}|$, and c) whether the correlation is positive or negative. For this experiment, we assume that 0 time lag refers to systems that represent "reactions" (*ie.* interconversion of species that must be conserved) and that non-zero time lags show inhibitory (negative correlation) or activating (positive correlation) effects. This allows a rule table to be drawn, as shown in Table 4-2. The corresponding graphical mapping is shown in Figure 4-3.

Table 4-2: Connections derived from time-lagged correlation

Rule			Calculated Time Lag
Glucose	Inhibits	cAMP exprs.	15
cAMP	Determines	cAMP*	0
Glucose	Sum of the last two rules above	cAMP*	15
cAMP*	Activates	lacZYA expr.	3
cAMP*	Activates	araBAD expr.	5
Lactose	Determines	lacr*	0
lacr*	Activates	lacZYA	10
Lactose	Sum of the last two rules above	lacZYA	12
Arabinose	Determines	ARAC*	0
ARAC*	Determines	ARAC	0
Arabinose	Sum of the last two rules above	ARAC	0
ARAC*	Inhibits	araC expr.	5
ARAC*	Activates	araBAD expr.	10

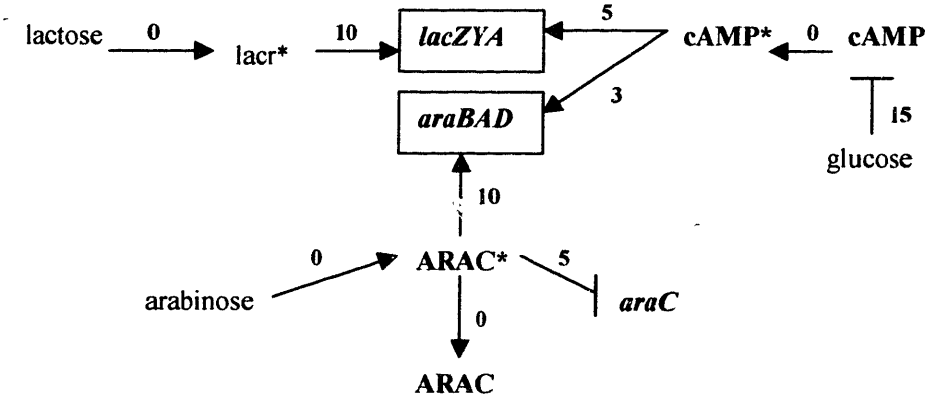

Figure 4-3: Graphical network derived from time lagged correlation

The time lags calculated reflect nearly exactly the underlying model. Note one exception where the effects of lactose on lacr* ($\tau = 0$) and lacr* on *lacZYA* ($\tau = 10$) do not exactly sum to the correct time lag for the entire cascade ($0 + 10 \neq 12$). These types of differences may be difficult to interpret, because it may be impossible to determine whether the discrepancy is due to an unmeasured species or due to small numerical artifacts (as is the case here).

4.1.3 Graphviz for correlation network visualization

For larger systems, drawing interactions by hand is generally impractical, especially if a multiple networks are to be generated with different significance cutoff values. The Graphviz program from ATT Research Labs[5] has been adapted for this purpose. This program has been optimized using heuristics to minimize such events as cross-over between edges in order to create easily interpreted output figures. For practical purposes, the output of analysis software (such as functions written in MatLab) can be written into simple text code that is re-interpreted by Graphviz to create jpeg images. For example, for the rules written in Table 4-2, the simple input file shown in Figure 4-4 was converted into Figure 4-5. Matlab functions for automating the creation of such files from time-lagged correlation analysis have been written for this purpose and are available from the author.

```
digraph arabalac {
glucose [label = "Glucose"]
camp [label = "cAMP"]
camp_star [label = "cAMP*"]
lacZYA [label = "lacZYA"]
lacr_star [label = "lacr*"]
lactose [label = "lactose"]
arabinose [label = "arabinose"]
arac [label = "AraC"]
arac_gene [label = "arac"]
arac_star [label = "AraC*"]
arabad [label = "araBAD"]

glucose -> camp [arrowhead = "tee", label= "15"];
camp -> camp_star [label = "0"];
camp_star -> lacZYA [label = "3"];
camp_star -> arabad [label = "3"];
lactose -> lacr_star [label = "0"];
lacr_star -> lacZYA [label = "10"];
arabinose -> arac_star [label = "0"];
arac_star -> arac [label = "0"];
arac_star  -> arac_gene [arrowhead = 'tee", label = "5"];
arac_star -> arabad [label = "10"];

}
```

Figure 4-4: Graphviz sample input

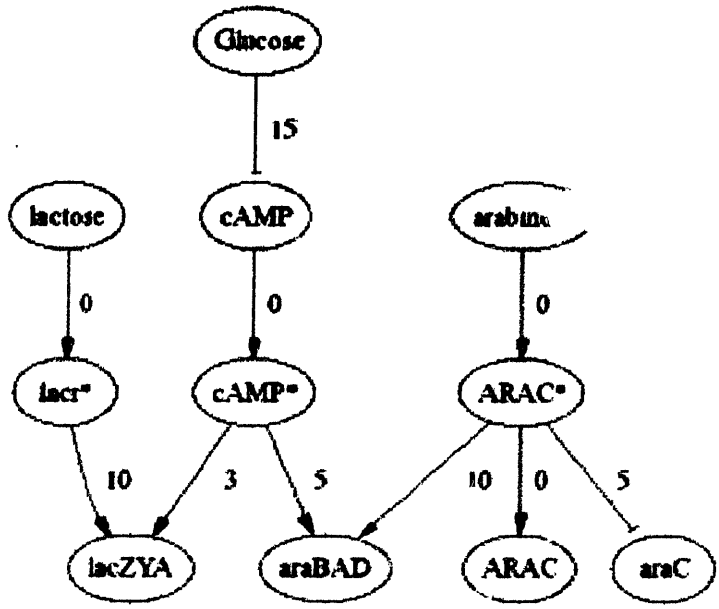

Figure 4-5: Graphviz representation of lactose/arbinose example reconstruction

4.2 ARX models

While time-lagged correlations help to discover potential relationships between chemical species in a dynamic system, they do not offer quantitative predictive value for the uncovered relationships. To take full advantage of diagrams such as Figure 4-5, it is desirable to create numerical models that forecast the degree of transcription expected for a gene in a given situation. For example, if the amount of glucose is changed from high concentration to low concentration, Figure 4-5 suggests that the *araBAD* operon will be upregulated some 20 time "units" later. On the other hand, time-lagged correlation diagrams do not address whether this upregulation will also be affected by the current arabinose concentration, or even whether the degree of upregulation is expected to be large or small. To answer these types of questions, a numerical modeling approach is required, where the expression level of a gene is written as a function of environmental variables such as extracellular chemical concentrations, other gene expression values, and even a gene's own expression values at some earlier time points.

AutoRegressive (AR) models are time-series models that assume that variable y is dependent on earlier values of that variable[6]

$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) = \varepsilon(t) \qquad 4\text{-}12$$

where *na* is the order of the model – that is, the number of earlier time points that *y(t)* is dependent on. Thus, *y(t-1)* refers to *y* one time-interval before *y(t)*, where the time-interval Δt is dictated by the experimental conditions. ε is a white-noise process, and the coefficients a_i indicate the dependence of *y(t)* on its own history. This equation can be expanded for vectors of *y* variables

$$Y(t) + A_1 Y(t-1) + \ldots + A_{na} Y(t-na) = \mathrm{E}(t) \qquad 4\text{-}13$$

were each A_i now represents a matrix of cross-interaction terms, and E is a vector of white noise processes. Note that if a first-order process is assumed and the white-noise term is ignored, this becomes

$$Y(t) = -A_1(t-1) \qquad 4\text{-}14$$

which is functionally the same as the model used by Holter *et al.* in earlier yeast studies[7] when $\Delta t = 1$.

If the variable of interest *y(t)* is also a function of a known external signal *u(t)*, then AutoRegressive with eXogeneous (ARX) models can be used. In this case, the expression becomes

$$y(t) + a_1 y(t-1) + \ldots + a_{na} y(t-na) = b_1 u(t-nk) + \ldots b_{nb} u(t-nb-nk+1) + \varepsilon(t) \qquad 4\text{-}15$$

where *nk* is the time-lag between *u* and *y*, and *nb* is the order of the model with respect to the input. Note that only *na* and *nb* change the complexity of the model: *nk* only shifts the time points in *u* to be used for calculation.

For this study, this relationship is usually re-cast for vectors of outputs (genes) but only a single input signal

$$Y(t) + A_1 Y(t-1) + \ldots + A_{na} Y(t-na) = B_1 u(t-nk) + \ldots B_{nb} u(t-nb-nk+1) + \varepsilon(t) \qquad 4\text{-}16$$

In the studies presented for *Synechocystis* in Chapter 6, we have focused on modeling pairs of gene groups: thus, the *Y(t)* values are 2x1 vectors, the A_i parameters are 2x2 matricies, the B_i

parameters are 2x1 vectors, and the *u(t)* are scalar values. For naming purposes, we shall call these models ARX[na nb nk] models, so that an ARX441 model uses the last 4 values before *Y(t)* and the 4 values before *u(t-1)* to predict *Y(t)*.

The ARX modeling framework has been applied extensively to a variety of dynamic systems, from industrial control[8] and fault detection[9], to prediction of greenhouse temperatures[10], and even to modeling of voice patterns[11,12]. These models are particularly attractive because of the relatively straightforward interpretation of their results, as regressed parameters are directly indicative of the relative importance of the model variables. ARX models are also easily adapted as the basis for more complex methodologies, and are often combined in neural network strategies to improve prediction performance[13,14].

A typical adaptation of ARX processes are AutoRegressive Moving Average with eXogeneous input (ARMAX)[6] processes, which assume moving averages over some time window instead of assuming stationary processes. The moving average assumption de-emphasizes the absolute value of a variable in favor of the *change* of variable relative to its recent expression level. Because an appropriate time-window for calculation of each variable's moving average must be selected, these models represent an increase in complexity over ARX models. For our study, we decided to simplify model calculations by using as few user-defined parameters as possible, so ARMAX models were not explored. Nevertheless, such an extension merits further consideration as it may be more directly applicable to relationships in transcriptional data.

4.2.1 Model building

Fitting of ARX models was done in Matlab's system identification toolbox. Average autoscaled profiles for groups of genes were fit to models of varying complexity. Parameters were fit to minimize sum of square error (SSE)[7] between the model and the training data.

In many DNA microarray experiments, missing data points (due to flaws on the particular chip used or poor RNA extraction efficiency) create "holes" in the sample profiles of some genes. For the time-lagged correlations discussed in section 4.1, these points are easily handled by eliminating all calculations involving a missing value, thus calculating correlation only for the remaining points. Provided that care is taken not to lose the placeholder status of the missing

data points, the calculation does not require that these points be replaced, and such points can be ignored. For dynamic modeling, however, some sort of estimation of the transcriptional profile at that point must be made. For this study, simple interpolations were performed to fill in the missing data points, and if two or more time points in a row were missing, that gene/gene group was eliminated from the modeling effort. More complex methods (*e.g.* splines[15]) could be used for this purpose, but for simplicity are not explored in this work.

4.2.2 Choosing between models

Goodness of fit is a metric often used to compare models of different type or complexity. However, calculating the least-squared error (or % of data variance explained) can be misleading for the data used to train the model, as the number of parameters used generally improves the goodness of fit. To rank models in terms of both the number of parameters used and the goodness of fit, Akaike's information criterion (AIC)[15,16] can be considered:

$$AIC = -\log likelihood + 2\frac{d}{N} \qquad 4\text{-}17$$

where d is the number of fit parameters and N is the number of observations in the estimation data used. "Log likelihood" is a measure of the likelihood that the regressed parameters are the "true" parameters (assuming the model form is correct), given the data at hand. The covariance of the error vector $E(t)$, which is assumed to be white noise, gives a measure of this unexplained data, and its determinant can be used as an inverse indicator of likelihood. Thus, for ARX models, the AIC criteria is calculated as:

$$AIC = -\log\left[\frac{1}{\det(\mathrm{cov}(E(t))}\right] + 2\frac{d}{N} = \log[\det(\mathrm{cov}(E(t))] + 2\frac{d}{N} \qquad 4\text{-}18$$

Models can therefore be compared by their ability to minimize this function, as a balance between model fit (the first term) offset by model complexity (the second term). For perfect fit ($E(t)$ is a matrix of zeros) AIC is equal to negative infinity.

Models with more than two outputs can also be used, but the number of parameters fit increases rapidly as the number of cross-terms in the parameter matrices A_i increases. However, small order models for three output variables can still be fit to the data generated for *Synechocystis* (Chapter 6) with some success (data not shown).

4.3 Conclusions

We have shown tools for both the analysis and modeling of time-series data. Time-lagged correlations have been applied successfully to recreate the basic interaction structure of networks of introverting chemical species, and we hypothesize that they will prove similarly useful for reconstruction of gene regulatory networks. With the basic network features suggested by this analysis, we then propose to use AutoRegressive with eXogeneous input (ARX) models to affix quantitative, predictive capabilities to the models of gene regulation. Application of these tools in tandem introduce to a model system, *Synechocystis* PCC6803 (Chapter 5) is discussed in Chapter 6.

4.4 References

1. D'Haeseleer, P., Wen, X., Fuhrman, S. & Somogyi, R. in *Information Processing in Cells and Tissues* (eds. Paton, R. C. & Holcombe, M.) 203-212 (Plenum Publishing, 1998).

2. Arkin, A. & Ross, J. "Statistical Construction Of Chemical-Reaction Mechanisms From Measured Time-Series." *Journal Of Physical Chemistry* **99**, 970-979 (1995).

3. Arkin, A., Shen, P. D. & Ross, J. "A test case of correlation metric construction of a reaction pathway from measurements." *Science* **277**, 1275-1279 (1997).

4. Chen, T., Filkov, V. & Skiena, S. S. "Identifying Gene Regulatory Networks from Experimental Data." *The Third International Conference on Computational Biology* Lyon, (1999).

5. AT&T Labs: Graphviz. http://www.research.att.com/sw/tools/graphviz/

6. Wei, W. *Time Series Analysis* (Addison-Wesley Publishing Company, Redwood City, CA, 1990).

7. Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V. & Banavar, J. R. "Dynamic modeling of gene expression data." *Proceedings of the National Academy of Sciences of the United States of America* **98**, 1693-1698 (2001).

8. Bobal, V., Sysel, M. & Dostal, P. "Self-tuning PID controller using delta-model identification." *International Journal of Adaptive Control and Signal Processing* **16**, 455-471 (2002).

9. Simani, S., Fantuzzi, C. & Beghelli, S. "Diagnosis techniques for sensor faults of industrial processes." *IEEE Transactions on Control Systems Technology* **8**, 848-855 (2000).

10. Frausto, H. U., Pieters, J. G. & Deltour, J. M. "Modelling greenhouse temperature by means of auto regressive models." *Biosystems Engineering* **84**, 147-157 (2003).

11. Fort, A., Ismaelli, A., Manfredi, C. & Bruscaglioni, P. "Parametric and non-parametric estimation of speech formants: Application to infant cry." *Medical Engineering & Physics* **18**, 677-691 (1996).

12. Ding, W., Kasuya, H. & Adachi, S. "Simultaneous Estimation of Vocal-Tract and Voice Source Parameters Based on an Arx Model." *IEICE Transactions on Information and Systems* **E78D**, 738-743 (1995).

13. Chong, A. Z. S., Wilcox, S. J. & Ward, J. "Prediction of gaseous emissions from a chain grate stoker boiler using neural networks of ARX structure." *IEE Proceedings-Science Measurement and Technology* **148**, 95-102 (2001).

14. Chen, X. H., Racine, J. & Swanson, N. R. "Semiparametric ARX neural-network models with an application to forecasting inflation." *IEEE Transactions on Neural Networks* **12**, 674-683 (2001).

15. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, 2001).

16. Akaike, H. "New Look at Statistical-Model Identification." *IEEE Transactions on Automatic Control* **AC19**, 716-723 (1974).

CHAPTER 5 BIOLOGICAL SYSTEM AND EXPERIMENTAL PROTOCOLS

The cyanobacteria *Synechocystis* PCC6803 was chosen as a model organism to test the methods discussed in Chapter 4. In this chapter, the attributes of *Synechocystis* and cyanobacteria in general are discussed, followed by procedures adapted by our lab for the growth and maintenance of these cultures. Finally, protocols are also presented for the creation and use of full-length cDNA microarrays to measure full-genome transcription levels in *Synechocystis*.

5.1 Cyanobacteria

Cyanobacteria are a group of gram-negative photosynthetic prokaryotes. They are distinguishable from other photosynthetic bacteria (purple or green) because of the nature of their photosynthethic pigment system, which consists of chlorophyll and phycobiliproteins. Furthermore, they are capable of oxygenic biosynthesis, which gives them a major role in the ecosystems of marine and freshwater systems. Their photosystem apparatus bears remarkable similarity to eukaryotic chloroplasts[1], which make them a model organism for the study of plant photosynthesis[2], especially due to their fast growth rate relative to most plants.

Studies of the physiology and molecular biology of these organisms have resulted in the elucidation of many of their biochemical processes[2]. Of particular interest is their ability of actively transport both CO_2 and HCO_3^-, allowing for autotrophic growth (*e.g.* without organic matter) through carbon fixation[3]. The major pathway of photosynthetic carbon fixation in cyanobacteria is via the Calvin cycle where the Rubisco (ribulose-1,5-bisphosphate carboxylase-oxygenase) enzyme adds CO_2 to ribulose-1,5-bisphosphate to create other metabolic species, with the net result of an additional carbon atom in the metabolic cycle and the generation of O_2 (see Figure 5-1).

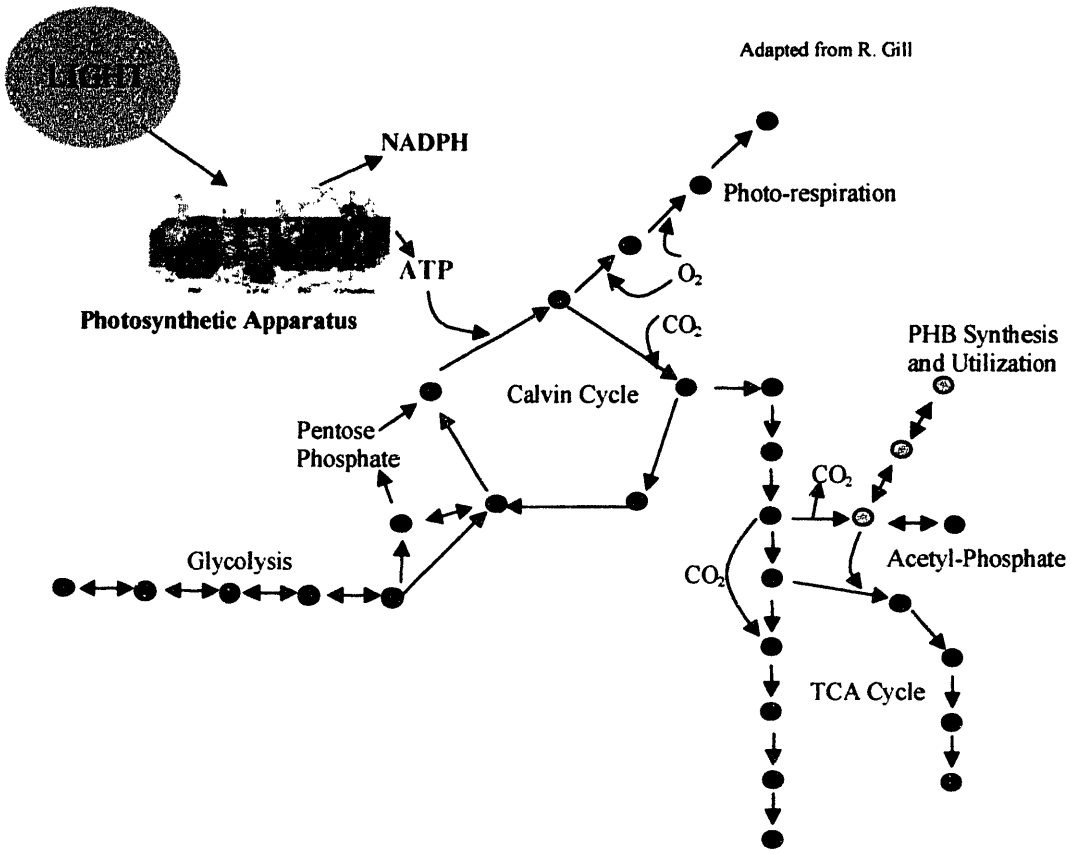

Figure 5-1: Simplified structure of *Synechocystis* PCC6803 central metabolism

5.1.1 *Synechocystis* PCC6803

Synechocystis PCC6803 is a unicellular, spherical, polyploid (having many copies of its genome in each cell) cyanobacterium that reproduces through binary fission. It is particularly useful among cyanobacteria because of its extremely efficient natural genetic transformation capability. Furthermore, it has the ability to grow photoheterotrophically on glucose, a characteristic that is necessary in order to study strains or mutants deficient in photosynthesis related functions[4]. Very few other cyanobacteria possess these characteristics. Finally, the recent sequencing of the entire genome of *Synechocystis*[5], has opened the door to genomic manipulation and measurement techniques such as DNA microarrays.

Synechocystis is also interesting for its production of biopolymers. A two-component polyhydroxyalkanoate (PHA) synthase has been identified and characterized in *Synechocystis*[6], as shown in Figure 5-1. Gas-chromatographic analysis has revealed that the PHA granules in

Synechocystis are composed of a poly(3-hydroxybutyrate) (PHB) homopolyester[6]. This biodegradable straight-chain polymer is a stiff and rather brittle polymer of high crystallinity, whose mechanical properties are not unlike those of polystyrene, though it is less brittle and more temperature resistant[7]. Coupled with the cyanobacerium's CO_2 fixation ability, *Synechocystis* is ideally suited to the bio-remediation of CO_2 gases from industrial applications into useful material. However, the PHB content in the cells under the most "optimal" conditions currently known (*i.e.*, nitrogen starved) amounts to only 5-10% (w/w) of the cell's dry weight. It is therefore advantageous to better understand *Synechocystis* regulation under a variety of conditions, in order to determine not only optimum conditions for fixation to polymers, but also to discover targets for genetic manipulation resulting in up regulation of biopolymer production.

5.1.1.1 *Synechocystis* photosystem operation

As shown in Figure 5-1, the photosynthesis apparatus is key to *Synechocystis* metabolism, and therefore a reasonable starting point for regulatory study. Here light energy is converted into chemical energy in the form of ATP and into reducing power in the form of the electron carrier NADPH. An expanded diagram (from James Barber[8]) of this system is shown below in Figure 5-2. Energy from light is transferred to Photosystems I and II through the phycobilisomes, a structural model (from Wendy Schluchter[9]) of which is shown in Figure 5-3. Photosystem II (PSII) essentially uses this energy to break down H_2O and reduce plastoquinone (PQ). The cytochrome b_6f complex (Cyt b_6f) utilizes this energy pool to pump H^+ from the stroma into the lumen, against the energy potential gradient maintained by the thylakoid membrane. Photosystem I, on the other hand, transfers electrons to ferrodoxin (Fd), which either reduces $NADP^+$ to NADPH for other metabolism requirements, or cycles them back to Cyt b_6f to reduce quinone (Q) and drive the H^+ pump. Finally, ATP synthase utilizes the energy of the H^+ ions trapped against the potential gradient to add a phosphate group to ADP, generating ATP for cellular energy needs.

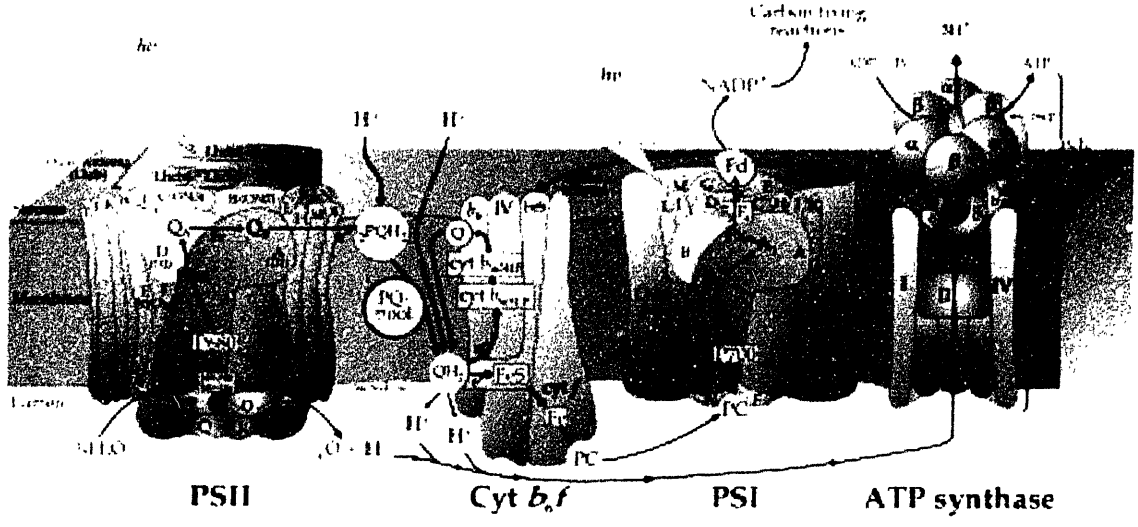

Figure 5-2: Overview of *Synechocystis* photosynthesis machinery (from J. Barber)

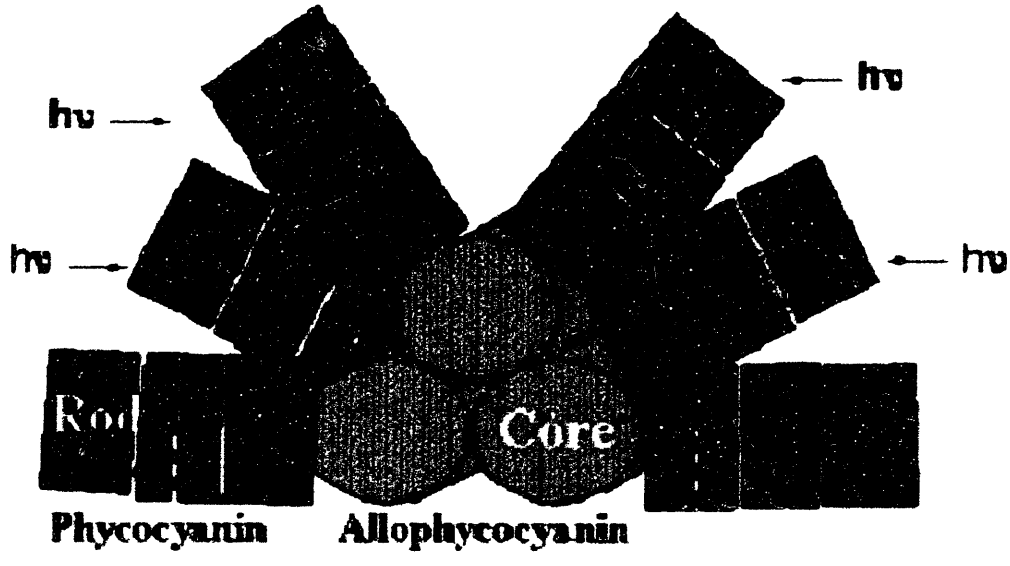

Figure 5-3: Model of *Synechocystis* phycobilisomes (from W. Schluchter)

5.2 Experimental protocols for *Synechocystis*

All of the protocols discussed here describe procedures used by our research lab and/or developed by the author. A list of other protocols for *Synechocystis* and cyanobacteria in general can be found at http://www-cyanosite.bio.purdue.edu/protocols/protocols.html.

5.2.1 Growth and maintenance

Since *Synechocystis* is a heterotrophic organism, it can be grown with glucose or an analogous carbon source, or grown autotrophically if CO_2 or HCO_3^- is present. For all of the studies conducted here, cells are grown solely on dissolved CO_2 as HCO_3^-. Of course, there are other cellular requirements, such as a source of nitrogen for amino acid synthesis and salts required for a variety of cellular functions. BG-11 medium (Sigma) is designed specifically to meet these minimal requirements for freshwater cyanobacteria, and was used in all cultures. Its composition is listed below in Table 4-2[10].

Table 5-1: BG-11 medium composition

Component	g l^{-1}	mM
$NaNO_3$	1.5	17.65
$K_2HPO_4.3H_2O$	0.04	0.18
$MgSO_4.7H_2O$	0.075	0.30
$CaCl_2.2H_2O$	0.036	0.25
Citric acid	0.006	0.03
Ferric ammonium citrate	0.006	0.03
EDTA (disodium magnesium)	0.001	0.003
Na_2CO_3	0.02	0.19
Trace metal mix A5+Co	1 ml	
Deionized water	to 1 l	
pH after autoclaving and cooling: 7.4		

All cultures were grown in an incubator at 30°C under florescent light. Light intensity in the incubator was determined to be approximately 6900 LUX, or a photosynthetic photon flux (PPF) of about 90 µmol /m^2 s at the surface of a culture. This flux is expected to drop significantly

inside of the cultures due to shielding by the outermost cells; therefore all cultures were continuously shaken or stirred to ensure homogeneity of light exposure.

