# Exploiting Biological Pathways to Infer Temporal Gene Interaction Models

by

Corey Ann Kemper

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2006]
August 2006



Author .....
Department of Electrical Engineering and Computer Science
August, 2006

Certified by .....
W. Eric L. Grimson
Bernard Gordon Professor of Medical Engineering
Thesis Supervisor

Certified by .....
John Fisher
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

# Exploiting Biological Pathways to Infer Temporal Gene Interaction Models

by

Corey Ann Kemper

Submitted to the Department of Electrical Engineering and Computer Science
on August, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

## Abstract

An important goal in genomic research is the reconstruction of the complete picture of temporal interactions among all genes, but this inference problem is not tractable because of the large number of genes, the small number of experimental observations for each gene, and the complexity of biological networks. We focus instead on the B cell receptor (BCR) signaling pathway, which narrows the inference problem and provides a clinical application, as B cell chronic lymphocytic leukemia (B-CLL) is believed to be related to BCR response. In this work, we infer population-dependent gene networks of temporal interaction within the BCR signaling pathway. We develop simple statistical models that capture the temporal behavior of differentially expressed genes and then estimate the parameters in an Expectation-Maximization framework, resulting in clusters with a biological interpretation for each subject population. Using the cluster labels to define a small number of modes of interaction and imposing sparsity constraints to effectively limit the number of genes influencing each target gene makes the ill-posed problem of network inference tractable. For both the clustering and the inference of the predictive models, we have statistical results that show that we successfully capture the temporal structure of and the interactions between the genes relevant to the BCR signaling pathway. We have confirmatory results from a biological standpoint, in which genes that we have identified as playing key roles in the networks have already been shown in previous work to be relevant to BCR stimulation, but we also have results that guide future experiments in the study of other related genes, in order to further the long term goal of a full understanding of how and why B-CLL cells behave abnormally.

Thesis Supervisor: W. Eric L. Grimson
Title: Bernard Gordon Professor of Medical Engineering

Thesis Supervisor: John Fisher
Title: Principal Research Scientist

# Acknowledgments

First, I would like acknowledge the support and guidance of my advisor, Eric Grimson, whose enthusiasm concerning the research in his Medical Vision group originally helped to convince me to choose MIT for graduate school. Although I ended up moving toward a bioinformatics application in my PhD project, his continued input was always beneficial. I also greatly appreciate the opportunity to work on this project, which was thanks to John Fisher, my co-supervisor on the thesis. He helped so much with the formulation and implementation that without his direction, I probably would still be stuck on the math. His advice on the thesis document itself was essential, and the additional perspective provided by Tommi Jaakkola on my thesis committee reminded me of the importance of stating results precisely. Finally, many thanks to Laurent Vallat, the biologist who not only provided the hours put in to collect the data on which this thesis is based, but also the understanding of the biological background and implications of the work.

All of those in the Vision Group at MIT have been important in making it through the last five years, both academically and personally. Lauren and Lilla, first labmates, then became my roommates and friends. They are terrific role models, and watching both finish up over the last several months kept me on track. I am so appreciative of everyone that I have worked with since I have been here, especially those who helped in the preparation of my defense.

Thanks so much to all of the rest of my friends and family for being so understanding about the craziness of graduate school. I am not always as good about keeping in touch as I should be, but I am thankful for all of them. I hate to start listing people individually because I really should have an acknowledgments section that is as long as the thesis, but I have to mention Christina, Meghan, Sirkka, Emily, Katie, and Diana, all special for different reasons.

Finally, thanks to the Whitaker Foundation for the financial support provided by a graduate fellowship.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

One of the fundamental problems in biology is understanding genomic function. While advances in experimental technology have made it possible to monitor the expression levels of tens of thousands of genes in parallel, interpretation of data on this scale requires sophisticated computational methods in order to gain insight into the inner workings of cells. Ultimately, the goal is a complete picture of the temporal interactions among all genes, but this is an inference problem that is not tractable because of the large number of genes, the small number of experimental observations for each gene, and the complexity of biological networks. We focus instead on the network of temporal interactions in a single signaling pathway which is initiated by stimulation of the B cell receptor (BCR). Not only does this narrow the inference problem, but it also has a clinical application, as B cell chronic lymphocytic leukemia (B-CLL) is believed to be related to BCR response.

B-CLL is the most prevalent leukemia affecting the aging Caucasian population. Not only is B-CLL currently incurable, but patients suffering from B-CLL often follow divergent clinical courses, with some patients living decades after diagnosis while others decline and die relatively rapidly, despite chemotherapy [16]. It is thus difficult to justify an aggressive treatment regimen early in the disease because it is not necessary for the indolent form and is not successful for the aggressive one. In order to do accurate diagnosis and eventually, targeted treatments, the goal is to understand as much about the underlying disease process as possible. In terms of recent B-CLL

research, many genomic studies focus on gene-specific changes that are related to prognosis, as in [13, 17, 19, 24], but a global picture of how and why the differences in cellular behavior occur does not exist. The key to distinctive B-CLL behavior appears to relate to the differential ability of the B cell receptor (BCR) to respond to stimulus [67]. Therefore, in this work, we infer population-dependent gene networks of temporal interaction within the BCR signaling pathway.

An important question is how to acquire the data used for model inference. Ideally, we would measure the levels of expression of proteins and their physical interactions in the cell, but there is currently no systematic way to directly measure protein expression levels or biochemical activity of gene products. Expression patterns of the corresponding mRNA sequence, however, provide an indirect measure [9], and though they contain a large amount of noise [39], DNA microarrays measure the transcripts of tens of thousands of genes simultaneously. Although the time scale of cellular reactions is on the order of seconds to minutes, acquisition of microarray data at that rate is not feasible because time series experiments require multiple arrays and large quantities of biological material. Based upon a pilot study [78], four time points (60, 90, 210 and 390 minutes) appear to capture the vast majority of the trends in temporal expression. While it is possible that important events in regulating the signaling pathway are missed because of restricting analysis to these time points, this is a necessary practical choice. Another practical choice was made in the number of subjects to include in the study, so in addition to 6 healthy subjects, 11 patients were selected on the basis of B-CLL prognosis and response of the cells to BCR stimulation. The entire data set, shown in Figure 1-1, is comprised of differential expression measurements for 54,613 genes, 4 time points, and 17 patients.

The scale of this data set and microarray data in general raises new challenges and new opportunities. Where it was previously possible to analyze the experimental results virtually by hand, microarrays provide an overwhelming amount of information. Keeping in mind the overall goal of interaction networks, we first identify and classify genes that are differentially expressed under BCR stimulation. Measuring stimulated expression with respect to a control distinguishes the genes affected by

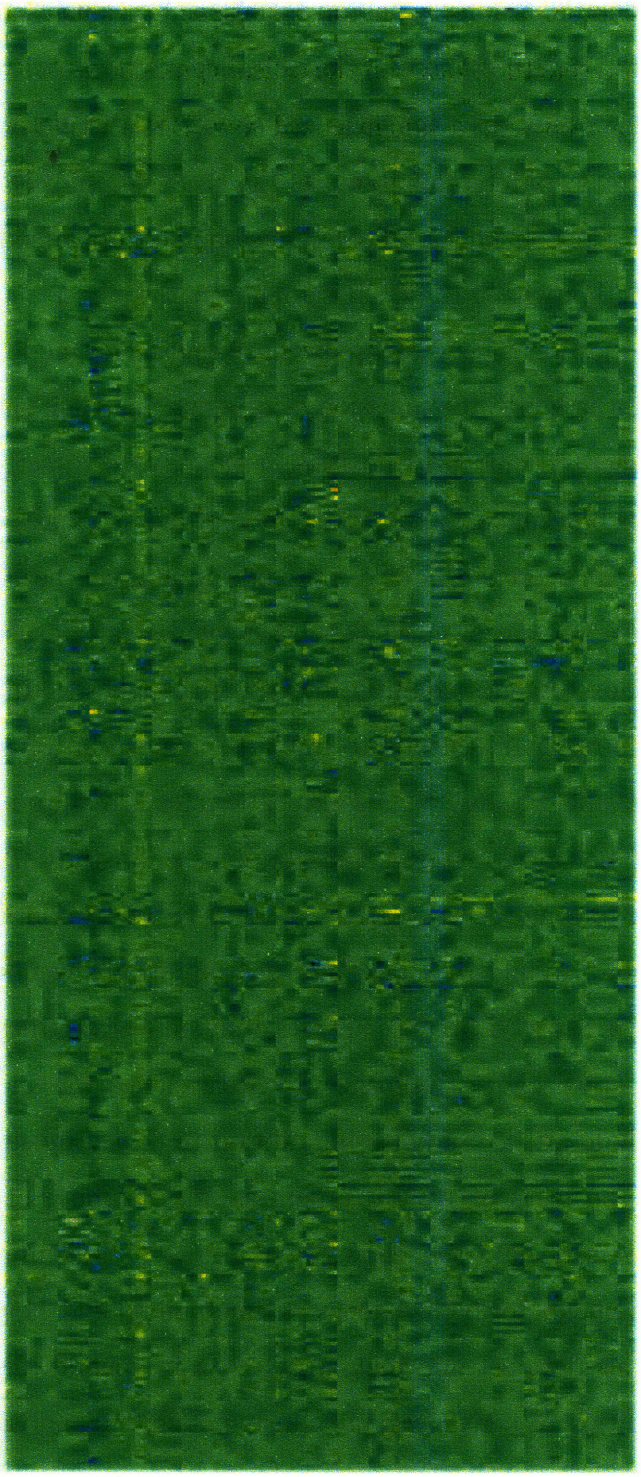


Figure 1-1: Log ratios of stimulated to unstimulated expression data. Each row of the image corresponds to a given gene ($N$ = 54,613). There are four columns, corresponding to time points, which each include data from the 17 subjects.

the BCR pathway from those that are expressed in the cell generally, and BCR stimulation essentially "resets the clock" on the cells, beginning the cascade of reactions in the BCR signaling pathway in a synchronized way. Because B-CLL is believed to be related to genetic differences in BCR signal competence, having subjects from healthy and B-CLL populations allows us to combine subject information into cluster labelings that are specific to the three subpopulations. Data from different patients and different time points are generally treated simply as more experimental conditions, so our approach to clustering temporal expression profiles is unique.

We develop simple statistical models that capture the temporal behavior seen in the pilot study, and then estimate the parameters in an Expectation-Maximization framework, resulting in clusters with a biological interpretation. We model genes that peak at localized time points, i.e. those that have large positive deflections at $t_1$,$t_2$,$t_3$, or $t_4$. Those that are not differentially expressed under BCR stimulation in that way are classified as the "background" genes. In Figure 1-2, we reorder the genes for each subject group according to the cluster labels and then, within a cluster, according to the maximum *a posteriori* assignment value (i.e. how well the gene fits the cluster). Visual structure is evident, as high values in yellow or red start at the top left (first wave, $t_1$) and move down the diagonal to the bottom right (fourth wave, $t_4$). Because individual validation experiments are simply not practical for 50,000 genes and would not be generalizable to multiple subjects or different experimental conditions, we use other measures to determine the success of the clustering, such as significance, robustness to noise and initialization, and quantification of how well the data is predicted by the models.

For the inference of temporal interaction networks, there are many times more samples (genes) than there are observations (time points and subjects), even given the small subset of genes identified as relevant to BCR stimulation by the clustering procedure. Therefore, modeling interactions between every possible pair of genes is not possible. For example, linear models would require inference of $N^2T^2$ parameters (where $N$ is the number of genes and $T$ the number of time points) from $N \times T \times P$ data points (where $P$ is the number of patients). Because we only have 5 or 6 subjects

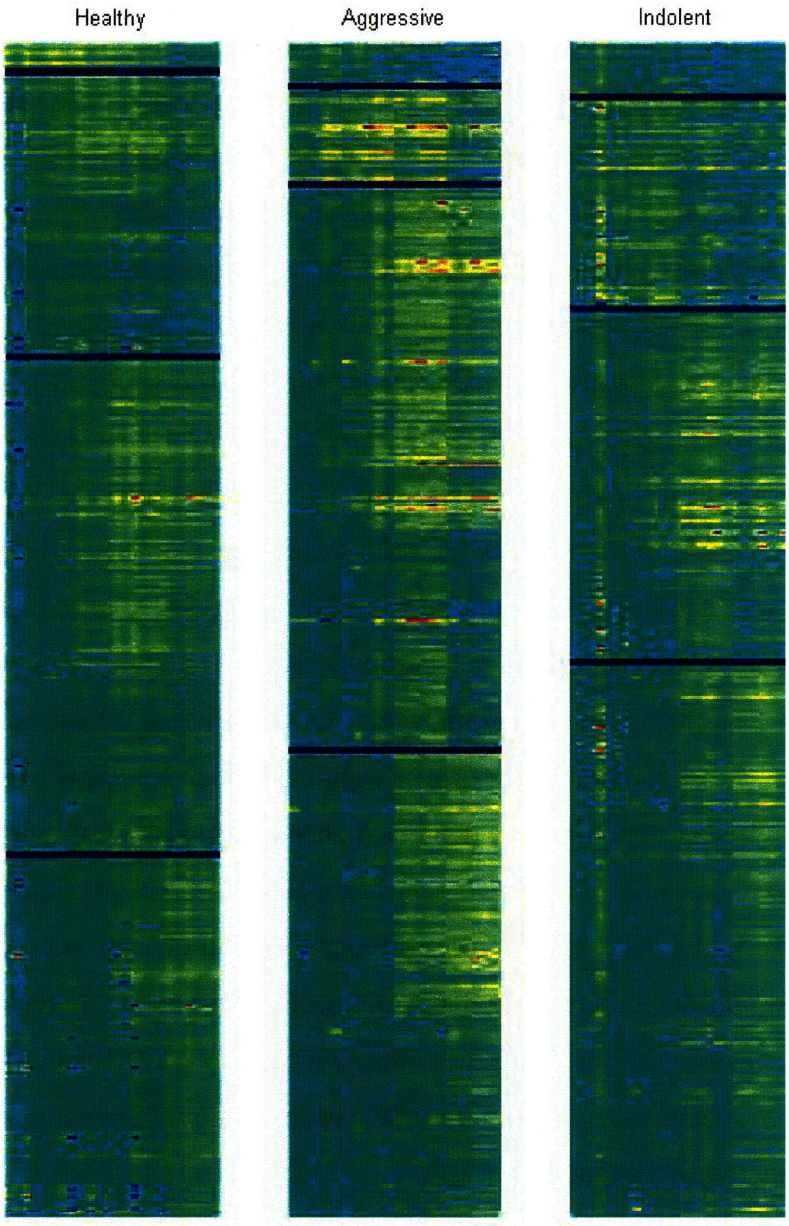



Figure 1-2: Log ratios of stimulated to unstimulated expression data, reordered according to wave clustering results. Only the 500 genes for each subject group with the highest *a posteriori* values for cluster membership are included.

per group, this inference problem is severely ill-posed. Thus we use the cluster labels, which have a consistent biological interpretation, to define a small number of modes of interaction between the genes. We also impose sparsity constraints which effectively limit the number of genes influencing each target gene. Both of these choices are made in order to make inference tractable, but they are also supported biologically.

In order to represent the dynamic networks of temporal interaction statically, we show in Figure 1-3 the progression through waves of expression, which is related to time. Though a single node still represents the expression profile over 4 time points for that gene, we use the cluster labels of the genes, which correspond to the time point at which they peak, to show which genes and which interactions dominate at a given time. From these figures, we make two observations about the networks, which are confirmed quantitatively in Chapter 5. While sparsity constraints imposed during inference of models of interaction limited the number of *incoming* edges to any one gene, the number of *outgoing* edges was not constrained. The networks, however, have a scale-free structure, where a small number of genes have a very large number of outgoing edges. This is consistent with what has been found in protein interaction networks [8], where a small number of proteins serve as hubs for very large numbers of interactions and the others participate in very few interactions. Of the 15 genes we identify as hubs in the three networks, *EGR1*, *NR4A1*, and *ZFP36* are already known to be expressed after BCR stimulation [64, 52]. Those three genes, in addition to the hubs *BTG2*, *CXXC5*, and *NR4A3*, are transcription factors [51]. It confirms the expectation that transcription factors would be expected to act as hubs in the network. The second observation is that "red" edges (those between adjacent waves) appear to dominate the network, which is consistent with the assumption that genes are most likely to affect genes in the immediately following wave, even though this was also not an explicit constraint.

In addition to visualizing the networks, we perform analysis of network structure, leave-one-out cross validation, and tests of robustness and statistical significance. These show that the modeling choices we have made are not arbitrary but successfully capture the relationships between genes. An obvious next step after the networks have

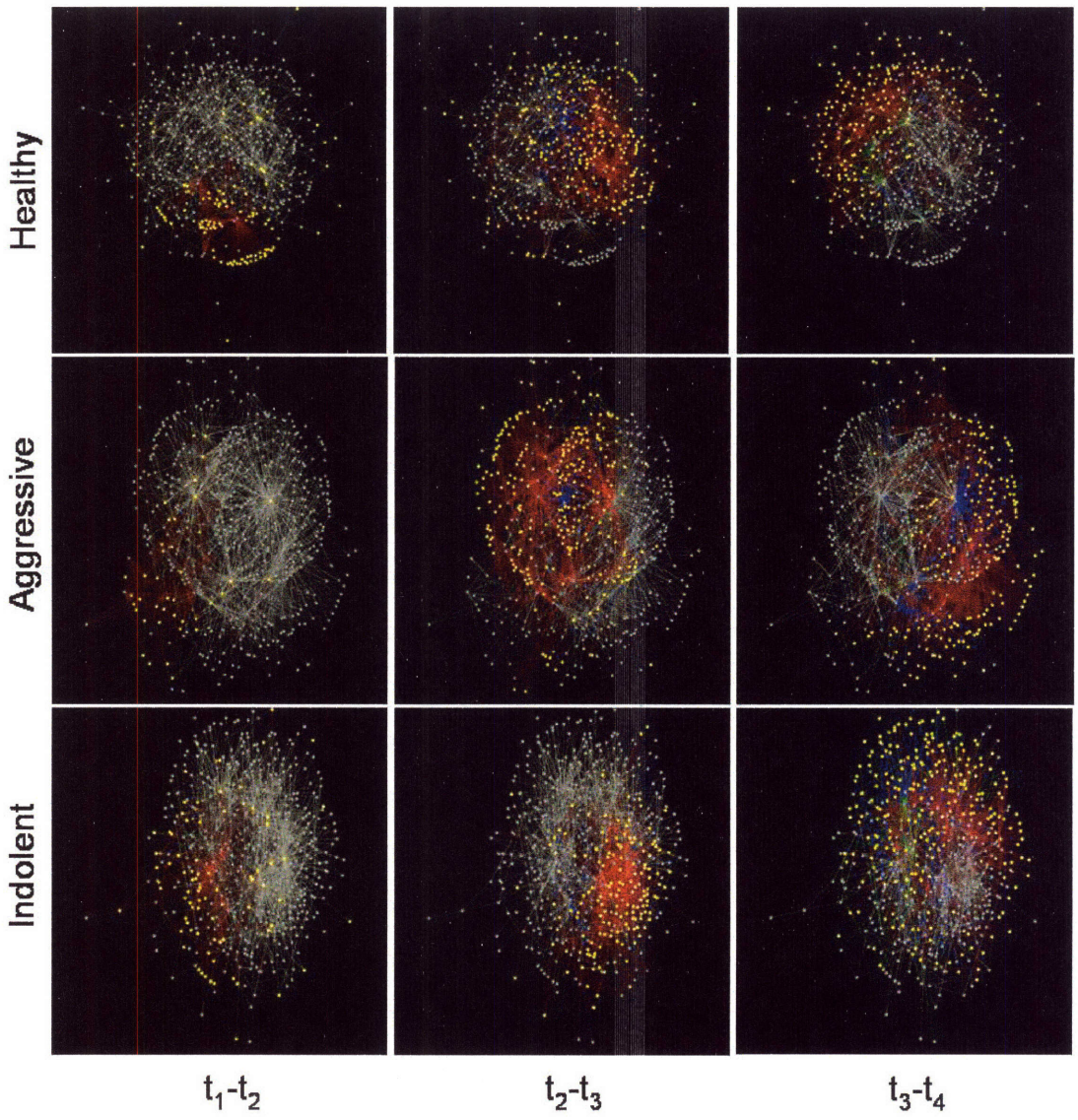



Figure 1-3: In the leftmost column, wave 2 genes are highlighted in yellow. They are influenced by wave 1 genes only. These connections are shown in red. In the center column, wave 3 genes are highlighted, and inputs that are direct (those from wave 2, which is only one time step removed) are shown in red, while inputs that are from wave 1 to wave 3 are shown in blue. Finally, in the rightmost column, wave 4 genes are highlighted, with direct connections from wave 3 genes in red, inputs from wave 2 in blue, and inputs from wave 1 in green.