For basis cultures used to maintain the cell line and inoculate new cultures, cells were grown in 250ml flasks with 100ml of H_2O that had been autoclaved for sterility beforehand. Flasks were stopped with cotton in gauze caps to ensure sterile airflow. 50X BG-11 media (2ml) and 0.38M Na_2CO_3 (300µl) were added to the cooled (30°C maximum) flasks as the carbon source and basic media, respectively. Such cultures were usually inoculated with < 2ml of cells from another culture, usually for the late (stationary growth) phase. Depending on the volume and density of the inoculating sample, these cultures generally survived for several weeks, and growth was often not visible to the naked eye for a few days if the inoculating sample was extremely small (< 100µl).

Inoculation cultures used to seed the larger reactor vessels were grown in 1L flasks with 300ml of H_2O. Again, BG-11 (6ml) and 0.38M Na_2CO_3 (300µl) were added to the culture, but here the additional buffer of sterile-filtered 1M HEPES (6ml) was added to create an environment more similar to the sparged reactor vessel (see below). ~10ml of a basis culture (as above) in late-exponential or stationary phase was used for inoculation. These cultures were grown to mid-exponential phase (A_{730} ~ 1.0, approximately 4 days) before use for inoculation of the large reactors.

The layout of the sparged-gas vessel is shown in Figure 5-4. 6L of H_2O was first autoclaved in the 10L reactor vessel with a large stir-bar placed in the center of the gas-sparging ring. Since CO_2 was bubbled through this reactor, no Na_2CO_3 was added, only BG-11 media (120ml). Dissolved CO_2 gas in the form of $(H^+)(HCO_3^-)$ increases the acidity (decreases the pH) of the culture, drastically inhibiting growth. To counteract this, 1M HEPES (120ml) was added as a pH buffer (pKa = 7.31 at 37°C). The CO_2 source gas (1-3% CO_2, ~16% O_2, balance N_2) was forced through a sterile filter into a plastic tubing ring with many small punctures at a rate of about 150ml/min. Sparged gas escaped through a pressure release valve at the top of the vessel. A final tube, extending deep into the liquid culture, was sealed with a quick-release clamp that could be quickly opened and closed for sampling.

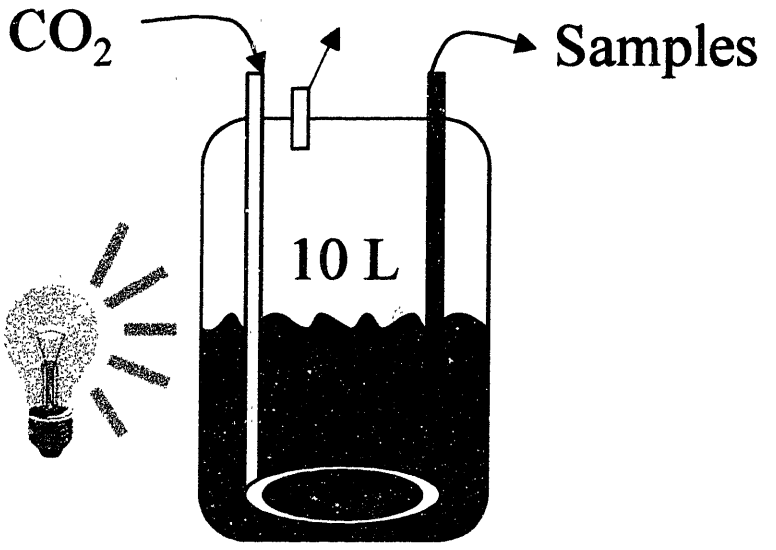

Figure 5-4: Setup of sparged-gas reactor vessel

5.2.2 Sample collection for RNA extraction

Because RNA tends to be unstable and is easily degraded by common RNAse enzymes found in cultures, it is imperative to not only stop new transcription by cells but also freeze them to minimize RNA degradation. However, the cells should be concentrated and stored liquid-free (or nearly so) to ensure efficient downstream processing. The following protocol is designed to stop transcription through inhibition by phenol, but care must be taken to keep the samples cool and complete centrifugation as quickly as possible. If ice forms in the sample, centrifugation to concentrate the cells will be inefficient at best, and may not allow for proper formation of a cell pellet.

1. Prepare sampling tubes by adding 10% by volume of EtOH/phenol mixture (95% EtOH, 5% phenol, by volume)
    a. For *Synechocystis*, an A_{730} reading of about 1 for cells growing in exponential phase tends to give enough RNA for 1-3 arrays in 50 ml of sample
    b. For test samples, 50 ml tubes were prepped with 5 ml of EtOH/phenol mixture
    c. For control samples, 250 ml tubes were prepped with 25 ml EtOH/phenol mixture

2. Store sampling tubes at -30°C until ready to take samples
3. Collect sample
    a. If using sparging vessel, collect sample by closing the gas outlet and opening the sampling tube directly into the collection tube. It may take time for enough pressure to build to force the culture up the sampling tube for the first time
    b. If collecting cells from a flask, simply pour the contents into the sampling tube
4. Quickly shake closed sample tube to ensure mixing and place tube into liquid nitrogen
    a. DO NOT FREEZE SAMPLE
    b. Manual shaking during cooling is recommended to ensure the sample is homogeneously cooled
5. Centrifuge sample at ~0°C for 5 minutes at highest rate possible without tube breakage
6. Discard supernatant liquid in hazardous waste container (due to the presence of phenol)
7. Place tube with solid cell pellet back into liquid nitrogen until completely frozen
8. Store sample at -80°C until RNA extraction step (section 5.2.3)

5.2.3 *Synechocystis* RNA extraction

Extraction of RNA is particularly difficult in *Synechocystis* because of its tough outer cell wall. In general, chemical or enzymatic means of lysing the cells are insufficient, and the cells must be lysed through physical means. The following protocol uses grinding to destroy the cell wall through shaking of the cultures with beads, but other methods (*e.g.* rapid pressure changes) may be used.

This protocol is an adaptation of the bacterial protocol in the RNeasy Midi/Maxi Handbook (Qiagen) and the required columns and buffers are included in the RNeasy Midi or Maxi kits (Qiagen). For 50ml sample tubes, the Midi columns were used with the steps listed below. For 250ml samples, Maxi columns were used with the same steps – however, when the protocols differ, the Maxi parameters are listed in brackets. In my experience, four 50ml sample tubes or 2 250ml sample tubes can be easily processed in parallel. Note that total RNA is extracted in this procedure, including both mRNA and tRNA, and that no DNAse step has been used to eliminate DNA from the sample.

1. Add 450 µl β-Mercaptoethanol to 45 ml of Buffer RLT (Qiagen)

2. Add 3.5 µl [7.5 µl] of the Buffer RLT solution directly to the still-frozen sample tube
    a. Vortexing may be required to fully dissolve the pellet
3. Pour contents into a single [a pair of] 7 µl flat bottomed glass vial
4. Add 3.5 µl of glass beads (0.1mm, B. Braun) to each vial and seal tightly
    a. The vial will not be completely full due to packing of the beads
5. Shake the tubes, in pairs, in a mixer mill for 2 minutes
6. Place tubes on ice for ~2 minutes to keep them cool
    a. A second pair of tubes may be placed in the shaker at this time, and alternated with the primary pair for the shaking steps
7. Shake for another 2 minutes
8. Ice again, 2 minutes
    a. Optional: the cycle may be repeated once more, but no major differences in RNA yield have been found
9. Pipette liquid from each vial into a clean 15ml tube
    a. Care should be taken to remove as much liquid as possible, even if beads are extracted as well
    b. Do not discard vials until after Step 10
10. Rinse each vial with an additional 1-2 ml of Buffer RLT solution
    a. Vortex the vials to rinse the beads thoroughly
11. Pipette the additional liquid into the corresponding 15ml tubes and discard vials
12. Centrifuge lysate for 5 minutes at 3200g
13. Pipette supernatant liquid from each vial into a new 15ml tube
    a. Make sure to avoid pipetting any cell material or beads into the new tubes
    b. [Recombine liquid from matching tubes for Maxi protocol]
14. Add 2.5ml [5.5ml] of pure EtOH to each tube and shake vigorously
15. Pour contents into a Midi [Maxi] column in a 15ml [50ml] tube
16. Centrifuge the tube for 3 min at 3200g and discard flow-through
    a. For Midi kits, sample must be added to the column in 2 steps because there is not enough room for the entire sample; the first centrifugation should therefore only last for a few seconds to clear enough space in the column to add the remaining sample

17. Add 4ml [15ml] Buffer RW1 to the column and centrifuge for 3 minutes at 3200g, discarding flow-through
18. Add 2.5ml [10ml] Buffer RPE to the column and centrifuge for a few seconds at 3200g, discarding flow-through
19. Add another 2.5ml [10ml] Buffer RPE to the column and centrifuge for 5min [10min] at 3200g
    a. The longer centrifuge time in this step is required to ensure the column is completely dry – otherwise, the EtOH in Buffer RPE may inhibit downstream reactions
20. Transfer column to a new 15ml [50ml] collection tube
21. Add 150 µl [0.8ml] of H_2O to the column and let stand for 1 minute
22. Centrifuge for 3 min at 3200g
23. Repeat the elution step
    a. If the sample will be concentrated with LiCl precipitation (less recommended), re-pipette original 150 µl [0.8ml] of liquid back into the column, let stand 1 minute, and centrifuge 3 minutes at 3200g
    b. If the sample will be dried in a vacufuge, apply a new 150 µl [0.8ml] aliquot of H_2O to the column, let stand 1 minute, and centrifuge 3 minutes at 3200g
24. Concentrate the sample
    a. LiCl precipitation procedure:
        i. Add 75 µl [0.4 ml] of 4M LiCl and the sample to a new 1.5ml microfuge tube
        ii. Refrigerate for >1/2 hour at ~-15°C
        iii. Centrifuge at 17,500g at ~2°C for 15 minutes
        iv. Pipette and discard supernatant liquid
    b. Vacufuge procedure
        i. Place sample into a new 1.5ml microfuge tube
        ii. Spin dry either without heating or with only minimal heating to minimize RNA degradation
        iii. Stop drying when < 1 µl remains, as completely dry RNA pellets do not seem to work well in downstream processes

25. Resuspend pellet in RNAse-free H_2O

i. For 50ml sample tubes of *Synechocystis*, 10 µl usually gives moderate RNA concentrations ~3 µg/µl

ii. For 250ml sample tubes, 40 µl usually gives higher concentrations of RNA ~5 µg/µl

iii. In either case, the vacufuge drying procedure usually gives higher final concentrations than the LiCl precipitation procedure

5.3 DNA microarray protocols

These protocols, while tested specifically on *Synechocystis* cultures, should be generally applicable to any samples of similar concentrations.

5.3.1 Printing of arrays

For the studies in this thesis, Dupont Co. provided plates of full-length cDNA that were spotted to create the microarrays. These samples represented 3078 unique cDNA sequences and a few sequences eventually eliminated from *Synechocystis* ORF databases[11]. These were shipped dry in 384 well (16x24) plates with v-bottomed wells (Genetix) and resuspended in 5 µl 50% (vol.) DMSO in H_2O and stored at -80°C until printing.

Arrays were printed on a MicroGrid II quill pin microarrayer (*Bio*Robotics) at 35-45% relative humidity at room temperature on Corning Gap slides. See Figure 5-5 for an example slide. A 4 by 4 array of quill pins was used to print with a 0.29 pitch (290 µm spacing) between spots, with spots printed in a 15 by 15 sub-grid in each of the 16 super-grids.

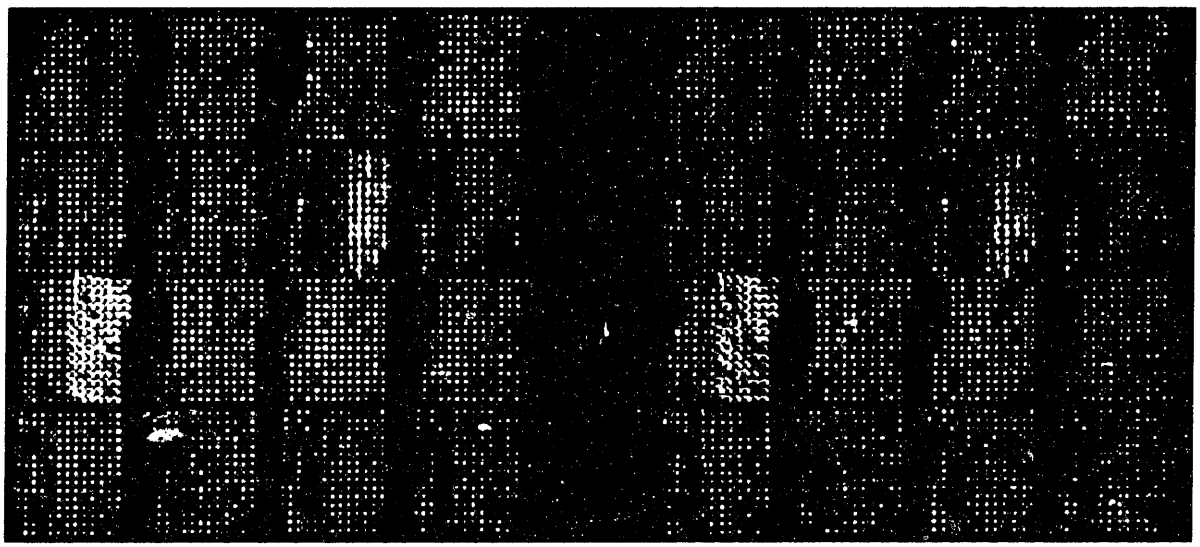

Figure 5-5: Example *Synechocystis* microarray

For each dip into the cDNA wells, 3 taps were first performed on each of 4 spare slides to remove excess liquid from the quills. Then printing was performed at one tap/slide for each of 104 slides. This entire procedure was repeated again on the bottom half of each slide, so a total of 2 dips into each well resulted in a total of 232 taps onto arrays. This entire procedure took about 20 hours to print all 3078 cDNA samples in duplicate on 104 slides. Slides were then crosslinked in batches of 18 slides using a Stratolinker set to the "Autocrosslink" option at 1200μJ, and were stored in the dark until use.

5.3.2 RNA processing into labeled cDNA

For this procedure, the samples were processed with random hexamer primers of the form (dNTP)$_6$ in an attempt to reverse-transcribe all of the transcripts present. Success has also been reported in our lab with primers specific to transcripts of interest.

5.3.2.1 Original procedure

This procedure was used for most of the arrays performed in this study, but was eventually changed as catalogued in section 5.3.2.2 with some improvement in the resulting arrays.

26. Start with 10-15 μg of RNA (3 μg/μl optimal concentration) + 5 μl

27. Add 1 µl of random hexamer primer
28. Place in a heating block at 65°C for 10 min (H_2O is usually added to the heating block to increase heat transfer)
29. Place sample on ice for 2 min to quench primer hybridization
30. Add sufficient H_2O to ensure a final total volume of ~20 µl
    a. Calculate this volume based on volume needed for later steps and the concentration of RNA in the sample
    b. If initial volume of RNA is large (~10 µl) add no H_2O. Add alternative amounts of the reagents in Step 31 as noted in brackets
31. Add:
    a. 2 µL DTT (10x) [2.5 µl]
    b. 4 µL 1st Strand Buffer (5x) [5 µl]
    c. 2 µL of dNTP (10x) [2.5 µl]
        i. 5 µl each of 100mM dCTP, dATP, dGTP
        ii. 2 µl 100mM dTTP
        iii. 83 µl of H_2O
        iv. Yields 100 µl of 10x dNTP
    d. 2 µL of Cy3 (sample) or Cy5 (control) [2.5 µl] (mix well)
    e. 2.2 µL of Superscript II [2.8 µl] (mix well)
32. Place in water bath at 42°C for > 2 hrs (close to 2 hr is most desirable)
33. Add 1.5 µl NaOH (1 N) [2 µl]
34. Place sample in water bath at 65°C for 10 min (destroys RNA)
    a. Simultaneously prepare DyeEx Column (Quiagen)
        i. Open cap ½ twist, break off tab and insert into collection tube
        ii. Centrifuge 3 min/700 g/room temp
        iii. Discard collection tube and put gel matrix into a new 1.5 µl tube
35. Place sample on ice for 2 min
36. Add 1.5 µl HCl (1 N) [2 µl] to sample (to neutralize)
37. Add control sample (Cy5) to its corresponding sample (Cy3) and mix well
38. Add mixture to DyeEx column

39. Centrifuge column for 3 min at 700 g at room temp (discard column matrix)
40. Add 100 µl cold (-20°C), pure EtOH + 16 µl sodium acetate (3M)
41. Place solution at –15°C for >1/2 hr
42. Centrifuge at 17,500g, ~0°C for 15 minutes
43. Pipette out supernatant liquid (be sure to remove ALL liquid, or hybridization will be impaired)
44. Samples may be stored at –30 °C, but usually samples are processed immediately as in section 5.3.3

5.3.2.2 Adjusted procedure

As in section 5.3.2.1, with the following changes to improve cleanup of unreacted dNTPs and dyes by replacing the DyeEx kit with the QIAQuick kit (Quiagen).

9. Eliminate sub-steps

13. QIAQuick processing
    a. Add 490 µl Buffer PN [580 µl] to sample
    b. Apply to QIAQuick column
    c. Centrifuge at 6000rpm for 1 minute and discard liquid
    d. Add 750 µl Buffer PE
    e. Centrifuge at 6000rpm for 1 minute and discard liquid
    f. Centrifuge at maximum for 1 minute to dry
    g. Place sample in new tube
14. QIAQuick Elution
    h. Add 30 µl H_2O
        i. Wait 1 minute
        ii. Centrifuge at maximum for 1 minute
    i. Add 20 µl H_2O
        i. Wait 1 minute
        ii. Centrifuge at maximum for 1 minute

Instead of EtOH precipitation, cDNA may be concentrated through drying in a vacufuge.

15. Spin dry in vacufuge until volume ~1 µL (dry is OK, but not preferred)
16. Eliminate this step

5.3.3 Hybridization of samples

Usually this step is performed immediately after processing of RNA to create labeled cDNA (section 5.3.2).

17. Add 32 µl of hybridization fluid (Clontech)
18. Place at 95 °C for 8-10 minutes
    a. If necessary, store samples in 50 °C heating bath until slides are ready
    b. Simultaneously prepare slides
        i. Boil slides for 2 minutes in H_2O
        ii. Quench slides in cold (-20°C), pure EtOH
        iii. Spin dry slides at 300-500g for 2 minutes
    c. Simultaneously prepare hybridization chambers
        i. Pipette ~20 µl of water into each well
19. Place dry slides face-up in chamber
20. Spin down sample (low g's) and pipette onto slide surface
    a. Optional – take 1 µl of sample and save for spectrophotometer reading
21. Lower slide cover over sample
    a. Make sure to wet entire microarray surface – generally, applying one edge of the slide cover to the surface and gently dropping the other edge onto the slide works well
    b. Watch for bubbles trapped under slide cover – in general, these will work their way out from under the cover due to surface tension if there is enough liquid, but if not, small amounts of pressure may be applied with forceps to the cover to force bubbles out
    c. All steps must be done quickly, or the hot sample will evaporate somewhat and the chances of the slide drying out are much greater
22. Close chamber, seal, and place gently into 50°C water bath

a. 55°C and 60°C baths have also worked, but drying becomes more of a problem
b. 40°C or 45°C will result in more cross-hybridization

23. Allow hybridization to occur over ~12-14 hours
    a. Longer than 20 hours frequently causes problems with drying
    b. Less time may be acceptable

5.3.4 Cleaning and scanning of arrays

This step is always performed immediately after hybridization is complete (section 5.3.3) to limit the possibility of samples drying before they are rinsed, because labeled cDNA tends to stick to the surface of slides when the surfaces dry. In general, after hybridization, rinsing, and drying, slides can be stored in the dark for an undetermined amount of time (at least one week) before scanning without any major degradation to the signal measured on the array.

24. Rinse samples in 3% SDS/ 20X SSC (175g NaCl and 88.2 NaCit • 2 H_2O in 1 liter of H_2O) solutions – 5-7 minutes each rinse
    a. 10ml 20X SSC, 10ml 3%SDS, in 100 ml H_2O
        i. Be sure to remove slide covers as soon as they slide naturally (without force) from slides placed vertically in the solution
    b. 5ml 20X SSC, 1ml 3%SDS, in 100ml H_2O
    c. 1ml 20X SSC in 100ml H_2O
    d. 3ml 20X SSC in 600ml H_2O
25. Spin dry slides at 300-500g for 2 minutes
26. Scan each slide in Axon scanner
    a. "Preview" each slide with default intensities of ~500mV for each channel
    b. Adjust intensities to get scanned intensity ratio of ~1 on the "Histogram" screen, noting that the histogram shows results only for on-screen section of slide
    c. Autoscale image as necessary to ensure a full range of spot intensities can be observed
    d. Adjust scan boundaries to cover spotted surface
    e. Perform full scan
27. Save images

a. Preferably as 2-channel tiff files, which retain the information needed to rescale each channel separately

28. Find microarray features
    a. Use automatic spot-finding options, followed by manual examination of spots to ensure all features on the array have been highlighted
    b. Eliminate obviously damaged spots
29. Save settings file
    a. A different settings file is recommended for each slide, in case slides need to be re-analyzed
30. Use the "Analyze" icon to collect data from the outlined spots
31. Save output data file
32. Normalize and process the data file as necessary (see Chapter 6)

5.4 References

1. Stanier, R. Y. & Cohen-Bazire, G. "Phototrophic Prokaryotes: The Cyanobacteria." *Annual Review of Microbiology* **31**, 225-274 (1977).

2. Silva, S. *Personal communication*, (2003).

3. Miller, A. G. & Colman, B. "Active-Transport and Accumulation of Bicarbonate by a Unicellular Cyanobacterium." *Journal of Bacteriology* **143**, 1253-1259 (1980).

4. Williams, J. G. K. "Construction of Specific Mutations in Photosystem-Ii Photosynthetic Reaction Center by Genetic-Engineering Methods in Synechocystis-6803." *Methods in Enzymology* **167**, 766-778 (1988).

5. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E. Y., N., N., M., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okomura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. & Tabata, S. "Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. 2. Sequence determination of the entire genome and assignment of potential protein-coding regions." *DNA Research* **3**, 109-36 (1996).

6. Hein, S., Tran, H. & Steinbuchel, A. "Synechocystis sp. PCC6803 possesses a two-component polyhydroxyalkanoic acid synthase similar to that of anoxygenic purple sulfur bacteria." *Archives of Microbiology* **170**, 162-170 (1998).

7. AZoM - the A-Z of materials. http://www.azom.com

8. Photosynthesis at Imperial College. http://www.bio.ic.ac.uk/research/barber/index.html

9. Schluchter, W. "UNO Biological Sciences faculty web site."

10. Pasteur Culture Collection of Cyanobacteria: Culture media. http://www.pasteur.fr/recherche/banques/PCC/Media.htm#BG11

11. CyanoBase: The Genome Database for Cyanobacteria. http://www.kazusa.or.jp/cyano/

12. Roberge, C. *Personal communication*, (2003).

CHAPTER 6 NETWORK DISCOVERY EXPERIMENTS

As discussed in Chapter 5, it has been previously reported that changing the level of light fueling the growth of photoautotrophic *Synechocystis* alters the transcriptional profile of the cell culture[1,2]. This provides an adjustable input stimulus for dynamically manipulating the cellular state. For our network discovery experiments, we altered this light level many times over a period of almost 17 hours, keeping a single large culture of cells from settling into a steady-state while collecting samples at 50 evenly-spaced time points. Full-genome DNA microarrays were used to determine the transcriptional profile at each time point. Time-lagged correlations[3], discussed in section 4.1, were then applied to find genes that were highly correlated to the input signal of light intensity. These genes were clustered and collated into a network of highly correlated groups.

The expression profiles of the genes of these groups were used to build AutoRegressive with eXogenous input (ARX) models. We then used these models to predict expression levels for a variety of potential follow-up experiments, and selected a "maximally informative" confirmatory experiment. This experiment of an additional 27 time points with a new pattern of light input was carried out to test the accuracy of the ARX models. The resulting models contain predictive information about transcriptional behavior and form the basis for formulating hypotheses to be tested in future experiments.

6.1 Experiment 1: network training data

In the first experiment, a series of shifts in the experimental conditions between three light levels ("light," "low light," and "dark") was used to ensure the cells did not settle into a steady-state. This forcing function was chosen in an attempt to allow time for major effects from each transition to be observed (as inferred from the analysis in Gill *et al.*[1]) and to include most of the potential transitions (*i.e.* dark to low light, light to low light, *etc.*).

6.1.1 Experimental details

This experiment was conducted according to the protocols discussed in Chapter 5. In brief, batch cultures of *Synechocystis* sp. strain PCC 6803 were maintained in BG-11 medium at 30 °C.

1% of a 4M solution of Na_2CO_3 was added as a carbon source for maintenance and inoculation cultures. The experiment was carried out in a 10 liter glass vessel with 6 liters of working volume sparged continuously with 1% CO_2 air at a rate of approximately 150ml/min. This sparged gas eliminated the need for Na_2CO_3. In this experiment, HEPES (10mM, pH 8.5) was added as a buffer to the BG-11 medium to counteract acidity caused by the dissolved CO_2 gas.

Cultures were typically grown at about 7000 lux under cool fluorescent bulbs, corresponding to photosynthetic photon flux (PPF) of ~90 μmol/m^2/s. Note that this light level well below the light level at which *Synechocystis* has been shown to grow without significant light damage[1,2]. For "low light" experimental conditions, some of the incubator bulbs were turned off, resulting in light readings of about 1200 lux (PPF ~16 μmol/m^2/s). In "dark" conditions, all bulbs were extinguished and a box was placed over the vessel, resulting in negligible light input (< 20 lux) compared to the other two conditions.

Cells were grown to mid-exponential phase ($A_{730} \sim 1.0$) with an approximate doubling time of 12 hours. A large reference sample was taken directly from a culture at this condition. For the time-series experiment the cells were then left in the "dark" conditions for 24 hours before the experiment initiation with the first switch to the "low light" condition. The culture density remained constant (with changes of < 10%) after the experiments were initiated (as in Gill *et al.*[1]).

Microarray quality, sample processing, and data filters were analyzed in aggregate by hybridizing to three microarrays 6 samples from cultures grown in parallel and labeled with different dyes. On average, the expression ratio of 8.8% of the genes differed by greater than 1.75, while only 5.1% of the genes differed by more than 2-fold. Both of these measures are consistent with other cDNA microarray experiments[1,4-6]. Based on these results, genes were deemed to be differentially expressed if they exhibited an expression ratio of 2-fold or greater with respect to the control. A compilation of all duplicate spots within the es gave a within-slide coefficient of variation of 0.18.

An example of such a control experiment is given in Figure 6-1, with the three diagonal lines representing (from top to bottom) 2-fold induction, equal induction, and 2-fold repression of the Cy3-labled sample relative to the Cy5-labled one. Contrast this figure with Figure 6-2, which

shows the high dissimilarity between the transcriptional profile of the reference (moderate light) sample and a sample from a "dark" condition.

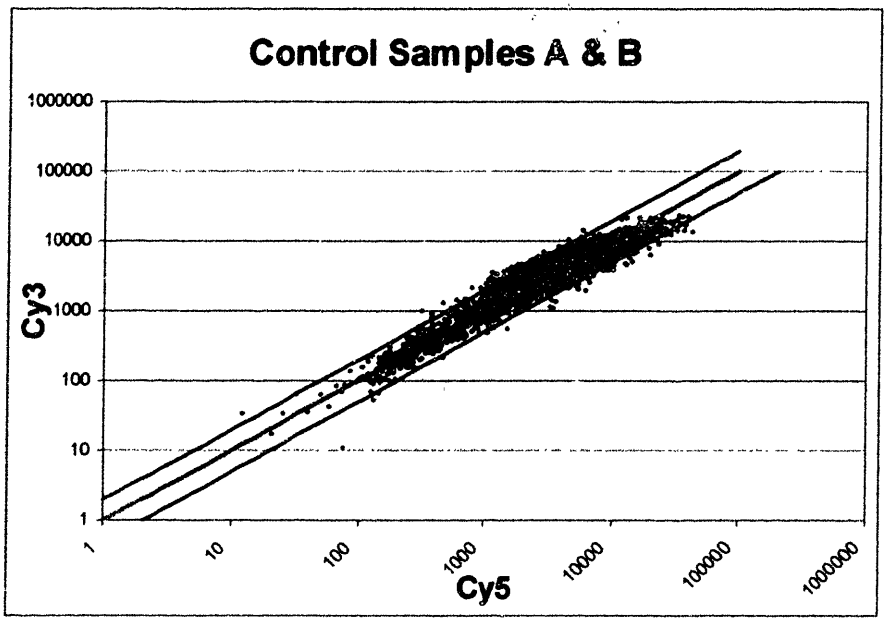

Figure 6-1: Example two-channel control DNA microarray experiment

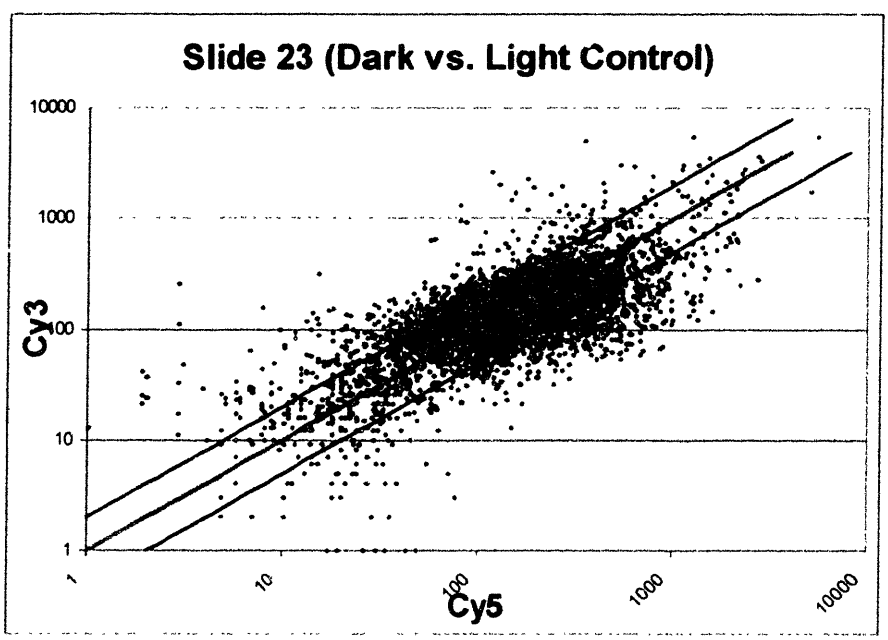

Figure 6-2: Example two-channel DNA microarray experiment

Samples were taken and stored as described in Gill *et al.*[1] except that samples were added directly to a 10% vol. mixture of ethanol (95%) and phenol (5%) before chilling with liquid N_2 (see section 5.2.2). RNA extraction and purification, labeled cDNA creation, and hybridization were all performed exactly as in Gill *et al.* for the first experiment. The Cy5 labeled sample was taken from a pool of reference RNA extracted from cells grown in an identical experimental setup to mid-exponential phase under moderate light conditions continuously from inoculation. Slides were scanned with an Axon Instruments 4000B scanner. Laser intensity was adjusted manually on each channel to achieve strong spot readings relative to the background (without saturating the detector) at both the 532nm and 635nm wavelengths. All data was filtered to eliminate spots not significantly expressed over background noise and normalized to set the average logarithmic expression ratio to one.

6.1.2 Time-lagged correlation implementation methodology

The step-wise procedure for implementation of time-lagged correlation analysis is enumerated below, adapted from section 4.1. Specifically, this algorithm has been designed for use with *Synechocystis*, but it has been created to be generally applicable to similar types of regulatory reconstruction problems.

Step 1: Filter data to remove genes without signal intensity above noise level, and cluster genes that may represent operons of co-expression

Step 2: Calculate time-lagged correlations between the input signal and each gene, and each cluster of genes created in Step 1

Step 3: Cluster genes found in Step 2 to combine genes with highly correlated expression (without time-lags)

Step 4: Expand the networks of connected groups by repeating Steps 2-4

Step 5: Draw the correlated network

6.1.2.1 Step 1: Filter genes and cluster genes into potential operons

Many genes on a cDNA microarray may not have significant expression for the conditions measured, or may be represented by spots on the array that are prone to cross-hybridization with cDNA from other transcripts. In either case, such genes muddle the computational picture unnecessarily, and are easily filtered by the application of filters for not only significant expression level but also significant expression change. Genes that were not determined to have a measurable signal experimentally (that is, spots on the array with intensities not significantly above background noise) for over half of the data points in our experiment were excluded from further consideration. Also, all genes without a significant expression change (here defined at 2-fold induction or inhibition, although this measure could differ for more accurate microarrays) for at least one time-point were eliminated from further analysis. For our studies, such filters typically eliminated less than 1/3 of the genome from further consideration.

The methods of clustering of data to reduce the scope of the problem and exclude uninteresting features has been explored by many researchers[7-12] for both DNA microarray and other data sources. Methods such as hierarchical clustering[8], principal component analysis[13,14], and self-organizing maps[11] provide methods for grouping genes into related profiles, but all are tightly linked to the parameter values chosen by the user. For this work, a simple system that is amenable to high-throughput analysis was desired, and therefore the number of parameters to be input by the user was minimized.

For *Synechocystis* (see section 5.1.1) as well as many other organisms, additional information for each gene is available in the form of its position and ordering within the genome[15]. This information, along with the experimental expression data, may suggest the existence of operons, or genes that are co-expressed due to a common upstream promoter system. For this work, instead of traditional clustering, genes were instead analyzed for correlation of expression (with zero time-lag) with other genes adjacent on the genome. Those that were correlated were grouped into clusters and their average autoscaled profile was calculated to represent the entire group of adjacent genes before proceeding with the rest of the analysis.

As an example, consider the sub-section of the *Synechocystis* genome shown in Figure 6-3[15]. The numbers 565-580 refer base-pair position in the genome, in thousands, and cs0076 refers to

the sequenced contig including the genes shown. Each of the arrows represents a gene and its orientation: genes found on the complementary strand of the DNA are shown at the bottom, pointing from right to left. The sequence starting at slr2033 and ending at slr2035 includes a number of interesting genes:

rubA (slr2033), a membrane-associated rubredoxin,

ycf48 (slr2034), which is essential for stability or assembly of photsystem II,

psbEFLJ(ssr3451, smr0006-8), sub-units of photosystem II and the cytrochrome b_6f complex, and

proB (slr2035), glutamate 5-kinase

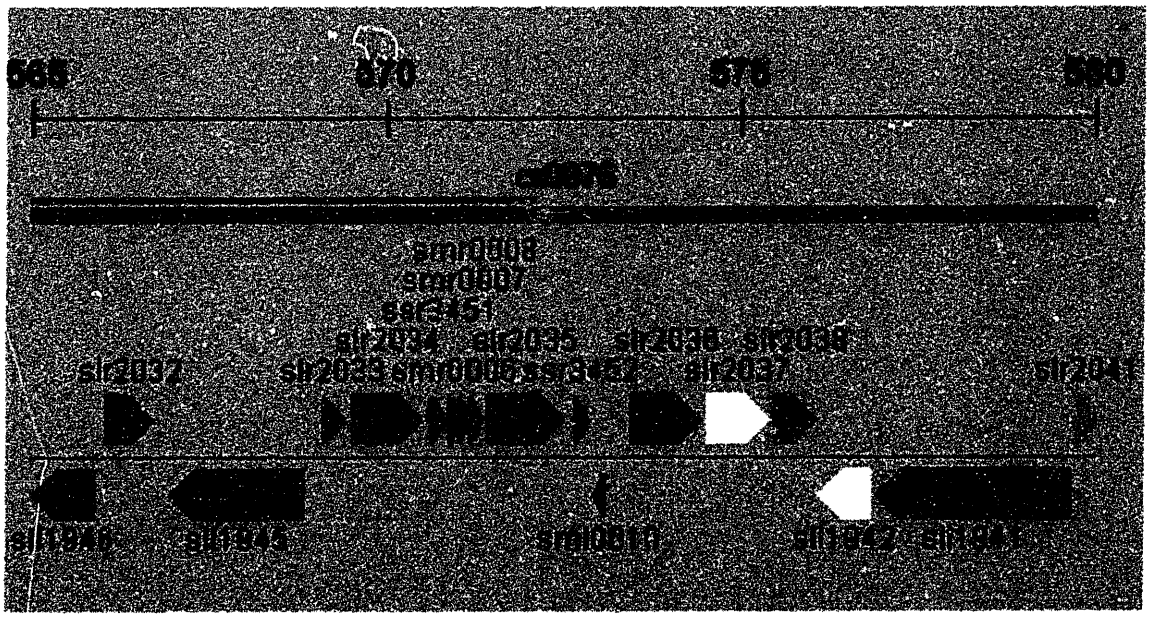

Figure 6-3: Schematic of *Synechocystis* genome section

The transcriptional data from Experiment 1 was used to calculate the correlation of expression for each gene in this sequence against the others. The results are shown in Table 6-1. Note that the middle 5 genes in this sequence, marked in bold, are all correlated at levels greater than 0.65, while the first and last genes are not well correlated with any of the others. This suggests that these middle 5 genes are co-expressed because they may be part of an operon, with common

regulatory elements upstream of the group. Thus, these 5 genes are clustered before continuing the analysis to reflect this possibility.

Table 6-1: Correlation between sequential genes

	slr2033	slr2034	ssr3451	smr0006	smr0007	smr0008	slr2036
slr2033	**1.0000**	-0.4895	-0.4114	-0.3835	-0.3894	-0.3468	-0.3082
slr2034	-	**1.0000**	**0.8617**	**0.7857**	**0.8075**	**0.6508**	0.4469
ssr3451	-	-	**1.0000**	**0.9190**	**0.8928**	**0.7842**	0.3980
smr0006	-	-	-	**1.0000**	**0.9216**	**0.7133**	0.2476
smr0007	-	-	-	-	**1.0000**	**0.8166**	0.3790
smr0008	-	-	-	-	-	**1.0000**	0.4401
slr2036	-	-	-	-	-	-	**1.0000**

6.1.2.2 Step 2: Identification of clusters correlated with the input

The remaining clusters were then analyzed with time-lagged correlations as presented in section 4.1. This technique aims to identify correlated gene expression patterns with allowance for time lags between the expression levels. Recall from section 4.1.1 the correlation formulation, comparing gene (or cluster of genes) *i* to gene *j*.

$$S_{ij}(\tau) = \left\langle (g_i(t) - \bar{g}_i)(g_j(t+\tau) - \bar{g}_j) \right\rangle \qquad 6\text{-}1$$

$$r_{ij}(\tau) = \frac{S_{ij}(\tau)}{\sqrt{S_{ii}S_{jj}}} \qquad 6\text{-}2$$

This correlation best identifies relationships of the type

$$g_1(t) = Ag_2(t - \tau_0) + B \qquad 6\text{-}3$$

Substituting this relationship into the correlation equations shows that a maximum S_{12} value will occur at $\tau = -\tau_0$. At this time-lag, $S_{12} = A \cdot \sigma_2$ where σ_2 is the variance of g_2. This condition corresponds to $r_{12} = 1$. However, as is shown in the case studies in Chapter 3, most gene expression values are likely to be the result of a combination of effects such as

$$g_1(t) = Ag_1(t-\tau_1) + Bg_2(t-\tau_2) + ... + aP_n(t-\tau_n) + ... \qquad 6\text{-}4$$

where P are protein concentrations in the system which we may be unable to measure directly. In this case the pairwise correlation will be less than one. However, if we assume that most of the genes and proteins present do not actually impact the gene during the experiment being studied, then we may reduce the relationship to a few key interactions, each of which shows imperfect correlation with the gene in question. By using reasonably low threshold values for r_{ij} a set of imperfect connections with appropriate time lags can be calculated.

Another difficulty of this method is the nature of the data used for the correlation calculation. A set of case studies using this technique (not presented here) shows that the accurate determination of time-lags and correlations is highly dependent on the existence of *features* in the data set, where *features* are defined as data points which vary around the signal mean. This has also been noted by Arkin *et al.*[3] who suggest continuous perturbation of the system away from steady-state to ensure that features exist to help uncover the underlying system structure. Because of this, experimental data that represents development of a single event in time (such as viral infection in the T7 bacteriophage example) is poorly suited for use with this technique. Thus, all experiments in our studies are run with a "large" number of changes relative to the amount of time required to reach steady-state after such a change.

The practical difficulty with this framework is that the complexity of calculation of time-lagged correlations for thousands of genes increases exponentially as the number of genes increases. To simplify the calculations further, in the first pass, an input signal representing experimental conditions (for the lactose/arabinose case in section 3.1, this variable would be sugar concentration; in these experiments conducted for *Synechocystis*, light intensity) was compared to each gene sequentially and all genes having at least one $r(\tau)$ value greater than the preselected cutoff were set aside for further consideration. In this way, a set of "first-order" interactions was

obtained in a computational time increasing linearly with the number of genes included, and further iterations were then run to expand the correlation network.

6.1.2.3 Step 3: Cluster groups found in Step 2

Regulons are similar to operons as they include co-expressed genes under the control of common promoter regions. However, regulons differ from operons because they are not necessarily sequentially oriented in the genome. Thus, it is important to consider similarly expressed genes for potential clusters of expression, as the expression of these genes may represent a common regulatory effect. Those genes with high time-lagged correlation at the same time-lag compared to the experimental conditions are good candidates for such clusters.

Here the retained genes were sorted by their time-lags with the input signal – all genes with lags of 1 were put into one category, lags of 2 into another, etc. Then a nearest-neighbor[9] clustering scheme was implemented with correlation (time-lag zero) as the definition of similarity. In the nearest neighbor approach, each gene/cluster is compared to all other genes/clusters, and the most-closely correlated two are paired. This procedure was repeated until the correlation between groups fell below the cutoff value selected for Step 2. In this way, the difference between these clusters is at least as strong as the differences between the correlated and uncorrelated genes found in Step 2.

6.1.2.4 Step 4: Expand the groups by repeating Steps 2 & 3

Each of the discovered groups can then be used as a "seed" node in the same way the input signal (*i.e.* light intensity) was in Step 2. In this way, the network can be expanded to arbitrarily large size (compare Figure 6-5 to Figure 6-7, below). In general, a new correlation cutoff could be selected here to either encourage inclusion of genes into the network or promote exclusion and focus on "core" interactions. If the cutoff is chosen too low, however, the network could expand to a completely incomprehensible degree for systems including thousands of genes such as *Synechocystis*. For these studies, a stringent cutoff was used.

6.1.2.5 Step 5: Draw network of interactions

The Graphviz program from ATT, presented in 4.1.3, was used to draw the sequences of lagged correlations derived from Steps 2-4. Each gene/cluster was fixed in a temporal hierarchy based on zero-lag for the input signal: if one group has a time-lag of 1 interval (20 minutes) after the input signal, and second group has a lag of 2 intervals from the first group, then the second group is placed on the third tier below the input signal. Other than this constraint, Graphviz was free to optimize the arrangement of the groups so as to minimize the overlap between their connecting lines, creating easily interpreted jpeg images. Examples can be found below in Figure 6-5 through Figure 6-7.

6.1.3 Results

Consider Figure 6-4. The pattern of input light is shown as a solid bar (with "light," "low light," and "dark" conditions represented accordingly), and the scaled expression pattern of each gene in the *psbEFLJ* operon are plotted with thin lines. All data has been autoscaled to show the expression pattern instead of absolute values. Note that all of the genes are closely correlated to the input signal (the solid line), but generally to lag it by about 1 time point, or 20 minutes in this experiment.

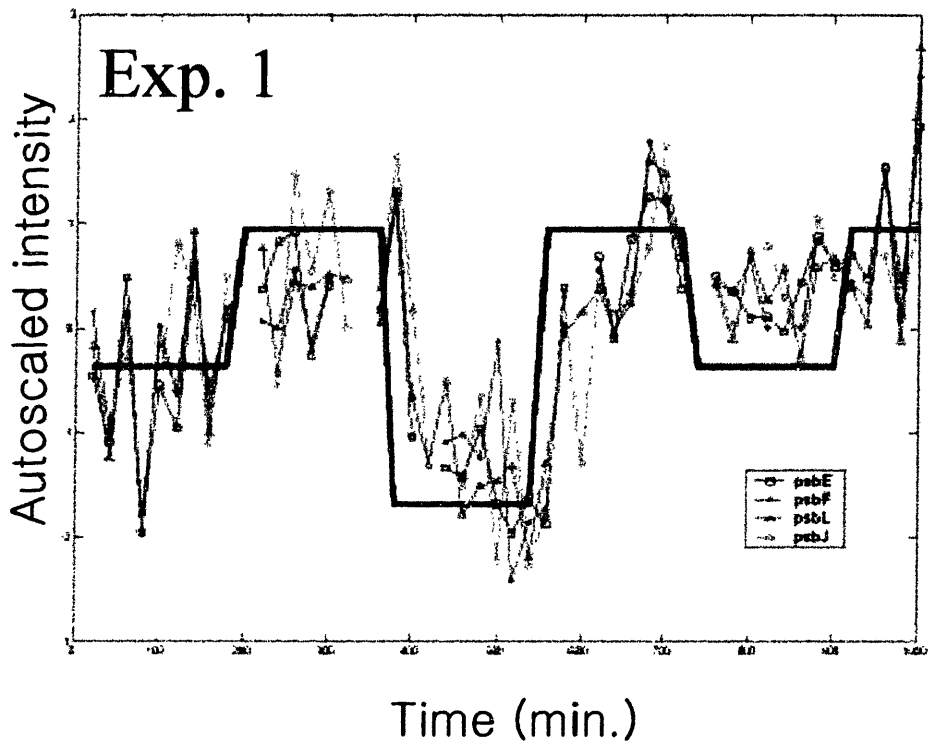

Figure 6-4: Input light (solid line) and gene expression profiles, *psbEFLJ* operon

All such correlated genes are then complied into a network, using the algorithm discussed in section 6.1.2. The genes found on the first iteration (*e.g.* those genes most directly correlated with the input signal) are shown in Figure 6-5. Here, clusters of co-expressed genes are shown in groups, where dark arrows indicate between-group lagged correlations. For example, the input light signal is highly correlated, with a lag of 1 interval (20 minutes), to the genes of Groups 2-6. Direct correlation is also observed with Group 7 at a 40 minutes lag. These groups were then expanded in the second iteration by searching for genes with high correlation with these groups, even if less significant correlation is observed against the input signal. In Figure 6-6, only those groups with a time-lag of zero relative to the primary groups identified by the first iteration have been added. Lines without arrowheads represent these zero-lag connections. Figure 6-7 includes all lagged and non-lagged correlations, with the appropriate time scales relative to the changes in input light intensity. Dashed lines show inverse correlations in all

figures in the same manner as direct correlations. For the first iteration, the cutoff correlation value was $R = 0.7$, while $R = 0.5$ was used for the second iteration.

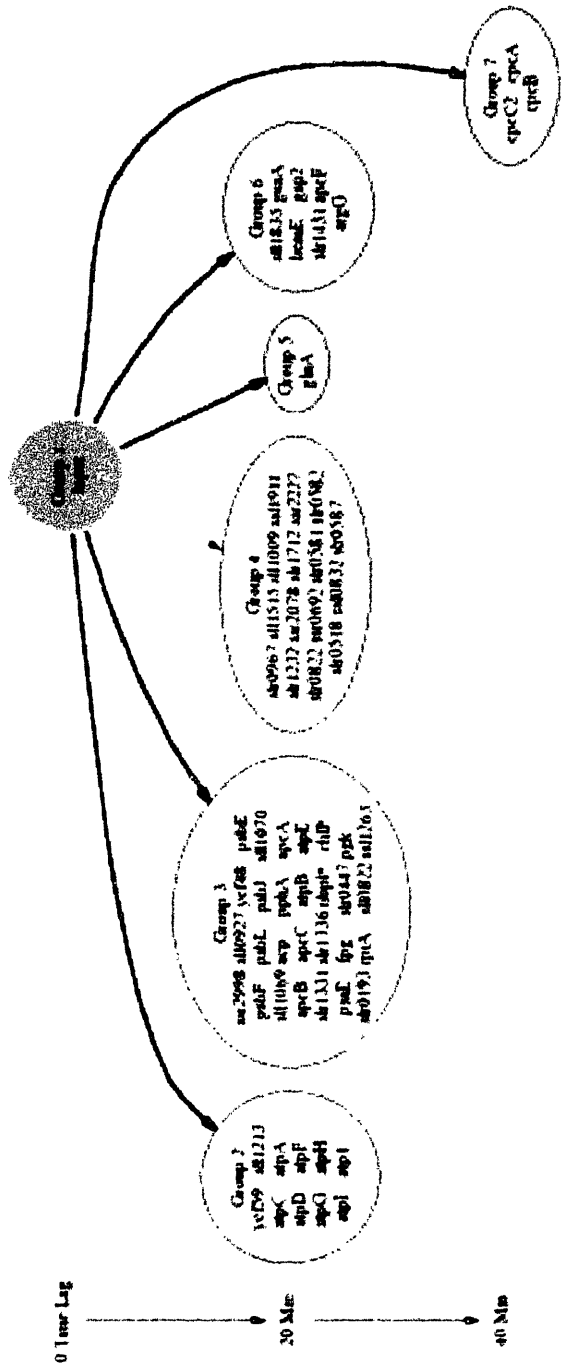

Figure 6-5: Network construction, first iteration

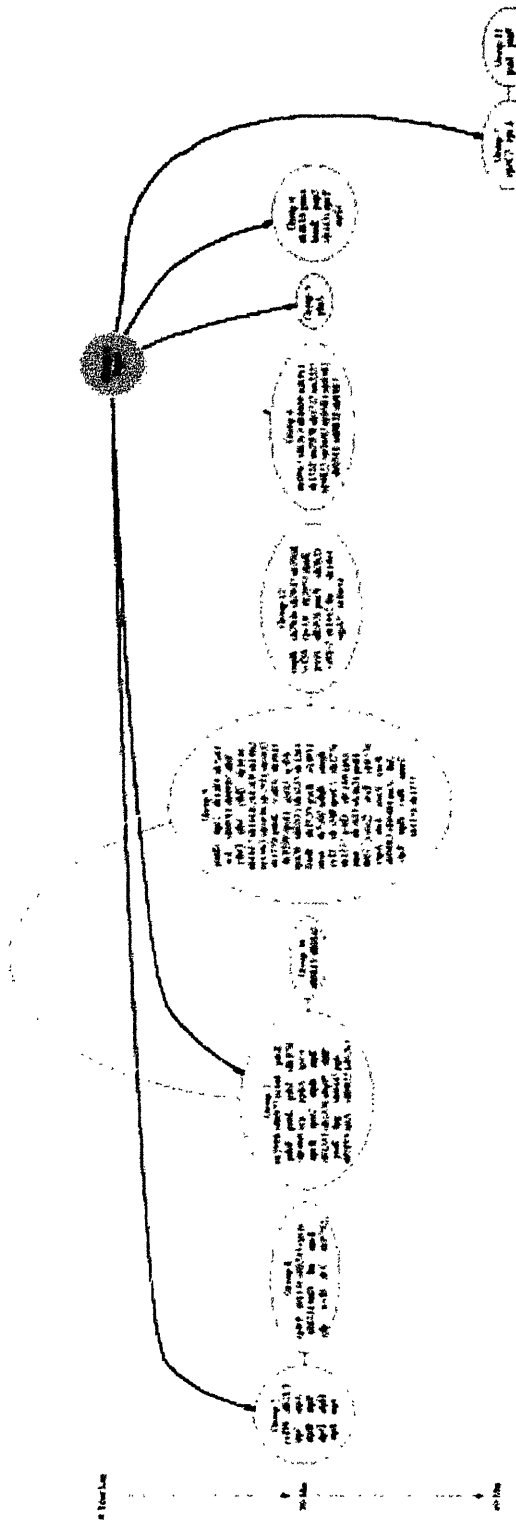

Figure 6-6: Network construction, second iteration, zero-lag clusters

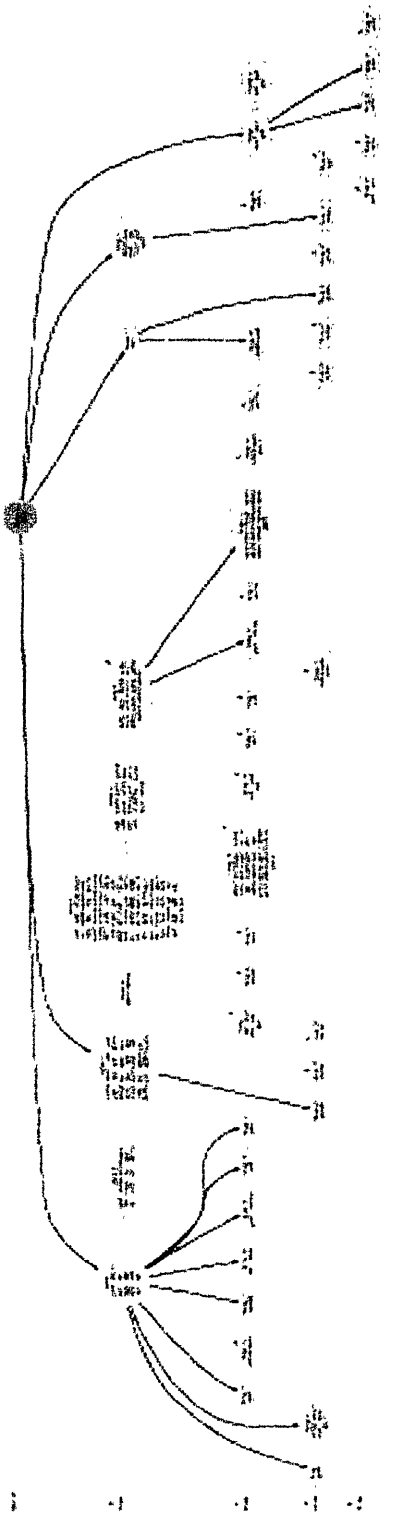

Figure 6-7: Network construction, second iteration, all clusters

For simplicity, the named genes in each time-delayed "wave" relative to the input signal are listed in Table 4-2 (hypothetical, unnamed genes have been excluded from this list). Particularly abundant are many of the transcripts associated with proteins in the *Synechocystis* photosystem complexes. For example, genes associated with photosystem II (such as *psbEFLJ* in Group 3 – see Figure 6-5) and photosystem I (such as *psaE* or *psaK*) seem to be activated at several different time-lags relative to the light intensity. On the other hand, many of the sub-units for ATP synthase (such as *atpCADFGHII* in Group 2 – see Figure 6-5) are best correlated with the light intensity at the smallest measurable time-lag of 20 minutes. At least one sub unit for the cytochrome b_6f complex (*petG)* is also identified. Interestingly, this analysis also finds both *ycf3* and *ycf48* have transcriptional expression coordinated with light exposure with the minimum time-lag of 20 minues. It has been suggested that they play a role in either assembly or stability of photosystem I[16] and II[17]. The fast response, at the transcriptional level, of these genes to changing light conditions is consistent with these hypotheses.

Table 6-2: Genes in correlation network at appropriate time-lags

20min:		
accB,efp		**acp**
apcA,B,C		**apcE,F,G**
atpBE,	**atpC,A,D,F,G,H,I,1**	
bioB	**chlP**	**clpP, trpB**
trpE, psaD		**ycf58, cpcG1**
crtQ-2	**ctpA, rbcL,X,S**	
cupB	**fus**	**fpg, psaE**
tufA, fus		**dnaK, glyA**
gap2	**glnA**	**gpx1**
guaA	**gyrB**	**hemB**
hemE	**icd**	**ilvC**
murC	**nbp1**	**ndhH**
nirA	**pacS**	**petH**
pgk	**ppa**	**pphA**
ycf48, psbE,F,L,J		**psbK**
purD	**rfbFGC**	**rfbE**
rpl19	**rpl36,rps11,rps13**	
rpoC1	**rps1a**	**rps20**
secDF	**serA**	**sigA**
thiC	**valS**	**ycf23**
ycf3	**ycf59**	

40min:		
ccmK	**gap1**	**glnB**
cpcA,B,C2,C1,D		**hemD**
natE	**nblA1**	**ndbB**
ndhD1	**psaC**	**ndhJ**
petF	**petG**	**pppA**
psaL,I	**psaF,J**	**psaK**
psbI	**psbX**	**ribF**
rpl21,27	**rps15**	**rps21**
rps4	**serS**	

60min:		
hspA	**psbB**	**psbM**

80min:		
hisD	**ycf46**	**pxcA**
rpl24		

Other genes which have been previously identified in our lab and elsewhere as being light-regulated, such as the allophycocyanin genes *apcF* (Group 6), *apcE* (Group8), and *apcABC* (Group 3)[1] are also found grouped with other highly correlated genes. No that these allophycocyanin genes are all found at time-lag of 20 minutes, while several phycocyanin genes, such as *cpcABC2C1D*, are found at a later time lag (Table 6-2 – note that *cpcC1* was filtered from the original analysis due to low measured expression ratios, but is in fact well-correlated with the other genes listed). Given that the allophycocyanin units make up the core of the phycobilisome structure while the phycocyanin genes make up the rod-like projections from this core (see section 5.1.1.1, especially Figure 5-3), a model of sequential activation seems plausible. Furthermore, *cpcG1*, found at the earliest measured time-lag, links the phycocyanin rods to the core allopycocyanin proteins of the phycobilisome[18].

As reported in earlier studies[1,2,19], the sub-units of the carbon-dioxide fixation complex rubisco (*rbcL, rbcS,* and the potential chaperone protein *rbcX*) are shown to be highly correlated, at the transcriptional level, with light intensity. Other findings, including a homologue of the carbon dioxide concentration unit *ccmK*[19] as well as a handful of genes related to metabolism (*icd, gap2,etc.)* are also catalogued in Table 4-2 and Figure 6-7.

Also interesting is the role of any of the hypothetical genes such as *slr0581* and *slr0582* in Group 4, which are in fact inversely correlated to the light intensity. A homology search using BLAST on these ORFs suggests no strong homologies with known proteins, so assigning functional role is difficult. However, *slr0582* has at least some homology with putative binding factors, and therefore may play some role in the transcription of genes regulated as a response to light. This pair of adjacent genes, with operon-like co-expression and a high correlation to light-regulated genes in *Synechocystis*, is worthy of further study.

Since some correlations may be expected by random chance (although for the levels of R chosen here, the probability of observing such correlations by chance approach one in millions, as shown in section 6.3.1) a second experiment, with a different input light signal, was conducted to confirm or contradict the correlations observed in the first experiment.

6.2 Experiment 2: network validation data

The second experiment was done with a different forcing function in an attempt to maximize the information content from both experiments (as described in Schmitt *et al.*[20]).