been inferred and validated mathematically is testing the predictions biologically. One way to do this is to manipulate the expression levels of a single gene, either by silencing or overexpression, in an intervention experiment. This is the only way to distinguish between relationships of causality versus correlation in the local regions of the network. Because intervention experiments are expensive and time-consuming to perform, particularly for time series acquisitions, much thought must go into selecting the target gene. It should peak early in the BCR activation response, so that it will have downstream effects in the cascade of reactions that can be measured in the time course of the experiment. It also should be expected to influence many other genes, so model predictions can be compared to actual measurements. Identification of genes with these two properties are precisely what wave cluster labeling and the inference of the predictive models provide. First wave genes are, by definition, those that peak early in the response. By propagating the effects of a change in a first wave gene to all the remaining genes in the network via direct and indirect connections, we identify those genes that appear to play important roles. In Figure 1-4, we show an example of this type of sensitivity analysis for all the first wave genes in the Aggressive network.

We see qualitatively that a few of the first wave genes have the most prominent effects on the genes in the network. Over the three subject groups, four genes (*FOS, IER2, JUN, EGR1*) identified as potential target genes were found previously to be expressed by inducing early response genes with BCR activation [52]. For the Aggressive network, column $i$ of Figure 1-4 corresponds to the probe set label for *DUSP1*, which was selected for the first set of intervention experiments on the basis of its expression profile and its role in cell cycle regulation [45] and apoptosis [53]. We examine the consequences of silencing of *DUSP1* using small interfering RNA after BCR activation in a patient with the aggressive form of B-CLL. In the Aggressive network, *DUSP1* is directly connected to 37 genes and indirectly connected to an additional 136. This subnetwork is shown in Figure 1-5, where *DUSP1* is white, second wave genes are red, third wave genes are blue, and fourth wave genes are green.

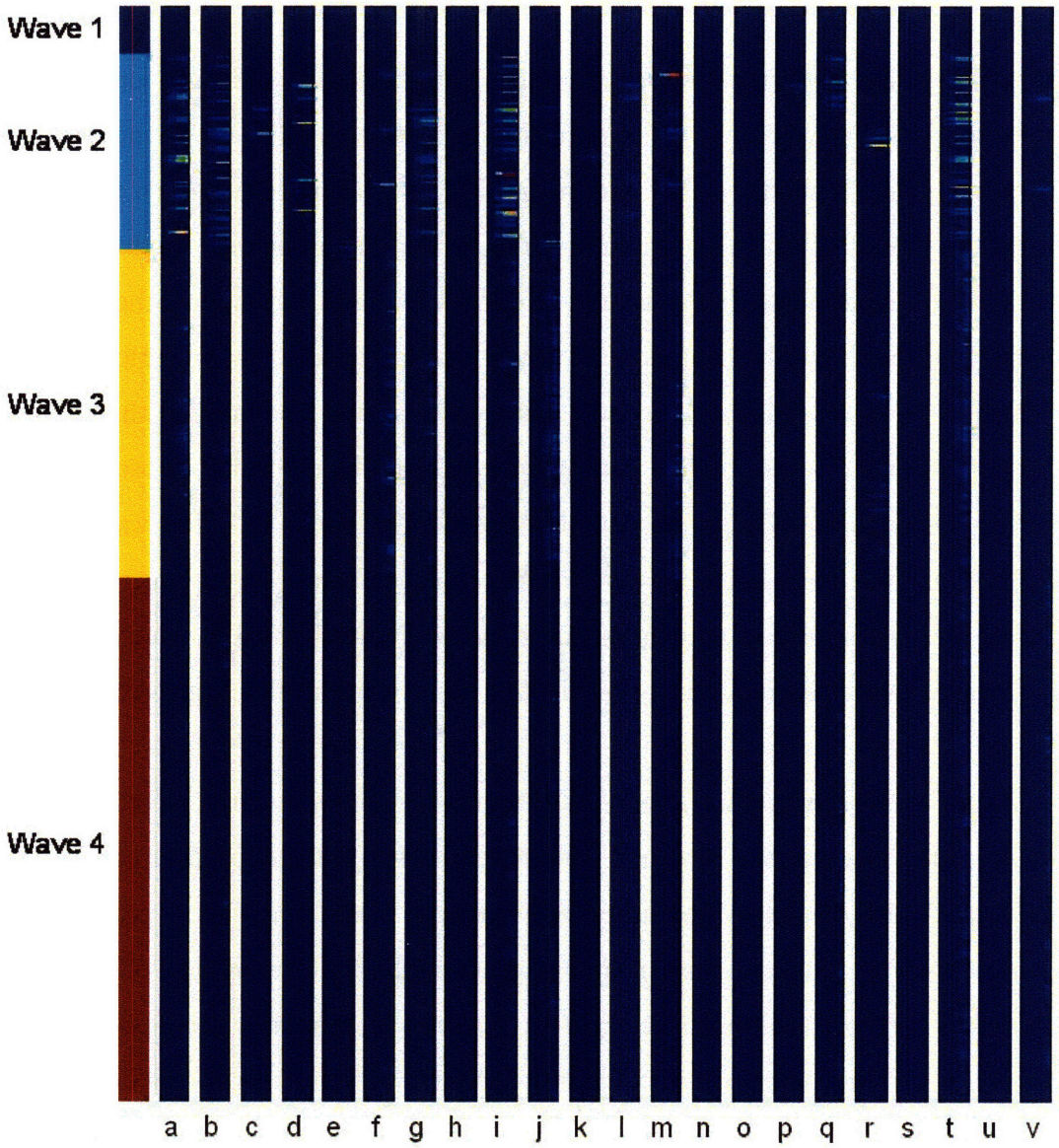



Figure 1-4: For the Aggressive network, the effects of uniformly perturbing each of the first wave genes on all other genes in the network over time. The leftmost column corresponds to the wave label.

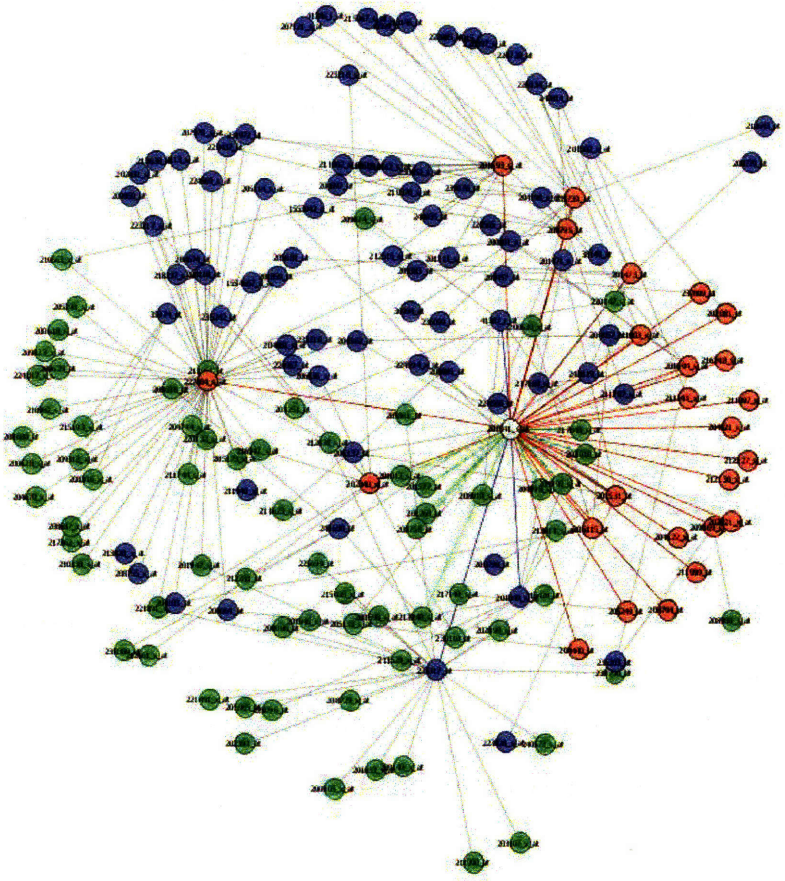


Figure 1-5: *DUSP1* subnetwork for the aggressive patient group. *DUSP1* is white, second wave genes are red, third wave genes are blue, and fourth wave genes are green.