6.2.1 Design of experiment

In order to determine how such an experiment should be conducted to ensure it would be maximally informative, it was necessary to construct a model of transcriptional behavior to allow for prediction of the validation experiment's results. We predicted the output profiles caused by a given input light intensity profile for each pair of connected groups in Figure 6-6 using ARX models as described in section 4.2. See Figure 6-8 for an example model fit for Group 2 from Figure 6-6. Here an ARX441 model (*e.g.* 4 prior measurements of the output variable, plus 4 prior measurements of the input variable lagged by 1 interval are used to predict the next value of the output variable) model is used for demonstration purposes only.

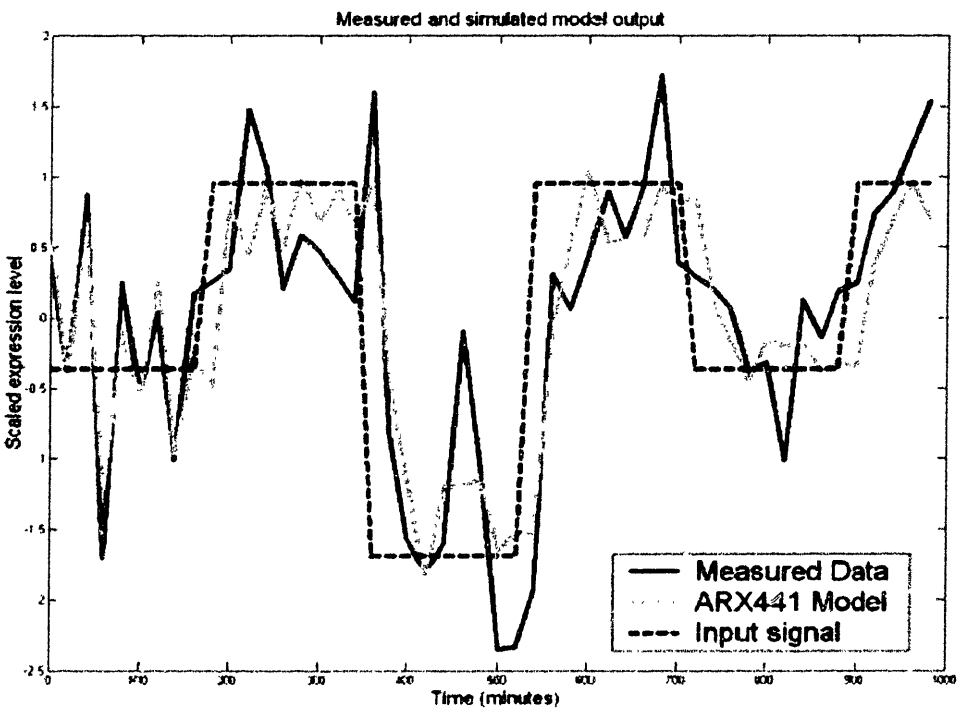

Figure 6-8: ARX model fit to the average expression profile of Group 2

In order to decide which ARX models were most appropriate for this study, a method for their comparison was needed. Goodness of fit is a metric often used to compare models of different type or complexity. However, calculating the least-squared error (or % of data variance explained) can be misleading for the data used to train the model, as the number of parameters used generally improves the goodness of fit. To rank models while considering both the number of parameters used and the goodness of fit simultaneously, Akaike's information criterion (AIC)[21,22] was employed, as discussed in section 4.2.2.

As a test case, the prediction sum-of-square-error (SSE) and AIC criterion for all possible models varying in complexity between ARX111 and ARX663 were calculated for the Group 2/Group 8 pair (shown in Figure 6-6). Figure 6-9 summarizes the results, with the minimum SSE and AIC value plotted for each model complexity. In general, models with larger numbers of parameters have improved values of SSE. On the other hand, the AIC criterion shows a minimum at 12 parameters, in this case corresponding to an ARX221 model. Repeating this exercise on other pairs of gene clusters within the data set give similar results with similar model orders (data not shown).

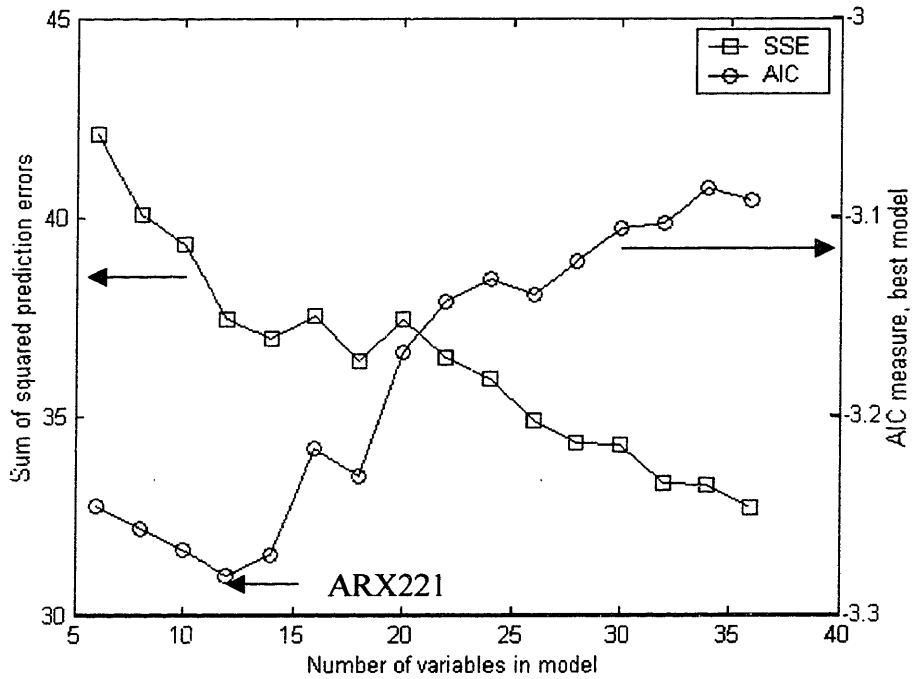

Figure 6-9: Prediction errors *vs.* AIC criterion for ARX models

These models were then used to predict the output of new experimental profiles – in this case, the level of light that the cells were exposed to. In order to choose an "optimal" validation experiment, our aim was to find a profile which was predicted to create the greatest output discrepancy between the clusters shown in Figure 6-6 (note that not all of the clusters found in Figure 6-7 were used, as it was deemed more important to accurately determine primary interactions than more secondary ones). The metric used to judge this was the square of the difference between each of the modeled output profiles – that is, the expected difference the between output variables representing each of the gene clusters.

All possible input profiles were generated subject to the following constraints:

1. the experiment should have a 20 minute sampling interval (to ensure Δt consistent with the time interval of the model)

2. the input light intensity should be limited to the three levels used in the original experiment (to avoid using the model for extrapolation)

3. changes in the input light signal should not to be performed more often than every 2 hours

4. 27 total time points should be included

The third constraint was purely based on observations from earlier experiments, as it can be difficult to resolve which input signal changes are most directly responsible for which output features in experiments with more frequent shifts[1]. The final constraint was imposed by the resources available for this study.

A summary of sum-squared differences is presented in Figure 6-10. We expect that the profile which maximized the differences between the groups can be used as a set of optimal experimental conditions for the follow-up experiment[20]. This optimal profile is shown in Figure 6-10. Other, similar profiles are predicted to give similar between-group differences, but these profiles are not significantly different (a transition between light levels shifted by one time point, for example). Contrast this with a minimally informative profile, also shown in Figure 6-10. Such a profile, with only one state change, is predicted to allow the transcriptional profile of most genes to reach steady-state (data not shown), limiting the experiment's effectiveness at validating or disproving the measured correlations from Experiment 1[3].

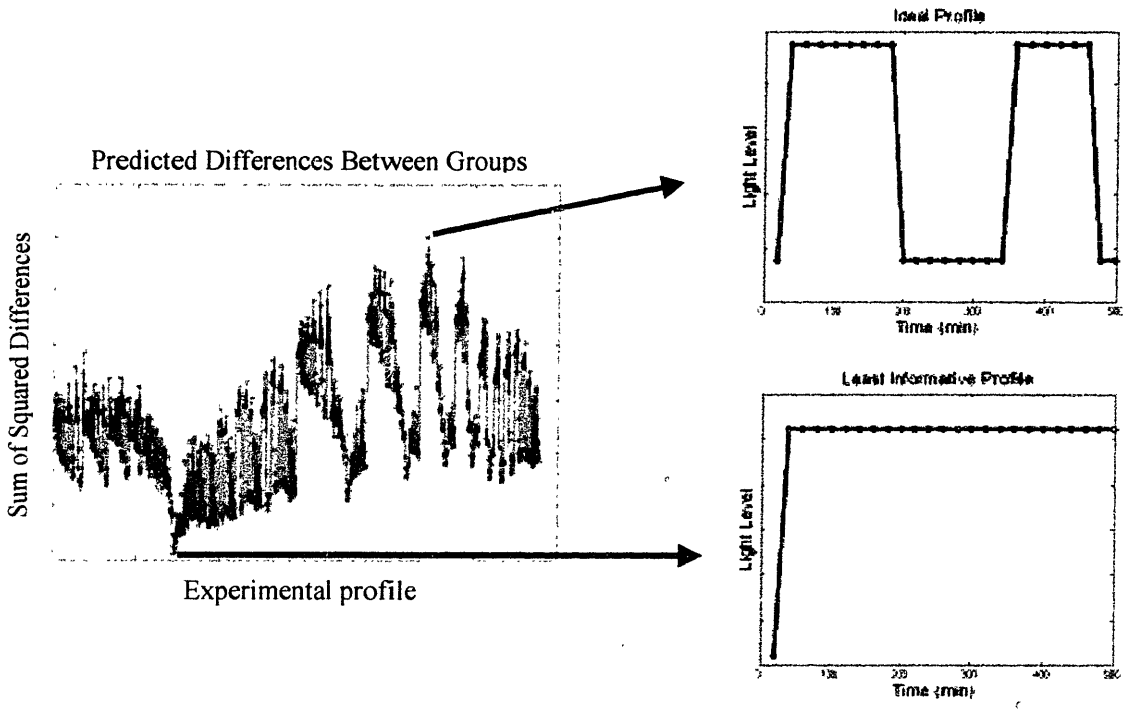

Figure 6-10: Group differences predicted by ARX models for a collection of inputs

6.2.2 Experimental details

All experimental procedures were carried out as described in Chapter 5 and elsewhere[20]. Processing of samples in this experiment differed from the first only in the way cDNA was handled. Labeled cDNA samples were further processed with a Qiaquick nucleotide removal kit (Qiagen) for elimination of unreacted species. Furthermore, instead of EtOH precipitation, samples were spun-dry in a Vacufuge to increase their concentration (see section 5.3.2.2). In all cases, the Cy5-labeled sample was taken from a reference pool of RNA extracted from cells in the identical experimental setup in mid-exponential phase grown at moderate light conditions continuously from inoculation.

6.2.3 Comparison of testing and validation data

Results of the second experiment for the genes shown in Figure 6-5 are shown in Table 4-2. Most of the genes match up well between the two experiments, but exceptions could then be used to "prune" or adjust the network shown in Figure 2. Consider, for example, the *cpc* genes found in Group7, which seem to correlate less well in the second experiment. Although the

correlation is still significant at this level for a time-lag of 2 units (40 minutes), a nearly equal correlation at a lag of 3 units (60 minutes) has also been observed (data not shown) and further experiments may be warranted to accurately plot these genes within the networks being drawn. However, unless exceedingly low correlation is observed, these connections cannot reasonably be rejected.

Table 6-3: Time-lagged correlation values, experiments 1 and 2

	Expt 1		Expt 2
Gene	R value	lag	R value
ycf59	0.7076	-1	0.8459
sll1213	0.6528	-1	0.6282
atpC	0.7104	-2	0.6834
	0.6954	-1	0.5538
atpA	0.7621	-1	0.7062
atpD	0.7656	-1	0.6996
atpF	0.7887	-1	0.7376
atpG	0.7923	-1	0.7986
atpH	0.7515	-1	0.726
atpI	0.7174	-1	0.7129
atp1	0.6957	-1	0.6956
ssr2998	0.7368	-1	0.5498
sll0927	0.7577	-1	0.7832
ycf48	0.6153	-1	0.617
psbC	0.773	-1	0.8076
psbF	0.6935	-1	0.7679
psbL	0.7051	-1	0.7225
psbJ	0.6534	-3	0.4751
	0.6454	-1	0.4812
sll1070	0.7236	-1	0.5628
sll1069	0.6787	-1	0.5806
acp	0.7625	-1	0.734
pphA	0.737	-1	0.3816
apcA	0.7081	-1	0.7729
apcB	0.7122	-2	0.7714
	0.7101	-1	0.7686
apcC	0.7757	-1	0.7301
atpB	0.6959	-1	0.7359
atpE	0.7519	-1	0.7067
slr1331	0.6175	-1	0.5313
slr1336	0.7095	-1	0.4099
nbpl*	0.7482	-1	0.6761
chlP	0.782	-1	0.7006

	Expt 1		Expt 2
Gene	R value	lag	R value
psaE	0.7258	-1	0.626
fpg	0.5831	-1	0.4718
slr0447	0.7872	-1	0.7425
pgk	0.7521	-1	0.7304
slr0193	0.7216	-1	0.6882
rpiA	0.6194	-1	0.4247
sll0822	0.7776	-1	0.7791
ssl1263	0.7031	-1	0.7185
slr0967	0.8524	-1	0.8587
sll1515	-0.7392	-1	-0.8205
sll1009	-0.7878	-1	-0.275
ssl1911	-0.7832	-1	-0.8907
slr1232	-0.7922	-1	-0.8773
ssr2078	-0.7062	-1	-0.7272
slr1712	-0.7016	-1	-0.6647
ssr2227	-0.8571	-1	-0.9117
slr0822	-0.7292	-1	-0.7255
ssr0692	-0.7791	-1	-0.8701
slr0581	-0.7666	-1	-0.5221
slr0582	-0.6686	-1	-0.4439
slr0518	-0.7619	-1	-0.9146
ssl0832	-0.7446	-1	-0.9486
slr0587	-0.7757	-1	-0.8577
glnA	0.7044	-1	0.7902
sll1835	0.7433	-1	0.7475
guaA	0.706	-1	0.6698
hemE	0.7058	-1	0.6841
gap2	0.8028	-1	0.7914
slr1431	0.7106	-1	0.7032
apcF	0.7222	-1	0.7767
argG	0.7397	-1	0.7682
cpcC2	0.6376	-2	0.2069
cpcA	0.7095	-2	0.4941
cpcB	0.7299	-2	0.593

Another comparison can be made by examining the composite transcriptional profiles in a lower-dimensional space. Data for hundreds of genes can be simplified into a few composite factors through Principal Components Analysis (PCA). PCA has been discussed by a number of researchers as a method for systematically reducing the dimensionality of data in an unsupervised, or data-driven, fashion[13,14,23,24]. Although there are a few different ways that this technique can be applied, we use the method of *singular value decomposition* (SVD) applied

directly to the original data. Essentially, SVD decomposes a matrix X with dimensions of n rows by p columns as follows:

$$\underset{(nxp)}{X} = \underset{(nxR)}{U}\ \underset{(RxR)}{T}\ \underset{(Rxp)}{V'} \qquad 6\text{-}5$$

where T is a diagonal matrix whose values are the singular values of the matrix X. The singular values of X are defined as the square roots of the nonzero eigenvalues of the square matrix $X'X$ as well as XX' (where X' is the transpose of X). The columns of U and V contain the eigenvectors of XX' and $X'X$, respectively. The maximum number of dimensions, R, is determined by the rank of the matrix X.

The *loadings* of the genes for each of the Principal Components is given by the column vectors of the matrix V. The *projection* of the samples, or the *scores* of the samples on the principal components, is given by:

$$S = X\ V' \qquad 6\text{-}6$$

In this study, only the first two principal components were retained, representing the eigenvectors of the two largest eigenvalues. Figure 6-11 shows the projection of all 74 samples into this two-dimensional space. Because some genes have missing values for some of the time-points, only genes represented by at least one unfiltered spot for all 74 samples (47 from the first experiment, 27 from the second) were considered. For these 113 genes, the two largest components account for approximately 68.7% of the variance information of the total data set. 95% and 99% confidence limits, shown as dashed circles in Figure 6-11, are based on the assumption that the variance of the 74 samples will follow a "t" distribution along each principal component. Examination of Figure 6-11 shows that this is obviously not the case – however, the confidence limits are still useful as a guide to the differentiation which exists between the samples.

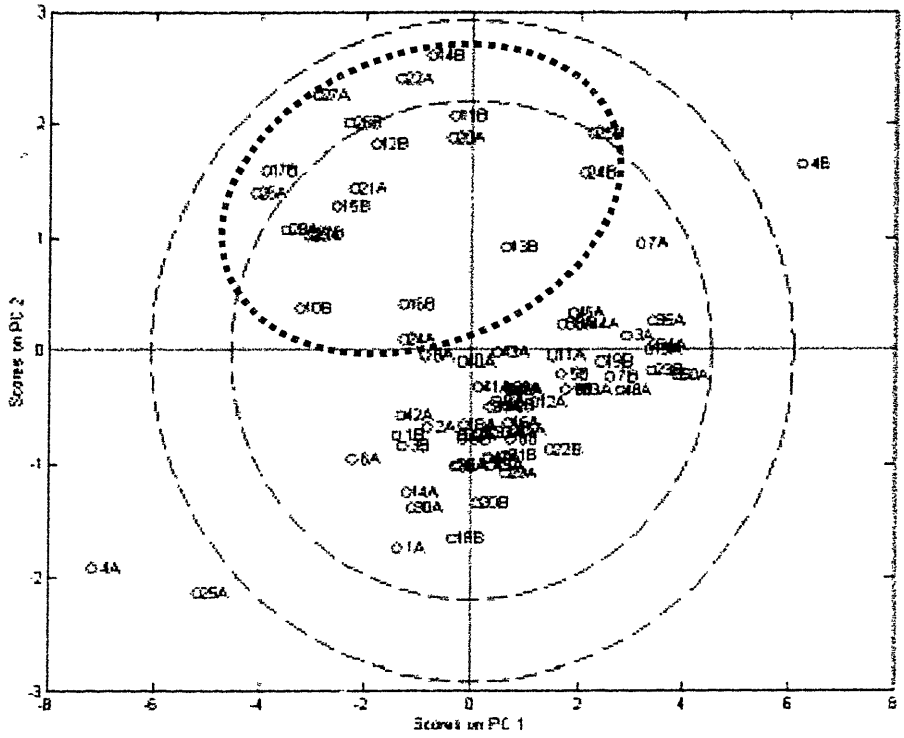

Figure 6-11: PCA projection of all 74 samples

Note that the area circled by the dotted line contains all but one of the samples taken under "dark" conditions (with a time-lag of one). Sample 25A, the one "dark" condition sample not within this cluster, is one of the 3 points that is most likely to be an outlier (see bottom-left quadrant of Figure 6-11). This unsupervised approach (PCA) shows clearly that results from the first experiment are consistent at a macro-level with the second experiment, as light and dark samples are systematically differentiated.

Some further insight can be gained by examining the loadings of the genes (the columns of the matrix *V*) in this PCA analysis, as shown in Figure 6-12. Here genes closest to the origin are least relevant to the position of the samples in Figure 6-11, while the genes farther away are more important. Note the abundance of genes also included in Figure 6-7, such as *psb* genes (*psbA2, psbA3, psbB, psbC, psbD, psbD2*), *psaB*, and *cpcG1* (all in the rightmost two quadrants of Figure 6-12). Additionally, note the position of *slr0581* (discussed in section 6.1.2), one of

the genes found to be inversely correlated with the input light signal. In Figure 6-12, it is located in the upper left-hand quadrant, opposite of the positively correlated light induced genes. This is consistent with the correlation map of Figure 6-7. Because Figure 6-12 represents only a fraction of the genes measured, not all of the genes in Figure 6-7 are shown. Nevertheless, the existence of significant loading values for a number of those genes found by time-lagged correlation analysis gives an independent indicator of their importance.

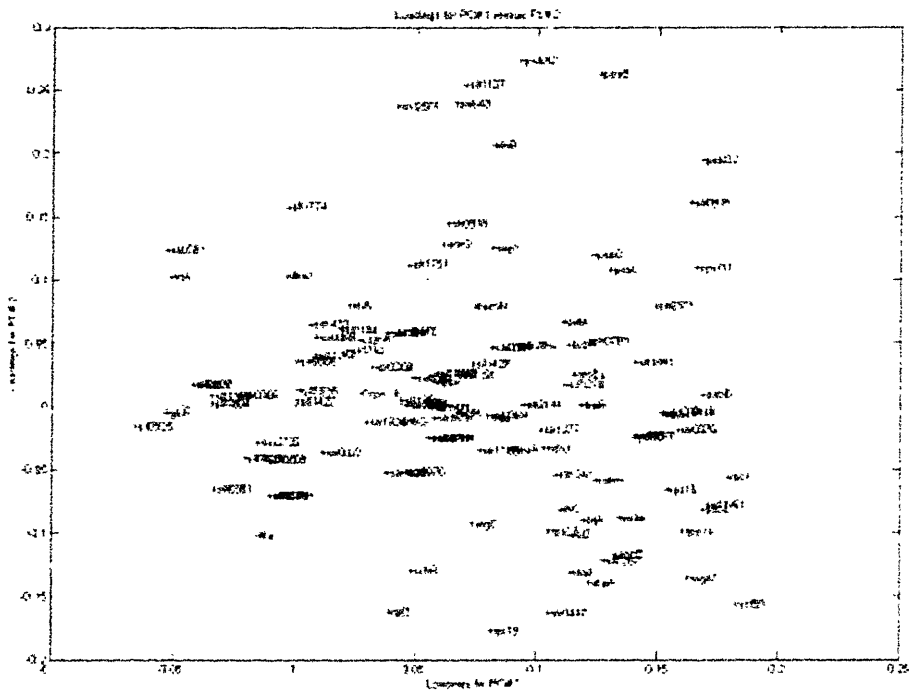

Figure 6-12: PCA loadings of all 113 genes used for PCA

The two experiments may also be compared by determining how well the ARX models developed in section 6.2.1 succeed in predicting the transcriptional profiles found in the second experiment. This is discussed separately in section 6.3.2 as part of the network validation studies.

6.3 Network validation

There are two ways to evaluate the robustness and validity of the network shown in Figure 6-7. The first is to determine how the parameters and measurements taken affect the networks drawn – in other words, how do small changes in the parameters or data change the conclusions of the experiment? Additionally, the data from the second experiment may be compared to the model predictions generated by analysis of the data from the first experiment. Both types of validation are discussed here.

6.3.1 Network robustness

In developing the time-lagged correlation algorithm that was used to generate Figure 6-5 through Figure 6-7, care was taken to consider the impact of the various parameters on the results. If changing the user-input parameters a slight amount drastically changes the network structure, then the results are called into question. Furthermore, it is important to have a metric for understanding the impact of experimental noise or chance observations. All of these issues were explored in this work.

6.3.1.1 The effect of *R* value selection

The most important input variable for construction of correlation networks as described here is the cutoff correlation R that determines the size of the network under consideration. Consider a shift in cutoff correlation from 0.7, as used in this study, to 0.75, the results of which are shown in Figure 6-13. Obviously, the number of genes under consideration is reduced from the original network shown in Figure 6-6, but this sub-set of genes merely indicates the most solid of the correlations observed in the original network. Furthermore, since there are fewer genes to consider, many less connections are observed.

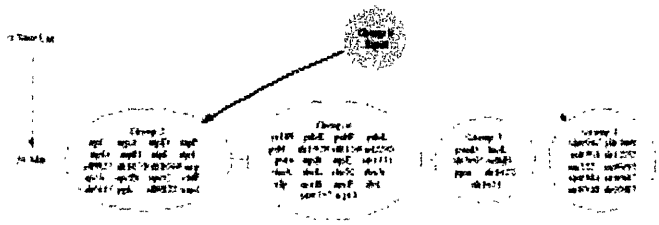

Figure 6-13: First-iteration correlations found with $|R| \geq 0.75$

On the other hand, decreasing the cutoff to 0.65 increases the number of genes included in the network. The general mapping of these genes compared to the original structure is shown in Figure 6-14. Arrows have been included in this figure to show the how the members of the gene clusters are mapped between the original diagram (top right) and the new diagram (bottom left). Since the cutoff value is lower, more genes are clustered into larger groups. Most important, however, is the fact that the overall network structure is robust to the change in R value. Thus the R value chosen offers a tunable parameter to include a greater or fewer number of genes, depending on the level of detail the researcher wishes to study, while only adding or subtracting components to the network structure and not changing significantly how they are arranged.

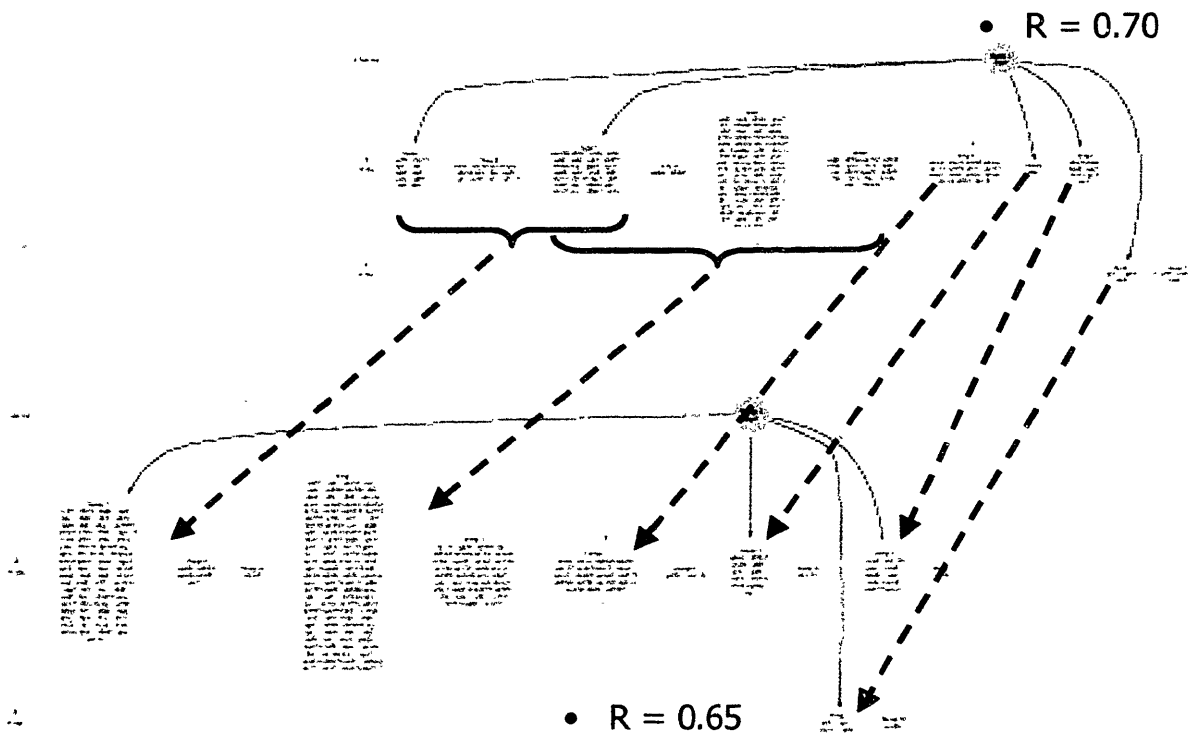

Figure 6-14: Effect of reduction of cutoff to $|R| \geq 0.65$

6.3.1.2 The possibility of correlations by random chance

Because R is the most important tunable parameter in this procedure, it is important to understand how spurious observations might be introduced at a given value of R. All correlation calculations are indicators of how improbable it would be to find correlated measurements for two otherwise independent variables. For Pearson correlations, we expect R to follow a t-distribution after the following transformation has been applied

$$t_{N-2} = \frac{R}{\sqrt{(1-R^2)/(N-2)}} \qquad 6\text{-}7$$

where N is the number of observations and t_{N-2} indicates a t-distribution with N-2 degrees of freedom. This metric can be used with standard distribution tables to determine the significance level of a given correlation (R) value.

For N = 47 (as in the first experiment), the significance of an R value of 0.7 is nearly 100% - that is, there is nearly no possibility of finding such a correlation by chance. For commonly available tables and even Matlab's statistics toolbox, no such probability can even be calculated to default machine precision. In fact, there is less than a 0.01% chance of finding spurious correlations with $|R| \geq 0.7$ for all $N > 24$.

To better quantify this, 10 million random profiles of size N=50 were generated using Matlab's "randn" function with seed values set to the computer's clock cycle. Allowing for time-lags of 2 or less (*e.g.* $-2 \leq \tau \leq 2$) only 2 profiles were found to have absolute correlation ($|R|$) values over 0.7. Since this test accounts for 50 million comparisons (5 lags per profile), it is exceptionally unlikely that any of the correlations found represent chance observations, and even more unlikely that a network as large as seen in Figure 6-7 could be observed.

One difficulty with this analysis, however, is the distributions of the variables being considered (*i.e.*, the profiles generated randomly through Matlab) do not necessary approximate actual DNA microarray data. One way to escape this is by shuffling the columns of the original microarray data as a control data set. Thus, the means and variances of the data set rows and columns are preserved, but the dynamic information content is destroyed to a greater or lesser degree, depending on the shuffling. Using this procedure, each "gene" profile could then be compared to the original, unshuffled light intensity profile to determine the frequency at which spurious time-lagged correlations are measured. Note that in this scheme, the correlations within groups will remain the same, since non-lagged correlation is independent of sample order. Therefore the important metric to consider is the number of connected groups; that is, the number of clusters of genes with lagged correlation compared to the input, instead of the number of individual genes found.

Ten thousand permutations of the original data set were explored in this fashion, as shown in Figure 6-15. Only nine of the ten thousand permutations exhibited any correlated groups, with only one example exhibiting more than 20 connections. Compare this histogram with the number of connections found using the actual data (shown in Figure 6-7), marked with an asterisk in Figure 6-15. This control also shows that it is exceptionally unlikely that the network developed in this work is a merely a result of an artifact in the DNA microarray data.

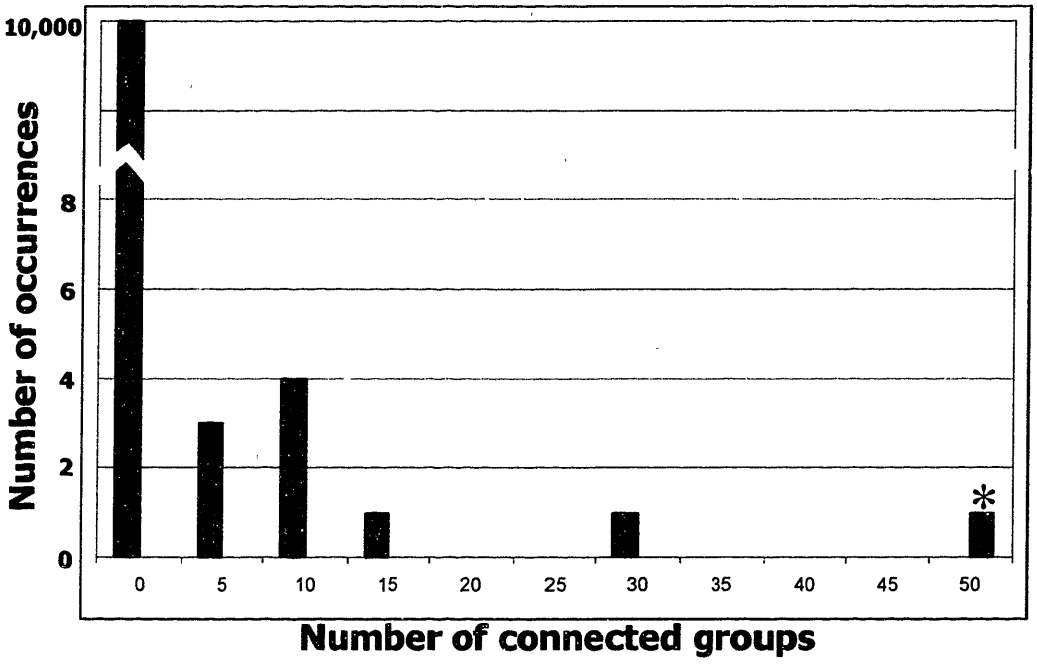

Figure 6-15: Correlations observed with shuffled time points

6.3.1.3 The impact of noise on network construction

As discussed in section 6.1.1 and shown in Figure 6-1, DNA microarray experiments have some measurable quantities of noise associated with not only the stochastic nature of the underlying biological populations[25], but also due to variability in the experimental techniques and even the microarray surface properties. By running control experiments such as those shown in Figure 6-1, a composite estimate for the effect of such errors can be made. For the arrays used in this study under these experimental conditions, it was determined that any changes greater than 2-fold control experiments data fell outside of the 95% confidence interval (see section 6.1.1).

In order to understand how this error could affect the network reconstruction effort, random noise with a normal distribution of similar limits (95% of noise below a factor of 2-fold in expression level measured) was added to increasingly greater fractions of the original data set. Consider Figure 6-16, which shows the percentage of genes from the original network that are retained when gradually increasing percentages of the original network are corrupted by randomly distributed noise. In the worst case, when all of the data has been corrupted by noise,

40% of the genes in the original network are retained, showing that noise can have a significant impact on the correlation calculations even though a core fraction of genes is robust even to severe data corruption.

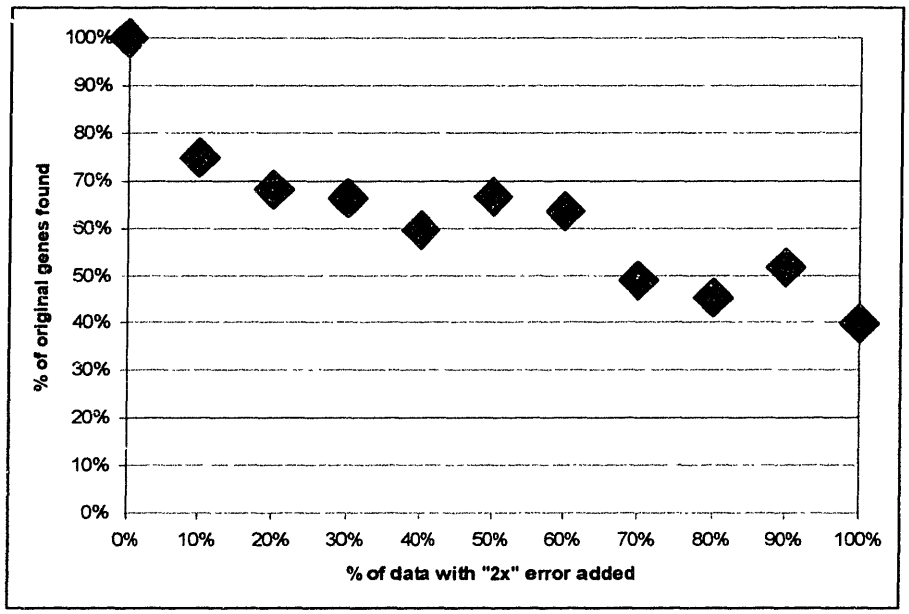

Figure 6-16: Affect of random noise on network identity

However, this figure assumes that network calculations are made at the same *R* cutoff value (0.7) used for the original, unadulterated data. Because randomly distributed profiles are expected to have an *R* value of zero when compared to any independent profile, adding random noise to the measurements is expected to reduce the *R* value which can be measured for even the most correlated variables. Figure 6-17 shows how a reduction of cutoff *R* value from 0.7 to 0.6 recovers completely the genes eliminated from the calculations when corrupting noise is added to 10% of the data. Greater increases in noise level obviously require even greater reductions in *R* to make up for the obscuring variability (data not shown).

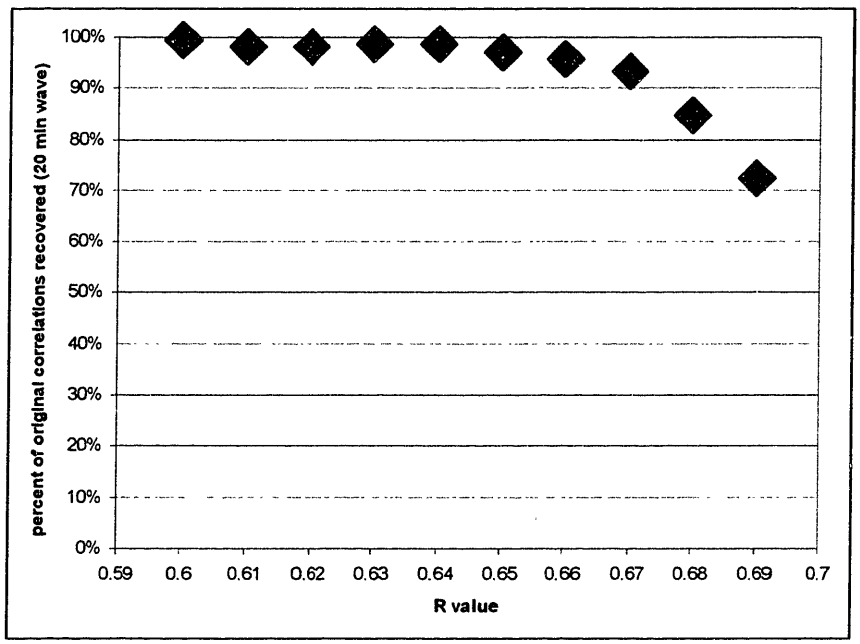

Figure 6-17: Reduction in cutoff *R* value required to compensate for random noise

Of course, any reduction in cutoff R is expected to increase the number of genes considered in the network (see section 6.3.1.1) and increase the chances that spurious correlations are included. Nevertheless, the robustness of core relationships in the network even when all of the data is corrupted by random noise increases our confidence in the derived network structure.

Although there will always be some variability in biological measurements due to molecular stochasticity, advances in DNA microarray technology promise to drive purely experimental error to increasingly negligible levels. This will enhance our ability to accurately determine correlation networks from transcriptional data.

6.3.2 Prediction of new profiles

The data from the second experiment was compared to all ARX models calculated from the original experimental data – see Table 6-4. The model selected by the AIC criterion for Group 2/8 (ARX221, see Figure 6-9 in section 6.2.1) has the twelfth-best fit to the validation data (of the 108 models tested) as measured by SSE. An overlay of the SSE values from both the training and validation data sets is shown with the corresponding AIC predictions in Figure 6-18.

As predicted by AIC, a model with 12 parameters minimized prediction error, validating the model order required for accurate simulation of the system.

Table 6-4: Validation of the 20 best ARX models by the AIC criterion

SSE(Training)	AIC	SSE(Validation)	& Rank	na	nb	nk	# of parameters
37.421	-3.2801	25.7276	12	2	2	1	12
36.933	-3.2697	30.289	27	2	3	1	14
39.32	-3.2669	22.9296	3	2	1	1	10
40.06	-3.2567	25.3088	10	1	2	1	8
42.076	-3.2447	23.4419	5	1	1	1	6
39.331	-3.2379	23.9466	7	1	3	1	10
36.392	-3.2302	31.9937	36	3	3	1	18
37.525	-3.2159	27.8692	16	3	2	1	16
38.013	-3.2076	28.4525	22	2	4	1	16
39.585	-3.198	23.8948	6	3	1	1	14
40.307	-3.194	22.6028	1	1	4	1	12
37.428	-3.1677	28.4419	21	2	6	1	20
38.703	-3.1614	28.1004	18	2	5	1	18
40.467	-3.144	23.1909	4	1	6	1	16
36.471	-3.1427	29.0695	25	4	3	1	22
38.255	-3.1407	28.3079	20	3	4	1	20
34.88	-3.1392	30.8786	29	5	3	1	26
35.955	-3.1314	30.8926	31	4	4	1	24
37.574	-3.13	26.9135	13	4	2	1	20
42.143	-3.1274	22.6248	2	1	5	1	14

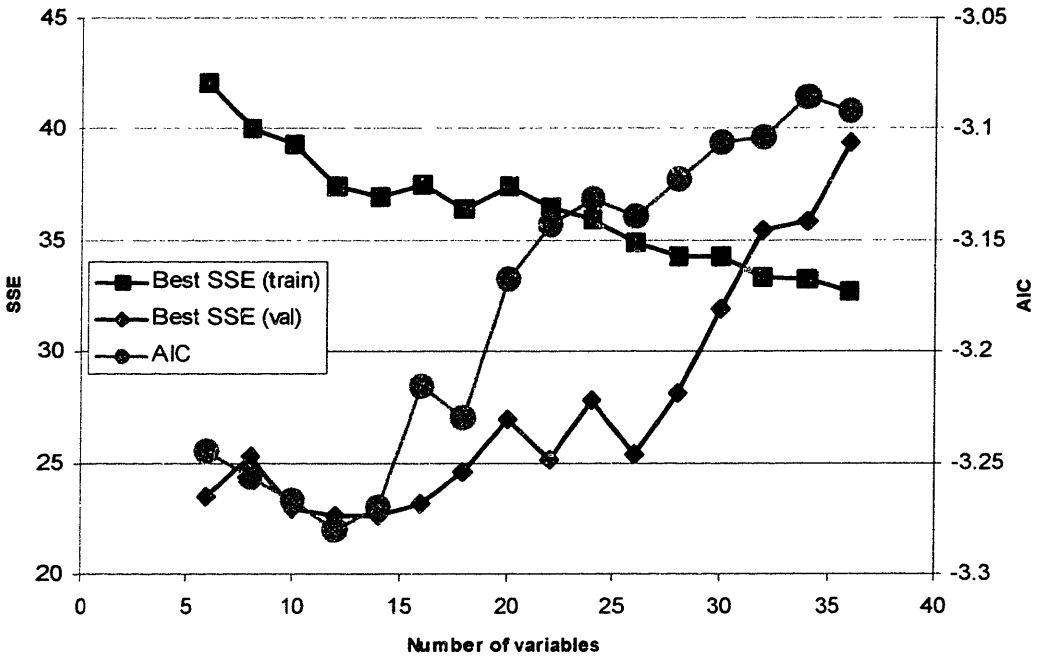

Figure 6-18: Prediction errors for both data sets *vs.* AIC criterion

Interestingly, the model with the best predictive capacity was an ARX141 model, which has a different form than the ARX221 models suggested by AIC, but still represents the same overall model complexity (*e.g.* both of these models have 12 parameters). As expected, the models with the largest number of parameters (ARX66x and ARX65x) do the worst job of fitting the validation data due to overfit (see Figure 6-18). Also important is that Table 6-4 shows that all of most accurate models include a time-lag of only one, by either AIC or SSE of the validation data. This fits the time-lagged correlation data for Groups 2/8 (see Figure 6-7).

Figure 6-19 shows the prediction of the validation data set for Group 2 with the ARX221 model constructed for both Groups 2 and 8. Figure 6-20 compares both groups' predicted values to the actual validation data. Note that the distribution in Figure 6-20 is somewhat biomodal despite the high correlation value. This is because the second experiment essentially measures transitions between 2 states (dark and moderate light). Similar results are observed for all other groups (data not shown) and show that models of this order are strong predictors of these groups

as well. The weakest fit to the validation data was seen for models built for Groups 7/11, which also have the lowest correlation coefficients with the input signal for the new data (see Table 6-3).

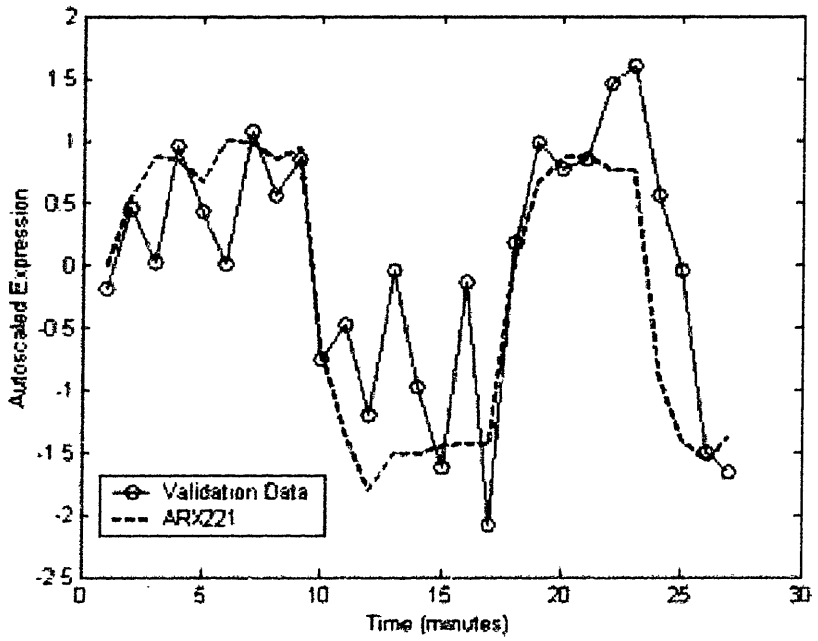

Figure 6-19: Predicted and validation data for Group 2

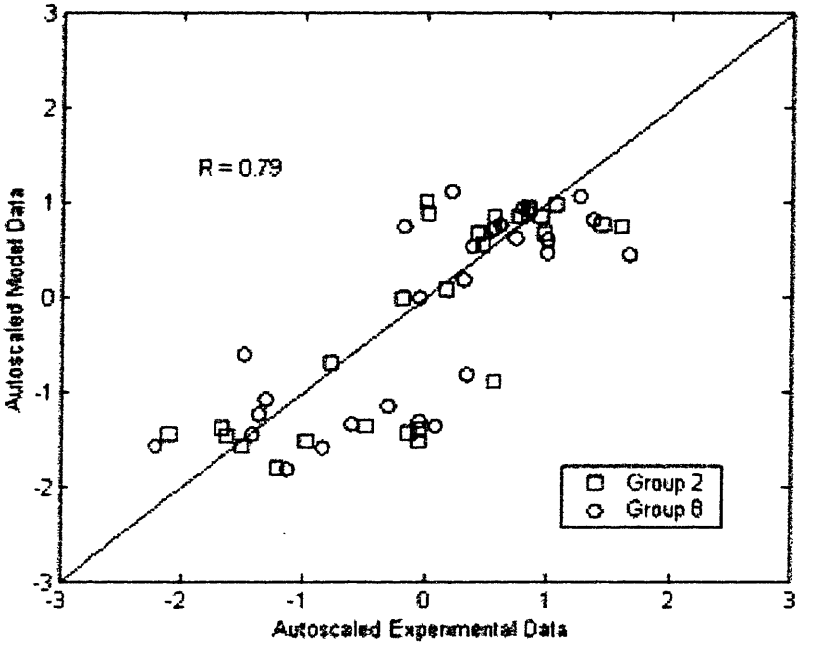

Figure 6-20: Comparison of predicted and validation data for Groups 2 and 8

6.4 References

1. Gill, R. T., Katsoulakis, E., Schmitt, W., Taroncher-Oldenburg, G., Misra, J. & Stephanopoulos, G. "Genome-wide dynamic transcriptional profiling of the light-to- dark transition in Synechocystis sp strain PCC 6803." *Journal of Bacteriology* **184**, 3671-3681 (2002).

2. Hihara, Y., Kamei, A., Kanehisa, M., Kaplan, A. & Ikeuchi, M. "DNA microarray analysis of cyanobacterial gene expression during acclimation to high light." *Plant Cell* **13**, 793-806 (2001).

3. Arkin, A. & Ross, J. "Statistical Construction Of Chemical-Reaction Mechanisms From Measured Time-Series." *Journal Of Physical Chemistry* **99**, 970-979 (1995).

4. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." *Molecular Biology Of the Cell* **9**, 3273-3297 (1998).

5. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herskowitz, I. "The transcriptional program of sporulation in budding yeast." *Science* **282**, 699-705 (1998).

6. DeRisi, J. L., Iyer, V. R. & Brown, P. O. "Exploring the metabolic and genetic control of gene expression on a genomic scale." *Science* **278**, 680-686 (1997).

7. Kamimura, R. T. in *Department of Chemical Engineering* (Massachusetts Institute of Technology, 1997).

8. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. "Cluster analysis and display of genome-wide expression patterns." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 14863-14868 (1998).

9. Dillon, W. R. & Goldstein, M. *Multivariate Analysis* (Wiley, New York, 1984).

10. Heyer, L. J., Kruglyak, S. & Yooseph, S. "Exploring expression data: Identification and analysis of coexpressed genes." *Genome Research* **9**, 1106-1115 (1999).

11. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **96**, 2907-2912 (1999).

12. Zhu, J., M.Q. Zhang. "Cluster, Function and Promoter: Analysis of Yeast Expression Array." *Pacific Symposium on Biocomputing* Hawaii, (2000).

13. Misra, J., Schmitt, W., Hwang, D., Hsiao, L. L., Gullans, S. & Stephanopoulos, G. "Interactive exploration of microarray gene expression patterns in a reduced dimensional space." *Genome Research* **12**, 1112-1120 (2002).

14. Raychaudhuri, S., Stuart, J. M. & Altman, R. B. "Principal components analysis to summarize microarray experiments:application to sporulation time series." *Pacific Symposium on Biocomputing* Hawaii, (2000).

15. CyanoBase: The Genome Database for Cyanobacteria. http://www.kazusa.or.jp/cyano/

16. Wilde, A., Lunser, K., Ossenbuhl, F., Nickelsen, J. & Borner, T. "Characterization of the cyanobacterial ycf37: mutation decreases the photosystem I content." *Biochemical Journal* **357**, 211-216 (2001).

17. Meurer, J., Plucken, H., Kowallik, K. V. & Westhoff, P. "A nuclear-encoded protein of prokaryotic origin is essential for the stability of photosystem II in Arabidopsis thaliana." *Embo Journal* **17**, 5286-5297 (1998).

18. Bryant, D. A., Delorimier, R., Guglielmi, G. & Stevens, S. E. "Structural and Compositional Analyses of the Phycobilisomes of Synechococcus Sp Pcc 7002 -

Analyses of the Wild-Type Strain and a Phycocyanin-Less Mutant Constructed by Interposon Mutagenesis." *Archives of Microbiology* **153**, 550-560 (1990).

19. Watson, G. M. F. & Tabita, F. R. "Regulation, unique gene organization, and unusual primary structure of carbon fixation genes from a marine phycoerythrin- containing cyanobacterium." *Plant Molecular Biology* **32**, 1103-1115 (1996).

20. Schmitt, W. A. & Stephanopoulos, G. "Prediction of transcriptional profiles of *Synechosystis* PCC6803 by dynamic autoregressive modeling of DNA microarray data." *Biotechnol Bioeng* **submitted** (2003).

21. Akaike, H. "New Look at Statistical-Model Identification." *IEEE Transactions on Automatic Control* **AC19**, 716-723 (1974).

22. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, 2001).

23. Alter, O., Brown, P. O. & Botstein, D. "Singular value decomposition for genome-wide expression data processing and modeling." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **97**, 10101-10106 (2000).

24. Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. "Fundamental patterns underlying gene expression profiles: Simplicity from complexity." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **97**, 8409-8414 (2000).

25. Arkin, A., Ross, J. & McAdams, H. H. "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." *Genetics* **149**, 1633-1648 (1998).

CHAPTER 7 OTHER TOOLS FOR ANALYSIS OF MICROARRAY DATA

Besides time-series experiments aimed at understanding of transcriptional regulation phenomenon, there are many applications of high-throughput DNA microarray data for other purposes. One particularly important example is that of statistically robust distinction between cells in two or more different states, in order to either diagnose new samples or search for patterns indicative of the underlying biology differentiating the states. Extensive efforts made during this work to tackle these problems are discussed in this chapter.

Oligo-nucleotide and cDNA arrays[1,2] are being increasingly employed for determining discriminatory genes and discovering new classes of disease subtypes that are differentiated at the level of transcription[3,4]. Data-driven hypotheses are developed from these types of measurements that suggest, in turn, novel experiments furthering biomedical research.

With the rapid increase of cDNA and oligonucleotide microarray data that has become available in recent years, the need to develop techniques to answer specific questions within the framework of massive amounts of information has become acute. The typical approach applied to these data sets is various types of simple clustering[5-7]. While clustering provides a framework for ordering expression data into groups, these methods do not directly address questions about the nature of sample differences or how to best categorize new samples. When samples from a set of distinct populations are considered, methods that specifically target distinguishing characteristics are preferable.

Differences between samples from different populations, manifested by variation in the expression levels of individual genes, can be captured by analysis focusing on those genes that exhibit significant changes in certain samples. This information can be used in class discovery to hypothesize the existence of distinct phenotypic sub-groups, or can be used for class distinction to search for biological insight into the genetic rules underlying known phenotypes. This effort seeks to close the gap between gene expression and physiological state. Of particular interest has been the distinction of cancers and cancer sub-types[3,6,8].

A related problem involves the use of microarrays to diagnose samples, such as disease states, metabolic state, or tissue types. Monitoring the genes identified as important for distinguishing classes can provide a means for differentiating between diseases or cellular states with similar physiological profiles but different underlying causes. Effort has been put into the problem of accurately classifying new samples based on the discriminating genes identified above[3].

This chapter seeks to establish straightforward approaches to both the identification of discriminating genes and the construction of classifiers based on these genes for identification purposes. We begin statistical techniques for finding variables with significant information content about the classification problem at hand. The focus is on techniques that will be robust even when there is significant overlap between expression levels, including mean hypothesis testing, Wilks' lambda criterion, and leave-one-out cross validation. Next, likelihood ratio testing and Fisher discriminant analysis (FDA) are presented as methods of applying gene expression information to the classification of new samples. Finally, the combination of these techniques to analyze the statistical robustness of conclusions drawn from DNA microarray data is discussed. Using power analysis, we suggest how to estimate the number of samples needed to reach statistically reliable conclusions with DNA microarrays.

7.1 Data sets

Publicly available data sets from outside of this lab have been considered for this study. The first data set comes from a study of human acute leukemias[3]. This data is well-suited to study the distinction between *acute myeloid leukemia* (AML) and *acute lymphoblastic leukemia* (ALL) and contains 72 total samples (25 AML and 47 ALL). Samples have been hybridized on Affymetrix chips with 7129 total features, including 6817 human genes. A further division of ALL samples into 38 B-lineage acute lymphoid leukemia (B-ALL) and 9 T-lineage acute lymphoid leukemia (T-ALL) was considered in extending the sample determination approach to the multi-class case of three disease subtypes (B-ALL, T-ALL, and AML). The sample classification among the three subtypes given in Golub *et al.*[3] was also used here, as it was based on both clinical information and validation through their pattern discovery technique.

The other data set contains 24 tissue samples from the HUGE Index generated on the same type of Affymetrix chips at Steve Gullans' lab at Brigham and Women's Hospital[9]. The data

consisted of 24 tissues, composed of 5 brain samples, 4 kidney samples (Kid), 3 vulva samples (VU), 2 samples each of proliferative endometrium (PE) and myometrium (Myo), and one sample each of lung, liver, skeletal muscle (Sk Musc), ovary, cervix (CER), placenta (Plac), spleen, and blood. The brain samples consisted of one each of the amygdala (BR-AG), the Hippocampus (BR-HC), the caudate putamen (BR-CA), the motor cortex (BR-MOT), and the level 4 sub-section (BR-L4). This data set therefore provides a challenge very different from the first set because there are many distinct sub-classes with few samples rather than two or three with many samples.

7.2 Discovery of interesting variables in known classes

When classes are known in advance it is necessary to consider the entire distributions of data for known classes rather than their means or maximum/minimum values. A reasonable approach to this problem has been undertaken by comparing the means of each class with the standard deviations simultaneously in the definition of a simple correlation metric[3]. This approach works well when there are large numbers of samples available for testing, but for small or intermediate numbers of samples, more robust measures are desirable.

7.2.1 *P* tests

Golub *et al*[3] proposed a measure P to distinguish between any two disease classes in their data set. For any gene g, the measure of its ability to distinguish classes 1 and 2 is calculated as:

$$P(g)=\frac{(\mu_1(g)-\mu_2(g))}{(\sigma_1(g)+\sigma_2(g))} \qquad 7\text{-}1$$

where μ_x is the mean and σ_x is the standard deviation of class x[9]. If class 1 is taken to be ALL and class 2 to be AML, then genes with the high positive P values will be characteristically high in ALL samples, while those with large negative values of P will be characteristically high in AML samples. In this way, a set of genes specific to both ALL and AML samples will be uncovered. Examination of the biological characteristics of these "top genes" should then lend insight into the different workings of the two cancers.

While this measure makes intuitive sense, as it considers not only the distance between the group means but also the distributions of the data in the two groups, this measure lacks a solid statistical basis for evaluating the importance of a given *P* value. This is because it is unclear how we might expect the distribution of *P* values to look for the case where there is no difference between the groups, and therefore we lack a control scenario as a frame of reference for further analysis. Furthermore, this measure fails to correct for greater uncertainty that may result from a small sample size (although some measure of this is built into the calculation of σ_i). This being said, the values of *P* calculated do not differ significantly in relative importance from the other calculations proposed here.

7.2.2 *t*-tests

In order to evaluate whether two populations are statistically different in the mean, a *2-tailed t-test* can be employed to compare the population means. If the two sample sets are drawn from populations *X* and *Y* with means μ_X and μ_Y and the same variance σ^2, then in the limit of infinite samples the difference between the measured means $\overline{X}$ and $\overline{Y}$ of the two sets will be normally distributed

$$\overline{X}-\overline{Y} \sim N\left[\mu_X-\mu_Y, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right] \qquad 7\text{-}2$$

where *n* and *m* are the number of samples measured from the two distributions. Applying a Z-transformation to this normal distribution reduces it to standard normal

$$Z=\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad 7\text{-}3$$

This Z statistic can be used to determine a confidence interval for the assumption that μ_X-$\mu_Y = 0$. For a given significance level α chosen by the user,

$$(\overline{X}-\overline{Y})-z(\alpha/2)\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}<(\mu_X-\mu_Y)<(\overline{X}-\overline{Y})+z(\alpha/2)\sigma\sqrt{\frac{1}{n}+\frac{1}{m}} \qquad 7\text{-}4$$

gives the confidence interval for μ_X-μ_Y. If this interval does not include zero, then we must reject the null hypothesis that $\mu_X = \mu_Y$, and we can conclude that the gene in question has discriminating power at the confidence level chosen.