To evaluate the predictive ability of the network, we compute scores for each gene at each time point based on whether the network is able to correctly predict the direction of change (increase or decrease) from the control experiment to the *DUSP1* silencing. For genes in the *DUSP1* subnetwork, the model predictions are correct nearly 70% of the time. Those genes that are unrelated to *DUSP1* should, in theory, have the same expression levels in both the silencing and the control experiments. Any differences that are seen in the data would be due to noise, which the model should be unable to predict. The direction of change for these genes is correctly predicted only approximately 40% of the time. This strengthens the results with respect to those genes that are related to *DUSP1*, as the model is uniquely able to predict changes due to gene silencing. In summary, for both the clustering and the inference of the predictive models, we have statistical results that show that we capture the temporal structure of and the interactions between the genes relevant to the BCR signaling pathway. We have confirmatory results from a biological standpoint, in which genes that we have identified as playing key roles in the networks have already been shown in previous work to be relevant to BCR stimulation, but we also have results that guide future experiments in the study of other related genes, in order to further the long term goal of a full understanding of how and why B-CLL cells behave abnormally. In this chapter, we have motivated the problem of inferring temporal gene interactions, introduced our approach, and shown a few sample results. We now provide a roadmap for the remainder of the thesis.

In Chapter 2, we begin with a basic overview of the background necessary to understand the biological context of the work in this thesis. We briefly describe how genes are expressed in the cell, and how these expression levels are measured with microarray technology. We also include a short introduction to B-CLL and its relationship to the BCR signaling pathway. Finally, we summarize the experimental procedure used by Vallat et al. [78] to acquire the microarray data set, including the selection of subjects from both healthy and B-CLL populations and the selection of time points that capture the behavior of genes differentially expressed under BCR stimulation.

In Chapter 3, we present a clustering technique that is able to incorporate prior biological knowledge and multiple subjects in an Expectation-Maximization framework. These results are compared to a naïve approach in terms of statistical significance, stability, and cluster structure. We show that a small number of representative temporal expression profiles, corresponding to waves of expression that peak at each time point, are able to successfully capture the underlying structure of the data. Additionally, for our application, clustering serves as a first step in order to make predictive modeling, as described in Chapter 5, tractable.

We then extend in Chapter 4 the EM formulation from Chapter 3 into a hierarchy comprised of two layers: an underlying distribution of subjects, which each have an underlying distribution of genes. This is analogous to a common challenge in microarray data analysis, referred to as biclustering, where the genes and the experimental conditions (in this case, healthy subjects and B-CLL patients) are simultaneously clustered in order to identify subsets of genes that behave similarly under a subset of the conditions. We apply the algorithm to a synthetic data set to show that it can, in fact, identify clusters embedded in the data, and then to our actual data set to evaluate the results of simultaneously clustering subjects and genes.

In Chapter 5, we model the interactions between the genes relevant to the BCR signaling pathway by inferring disease-dependent temporal interaction networks. Because of the number of subjects that we have available, inferring a network of all possible interactions is ill-posed. We therefore impose sparsity constraints that are consistent with biological expectations. We take advantage of the results of the clustering in Chapter 3 in order to infer networks from only those genes that were clustered as waves of expression, meaning that we reduced the number of genes from more than 50,000 to 500 per subject group. We also use these cluster labels to limit the modes of interaction possible between genes. Finally, we impose sparsity in terms of the number of genes that can influence any one output gene. In addition to network visualization, we provide quantitative analysis of statistical significance, robustness to noise, and cross validation.

In Chapter 6, we interpret the results from Chapters 3 and 5 from a biological

standpoint and provide a way to use the networks to design future experiments. We discuss how the structure of the networks corresponds to what is already believed about networks of protein interaction, despite the fact that these were not constraints that were imposed. We describe genes that play key roles in the network, and show experimental results on the *DUSP1* silencing experiment which confirms the ability of our models to predict changes in genes in the *DUSP1* subnetwork.

We summarize the work of the thesis in Chapter 7 and briefly discuss the broader implications of combining machine learning and large-scale biological experiments to better understand the inner workings of cells and disease processes.

# Chapter 2

# Biology

## 2.1 Background

In order to understand the clinical motivation of the work in this thesis, we provide a brief introduction to the biological context of the data set and the disease. We describe how microarrays measure gene expression levels and how gene expression levels relate to cellular function. We also discuss what is known currently about B cell chronic lymphocytic leukemia (B-CLL), the genomic differences between patient subgroups, and the relationship to the B cell receptor (BCR) signaling pathway.

### 2.1.1 Microarrays

With the advent of DNA microarray technology, the goal of a global understanding of the inner workings of transcriptional regulatory networks is becoming feasible. Instead of performing experiments on single genes, it is now possible to measure expression levels from tens of thousands of genes in parallel. The large increase in scale, however, raises new challenges and new opportunities. Where it was previously possible to analyze the experimental results virtually by hand, microarrays provide an overwhelming amount of information. The question is how to interpret the data in a meaningful, principled, and useful way.

Treating microarray data as a large collection of traditional experiments that sim-

ply need to be processed in an automated way ignores one of the key advantages. Global measurements allow us to search out patterns of expression and relationships between genes. By combining confirmatory results, which simply show that microarray analysis is consistent with previous biological experiments, and new results, we begin to form a more complete picture of what is occurring in the cell. We discuss recent work in microarray analysis in the Background sections of Chapters 3, 4, and 5, but first, a basic understanding of how gene expression works is necessary to understand how microarrays are used to measure the behavior of the cell.

### Central dogma of molecular biology

For a brief, oversimplified view of gene expression, we summarize the central dogma of molecular biology [18]. DNA provides the genetic code for the cell, and transcription is part of the method in which the genetic code is used to produce proteins. Proteins have many important functions in the cell, such as catalyzing chemical reactions, playing structural or mechanical roles, influencing immune response, and storing and transporting various ligands. In transcription, a DNA sequence is copied to produce a sequence of ribonucleotide bases that is referred to as messenger RNA (mRNA) and carries genetic information to the ribosomes. The ribonucleotides are read by translational machinery in sequences of triplets (codons) that specify amino acids, which then make up the protein. Translation of mRNA to protein is followed by folding, post-translational modification, and targeting.

While DNA microarrays enable measurement of the transcripts (or mRNA sequences) of every gene, we cannot measure protein expression levels or biochemical activity of gene products. These are, generally, what are actually physically interacting with one another and functioning in the cell. A comprehensive picture of the transcriptional regulatory network would require direct measurements. There is, however, a sensible link between the expression pattern and function of the gene product [9]. One reason we are able to infer chemical activity of the gene products is that they may influence the transcription of other genes. Regulatory proteins are classified either as activators, which increase the rate of transcription, or repressors, which

decrease it. Also, assuming that cells behave efficiently, transcription of a DNA sequence occurs when the corresponding protein is necessary for the cell to function. As more or less of the protein is needed, the amount of mRNA transcribed is adjusted as well. Thus, for our purposes, we use mRNA expression levels as an indirect measure of the behavior of the cell.

### Measuring mRNA expression levels

For the data collected for this work, Affymetrix oligonucleotide arrays are used, but other microarrays work along similar principles. We summarize the acquisition and processing described by Lockhart et al.[46]. Arrays of thousands of discrete oligonucleotide sequences are synthesized *in situ* on a slide. Each oligonucleotide relates to a DNA sequence specific to a particular gene in the human genome. These are referred to as probes because they serve to probe, or interrogate, the composition of the population of RNA sequences.

The arrays contain collections of pairs of probes for each of the RNAs to be monitored. Each probe pair consists of a short sequence that is perfectly complementary (referred to as a perfect match, or PM probe) to a particular subsequence, and a corresponding sequence that is identical except for a single base difference in a central position. The mismatch (MM) probe of each pair serves as an internal control for hybridization specificity. The analysis of PM/MM pairs allows low-intensity hybridization patterns from rare RNAs to be sensitively and accurately recognized in the presence of cross-hybridization signals. The target RNA population, which is comprised of the samples we want to measure, is prepared by incorporating fluorescently labeled ribonucleotides in an *in vitro* transcription reaction and then randomly fragmenting the RNA to an average size of 50 to 100 bases. After the samples are hybridized to arrays, fluorescence imaging of the arrays is accomplished with a specially designed scanning confocal microscope. The resulting images are processed to compute expression levels based on the amount of fluorescence at each grid cell in the array.

### 2.1.2 B-CLL

CLL represents 30% of adult leukemias, with the median survival varying from 2.5 to 14 years depending on the clinical stage. This form of leukemia, despite advances in treatment, is still incurable. It is characterized by the accumulation of CD5+ monoclonal B lymphocytes of mature cytological aspect, but the mechanism of the disease remains unknown. The mutational status of the VH genes encoding for immunoglobulin (Ig), constitutive of the B cell receptor (see Section 2.1.3), differentiates mutated (M-CLL) or unmutated CLL (UM-CLL) and is the major prognosis factor of this leukemia. There are two types of CLL: one with a better outcome, in which mutated VH genes are expressed, and another with a poor outcome, in which unmutated VH genes are expressed [20, 54].

The recent technological advances in gene expression analysis by microarrays allow the description of many molecular differences and similarities between malignant and healthy cells. Those analyses revealed for all the subgroups of CLL a distinguishable gene expression pattern for healthy B cells and other lymphoid malignancies, confirming that CLL is a unique entity with heterogeneous aspects [33]. Nevertheless, some genes are differentially expressed with respect to the mutational status that discriminates UM-CLL with a poor outcome [42, 59]. Among them, *ZAP-70* seems to enhance the signaling of the BCR in the more aggressive form of CLL cells [14, 13]. We hypothesize that downstream abnormality of the BCR signaling pathway conditions a temporal genetic program specific for aggressive leukemia after BCR stimulation. The inference of temporal transcriptional networks should provide insight into the genetic program used by these different categories of cells.

### 2.1.3 BCR stimulation

The key to distinctive B-CLL behavior likely relates to the differential ability of the B cell receptor (BCR) to respond to stimulus. Analysis of signal competence *in vitro* reveals that unmutated CLL generally continues to respond, whereas mutated CLL is anergized [67]. Therefore, in order to better understand the B-CLL disease process,

we investigate the B cell receptor and the reactions that comprise the BCR signaling pathway.

B cells are lymphocytes that play a large role in the humoral immune response. The human body makes millions of different types of B cells each day, and each type has a unique receptor protein referred to as the B cell receptor (BCR) on its membrane that will bind to one particular antigen. The BCR on mature B cells consists of an antigen binding subunit, surface immunoglobulin, and a signaling subunit. Cross-linking (which we refer to more generally here as stimulation) of the BCR initiates a tyrosine kinase cascade. During this process, positive and negative feedback modulates the signal. Activators or repressors, which are mainly transcription factors, arrive in the nucleus, and the output is the transcriptional program that defines the cellular behavior. Regulation of the BCR signaling pathway determines whether BCR engagement leads to activation, anergy or cell death by apoptosis [30, 50, 11], so the goal is to understand the population-specific core transcriptional programs.

From an experimental standpoint, BCR stimulation provides two important advantages over most of the other microarray data that is collected. The first is that by collecting data from stimulated and unstimulated samples simultaneously, we measure differential expression relative to a control. This should, at least to the extent possible, distinguish between those genes that are part of the BCR pathway and those that are expressed in the cell generally. It allows us to narrow our focus, and because it is the BCR pathway that is expected to be most relevant to B-CLL, that focus will be on the most important genes.

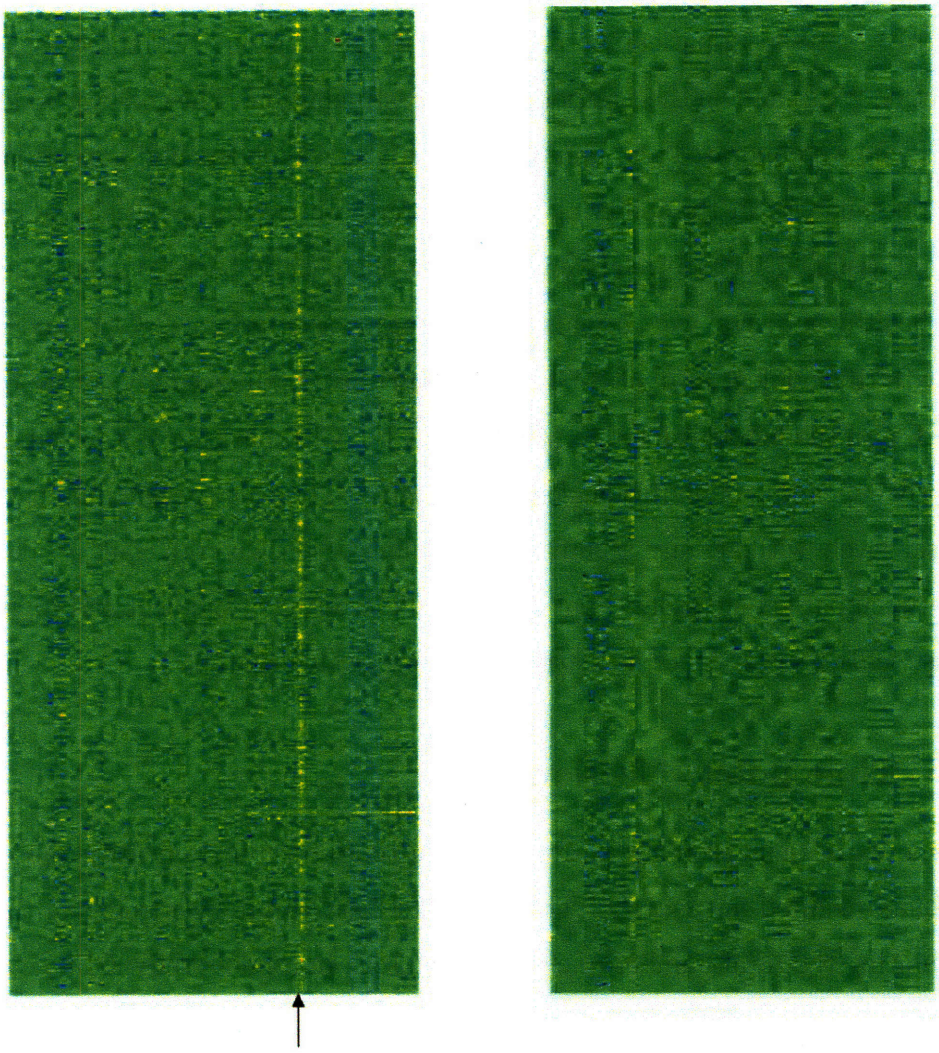
The second is that most gene expression data is collected at steady state, and interaction networks that are inferred represent dynamic systems. Steady state behavior constrains the dynamic behavior of the network, but does not determine it. Therefore, building a dynamic model from steady state data is a severely ill-posed problem [25]. Because cells in the sample are at various stages of the cell cycle, it is relatively easy to determine which genes are co-expressed, but not which genes have temporal influence over one another. In contrast, with BCR stimulation, one can essentially "reset the clock" on the cells, beginning the cascade of reactions in the

BCR signaling pathway in a synchronized way. Then, as data is collected at time points following stimulation, we observe the patterns of transcription that occur.

## 2.2 Data collection

For the data set used in nearly all the analysis in this thesis, details of the experimental procedure are found in Vallat et al. [78]. These are complex and time consuming experiments, but only a few points of particular importance to this work are presented here. Peripheral blood was obtained from twelve previously untreated patients diagnosed with B-CLL, and six healthy blood donors. B cells were divided in two in order to control for any effect of *ex vivo* handling of cells over the time period of the experiment. One set was used as an unstimulated (US) control, and the other was stimulated (S) by an antigen. The target cRNA was prepared in accordance with the Affymetrix protocol and hybridized to the HG-HU133 plus2.0 microarray, which contains 54,675 probe sets. Microarrays for the 18 subjects were analyzed. Data were normalized with the invariant set method and the model based expression index (MBEI) obtained by the pm-mm model using dChip software (dChip) [44].

We calculate differential expression by subtracting the unstimulated expression level from stimulated expression level. For a clearer visualization, the log ratio of stimulated to unstimulated is shown in Figure 2-1(a). Each row of the image corresponds to a given gene. There are four columns, corresponding to time points, for each of the 18 subjects. Neither the genes nor the subjects are in any particular order, and it is difficult to see any kind of structure in the data. Most of the genes are near zero, which means they are not differentially expressed under BCR stimulation. For one patient, the expression at the third time point was uniformly higher than the result of the data. In Figure 2-1(a), this is the bright yellow line marked with an arrow. This was either a problem with the experiment or the processing, so this outlier patient is removed from all further analysis. The new data set is displayed in Figure 2-1(b).



(a) 18 subjects

(b) 17 subjects, outlier removed

Figure 2-1: Log ratios of stimulated to unstimulated expression data. Each row of the image corresponds to a given gene. There are four columns, corresponding to time points, for each of the 18 subjects. Most of the genes are near zero (light green), which means they are not differentially expressed under BCR stimulation.

### 2.2.1 Patient selection

The role of BCR stimulation in apoptosis versus survival of CLL cells was evaluated in 23 patients and compared to B cells from 9 healthy donors. For the detailed microarray analysis, 12 patients were selected on the basis of their mutational status (see Section 2.1.2) and on their response to BCR stimulation. The goal was to identify patients who were mutated and unmutated, and also responders and non-responders.

For the analysis in the following four chapters, we use three patient subgroups, which are assumed to have relatively homogeneous genetic backgrounds. These are the six healthy subjects, five patients with the aggressive form of B-CLL, and six patients with the indolent form of B-CLL. The aggressive form generally corresponds to UM-CLL and the indolent form to M-CLL, but the CLL patients can be subdivided differently depending on mutational status, prognosis, or gene-specific differences. At the basal level, without any stimulation, unsupervised clustering fails to discriminate between patient subgroups, but supervised clustering of Indolent and Aggressive samples yields a set of genes do separate the two groups [42, 59].

### 2.2.2 Time point selection

Any gene expression profiling performed over time must first identify the critical time points required for analysis. The goal is to best capture the trends in the data while limiting the total number of chips being used. In a pilot study, CLL cells were isolated and RNA extracted after BCR stimulation at nine time points (baseline, 20, 30, 60, 90, 150, 210, 390 min, and 24 hr). Three patients were selected on the basis of their functional response to BCR stimulation: one low-responder, one median responder and one high responder. Using a self-organizing map (SOM) [72], a type of unsupervised mathematical cluster analysis, eight possible template patterns of gene expression over time were selected. Based upon this analysis, four time points (60, 90, 210 and 390 minutes) were identified that appeared to capture the vast majority of these trends. These time points were therefore selected for all subsequent analyses. Even with only four time points examined and analysis of a relatively small number

of subjects, this experiment required analysis of 144 arrays. While it is possible that important events in regulating BCR mediated signaling pathways are missed because of restricting analysis to these time points, it is a necessary practical choice.

# Chapter 3

# Clustering

In general, clustering provides a method for analyzing the results of microarray experiments, which are a substantial increase in experimental scale as compared to previously available technology. While tens of thousands of measurements can now be acquired simultaneously, the data is virtually useless until it can be summarized in a meaningful way. Gene-based clustering groups together genes that are coexpressed under various experimental conditions, and such coexpression is believed to indicate cofunction and coregulation. Results of these clustering algorithms are characterized by the existence of stable clusters, the shapes of the cluster expression profiles, and the actual groupings of the genes.

The goal of clustering in this work is to identify and label genes that are differentially expressed under BCR stimulation. From an experimental standpoint, BCR stimulation of cells from multiple subject groups provides three important advantages for clustering over much of the gene expression data that is collected. The first is that we measure differential expression due to BCR stimulation relative to a control. This should, at least to the extent possible, distinguish those genes affected by the BCR pathway from those that are expressed in the cell generally. This focuses the clustering problem to coexpression within the pathway. Second, BCR stimulation essentially "resets the clock" on the cells, beginning the cascade of reactions in the BCR signaling pathway in a synchronized way. Therefore, we cluster based on temporal expression profiles instead of steady state response. Finally, because B-CLL is

believed to be related to genetic differences in BCR signal competence (see Section 2.1.3), having subjects from healthy and B-CLL populations allows us to combine subject information into subpopulation-specific cluster labelings. Data from different patients and different time points are generally treated simply as more experimental conditions, so our approach is unique.