This formulation only holds if the sample sizes are large and normally distributed (n, $m > 30$). However, large numbers of samples are not available for many applications. For real data with a small number of samples, the normal distribution should be approximated as *Student's t distribution*, which is a function of the number of samples taken (as $n \rightarrow \infty$, the t distribution becomes the z distribution). Furthermore, to estimate the population variance σ^2 the individual measured variances s_X and s_Y should be combined into the *pooled sample variance*[10]

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2} \qquad 7\text{-}5$$

which is an average of the individual variances weighted by their degrees of freedom. This approximation is useful because truly indistinguishable distributions will indeed have the same variation and the difference in the group means will follow the t-distribution (alternatively, the *mean hypothesis test* formulation of the *t-test* may be used[11,12] which allows for approximate comparison of two populations with different variances. However, we have found little practical difference between the two and thus discuss only the simpler, more common version in this discussion). The confidence interval is then re-written as:

$$\left(\overline{X}-\overline{Y}\right) - t_{m+n-2}(\alpha/2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}} < (\mu_X - \mu_Y) < \left(\overline{X}-\overline{Y}\right) + t_{m+n-2}(\alpha/2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}} \qquad 7\text{-}6$$

where values of the t statistic for m+n-2 degrees of freedom can be readily obtained from tables in basic statistic texts[13].

As an example of the discovery of discriminating features, the leukemia data presented by Golub et al.[3] was examined to first find genes which discriminate ALL samples from AML samples. Subsequently, the ALL class was sub-divided in to B-ALL and T-ALL samples and discriminating genes were again identified (alternatively, each class can be considered individually. If sub-classes exist in the data, then the user may choose to analyze and remove the most unique classes in an iterative fashion, or consider each class relative to the whole). At the

99.99% confidence level using the two-tailed t-test, 96 genes were distinguished as discriminatory between ALL and AML (data not shown), while 49 genes were identified at that confidence level as distinguishing T-ALL from B-ALL (see Table 7-1).

Table 7-1: Discriminatory genes for the distinction of T-ALL from B-ALL samples

Gene	Gene Discription
X03934_at	GB DEF = T-cell antigen receptor gene T3-delta
U23852_s_at	GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA
M23323_s_at	T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN PRECURSOR
M37271_s_at	T-CELL ANTIGEN CD7 PRECURSOR
X69398_at	CD47 CD47 antigen (Rh-related antigen, integrin-associated signal transducer)
M12886_at	TCRB T-cell receptor, beta cluster
X00274_at	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR ALPHA CHAIN PRECURSOR
X76223_s_at	GB DEF = MAL gene exon 4
X59871_at	TCF7 Transcription factor 7 (T-cell specific)
U14603_at	Protein tyrosine phosphatase PTPCAAX2 (hPTPCAAX2) mRNA
X60992_at	T-CELL DIFFERENTIATION ANTIGEN CD6 PRECURSOR
M37271_s_at	T-CELL ANTIGEN CD7 PRECURSOR
M26692_s_at	GB DEF = Lymphocyte-specific protein tyrosine kinase (LCK) gene, exon 1, and downstream promoter region
D63878_at	PROBABLE PROTEIN DISULFIDE ISOMERASE ER-60 PRECURSOR
M37815_cds1_at	CD28 gene (glycoprotein CD28) extracted from Human T-cell membrane glycoprotein CD28 mRNA
L05148_at	Protein tyrosine kinase related mRNA sequence
D30758_at	KIAA0050 gene
X04391_at	CD5 CD5 antigen (p56-62)
M12886_at	TCRB T-cell receptor, beta cluster
U93049_at	GB DEF = SLP-76 associated protein mRNA
J03077_s_at	PSAP Sulfated glycoprotein 1
M32886_at	SRI Sorcin
U18009_at	Chromosome 17q21 mRNA clone LF113
X99584_at	SMT3A protein
J04132_at	CD3Z CD3Z antigen, zeta polypeptide (TiT3 complex)
HG4128-HT4398_at	Anion Exchanger 3, Cardiac Isoform
U05259_rna1_at	MB-1 gene
U59878_at	Low-Mr GTP-binding protein (RAB32) mRNA, partial cds
X69433_at	IDH2 Isocitrate dehydrogenase 2 (NADP+), mitochondrial
U90426_at	Nuclear RNA helicase
S78187_at	M-PHASE INDUCER PHOSPHATASE 2
L10373_at	MXS1 Membrane component, X chromosome, surface marker 1
X95677_at	GB DEF = ArgBPIB protein
U67171_at	GB DEF = Selenoprotein W (selW) mRNA
X62535_at	DAGK1 Diacylglycerol kinase, alpha (80kD)
X73358_s_at	HAES-1 mRNA
D83920_at	FCN1 Ficolin (collagen/fibrinogen domain-containing) 1
D11327_s_at	PTPN7 Protein tyrosine phosphatase, non-receptor type 7
X04145_at	CD3G CD3G antigen, gamma polypeptide (TiT3 complex)
U18422_at	DP2 (Humdp2) mRNA
M28826_at	CD1B CD1b antigen (thymocyte antigen)
L76200_at	Guanylate kinase (GUK1) mRNA
D87292_at	Rhodanese
Z50853_at	CLPP
U50743_at	Na,K-ATPase gamma subunit mRNA
U01691_s_at	Annexin V (ANX5) gene, 5'-untranslated region
U50327_s_at	Protein kinase C substrate 80K-H gene (PRKCSH)
X58529_at	IGHM Immunoglobulin mu

In this case there are only 11 AML samples, 8 T-ALL samples, and 19 B-ALL samples under consideration. Because of the relatively small number of samples ($n < 30$ within each class) in

this case it is important to consider the uncertainty in measurement due to sample size. The use of the t-test to compare the two sample means is thus perfectly suited to this consideration.

Another example is that of comparison of tissue samples. The 24 tissue samples from the HUGE Index generated at Steve Gullans' lab at Brigham and Women's Hospital[9] were separated into two categories: brain and non-brain. Even though the five brain samples are from a variety of regions within the brain, we expect some sub-set of the genes to have "brain-specific" characteristics, with a corresponding distribution of relative inactively in all other samples. Using a confidence interval of 99.99% we found 386 genes were expressed in a discriminatory manner between the two classes (data not shown). Use of even more stringent criteria, or the addition of new samples from both the brain and from other regions of the body to help resolve the shape of the distributions, can help to narrow the list of genes with interesting functions that seem specific to the brain.

7.2.3 Wilks' lambda

Parametric tests such as the P-value test and the *t*-test are based on differences of group means and variance calculations, and these tests may perform poorly if their underlying assumption are violated. Such assumptions include as normality (for high N) and equal variance in the various groups (even in the form of pooled sample variance s_p^2, which is a combination of two variances). A non-parametric test (such a Mann-Whitney test) does not rely on these assumptions and works well with a small sample size, but the results may be more critically sensitive on the nature of the samples used for the training of the classifier than those in parametric tests. No method is unanimously optimal for all kinds of data. Selection of a method for application to a certain data set should depend on the characteristics of the data, the degree of violation of the underlying assumptions, and the sample size. Our experience leads us to a well-characterized alternative measure, called Wilks' lambda score[10,14] to assess discriminatory powers of the individual genes. Wilks' lambda, which originated from ANOVA, is not limited only to two-class comparisons but can also be used for multi-class distributions. It produces more robust test results than multiple two-class comparisons using t-test because the Wilks' lambda is based on total group variance calculations instead of differences between individual group means.

Genes whose expression distribution has high between-group variance (the groups are well-separated) and small within-group variance (the samples inside each group are relatively similar) are deemed to be discriminatory for the sample classes[10,15]. The between-group variance (B_i) of the expression of a certain gene i is proportional to the sum of the differences between group means of expression levels. The within-group variance of the expression of gene i (W_i) is the sum of group variances of the expression levels of the gene in a single class. Given the total variance of expression levels of gene i, $T_i = (\mathbf{x}_i - \mathbf{1}\bar{\mathbf{x}}_i)^T(\mathbf{x}_i - \mathbf{1}\bar{\mathbf{x}}_i)$, the within- and the between-group variances are shown in Figure 7-1 and defined respectively as follows

$$W_i = \sum_{j=1}^{c} W_i^j = \sum_{j=1}^{c} (\mathbf{x}_i^j - \mathbf{1}\bar{x}_i^j)^T(\mathbf{x}_i^j - \mathbf{1}\bar{x}_i^j) \qquad 7\text{-}7$$

$$B_i = T_i - W_i \qquad 7\text{-}8$$

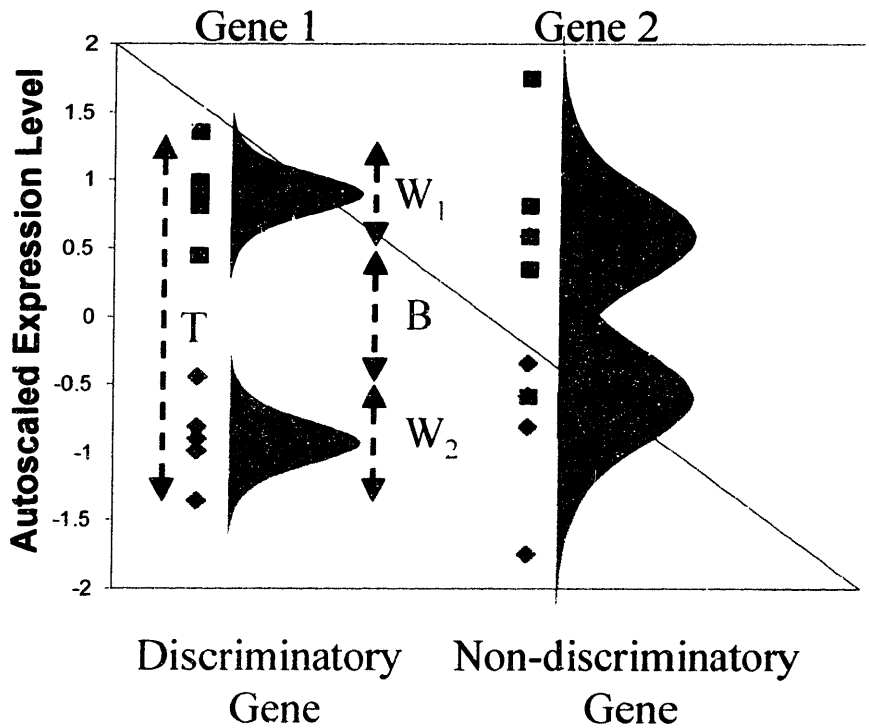

Figure 7-1: Theoretical discriminatory and non-discriminatory gene distributions

The vector, $\mathbf{x}_i$ (Nx1), contains the expression level of gene i in N samples and $\bar{x}_i$ is the mean expression of gene i in all N samples. The superscript j represents class j among the c classes. For the two genes shown schematically in Figure 7-1, Gene 1 has a large between-group variance and a small within-group variance while Gene 2 has a small between-group variance (overlapping distributions across the classes) and a large within-group variance. For Gene 1, the large ratio of the between-group variance to within-group variance indicates a gene with a discriminatory expression pattern. Without loss of information, the above procedure is implemented through a statistical test based on Wilks' lambda (Λ_i) that allows one to establish a formal boundary between discriminatory genes and non-discriminatory genes:

$$\Lambda_i = \frac{W_i}{T_i} \qquad 7\text{-}9$$

In order to compare the Wilks' lambda (Λ_i) score to a distribution with known parameters, it is transformed to the F distribution as follows[10.16]:

$$F_i = \frac{(1-\Lambda_i)}{\Lambda_i}\frac{(N-c)}{(c-1)} \sim F_{\alpha(c-1,N-c)} \qquad 7\text{-}10$$

where N is the total number of samples and c is the number of classes. In this form, discriminatory genes are selected by applying a statistical cutoff determined from the F distribution using some level of significance (in this case α=0.01). Note that a high F value signifies a more discriminatory gene relative to one with a low F value.

7.2.4 Cross-validation

As all of the methods presented are parametric, they might produce a high false positive error due to violation of the underlying assumptions, particularly normality. This is especially true for genes that have only a small difference in their expressions between groups. In order to improve the false-positive rate, we can incorporate an error rate calculation through a leave-one-out cross-validation (LOOCV)[10.17] procedure into the discriminatory gene analysis. In this procedure, a series of many LOOCVs are performed to get a good error estimate. The first step in this iterative procedure consists of randomly dividing the data set being considered into c test

samples (*i.e.* one test sample for each class) and *N-c* training samples. The training samples are used to generate an initial set of discriminatory genes using any of the gene selection methods presented (we will use Wilks' lambda score/F statistic values for examples given here). Using the gene with highest discriminatory score, a classifier (see section 7.3) is constructed and the error rate calculated for the *c* test samples. A second classifier is then constructed using the top two discriminating genes, which is again applied to the test samples. The number of genes included in the classifier is thus sequentially increased to form more complex classifiers until all genes selected by one of the parametric methods have been included. At each step, the number of misclassified samples is determined for calculation of the misclassification error rate (see next paragraph). A new division of the samples into training and test sets is then considered, and the procedure is repeated.

Figure 7-2 provides a schematic of the leave-one-out cross-validation algorithm for error rate estimation. Each LOOCV first splits the data set into randomly selected sets of *N-c* training and *c* test sets. Then, discriminatory genes are selected on the basis of their Wilks' lambda score. The number of genes included in the classifier is increased by one in order of decreasing magnitude of their *F* value from Wilks' lambda. The number of misclassified *c* test samples is counted as a function of the number of genes. The above procedure is repeated *g* times for different randomly selected training and test sets. The average error rates calculated by $e(p)=m_p/(c\times g)$ are then plotted vs. the number of genes included in the classifier and the discriminatory genes are selected based on the number of times they are identified as "discriminatory" in all the iterations (see section 7.4.3 for examples).

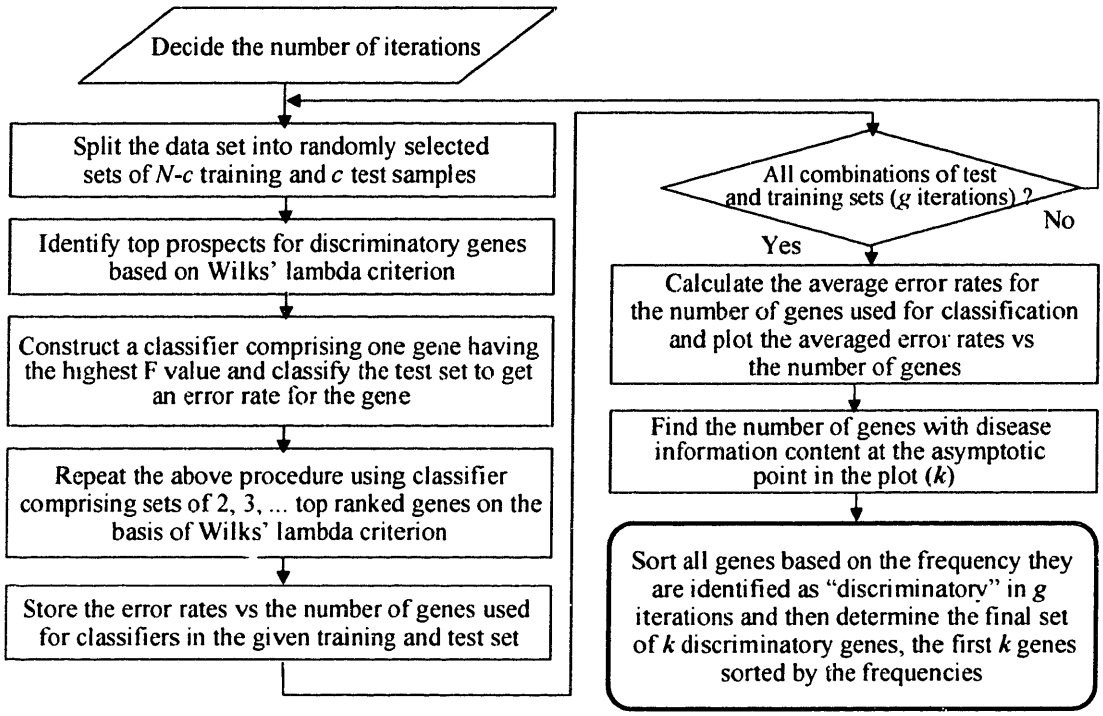

Figure 7-2: Cross-validation scheme for selection of genes

For the estimation of error rates, the entire LOOCV procedure is repeated g times using different test and training sets, until all samples have been withheld in the test set at least once. If we denote by m_p the number of misclassified samples in the g cross-validations for a given number of discriminatory genes (p) used in the classifiers, the averaged error rate is given by $e(p) = m_p/(c \times g)$. Then, the error rates from the g cross-validation iterations can be computed as function of the number of discriminatory genes. Using the error rate curve, the number of discriminatory genes can be determined at the point where the averaged error rates show an asymptotic behavior (see Figures in section 7.4). Then, a final set of discriminatory genes is determined based on the frequencies by which they appeared as discriminatory genes during the g LOOCVs. The final list of genes is generally shorter than the original list of discriminatory genes selected by parametric tests, thus enabling us to reduce the false positive error by identifying a small set of genes robust to sample variation. If a gene with small expression difference between the two classes of samples shows up consistently in the LOOCV procedure, it indicates that the observed difference, even though small, is statistically reliable.

Other methods have also introduced to reduce false positives in identifying discriminatory genes[18,19] and may be used instead of LOOCV for any given application, at the user's discretion. Indeed, such methods may be required for some data sets, if the presence of one or more particularly poor samples (that is, obvious outliers) causes LOOCV to give inconsistent conclusions. For the data sets discussed here however, LOOCV has proven to give consistently robust results.

7.3 Classification of samples

The use of microarray data to diagnose the state of a cell population assumes that the transcriptional activity of all genes may provide more insight into the distinction between disease states than could be uncovered through other diagnostic tools alone. Because some genes may be more informative for classification than others for a certain sample, earlier authors have adopted a voting scheme for diagnosing new samples[3]. In this method, each gene's expression level suggests one of the two classes, and that gene's *vote* is weighted by how much "closer" the expression level is to one class than the other. If possible, it is desirable to have a set of classifiers which are optimum relative to some statistical criteria. These classifiers should consider many variables simultaneously and assign automatically the importance of each gene's expression level depending on the sample.

7.3.1 Likelihood ratio tests as classifiers

One such "optimum" test exists in the form of the *likelihood ratio test*[13]. According to the Neyman-Pearson Lemma, no comparison of two hypotheses "null (0)" and "alternative (A)" (*e.g.*, a sample is cancerous vs. that it is not) has a higher significance value (lower probability of misclassification) than the ratio

$$\frac{f_0(\mathbf{x})}{f_A(\mathbf{x})} = \frac{p(\mathbf{x}\,|\,0)}{p(\mathbf{x}\,|\,A)} \qquad 7\text{-}11$$

where $p(x|0)$ is the probability of observing x if the null hypothesis is true, and $p(x|A)$ is the probability of observing x if the alternative is true. If this ratio is large, then the null hypothesis is accepted, and if small it must be rejected.

If there are many variables to be considered independently, then the probability can be re-written as

$$\frac{f_0(x)}{f_A(x)} = \frac{p(x_1 \mid 0)p(x_2 \mid 0)\ldots p(x_N \mid 0)}{p(x_1 \mid A)p(x_2 \mid A)\ldots p(x_N \mid A)} \qquad 7\text{-}12$$

A difficulty in the use of this formulation for microarray data is the assumption of independence between the genes. Presumably, there is significant interaction between genes; however, we adapt the same convention as other authors[3] by assuming that the expression level of each gene can be considered independently of the other genes.

For ease of interpretation, we take the log of this ratio

$$\log\frac{f_0(\mathbf{x})}{f_A(\mathbf{x})} = \log\frac{p(x_1 \mid 0)}{p(x_1 \mid A)} + \log\frac{p(x_2 \mid 0)}{p(x_2 \mid A)} + \ldots\log\frac{p(x_N \mid 0)}{p(x_N \mid A)} \qquad 7\text{-}13$$

Thus large positive numbers indicate hypothesis 0, while large negative numbers favor hypothesis A.

7.3.1.1 T-distributions/assumptions

In order to calculate the probability that the sample came from either of the classes being considered, we can compare the sample expression level to the known data using the single-tailed *t-test*. As with the two-tailed t-test, the statistic is set up to deal with distributions that approximate normal as the number of samples becomes large. Instead of comparing two distributions, however, in this case we compare a value to a distribution it "could" have been sampled from

$$Sample\,Value = \overline{X} \pm t(\alpha/2)s\sqrt{\frac{1}{n}} \qquad 7\text{-}14$$

By finding the significance value "α" which solves this equation we have a measure of the confidence that the sample value could be observed from a process with the same distribution as the data X.

Consider, for example, the idealized distributions shown in Figure 7-3 compared to the distribution for an actual gene X95735 (Zyxin) from the Leukemia case study (Figure 7-4). Because the distribution of AML samples is much larger than that of the ALL samples, the ratio of t-test significance levels will reflect the wide distribution of AML expression levels for this gene relative to the narrow distribution of ALL expression.

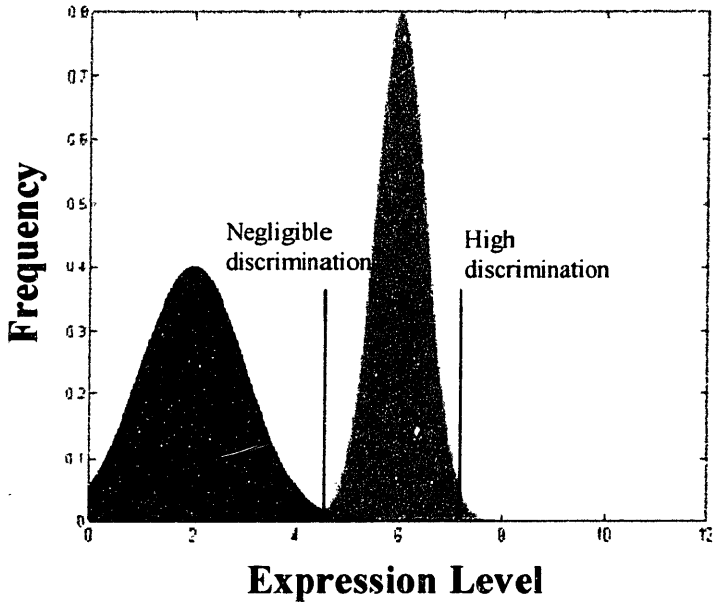

Figure 7-3: Idealized, normal gene distribution, 2 classes

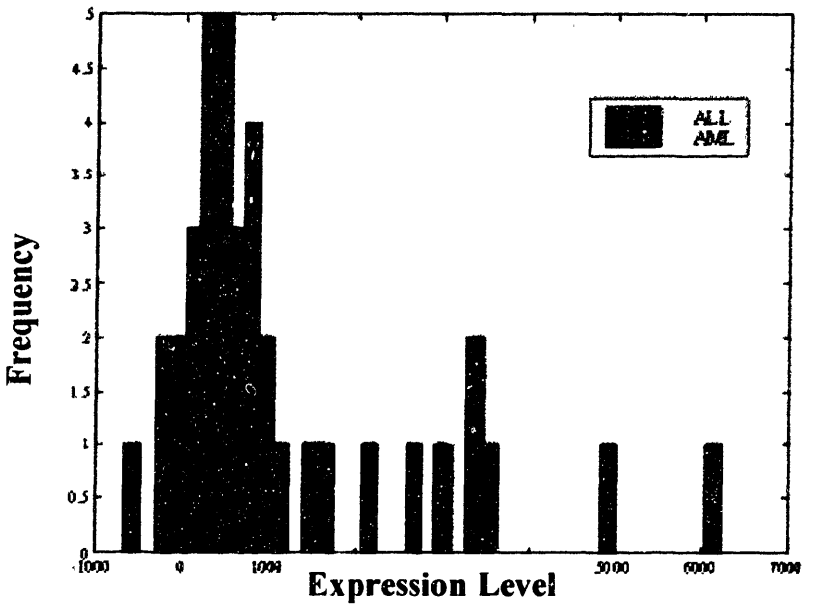

Figure 7-4: Expression distribution for gene X95735 (Zyxin)

As a practical matter, it is useful to fix an upper and lower limit on the ratio of scores, because computer rounding of values close zero can give log ratios of $\pm\infty$. By selecting a cutoff (for example, probabilities greater than 99% or 1% are reduced to these "maximum" values) this problem can be avoided.

7.3.1.2 Leukemia example

To demonstrate the use of log mean ratio scoring, the 96 genes identified through t-tests to discriminate AML from ALL leukemia samples were used with the 38 training samples to give 96 pairs of distributions. These were then used to calculate log ratios for those 96 genes in each of the 34 training samples from Golub *et al.*[3] The results are shown in Figure 7-5. Each line represents the cumulative score for an AML (dark) or ALL (light) sample, with the final score for each sample represented by a circle at the right side of the diagram. The gray box shows the region within which some of the samples have ambiguous classification – the region between the lowest-scoring AML sample and the highest-scoring ALL sample. In this figure the impact of each gene on the overall score is shown, ranked from most informative genes (scoring ratios with high absolute value) to the least informative (values close to zero). Because of this ranking, the

order in which the genes are plotted is different for each sample, depending on which genes contributed most to the overall score.

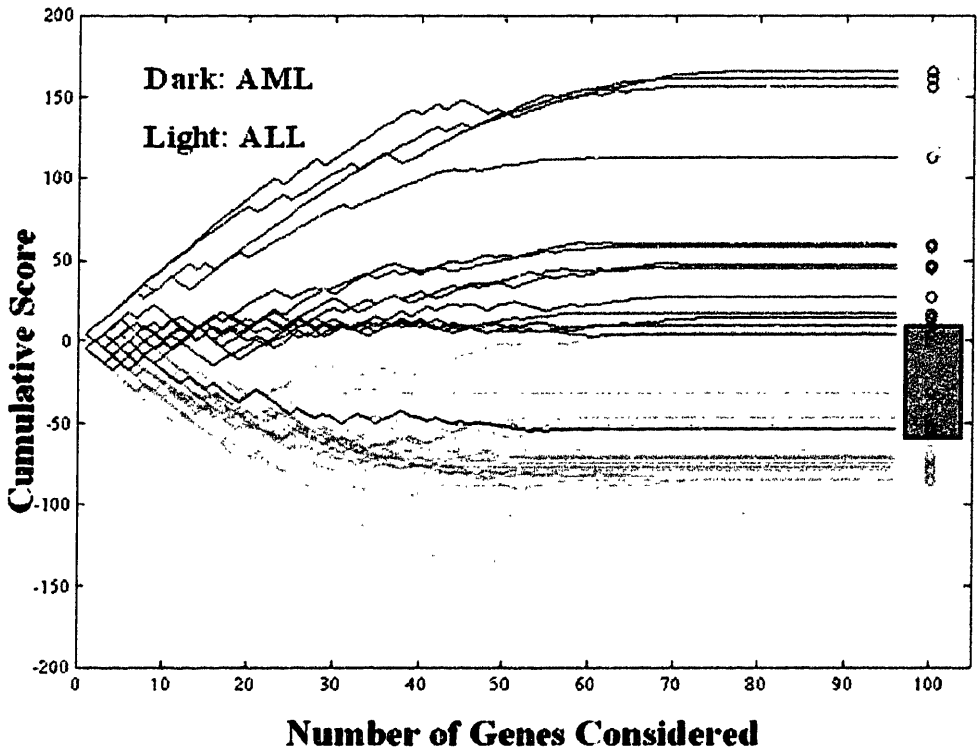

Figure 7-5: Discrimination of leukemia samples with log likelihood scoring

This example shows that the classification through the log ratio testing of probabilities gives not only effective classification, but also an ordering for confidence in the results. Those samples at the extreme of the chart are mostly clearly classified. As is shown in earlier studies[3], classification can be achieved successfully in all cases that deviate significantly from a "uncertainty region" as chosen by the user. If this region is chosen to be large, there will be large amounts of type 2 (missed detection) error while an exceedingly small region will cause an abundance of type 1 (incorrect detection) error. In this study no misclassification occurs for samples outside of a reasonably tight region: that is, strongly scoring samples are 100% correctly identified while weakly scoring samples may include only 2 misdiagnosed samples (*i.e.*, one AML classified as ALL and vice-versa) if only type 1 error is accepted and no uncertainty region is selected.

7.3.1.3 Brain example

As a final example, we used the data from the HUGE index to classify 4 additional samples collected from the Gullans laboratory (see section 7.1) as "brain" or "other". If enough samples of any given tissue type exist, it should be possible to test new samples against each of the known classes to see if that sample can be classified. For simplicity, the 50 most discriminatory genes selected by comparing brain samples against the remainder of the population (section 7.2.2) were used for the classification. See Figure 7-6: the classifier clearly separates the new brain sample from the other tissues.