Our clustering provides the shape of the cluster profiles, which provides insight into the structure of BCR response of the cell. We also obtain classifications for all of the genes, which identifies those that appear to be relevant to BCR (and at what time point they peak), those that are grouped together, and how the labelings change depending on the subject group (see Section 6.5). For validation, there is no ground truth of an exhaustive list of genes that are differentially active, nor which should be clustered together, and for more than 50,000 genes, these individual experiments are simply not practical and may not be generalizable to multiple subjects or different experimental conditions. Thus we have to use other measures to determine the success of the clustering, such as significance, robustness to noise and initialization, and quantification of how well the data is predicted by the models.

Additionally, for our application, clustering serves as a first step in order to make predictive modeling, as described in Chapter 5, tractable. This requires drastically reducing the number of genes. There is no way to infer all relationships between 50,000 genes at 4 time points with only 17 patients. For linear models, this would require inference of $N^2T^2$ parameters from $N \times T \times P$ data points. Even as $N$ is reduced by excluding genes not relevant to the BCR pathway ("background" genes), allowing unconstrained interactions between all pairs of genes leads to a problem that is still under-determined. Therefore, we use the cluster labels, which have a consistent biological interpretation (as explained in Section 3.2.2) to define a small number of modes of interaction between the genes. The details of this formulation are presented in Section 5.2.

## 3.1 Background

Jiang et al. [39] provide a review of cluster analysis of gene expression data, presenting specific challenges pertinent to categories of clustering and several representative approaches, as well as various methods to assess the quality and reliability of clustering results. We discuss these challenges and how they relate to our application. The first challenge is that prior knowledge of the structure of the data is usually not available because cluster analysis is often the first step in mining the data. In this work, we actually have some insight into the shapes of the cluster profiles because of what is known about BCR activation, but we also use an approach that requires no prior knowledge as a basis for comparison of the results. While our method may not be as useful for a problem where nothing of the structure of the data is known, it does allow us to obtain more consistent, meaningful clusters than the naïve approach. Second, because of the nature of microarray experiments, gene expression data often contains a large amount of noise. This is a challenge that we cannot avoid and is the reason that we present stability and significance results to show that the clusters that are identified are more than random effects. Third, clusters may overlap one another as genes may play more than one role under different conditions or for different patients. For example, a set of genes may be involved in apoptosis and another set in cell division. Some of these genes may be involved in both. We address this with BCR stimulation, time series acquisitions, and clustering of each subject group independently. Instead of having an expression profile which incorporates every possible condition, we take advantage of the fact that we have labels on subject and stimulation versus control in order to cluster based on disease-specific differential expression profiles over time. Finally, it is difficult to present the results in a way that makes it possible to interpret how clusters compare to one another and how genes within the cluster are related. Graphical representations are suggested, which we have addressed in several ways: cluster profiles, reordering of the original data, and affinity matrices showing how often genes are clustered together. We also provide quantitative results to confirm what is evident visually.

As would be expected, Jiang et al. found no single best algorithm which is the winner in every aspect. There are also no existing standard validity metrics. In fact, the performance of different clustering algorithms and different validation approaches is strongly dependent on both data distribution and application requirements. Therefore, as we made decisions in the design of the clustering algorithm, we investigated several approaches. Clustering techniques such as K-means [23], nonnegative matrix factorization [10], and factor analysis [84] have all been used as first steps in creating networks of genes. Sturn et al. [70] implement software tools for standard, general clustering methods such as hierarchical clustering, K-means, self-organizing maps, principal component analysis, and support vector machines. What none of these methods provide, often because they are the first steps in data mining, is a way to effectively incorporate prior knowledge about the expression profiles. However, Pan et al.[55] use an Expectation-Maximization algorithm to identify differentially expressed genes for model-based clustering. With a similar EM approach, we are able to design models according to the previous biological experiments in a flexible framework.

The issue of how to best model time series data, as opposed to observations under various experimental conditions, is still not addressed. Where time series data is available, it is often comprised of very few data points, especially compared to the number of genes. Time series experiments require multiple arrays and large quantities of biological material, so although the time scale of cellular reactions is on the order of seconds to minutes, acquisition of microarray data at that rate is not feasible. Bar-Joseph et al. [2] present a unified model that uses statistical spline estimation to represent gene time-series expression profiles as continuous curves. This algorithm successfully models cyclical patterns of dynamic genetic behavior for a relatively small number of genes, with a relatively high sampling rate (between 7 and 20 minutes between time points). We have only 4 time points, and even when only two spline segments are used to estimate the curve, this algorithm requires the estimation of five parameters for each gene in order to align the profiles, in addition to the spline parameters for the class [26]. This would overfit our data set. Also, our sampling rate is on the order of hours, and the reactions are triggered by BCR

stimulation, so we do not expect to see the cell-cycle related patterns that appear for the steady-state acquisitions in yeast. Fleury et al.[27] implement a robust method for detecting evolutionary trends of expression of mouse retinal genes acquired at different points over the lifetimes of a population of mice. The temporal structure in that analysis is relatively simplistic (i.e. monotonic increase in expression), whereas the temporal structure we expect to encounter in the BCR pathway will be more complex. Additionally, the time-scale of mouse development is on the order of months, while the cascade of reaction in the BCR pathway is at a much finer scale. The most closely-related approach to the one in this work addresses the problem of short time series data by assigning genes to a pre-defined set of model profiles that capture the potential distinct patterns that can be expected from the experiment [26]. It does not address several key issues that we face: combining multiple subjects from a group, incorporating priors on cluster profile shapes, and the classification of all genes, instead of only those from statistically significant clusters.

## 3.2 Modeling expression profiles

We implement model-based clustering in two ways. The first is a general Expectation-Maximization based clustering algorithm, not designed to take advantage of any of the expected structure of the data. Wave clustering, however, is influenced by previous work in biology, which provides a set of shapes for the expression profiles for relevant genes. Both methods are set up similarly in an EM framework, where the primary difference is in how the distributions at each time point are modeled. Section A.2 of Appendix A provides very simple example of how the EM algorithm can be used for clustering.

Genes are clustered on the basis of their differential expression profiles, which are computed by subtracting the unstimulated expression levels from the stimulated expression levels at each of the four time points. The data set $X$ (containing $N$ genes, $P$ patients, and $T$ time points) is assumed to come from a finite mixture of probability distributions, with each component corresponding to a different cluster.

The goal is to estimate the cluster memberships and the associated parameters $\Theta$, which maximize the likelihood:

$$L(\Theta) = \sum_{m=1}^{M} p(X; m, \Theta) p(m|\Theta), \tag{3.1}$$

$\Theta$ represents the set of parameters associated with *all* $M$ clusters, and $m$ specifies which parameters in $\Theta$ are used for a particular cluster. The mixture proportions for each cluster are $p(m|\Theta)$. For a single subject, the conditional probability for a given gene $\vec{x}_n$ in a given class is defined as:

$$p(\vec{x}_n|m, \Theta) = \prod_{t=1}^{T} p(x_{nt}|m, \Theta) \tag{3.2}$$

The subscripts on $x$ specify the gene $n$ and the time point $t$. The product model implies that the time points are independent, which is reasonable given the low sampling rate. In order to combine data from different subjects within a relatively homogeneous group, we assume a common labeling of the genes that incorporates the expression profiles of each of the subjects in that group. This common labeling means that for every gene, each patient $p$ is considered to be an independent draw from the same model type, which is reasonable because the disease-related genetic backgrounds of the subjects are expected to be consistent within a group. Therefore, we compute the conditional probability for each gene as:

$$p(\vec{x}_n|m, \Theta) = \prod_{p=1}^{P} \prod_{t=1}^{T} p(x_{npt}|m, \Theta) \tag{3.3}$$

In order to estimate the parameters in $\Theta$, we use the Expectation-Maximization algorithm [22], which iterates between Expectation (E) steps and Maximization (M) steps. In the E step, cluster memberships are conditionally estimated from the data with the current estimate of $\Theta$. In the M step, model parameters $\Theta$ are computed so as to maximize the likelihood of complete data given the estimated cluster memberships. When the EM algorithm converges, each gene is assigned to the cluster with the maximum *a posteriori* probability. For details of the EM derivation, see Appendix A.

### 3.2.1 Gaussian clustering

We consider a naïve approach to clustering in order to investigate the underlying structure of the data and to have a basis for comparison with the wave clustering. We model the clusters as 4D Gaussian distributions, with each cluster having an associated mean $\mu$ and variance $\sigma^2$ in each of the dimensions, which here correspond to time points. To compute the conditional probability assuming Gaussian distributions, we define:

$$P(x_{npt}|m,\Theta) = \frac{1}{\sqrt{2\pi\sigma_{mt}^2}} \exp\left(\frac{-(x_{npt}-\mu_{mt})^2}{2\sigma_{mt}^2}\right), \tag{3.4}$$

where $m$ specifies which of the parameters in $\Theta$ to use for that particular cluster.

#### EM iterations

On each iteration $k$, the current values of all the cluster parameters are $\Theta^k$. For the Expectation step, the posterior probability of a cluster, given a gene and the current cluster parameters, is:

$$P(m|\vec{x}_n,\Theta^k) = \frac{P(\vec{x}_n|m,\Theta^k)P(m|\Theta^k)}{\sum_{m'=1}^{M} P(\vec{x}_n|m,\Theta^k)P(m|\Theta^k)} \tag{3.5}$$

$$= w_{nm} \tag{3.6}$$

These are weights, $w_{nm}$, which are used in the Maximization step. Given these weights at iteration $k$, we derive a closed form solution for the best estimate of the cluster parameters for iteration $k+1$. The derivation for a simple example is in Appendix A. Solving for $\mu_{mt}$:

$$\mu_{mt} = \frac{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}x_{npt}}{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}} \tag{3.7}$$

This is equivalent to taking a weighted average of the data, based on cluster memberships, to compute the cluster means. In a similar way, the variance for a cluster is :

$$\sigma_{mt}^2 = \frac{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}(x_{npt}-\mu_{mt})^2}{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}}, \tag{3.8}$$

using the value for $\mu_{mt}$ just computed. Finally, the mixture proportions for each cluster are:

$$P(m|\Theta^k) = \frac{1}{N}\sum_{n=1}^{N} w_{nm} \tag{3.9}$$

**Initialization**

Because the EM algorithm is only guaranteed to converge to a local maximum, the result is sensitive to the initialization. As this method is intentionally as general as possible, we initialize by randomly assigning the genes to $M$ clusters. The initial cluster parameters are the Maximum Likelihood estimates of the means and variances given those assignments. The mixture proportions are set to $\frac{1}{M}$.

### 3.2.2 Wave clustering

From a biological standpoint, it has been conjectured that important genes (e.g. transcription factors) will peak at localized time points [78]. As was explained in Section 2.2.2, time series experiments were designed to show waves of expression. A pilot study was done to determine the time points at which data should be collected in order to capture the maximum amount of information while limiting the total number of chips. Significant time templates of BCR stimulated effects identified four time points: 60, 90, 210 and 390 min that appeared to capture the vast majority of these trends.

We develop simple statistical models that capture such behavior, resulting in clusters with a biological interpretation. Given the templates of the differential expression profiles from the pilot study, we intend to model genes that peak at localized time points, i.e. those that have large positive deflections at $t_1$,$t_2$,$t_3$, or $t_4$. Those that are not differentially expressed under BCR stimulation in that way are classified as the "background" genes. The EM formulation is extremely flexible as to incorporating different types of models (exponential distributions, gamma distributions, etc.) depending on the actual structure of the data. It is also possible to model large negative deflections as well as large positive deflections, as well as genes that peak at more

than one time point. These particular models were chosen based on the template shapes in the previously discussed pilot study and work well for this particular data set, but are not the only ones possible for this clustering method. It is also a simple matter to add additional clusters or re-adjust the cluster definitions to incorporate different numbers of time points.

This type of modeling provides a basis for constraining the predictive model inference step (see Chapter 5). The cluster labels first define the subset of genes that comprise the models. Because only those genes that are differentially expressed under BCR stimulation are expected to be relevant to the BCR pathway, all background genes are excluded from further analysis. Selection of the relevant subset is necessary because inferring predictive models of more than 50,000 genes is not tractable. Additionally, it is impractical to allow every possible pair of genes to interact in a unique way, so the cluster labels are used to compute common modes of interaction between classes of genes.

We define 5 classes of genes. Classes 1 through 4 are composed of the genes that are predominantly differentially expressed at the corresponding time point and will be referred to as waves of transcription, e.g. wave 1 genes have high expression at the first time point and relatively lower expression at the other three time points. Expression is modeled as a one-sided Laplacian distribution at the time point corresponding to the genes class label and as a two-sided Laplacian distribution elsewhere. Having heavier tails than a Gaussian, these distributions better model the considerable variation we see in genes that are differentially expressed under BCR stimulation.

As in the previous section, we assume the time points are independent, and for every gene, each subject in a group provides an independent draw from the same gene model. The probability for a given gene $\vec{x}_n$ in a given class is defined by:

$$p(\vec{x}_n|m,\Theta) = \prod_{p=1}^{P}\prod_{t=1}^{T} p(x_{npt}|m,\Theta) \tag{3.10}$$

Depending on the class, the formulation of $p(x_{npt}|m,\Theta)$ changes. For waves 1 through 4, the "dominant" time point is defined as being strictly positive, which corresponds to an increase in gene expression at that time point due to the initial stimulation.

When $m$ (the class label) corresponds to $t$ (the time point), the equations are:

$$p(x_{npt}|m,\Theta) = \lambda_{m*} \exp\left(-\lambda_{m*} x_{npt}\right) \text{ for } x_{npt} > 0 \tag{3.11}$$

$$p(x_{npt}|m,\Theta) = 0 \qquad \text{for } x_{npt} \leq 0 \tag{3.12}$$

At the other time points, because we expect negative deflections or small positive deflections, two-sided Laplacians with different positive and negative deflections are used:

$$p(x_{npt}|m,\Theta) = \frac{\lambda_{mt+}}{2} \exp\left(-\lambda_{mt+} x_{npt}\right) \quad \text{for } x_{npt} > 0 \tag{3.13}$$

$$p(x_{npt}|m,\Theta) = \frac{\lambda_{t-}}{2} \exp\left(\lambda_{t-} x_{npt}\right) \qquad \text{for } x_{npt} \leq 0 \tag{3.14}$$

The parameters for the distributions at the non-dominant time points are $\lambda_{mt+}$, which are unique to each model and are expected to be larger (thus modeling smaller deflections) than $\lambda_{m*}$, and $\lambda_{t-}$. The parameters for the negative deflections, $\lambda_{t-}$, are shared across the four wave models. Because these parameters are shared, the negative deflections will be equally likely under each of the models and will not have an impact on the assignment of a gene to a particular class. Given the expression profile templates in the pilot study described previously and the fact that we select genes that are associated with BCR stimulation on the basis of increased response, we focus on positive deflections. This is a modeling choice, which for other applications could be easily changed.

The final class is the background, which is composed of genes that are not differentially activated with BCR stimulation and are modeled by a zero-mean Gaussian distribution at each of the four time points.

$$P(x_{npt}|m,\Theta) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{-x_{npt}^2}{2\sigma_t^2}\right) \tag{3.15}$$

We use the EM Algorithm again to estimate the parameters of the distributions and the class memberships of each gene.

### EM iterations

In exactly the same way as the Gaussian clustering method, the Expectation step is:

$$P(m|\vec{x}_n, \Theta^k) = \frac{P(\vec{x}_n|m, \Theta^k)P(m|\Theta^k)}{\sum_{m'=1}^{M} P(\vec{x}_n|m, \Theta^k)P(m|\Theta^k)} \tag{3.16}$$

$$= w_{nm} \tag{3.17}$$

In the Maximization step, given these weights that were computed, we again have closed form solutions for the parameter values. Beginning with the background class, the variance at each time point is:

$$\sigma_t^2 = \frac{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}x_{npt}^2}{\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nm}} \tag{3.18}$$

The remainder of the parameters are for the wave classes. For $\lambda_{m*}$, the peak time point, when $m = t$:

$$\lambda_{m*} = \frac{\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nm}x_{npt}}{\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nm}} \tag{3.19}$$

Instead of summing over all $N$ genes, only those with $x_{npt} > 0$ are included. Currently, we impose a hard constraint on $\lambda_{m*}$, which guarantees large positive deflections at the cluster's peak time point. For the non-dominant time points, when $m \neq t$:

$$\lambda_{mt+} = \frac{\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nm}x_{npt}}{\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nm}} \tag{3.20}$$

For the shared parameter $\lambda_{t-}$:

$$\lambda_{t-} = \frac{\sum_{m=1}^{M-1}\sum_{p=1}^{P}\sum_{n\in x_{npt}\leq 0} w_{nm}|x_{npt}|}{\sum_{m=1}^{M-1}\sum_{p=1}^{P}\sum_{n\in x_{npt}\leq 0} w_{nm}} \tag{3.21}$$

The summations are over the four wave classes (from $m = 1$ to $m = M - 1$) and the summation over the genes is for all $x_{npt} \leq 0$. The formula for the mixture proportions is also the same as it was previously:

$$P(m|\Theta^k) = \frac{1}{N}\sum_{n=1}^{N} w_{nm} \tag{3.22}$$

Because the mixture proportion of the background class is expected to be much larger than that of the other classes, one could add a prior here. However, because of the initialization (described in the following section), and the constraint on $\lambda_{m*}$, we have found that $p(m|\Theta^k)$ converges quickly without a prior, as is shown in Figure 3-1.

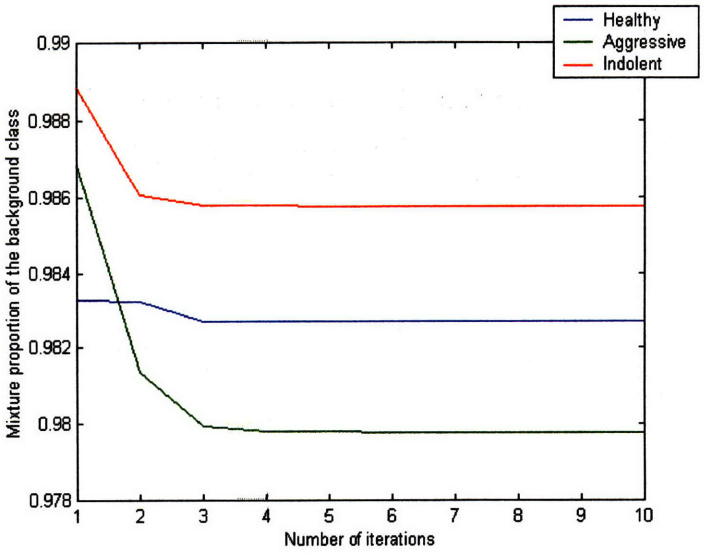



Figure 3-1: Mixture proportion for the background class as the EM algorithm converges

### Initialization

Instead of randomly assigning initial class labels, we initialize the parameters based on the types of genes we expect to see in each class. We compute the ML-estimates of each $\lambda_{mt+}$ from all the genes at that particular time point.

$$\lambda_{mt+} = \frac{NP}{\sum_{p=1}^{P}\sum_{n=1}^{N}|x_{npt}|} \tag{3.23}$$

The $\lambda_{m*}$ are proportionally smaller (about $0.01\lambda_{mt+}$) because smaller values of $\lambda$ allow for larger deflections, which will correspond to genes that peak at that time point. $\lambda_{t-}$ is set initially equal to the $\lambda_{m*}$ at that time point, in order to allow for large negative deflections at the non-dominant time points. In Figure 3-2, we show the probability distribution functions for this initialization for the healthy subjects. The flatter the distribution (smaller $\lambda$), the more likely large deflections. For the initial mixture proportions:

$$p(m|\Theta^0) = \epsilon \qquad \text{for } m < M \tag{3.24}$$

$$p(m|\Theta^0) = 1 - (M-1) \times \epsilon \qquad \text{for } m = M, \tag{3.25}$$

where $\epsilon$ is the approximate fraction of genes in each wave class.