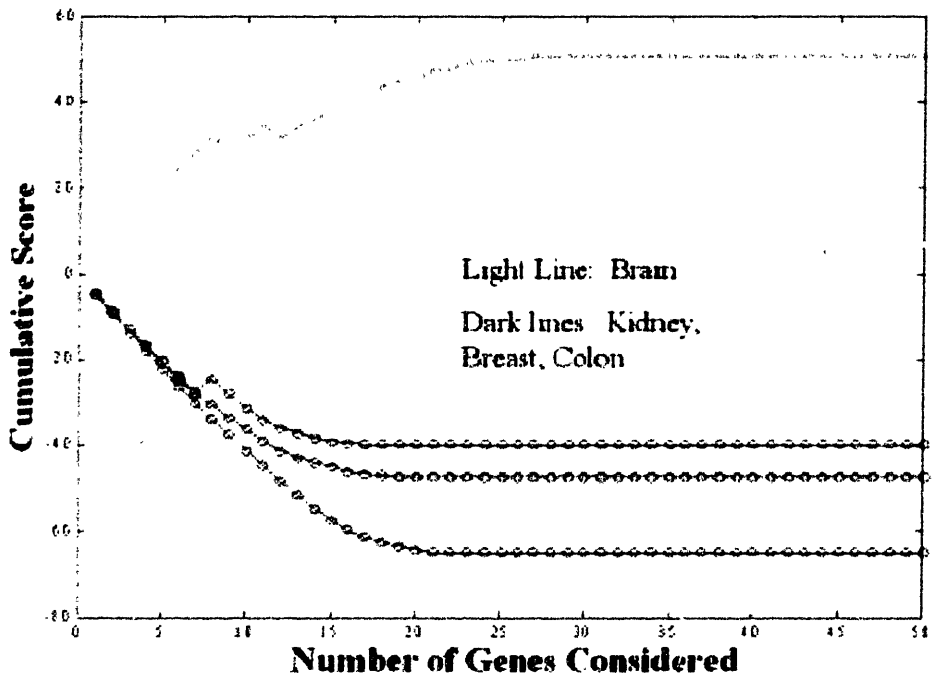

Figure 7-6: Discrimination of tissue samples with log likelihood scoring

Applications for such a classifier, besides diagnosis of disease states, may include identification of the homogeneity of samples from patients. For example, if good samples can be collected for sub-sections of a given tissue (for example, regions of the brain or kidneys) then new samples can be tested to observe if they contain more of one sub-set or another, or if they are a more even blend of multiple regions.

7.3.2 Multi-dimensional discriminant analysis

Fisher Discriminant Analysis (FDA) is a linear method of dimensionality reduction from the expression space comprised of all selected discriminatory genes to just a few dimensions where the separation of sample classes is maximized. FDA is similar to Principal Component Analysis (PCA)[20-22] in the linear reduction of data[10,14]. The major difference is that the discriminant axes of the FDA space are selected such as to maximize class separation in the reduced FDA space, instead of variability as in the case of PCA. The discriminant axes of FDA, termed as discriminant weights ($\boldsymbol{V}$), which maximize the separation of sample classes in their projection space can be shown to be equivalent to the eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$, the ratio of between-group variance ($\boldsymbol{B}$) to within-group variance ($\boldsymbol{W}$):

$$\mathbf{W}^{-1}\mathbf{BV} = \mathbf{V}\Lambda \qquad 7\text{-}15$$

where $\boldsymbol{B}=\boldsymbol{T}-\boldsymbol{W}$, $\boldsymbol{W}=\sum_{j=1}^{c}\left(\boldsymbol{X}_j-\boldsymbol{I}\bar{\boldsymbol{x}}_j^T\right)^T\left(\boldsymbol{X}_j-\boldsymbol{I}\bar{\boldsymbol{x}}_j^T\right)$, $\boldsymbol{T}=\left(\boldsymbol{X}-\boldsymbol{I}\bar{\boldsymbol{x}}^T\right)^T\left(\boldsymbol{X}-\boldsymbol{I}\bar{\boldsymbol{x}}^T\right)$, and c is the number of classes being considered. The eigenvalues (Λ) indicate the discrimination power for the corresponding discriminant axes. Further details of FDA and its application in classification of microarray data are described in Stephanopoulos *et al.*[1].

Figure 7-7 and Figure 7-8 show the projection through each vector of weights V_i into a new dimensional space y_i of the expression data in the 2-class (AML and AML) or the 3-class (B-ALL, T-ALL and AML), respectively. In general, c-1 dimensions may be used for classification using this method. Note that for the 2-class case in Figure 7-7, only one discriminating dimension is possible, so data is plotted with y_1 representing both axes. For 3 classes, as in Figure 7-8, up to 2 dimensions may be used (hence y_1 vs. y_2) although dimension y_1 alone has some classification power.

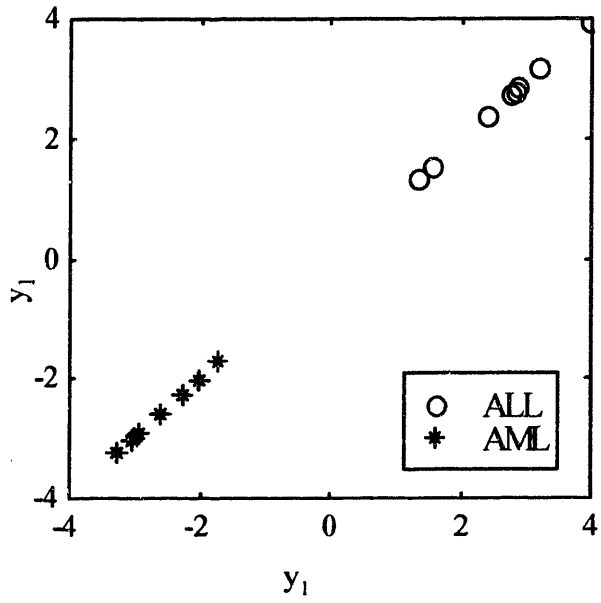

Figure 7-7: 2-class discrimination of leukemia samples using FDA

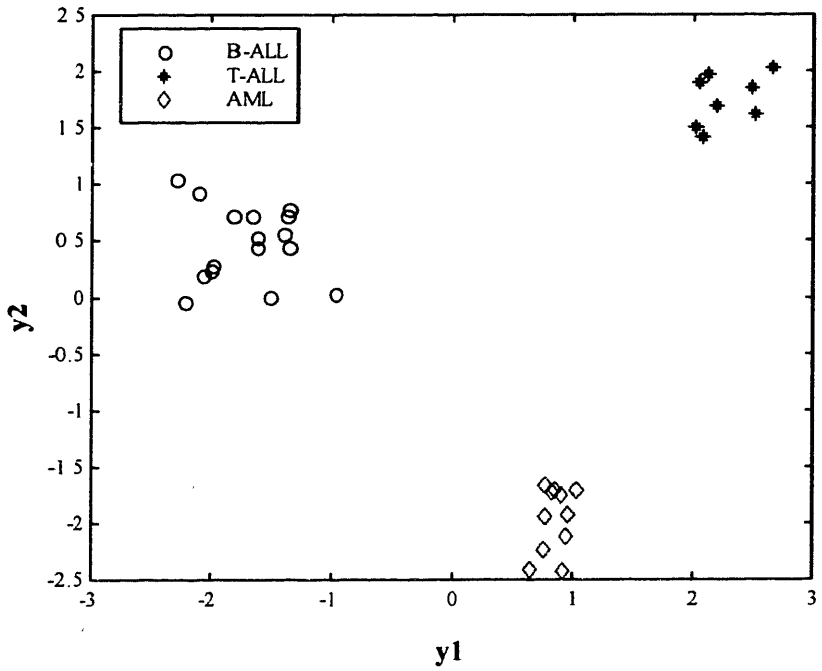

Figure 7-8: 3-class discrimination of leukemia samples using FDA

A classification rule can be built in this FDA space. A new sample is projected into the FDA space using the discriminant weights (V). Then, the new sample will be assigned to the predefined class whose mean is closest to the projection of the new sample[14]: a new sample ($\mathbf{x}$) will be allocated to class j if

$$\left\|\hat{\mathbf{y}} - \bar{\mathbf{y}}_j\right\|^2 = \left\|(\hat{\mathbf{x}} - \bar{\mathbf{x}}_j)\mathbf{V}\right\|^2 \le \left\|(\hat{\mathbf{x}} - \bar{\mathbf{x}}_k)\mathbf{V}\right\|^2 \text{ for all } k \ne j \qquad 7\text{-}16$$

where $\hat{\mathbf{y}}$ is a projection of the new sample into the discriminant axes (V). See Figure 7-9 for a graphical representation of the decision boundaries implied by this rule, where the solid circles represent the means of each class in the reduced dimensional space.

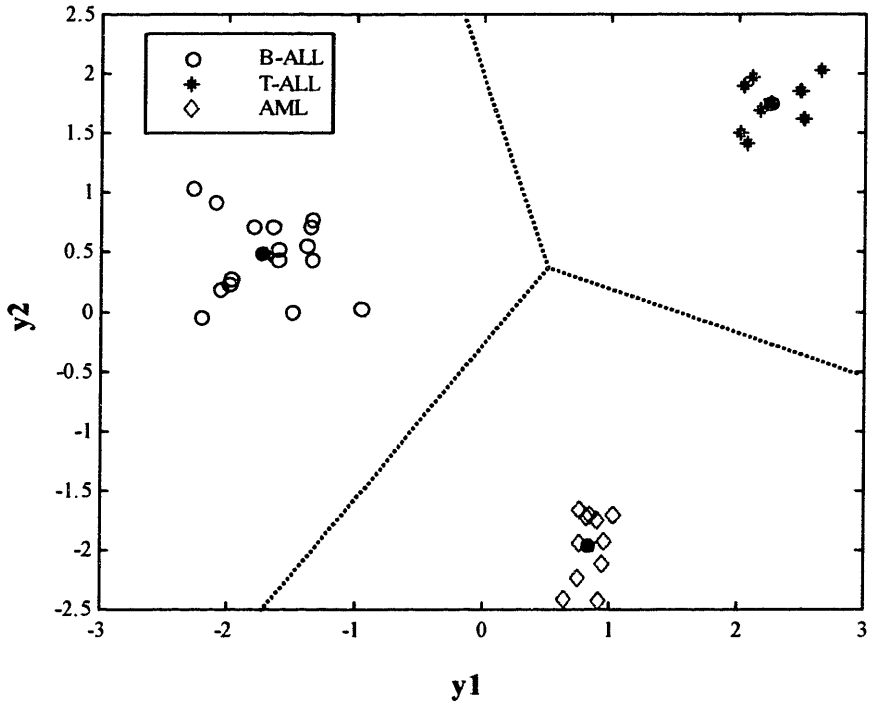

Figure 7-9: 3-class discrimination boundaries for leukemia samples using FDA

It has been shown[14] that FDA is an optimal classification procedure in the sense of misclassification error rate under two assumptions: 1) multivariate normality of the p discriminatory genes, and 2) equal $p \times p$ covariance matrices for each of the c classes. Violation of the assumptions affects several aspects of FDA. For instance, with unequal covariance matrices,

a quadratic classification rule in the FDA projection space performs better than the linear classification rule shown here. Agreement between the quadratic rule and the linear one will decline as the sample sizes decreases, the differences in class covariance matrices increase, the class means become closer, or the number of discriminatory genes increases. In this case study, we employed the linear FDA classifier for simplicity and applicability to the current study. However, we tried to minimize the effect of violations of the assumptions: false positives have been minimized by 2-step selection of a small sub-set of genes (Wilks' lambda as in section 7.2.3 followed by LOOCV as in 7.2.4), and we ensured sufficient mean difference among classes using power analysis (see section 7.4). Other classifiers have also been introduced which can be applied to various microarray data[15,23], and in general these may be substituted for the FDA classifier for application to other data, at the user's discretion.

7.4 Statistical robustness

There are certain issues of statistical reliability that need to be addressed in the implementation of array technologies. Microarray data are typically subjected to analyses such as hypothesis testing, classification, clustering, and network modeling that rely on statistical parameters in order to draw conclusions[3,4,8]. However, these parameters cannot be reliably estimated with only a small number of array samples and poor sample distributions of gene expression levels. Since the statistical reliability of conclusions largely depends on the accuracy of the parameters used, a certain minimum number of arrays is required to ensure confidence in the sample distribution and accurate parameter values.

This section is concerned with the determination of the minimum number of gene expression arrays required to ensure statistical reliability in disease classification and identification of distinguishing expression patterns. This is an important issue considering the scarcity of tissue samples that can be used for transcriptional profiling and the fact that microarray measurements are rather costly in terms of time and reagents required. As a result, there is a tendency to carry out only a small number of microarray measurements that in many cases are inadequate for the intended purpose. Conclusions based on an inadequate number of arrays will not be statistically sound.

The method proposed here first identifies differentially expressed genes across disease subtypes, hereafter called discriminatory genes, using Wilks' lambda score[4,10,14] (see section 7.2.3) and leave one out cross-validation (LOOCV)[17] (see section 7.2.4). Then, Fisher Discriminant Analysis (FDA)[4,10,24] (see section 7.3.2) is invoked to define linear combinations of these discriminatory genes that form a lower dimensional discrimination space where disease subtypes (classes) are maximally separated. Finally, the minimum number of array samples necessary is estimated to ensure satisfactory separation of the linear combinations (*i.e.* the projections) of the discriminatory genes in this lower-dimensional discrimination space. It should be noted that the minimum number of array samples is estimated only in the reduced dimensional space, and therefore the composite expressions of the genes are well characterized and not necessarily the individual genes themselves.

7.4.1 Power analysis

Determining the number of microarray samples has been presented as an important issue previously[25,26] and is one of the first things to be considered when attempting classification of samples through microarrays. We present power analysis for determination of the minimum sample size required for accurate classification. Instead of using individual genes, we used the c-1 dimensional FDA projections (y in Figure 7-7 through Figure 7-9) in our analysis, because the FDA classification is based on those projection variables. Then, we validated the estimated minimum sample size by testing the entire methodology presented in this section: selecting discriminatory genes, building a FDA classifier, and finally calculating the actual power through power analysis.

Power analysis[27-29] has been used in many applications and is based on two measures of statistical reliability in the hypothesis test, confidence level ($1-\alpha$) and power ($1-\beta$). The test compares the null hypothesis (H_0) that the means of classes are the same against the alternative hypothesis (H_1) that the means of classes are not same. While the confidence level of a test is the probability of accepting the null hypothesis, when the means of classes are in fact same, the power of a test is the probability of accepting the alternative hypothesis, when the means of classes are in fact different[30]. Alternatively, the type I error (false positives, α) is the probability of accepting the alternative hypothesis, when the means of classes are in fact the same, while the

type II error (false negatives, β) is the probability of accepting the null hypothesis, when the means of classes are in fact different[30]. The estimation of the sample size in power analysis is done in such a way that the two statistical reliability measures, the confidence and the power, in the hypothesis test are required to meet or exceed predefined values. Typical analyses may require as 95% confidence and 95% power, for example.

The confidence level and the power are calculated from the distributions of the null hypothesis (H_0) and the alternative hypothesis (H_1). Defining these distributions depends on the statistical measure being used in the hypothesis test. In the case of two-class distinction with a one-dimensional FDA projection, the normalized mean difference follows the *t* distribution in the FDA space. As discussed in section 7.2.2, the *t* statistical measure for the hypothesis test is defined as[27,29]

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2 \qquad 7\text{-}17$$

$$t = \frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \qquad 7\text{-}18$$

where, in this instance, μ_i and $\bar{\mathbf{y}}_i$ are the actual mean and the sample mean of the one-dimensional projection variable ($\mathbf{y}_i$) in class *i*. S_p is the pooled standard deviation of the projected samples of the two classes, n_i is the number of samples in class *i*, *N* is the total number of samples, and *N*-2 is the degrees of freedom for the *t* distribution.

While the distribution of H_0 with all classes having the same mean is defined as a central distribution, the distribution of H_1 with all classes having different means is non-central. The effect size (Δ_e) is a critical mean difference that can be considered important enough to warrant attention, and should be set before power analysis is conducted. Power analysis estimates the minimum sample size to ensure the power in the test for the effect size. The non-central distribution H_1 is defined by the non-centrality parameter (Δ), which is defined by the effect size (see below). For the case of two-class distinction, the effect size (Δ_e) is the critical mean difference normalized by the pooled standard deviation (S_p). Thus, the distributions of H_0 and H_1 are defined as follows.

$$H_0 : t = \frac{(\bar{y}_1 - \bar{y}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(N-2) \qquad 7\text{-}19$$

$$H_1 : t = \frac{(\bar{y}_1 - \bar{y}_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t\left(N-2; \Delta = \frac{\Delta_e}{\sqrt{1/n_1 + 1/n_2}}\right) \qquad 7\text{-}20$$

$$\text{with } \Delta_e = \frac{(\bar{y}_1 - \bar{y}_2)_{crit}}{S_p} \qquad 7\text{-}21$$

The confidence level and the power are calculated using the defined distributions of H_0 and H_1 for a given sample size and an initial guess for the effect size determined on the basis of engineering judgement or prior knowledge of the system. The critical value of the inverse *t* distribution at the probability of 1-α/2, shown by the dotted line in Figure 7-10, is first identified for the distribution of H_0 (here, α = 0.05 to give a 95% confidence level). For this confidence level, the power is determined from the distribution of H_1 in the region from this critical *t* value to positive infinity (indicated in Figure 7-10 by the area under the H_1 distribution after the critical value). If the power calculated is below the predefined value 1-β (here, 95%), the sample size is increased until the power reaches this threshold. Figure 7-10 shows the confidence level, power, type I error and type II error in the distributions of H_0 and H_1 defined by the determined sample size. The sample size estimated from this power analysis is the total number of samples, so that the number of samples required in each class is obtained by dividing the total sample size by the number of classes (*c*). This assumes that the standard deviation matrix is approximately similar for each class, implying that equal numbers of samples are needed for each class. For the case of 2-class distinction in leukemia samples (AML vs. ALL) the number of samples suggested by this analysis is shown in Figure 7-11 (see sections 7.4.2 and 7.4.3 for application specifics).

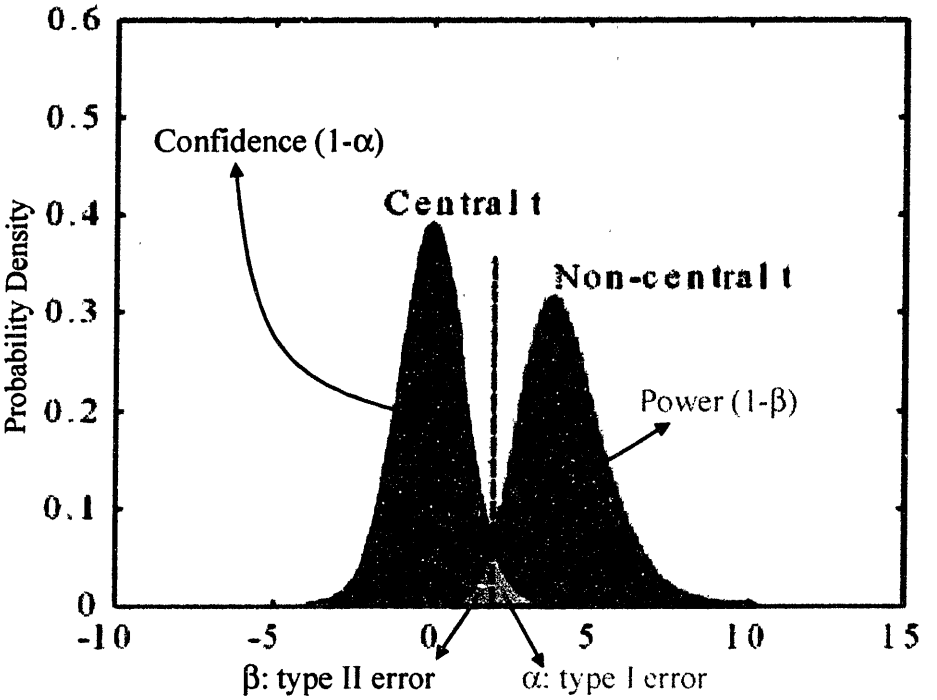

Figure 7-10: H_0 and H_1 for 2-class distinction

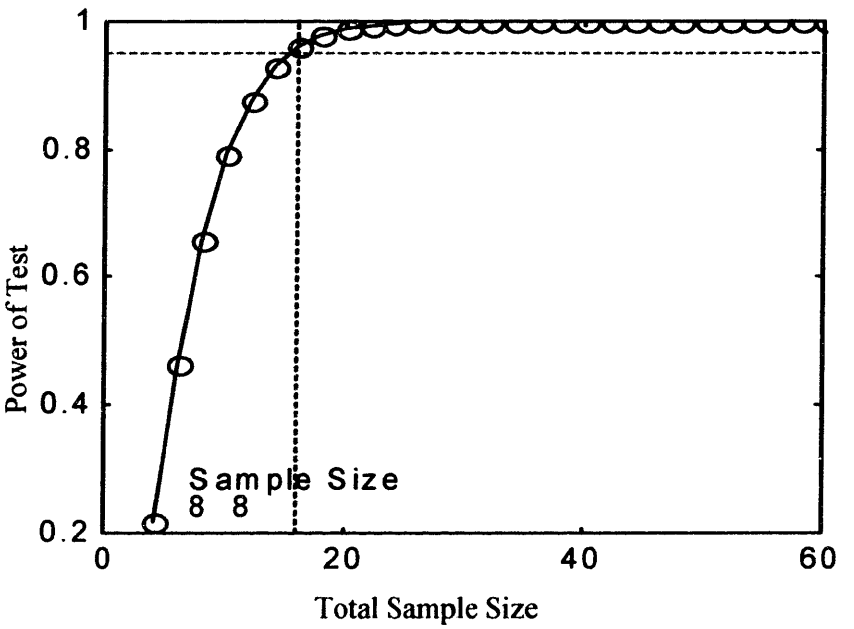

Figure 7-11: Sample size determination for 2-class distinction (AML/ALL)

In the case of distinguishing $c > 2$ classes, instead of the t statistic, the F statistic measure derived from Pillai's V is used for the estimation of the sample size[31]. Pillai's V is the trace of the matrix defined by the ratio of between-group variance ($\boldsymbol{B}$) to total variance ($\boldsymbol{T}$), and is a statistical measure often used in multivariate analysis of variance (MANOVA)[16,31].

$$V = trace\left(\boldsymbol{B}\boldsymbol{T}^{-1}\right) = \sum_{i=1}^{h} \frac{\lambda_i}{1+\lambda_i} \qquad 7\text{-}22$$

where λ_i is the i^{th} eigenvalue of $\boldsymbol{W}^{-1}\boldsymbol{B}$ and h is the number of factors being considered in MANOVA, defined by h=c-1. When $\boldsymbol{W}$ and $\boldsymbol{B}$ are computed, the c-1 dimensional FDA projections are used, because they are the test variables for this analysis. A high Pillai's V means a high amount of separation between the samples of classes, with the between-group variance being relatively large compared to the total variance. The hypothesis test can be designed as shown below using the F statistic transformed from Philai's V[32].

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_c \text{ and } H_1 : \mu_i - \mu_j \neq 0 \; \exists i, j \qquad 7\text{-}23$$

$$H_0 : F = \frac{(V/s)/(ph)}{(1-V/s)/[s(N-c-p+s)]} \sim F[ph, s(N-c-p+s)] \qquad 7\text{-}24$$

$$H_1 : F = \frac{(V/s)/(ph)}{(1-V/s)/[s(N-c-p+s)]} \sim F[ph, s(N-c-p+s), \Delta = s\Delta_e N] \qquad 7\text{-}25$$

$$\text{with } \Delta_e = \frac{V_{crit}}{(s - V_{crit})} \qquad 7\text{-}26$$

where p and c are the number of variables and the number of classes, respectively. s is defined by min(p,h). The confidence level and the power can be calculated using these defined distributions of H_0 and H_1 for a given sample size and an effect size. The same procedure used in the case of two-class distinction is used here to estimate the minimum sample size for statistical reliability whereby the sample size is increased until the calculated power reaches the predefined threshold value of 1-β (95% for the cases shown here). Figure 7-12 shows the distributions of H_0 and H_1 for the case of leukemia samples from three classes, and Figure 7-13 shows the resulting sample size needed for power of 95% (see sections 7.4.2 and 7.4.3 for application specifics).

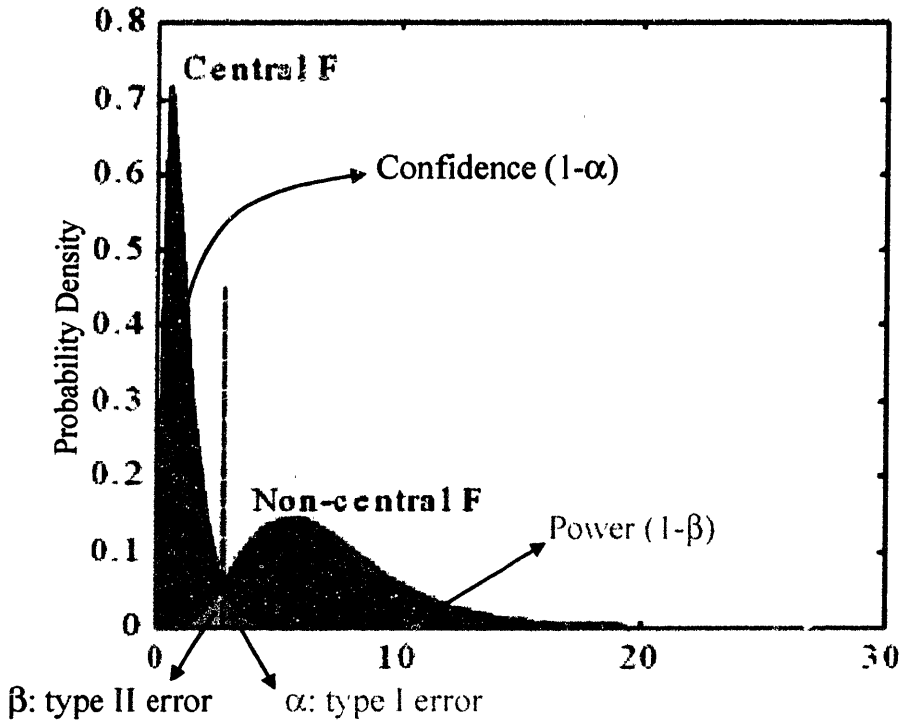

Figure 7-12: H_0 and H_1 for 3-class distinction

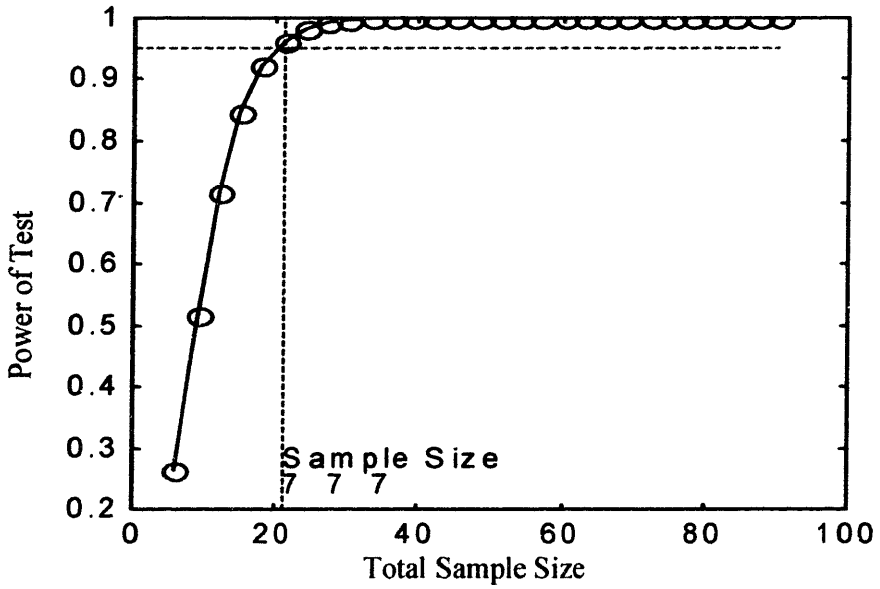

Figure 7-13: Sample size determination for 3-class distinction (AML/B-/T-ALL)

The above approach is applicable only to FDA projection variables and not to the expression data from a large number of individual genes, because the denominator in the F statistic above, which generally has a positive value, may become negative due to the large number of genes (p) typical in microarray experiments. PCA can be used to reduce the number of variables (p) to resolve such a problem. There is, however, a limitation that the number of PCs (p) cannot be larger than $N\text{-}c+s=N\text{-}1$ and in most array cases the maximum number of PCs ($p=N\text{-}c+s$) does not capture enough discriminating characteristics among the classes. Thus, we use only the projections through FDA in our analysis. This analysis may produce a misleading sample size estimate when the real gene expression data are not consistent with the assumptions underlying the statistics used in power analysis (*i.e.* normality and equal variance). To check the effect of possible violations of the assumptions on the estimated sample size, the actual power and mean differences between classes are compared to the pre-defined values (see section 7.4.3 for examples). The actual values in both cases studied were sufficiently large that we need not be worried about the impact of data which does not perfectly match the normality or equal variance assumptions.

7.4.2 Algorithm

See Figure 7-14 for a schematic representation of the power analysis algorithm. It is first necessary that the type I and type II errors, an initial sample size, and a reasonable effect size are selected for the initiation of the algorithm. Then, after the test is designed in terms of the null hypothesis, the alternative hypothesis, and an appropriate statistic measure (t-test or F test), the distributions of H_0 and H_1 are determined using the degrees of freedom and the non-centrality parameter as described above. Next, the inverse of the F distribution at the value of $1\text{-}\alpha$ in the probability distribution is identified and the power is calculated using the distribution of H_1. If the calculated power is less than the predefined power, $1\text{-}\beta$, then the sample size is increased and the power is recalculated using the same α, β, and effect size but a new sample size until it reaches the preset power value. Following determination of the number of samples from power analysis, the actual effect size and power are computed and their values compared to the initial guesses. The actual effect size and power should be larger than those used/calculated in the original analysis so as to not underestimate the sample size.

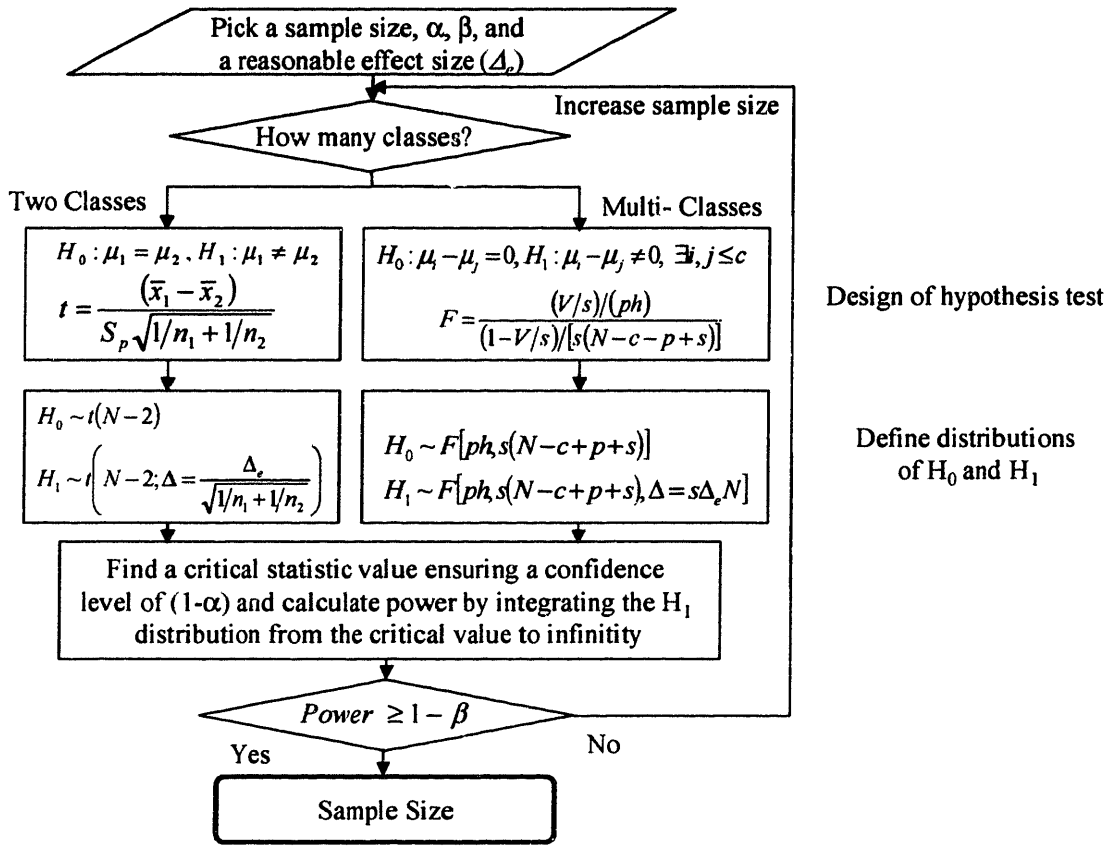

Figure 7-14: Power analysis scheme for determination of sample size

7.4.3 Implementation and results

Power analysis was applied to two-class distinction between ALL and AML subtypes of leukemia. The null hypothesis (H_0) was that the two group means (*i.e.* the group averages in the FDA space) were the same, with the alternative hypothesis (H_1) that the two group means were not the same. The mean difference normalized by the pooled standard deviation was used as the *t* statistic measure. The effect size was preset to 2, which corresponds to a mean difference 2 times larger than the pooled standard deviation and the predefined confidence and power were set to 95% (equivalent to α=0.05 and β=0.05). Figure 7-11 shows the dependence of the power calculated from the H_1 distribution on the sample size. 8 samples from each class (16 total) are required for the FDA projection to establish a sufficient base for the H_1 to be accepted. This indicates that with these 8 samples from each class, there is a large enough mean difference

between ALL and AML so that an accurate classifier can be constructed in the FDA projection space with statistical reliability.

In order to validate this minimum sample size, the proposed procedures for discriminatory gene selection and FDA classification were applied to 8 randomly chosen samples from each class and then the actual effect size and the actual power were calculated. The procedure of discriminatory gene selection through Wilks' lambda (section 7.2.3) and LOOCV (section 7.2.4) was used to identify the 50 most discriminatory genes (Figure 7-15). This final list of 50 genes is shorter than the 388 discriminatory genes obtained by using a simple Wilks' lambda score, thus enabling us to reduce the errors due to false positives. Then, using the 50 discriminatory genes, FDA classification was performed (as shown in Figure 7-7). In the FDA projection space, the actual normalized sampled mean difference was computed to be equal to 7.2453. This is more than three times larger than the effect size used for power analysis, confirming that the effect size chosen was reasonable enough not to underestimate the sample size. There are two potential sources of the difference between the sampled mean difference and the effect size: 1) only the most discriminatory genes were selected with a stringent level of significance in Wilks' lambda and by the LOOCV, and 2) the FDA further screens out the maximal discriminating information from the most discriminatory genes. The actual confidence and power were also close to 100%.

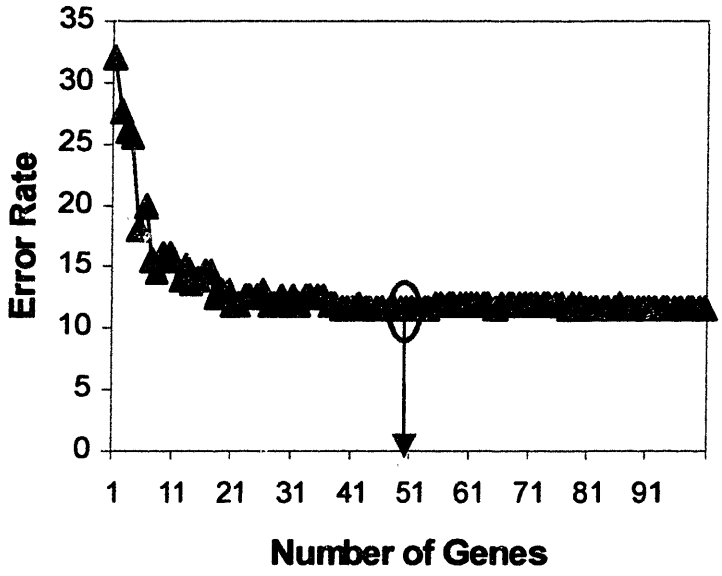

Figure 7-15: Cross-validation results for gene selection, AML/ALL classification

As a multi-class case study, the distinction of three subtypes, B-ALL, T-ALL, and AML was considered. The null hypothesis H_0 is that the three group means are same, while the alternative H_1 is that at least one of the group means is different from the rest. The F statistic measure was used for power analysis and the effect size was chosen to be 0.538. This effect size is equivalent to 0.7 critical Pillai's V for three classes, meaning that the between-group variance is 0.7 of the total variance. The predefined confidence and power were set to be 95%, equivalent to α=0.05 and β=0.05. The minimum sample size was computed to be 7 samples from each class from the power curve shown in Figure 7-13. The distributions of H_0 and H_1 are shown in Figure 7-12. After the gene selection procedure was applied to seven randomly selected samples from each class, the final set of 80 discriminatory genes was identified (data not shown, but analogous to Figure 7-15). With those genes, the FDA was performed as shown in Figure 7-16. In the FDA space, the actual measure of effect size defined by $V/(s-V)$ was computed to be 1.7552, which is about three times larger than the one used for power analysis for the same reasons given in the previous case. The actual confidence and power were also close to 100%.

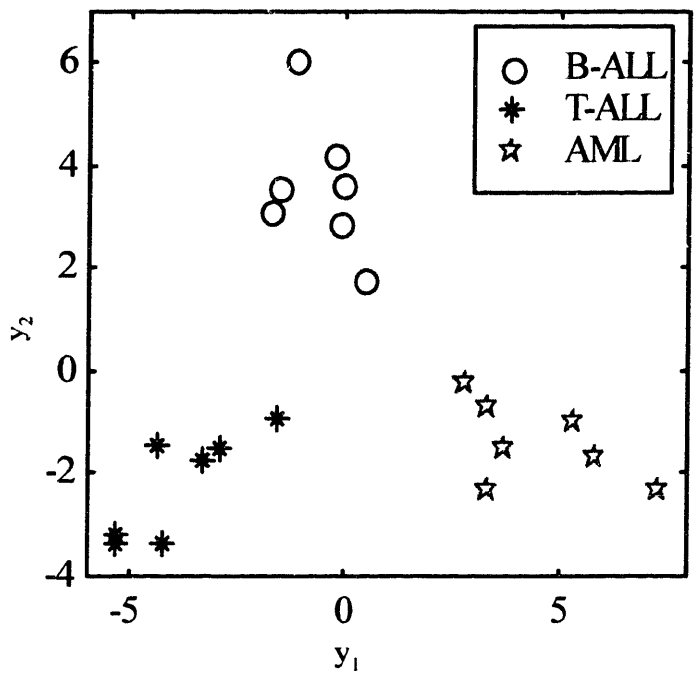

Figure 7-16: FDA results for sample-size testing, AML/B-/T-ALL classification

7.4.4 Reliability discussion and conclusions

This study addressed the issue of statistical reliability for the classification of disease subtypes on the basis of the sample size. The appropriate statistical measures have been defined for two-class and multi-class problems, and these statistics have been applied in a power-analysis framework to determine the minimal sample size based on the distributions of the statistic measures. This framework has a been applied in earlier studies[33] for determining the minimum number of subjects required in clinical trial studies, when a new drug is discovered and its efficacy is being evaluated. In this case, the minimal sample size determined from power analysis is used ensure statistical reliability of an efficacy measure.

This reliability issue can also be central in other applications involving any statistical analysis, with this study giving only one example. For instance, correlations between genes are often considered in microarray studies in the search for co-regulated genes, an example of which is a central focus of this work (Chapter 4 and Chapter 6). A small number of samples will result in unreliable correlation coefficients, so when additional samples are included, the estimated

correlation coefficients will show a high degree of variability. Thus, the appropriate sample size can be determined by power analysis to ensure that the distribution of correlation coefficients is reliable. Another application is the construction of a regression model using gene expression data to estimate the level of an important cellular variable. For instance, gene expression regression models of urea level in liver tissues should also be supplemented by power analysis to determine the sample size for the model to have statistically reliable regression parameters. Although the range of uses is broad, the appropriate statistical measures and their distributions should be carefully chosen in these sorts of applications.

Power analysis determines the sample size based on the assumption of homogeneous sampling from the entire population of each class (*ie.*, a disease subtype as in this study). Therefore, during sample collection, if the number of samples suggested by power analysis doesn't cover the broad population of each subtype to capture the inherent variance of the population, the distributions of parameters will be biased toward the type of samples collected. As a result, a poor sampling can make power analysis appear to underestimate the necessary sample size. Furthermore, statistical inference based on the calculated parameters can be misleading. The FDA has recently noticed the importance of broad sampling and requested pharmaceutical industries to include clinical trial studies on pediatric patients in order that the efficacy measure should not be biased to adults. As a result, a well-designed sampling strategy is required together with a reasonable estimate of sample size calculated from power analysis to ensure statistical reliability.

This study uses linear combinations of individual genes as variables in the classifier instead of classifying using the individual genes themselves. Although the discriminatory genes used for the classifier are chosen based on Wilks' lambda score and the error rate calculated through LOOCV, the number of selected genes is usually still large (50 or more depending on the situation). If all individual genes are considered independently in constructing a classifier, and new samples are classified using the sum of all gene contributions to the classifier (as in the "voting" schemes discussed in section 7.3), the classifier will not capture the interaction of the genes and may be biased to redundant characteristics. In addition, the parameters in the classifier will be subject to statistical variations of the individual genes. If all the genes are considered together as seen in multiple discriminant analysis (MDS), it may be difficult to

estimate the model parameters due to the large number of discriminatory genes and singularity in the data. On the other hand, the linear combinations of individual genes obtained from FDA capture the important discriminating characteristics at the outset because the algorithm seeks the most relevant directions (weights) for separation of classes. Thus, the number of variables used for the classifier is significantly reduced to several FDA projection variables (the number of classes minus one), while capturing in a large degree the discriminating characteristics in data. This reduction in variables is achieved without significant accuracy cost in discrimination.

The use of FDA also reduces the amount of noise obscuring the information content of the data. Signals that merely appear to be random noise will be filtered out during the process of obtaining the weights for the linear combinations. Just as the first few PCs in PCA usually capture the important patterns and the last few PCs only random noise, the first few discriminant functions in FDA capture the important discriminating characteristics in the data. Only systemic noise that happens to have similar patterns to the real signals may be retained in the data projected through the linear combinations.

Finally, the interactions and relative contributions of the individual genes to the classification can be interpreted from the discriminant weights in the linear combinations, improving our understanding the discriminant features in the data. As a result, the FDA classifier using linear combinations as variables can provide many preferable aspects of classification relative to other techniques, including robustness in performance, non-complexity in modeling, and improvement in interpretation.

7.5 Conclusions

We have demonstrated statistically justifiable tools for analysis of differently expressed samples of gene-chip data. Through the use of such measures as the t-test, Wilks' lambda criterion, and cross-validation, methods have been shown to give statistically sound indicators of the most relevant genes for a given classification problem. These tools include appropriate adjustments to account for the small numbers of samples typically found in microarray data. We have shown that comparing a sample population from one type of tissue against a varied set of other tissue samples we can identify discriminatory genes that are likely to represent some specialized function for the tissue type under study. We have made a similar comparison for the more

straightforward case of different disease states. These analyses provide for quick identification of targets for further research while employing a level of sophistication no more complicated than typical laboratory statistical techniques.

A framework for diagnostic evaluation of new expression samples using statistically optimal tests, such as the likelihood ratio test and Fisher's Discriminant Analysis, have also been demonstrated. These techniques are shown to be reasonably robust even when only a small number of samples are available, through the use of *t*- and *F*-tests on the distribution of the measurements. Use of these frameworks provides a measure of how strongly a new sample is classified, and whether or not the classification should be trusted. Application on leukemia data gives 100% classification under most conditions. These tools are equally applicable to the study of tissue identification and categorization.

Finally, we have included power analysis to create an algorithm for testing not only for robust gene identification and sample classification, but also to give a measure of the number of samples required to reach a given confidence level in the results of the test. Both 2-class and 3-class distinction examples have been demonstrated for leukemia microarray data. This methodology allows researchers to balance, in an iterative fashion, the need for reliable conclusions against the expense of running too many DNA microarrays.

7.6 References

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. "Quantitative Monitoring Of Gene-Expression Patterns With a Complementary-Dna Microarray." *Science* **270**, 467-470 (1995).

2. Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C. W., Kobayashi, M., Horton, H. & Brown, E. L. "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nature Biotechnology* **14**, 1675-1680 (1996).

3. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring." *Science* **286**, 531-537 (1999).

4. Stephanopoulos, G., Hwang, D. H., Schmitt, W. A. & Misra, J. "Mapping physiological states from microarray expression measurements." *Bioinformatics* **18**, 1054-1063 (2002).

5. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. "Cluster analysis and display of genome-wide expression patterns." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **95**, 14863-14868 (1998).

6. Perou, C. M., Jeffrey, S. S., Van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O. & Botstein, D. "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **96**, 9212-9217 (1999).

7. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. & Golub, T. R. "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **96**, 2907-2912 (1999).

8. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. G., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L. M., Marti, G. E., Moore, T., Hudson, J., Lu, L. S., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. & Staudt, L. M. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* **403**, 503-511 (2000).

9. Hsiao, L. L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee, K. F., Clark, K. E., Haverty, P., Weng, Z. P., Mutter, G. L., Frosch, M. P., MacDonald, M. E., Milford, E. L., Crum, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A., Stephanopoulos, G. & Gullans, S. R. "A compendium of gene expression in normal human tissues." *Physiological Genomics* **7**, 97-104 (2001).

10. Dillon, W. R. & Goldstein, M. *Multivariate Analysis* (Wiley, New York, 1984).

11. Welch, B. L. "The generalization of Student's problem when several populations are involved." *Biometrika* **34**, 28-35 (1947).

12. Kamimura, R. T. in *Department of Chemical Engineering* (Massachusetts Institute of Technology, 1997).

13. Rice, J. A. *Mathematical Statistics and Data Analysis* (Duxbury Press, Belmont, California, 1995).

14. Johnson, R. A. & Wichern, D. W. *Applied Multivariate Statistical Analysis* (Prentice Hall, Englewood Cliffs, New Jersey, 1992).

15. Dudoit, S., Fridlyand, J. & Speed, T. P. "Comparison of discrimination methods for the classification of tumors using gene expression data." *Journal of the American Statistical Association, Technical Report* **576** (2000).

16. *SAS/STAT User's Guide* (SAS Institute Inc., Cray, N.C., 1989).

17. Lachenbruch, P. A. & Mickey, M. R. "Estimation of Error Rates in Discriminant Analysis." *Technometrics* **10**, 1-11 (1968).

18. Tusher, V. G., Tibshirani, R. & Chu, G. "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121 (2001).

19. Storey, J. D. & Tibshirani, R. (Stanford Tech Report, 2001).

20. Alter, O., Brown, P. O. & Botstein, D. "Singular value decomposition for genome-wide expression data processing and modeling." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **97**, 10101-10106 (2000).

21. Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. "Fundamental patterns underlying gene expression profiles: Simplicity from complexity." *Proceedings Of the National Academy Of Sciences Of the United States Of America* **97**, 8409-8414 (2000).

22. Misra, J., Schmitt, W., Hwang, D., Hsiao, L. L., Gullans, S. & Stephanopoulos, G. "Interactive exploration of microarray gene expression patterns in a reduced dimensional space." *Genome Research* **12**, 1112-1120 (2002).

23. West, M. "DNA Microarray Data Analysis and Regression Modeling for Genetic Expression Profiling." *Critical Assesment of Microarray Data Analysis* Duke University, (2001).

24. Zhao, G. & MacLean, A. L. "A comparison of canonical discriminant analysis and principal component analysis for spectral transformation." *Photogramm Eng. Rem. S*, 841-847 (2000).

25. Zien, A., Fluck, J., Zimmer, R. & Lengauer, T. "Microarrays: How many do you need?" *Research in Computational Biology (RECOMB)* Washington, D.C., (2002).

26. Pan, W., Lin, J. & Le, C. (Biostatistics, University of MN tech report, 2001).

27. Mace, A. E. *Sample-size determination* (Krieger, Huntington, NY, 1974).

28. Cohen, J. *Statistical power analysis for the behavioral sciences* (Erlbaum, Hillsdale, NJ, 1988).

29. Kraemer, H. C. & Thiemann, S. *How many subjects? Statistical power analysis in research* (Sage, Newbury Park, CA, 1987).

30. G power reference material. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/reference/reference_manual_02.html#noncentral.

31. Olson, C. L. "Comparative Robustness of 6 Tests in Multivariate-Analysis of Variance." *Journal of the American Statistical Association* **69**, 894-908 (1974).

32. Other F-tests. http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/reference/reference_manual_09.html#t3

33. Thall, P. F. *Recent Advances in clinical trial design and analysis* (Kluwer Academic Publishers, Boston, 1995).

CHAPTER 8 SUMMARY AND SIGNIFICANCE OF WORK

8.1 Summary of thesis results

As stated in the introduction, this thesis aimed to:

- Explore the use of statistical methods for DNA microarray data to identify a subset of maximally informative genes for a given experiment,
- Develop methodologies for identifying genes that have significant changes in expression pattern which appear to be related to changes in environmental conditions,
- Construct hypothetical regulatory networks from the relationships suggested by correlational analyses,
- Apply these methods to elucidate the transcriptional programming of *Synechocystis* sp. PCC6803,
- Identify what information or experimental conditions are required to distinguish between hypothesized networks or establish their existence,
- Address all of these issues in a manner that is compatible with future high-throughput experimental data, from both DNA microarrays and other sources.

Statistical methods for not only gene identification but also other DNA microarray issues have been discussed at length in Chapter 7. *t*-tests, Wilks' lambda criterion, and cross-validation have proven to provide robust identification of the most discriminatory genes, even in the face of the relatively high amount of noise typical to DNA microarray experiments. Techniques including likelihood ratio tests and Fisher discriminant analysis (FDA) have been shown to provide strong classification ability based on such discriminatory genes. Finally, the use of power analysis in a multidimensional classification scheme has been explored to create a simple metric indicating whether or not a sufficient number of samples has been taken to support a given hypothesis, and if not, estimate the number of additional samples that may be required. All of these tools have been programmed into Matlab codes for use with the data formats typical to DNA microarrays.

Chapter 4 discusses other computational tools more directly aimed at the application of DNA microarray data to understanding transcriptional regulatory networks. Specifically, the technique of time-lagged correlations has been explored thoroughly to not only to identify sets of co-expressed genes, but also to put them into a temporal ordering. Such temporal patterns give insights into the cellular response to a given stimuli, but more importantly allow for the formation of cause-and-effect hypotheses. A practical algorithm for the application of these correlations to DNA microarray data, where the existence of thousands of genes makes correlation calculations exceedingly difficult computationally, has also been developed. Programs have been written in Matlab for this entire procedure, including not only the time-lagged correlation network algorithm but also a set of network visualization tools for use with AT&T's Graphviz program.

In order to model these correlation networks, the use of AutoRegressive with eXogenous input (ARX) models are also discussed in Chapter 4. In this formulation, the transcript abundance of a gene or group of genes is expressed as a function of prior measurements at earlier time-points: measurements of its own transcript levels, those of other genes, and an input forcing function of external conditions. These models can vary in complexity and expected prediction bias, so the use of Akaike's Information Criterion (AIC) has also been explored to make assessments of a model's relative predictive capacity while taking into account the possibility of overfit due to an excess of parameters.

The model organism for application of these techniques, *Synechocystis*, was introduced along with appropriate experimental techniques in Chapter 5. This cyanobacterium is a uniquely useful prototype system as it has been fully sequenced, has industrial application through fixation of CO_2 waste to biopolymers, and is photoautotrophic. This ability to process light energy into cellular energy has been shown to be regulated at the transcriptional level by experimental light conditions, and therefore light intensity provides an optimum, dynamically adjustable input parameter for the type of network reconstruction problem posed by this thesis.

The application to *Synechocystis* of the experimental and computational tools developed in Chapter 4 and Chapter 5 is shown in Chapter 6. A series of 50 time-points have been collected for a single, continuous series of light intensity shifts and the extracted RNA from these samples

has been hybridized to DNA microarrays. The microarray data has been used to create hypothetical network maps of varying complexity, which have been tested for robustness by a variety of statistical measures. ARX models of the gene groups in this mapping allowed for prediction of how new experimental conditions might lead to different measured output data. These predictions were then used to suggest the optimal validation experiment (*i.e.* the experimental profile which was predicted to create the biggest between-group differences in the network). This additional experiment of 27 DNA microarrays has been used to confirm most of the network connections proposed by the analysis of the first experiment, while suggesting that alternative relationships for other genes may need to be explored. Furthermore, this confirmatory experiment validated the constructed ARX models as well as the conclusions drawn by AIC about model complexity and predicted fit.

8.2 Significance of Results

Ultimately, it is a goal of biological engineering to achieve mechanistic, predictive models of cellular behavior for the purpose of determining the system response to applied perturbations. Examples include predicting the impact of environmental changes, application of novel drugs, or genetic modifications to the cell itself. Such models will allow for the directed design of new experiments, novel therapies, and new cell strains.

At this time, these types of mechanistic models are not generally possible, due to a lack of good measurements at a cell-wide scale with which to fit models. However, such data is becoming increasingly available as the promise of high-throughput proteomics and other genomic-based technologies are realized. In the meantime, only transcriptional data is available at the scale necessary for these modeling efforts. Since such data lack sufficient information about other molecules in the cell, practical statistical modeling efforts provide the predictive capacity needed to drive further experimentation. This thesis work has resulted in a suite of statistical tools for analyzing and modeling DNA microarray data, particularly for time-series experiments aimed at elucidating underlying transcriptional regulation.

As an example, Figure 8-1 shows the relationship between the functional form of the ARX models and the underlying biochemical interactions. Each of the parameters a_{ii} and b_j indicate a potential relationship between the transcriptional characteristics of the genes under

consideration, or between the environmental conditions and these genes. For strong parameters, we hypothesize the existence some underlying mechanism for interaction, whether it is direct or through some unmeasured intermediaries. Weak parameters, on the other hand, suggest that fundamental links corresponding to the parameter in question may not exist between these genes, or at least not in a form measured by the experimental conditions.

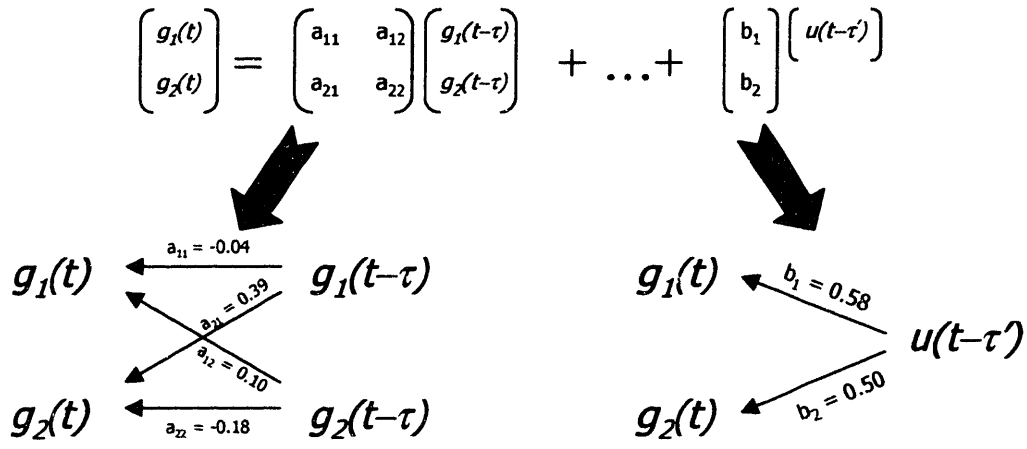

Figure 8-1: The relationship between ARX coefficients and underlying biology

As a whole, the success of this modeling endeavor for *Synechocystis* can be evaluated by:

1. The function of the factors present in the model. In this case, the biological identity of the genes in groups found through time-lagged correlation analysis indicates strong presence of photosystem-related components (phycobilisome components, atp synthase sub-units, photosystem I and II proteins, *etc.*), as discussed in Chapter 6.

2. The ability of the model to give predictive, forward-looking estimates of variables for independent experiments. ARX models generated by this study produced purely predictive outputs for any number of possible input profiles, although extrapolation far outside of the range of the original experiment is expected to perform weakly for these models.

3. The validation of the models through independent data sets. The good match between the predictions made and the data from the validation experiment, as shown in Chapter 6, give a measure of model robustness.

By these criteria, the statistical models shown have utility in providing useful information - even accounting for experimental limitations of unobservable variables and relatively high noise content typical to DNA microarrays.

Furthermore, we have shown that straightforward statistical models of dynamic transcriptional data not only provide predictive capabilities for gene transcription, but also can be used to suggest how additional experiments should be carried out. In general, such designs can elucidate not only which parts of the models are accurately understood, but also which aspects of the model are less strong. For example, if a coefficient in a model suggests strong correlation between the expression levels of two genes, but the validation experiment indicates that the hypothesis was wrong, then the coefficient in question may have been inaccurately assigned or there may in fact be no straightforward relationship present.

As greater experimental capacity becomes available, incorporation of new data into such statistical models can be handled directly through expansion of either the training or testing data sets. No model can be reasonably expected to be robust to every type of environmental insult, as models are created specifically to explain and predict the response to a sub-set of situations the system might encounter. Nevertheless, more complete sampling of the experimental space (*e.g.* the more experimental conditions measured) the better our predictive capacity for future situations. For example, having studied light response dynamically in *Synechocystis* for this set of conditions, it makes sense to turn attention to a wider array of light conditions, other factors such as nutrient concentrations, or the transcriptional response profiles of genetically modified strains. The synthesis of such data will lead to ever-increasing predictive accuracy for a wider range of situations.

Another benefit of improved experimental capacity will be the ability to reduce the time interval between measurements arbitrarily close to zero. In the current experiment, all phenomena that occur at time scales faster than 20 minutes are temporally indistinguishable, although more slowly evolving trends can be measured. Every reduction in time scale improves the resolution of the measured interactions, to the point where it may become possible to directly observe the dynamic biochemical interactions that drive cellular behavior.

As mechanistic models also gain in accuracy and complexity, it will become important to replace statistically derived relationships with those that can be justified physically. As long as some parts of cellular physiology remain unknown, however, statistical models will fill the gap between practical understanding and the underlying phenomena.

THESIS PROCESSING SLIP

FIXED FIELD: ill. _____ name _____

index _____ biblio _____

► **COPIES:** (Archives) Aero Dewey Barker Hum

Lindgren Music Rotch (Science) Sche-Plough

TITLE VARIES: ► ☐ _____

_____

_____

NAME VARIES: ► ☑ Anthony _____

_____

IMPRINT: **(COPYRIGHT)** _____

_____

► **COLLATION:** _____

_____

► **ADD: DEGREE:** _____ ► **DEPT.:** _____

► **ADD: DEGREE:** _____ ► **DEPT.:** _____

SUPERVISORS: _____

_____

NOTES:

cat'r: date:

page: S41

► **DEPT:** Chem. Eng

► **YEAR:** 2003 ► **DEGREE:** Ph.D.

► **NAME:** SCHMITT, William A.

_____