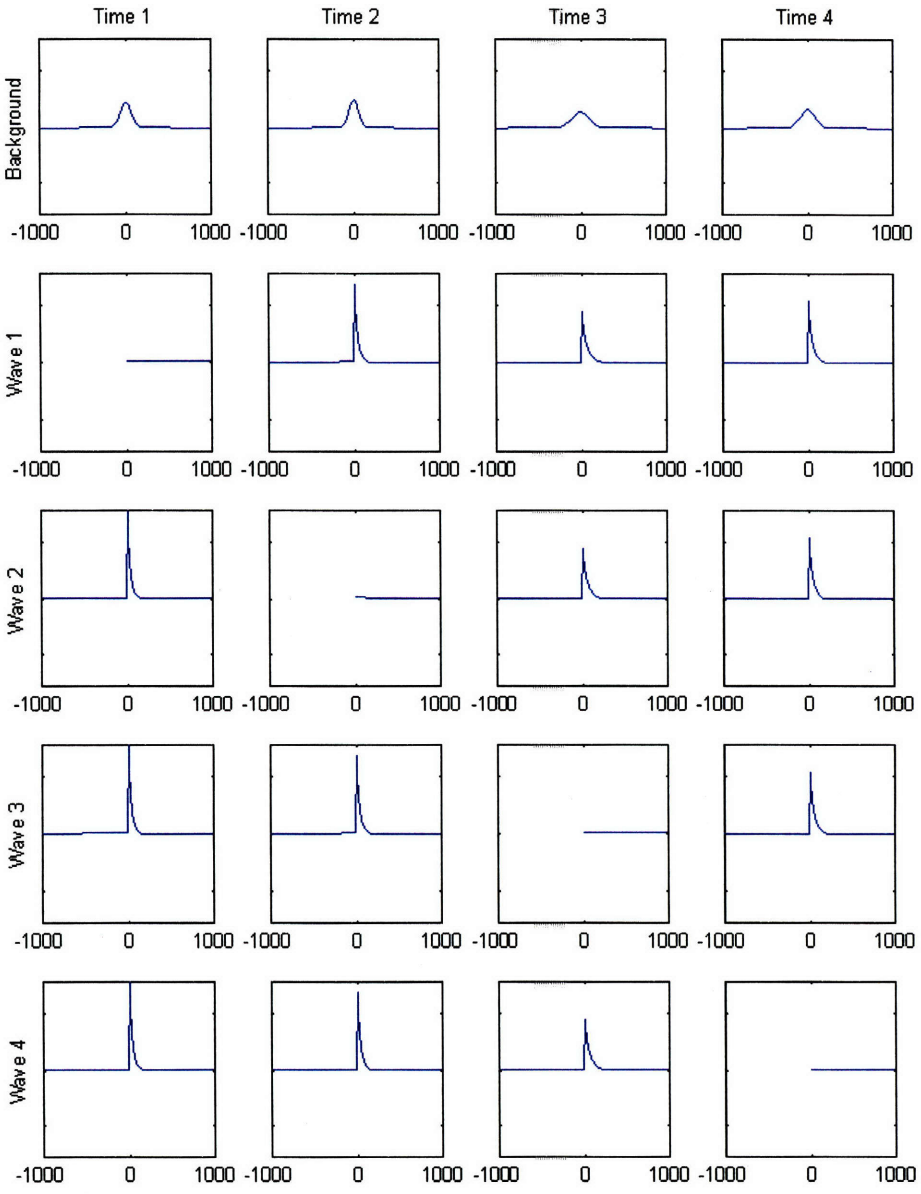



Figure 3-2: Probability distribution functions for the background class and four wave classes at each of the four time points, given the parameter initialization (healthy subjects)

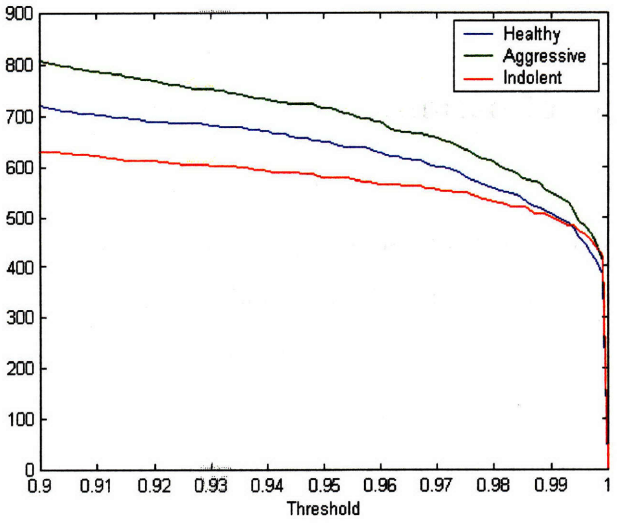



Figure 3-3: Number of genes selected as a function of MAP value threshold

**Threshold selection**

After the EM algorithm converges, each gene is assigned to the class according to its maximum *a posterior* (MAP) probability under the model. We choose to threshold based on this MAP value in order to include the top 500 genes from each subject group. There are several reasons for this. In order to compare results across groups, we work with a consistent number of genes per group. It also allows us to further pare down the number of genes for visualization and predictive model inference in a principled way. In Figure 3-3, we see that by varying the threshold, the number of genes that would be included changes. As it becomes more practical to work with a larger number of genes (or if one were working with a different data set altogether), the threshold can be adjusted.

## 3.3 Model comparison

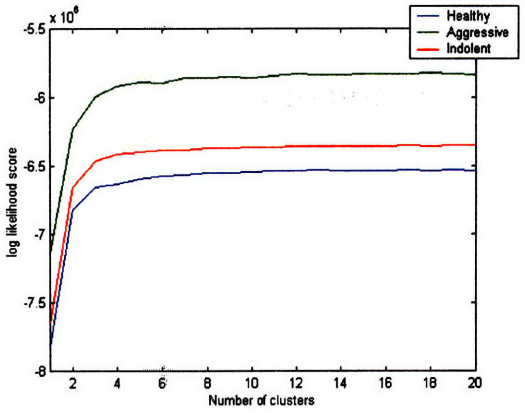
We first evaluate the quality of the clustering result quantitatively. In this section, we compare the number of parameters for the models, cluster labelings, shapes of the cluster profiles, likelihood scores, statistical significance, and stability. The biological

interpretation of the results is presented in Chapter 6.

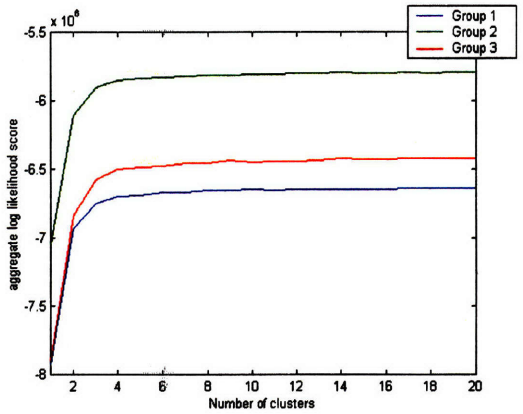
### 3.3.1 Numbers of parameters

For the Gaussian clustering, there are $2 \times T$ parameters associated with each of the $M$ clusters: a mean and variance corresponding to each time point. In Figure 3-4(a), we vary the number of clusters, $M$, from 1 to 20, and record the log likelihood score of the clustering. The scores of the three subject groups are separated consistently. While this would appear to show a difference between the groups, it is an effect of using the same set of patients over the trials. A random re-grouping of the subjects, shown Figure 3-4(b) yields the same type of separation. Also, the score for the Aggressive group is higher than that of the other two groups, but this is because it has only 5 patients, as compared to 6 in each of the other two groups. After we normalize over the number of patients, time points, and genes, it actually has the lowest score in Figure 3-4(c). The only conclusion to be drawn from these figures is the trend we see in how the score changes with increasing $M$. It converges quickly, and at approximately $M = 4$ or $M = 5$, there is little improvement in the score. All the further analysis of the Gaussian clustering will use $M = 5$, in which 40 parameters are estimated.

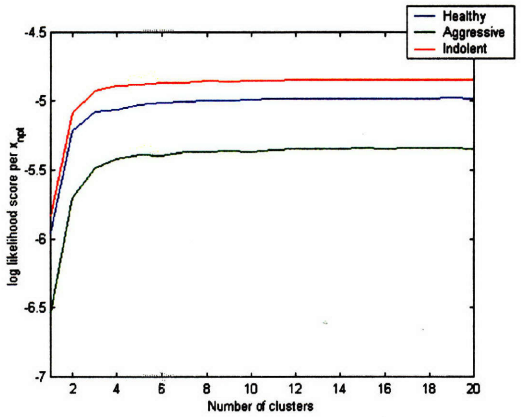
By definition, the wave clustering has $T$ parameters associated with the background class, as there is a zero mean Gaussian distribution with an associated variance, $\sigma^2_{nt}$,at each time point. For each wave class, there is one corresponding $\lambda_{m*}$, which represents the distribution of the positive values for the dominant time point, and $T - 1$ $\lambda_{mt+}$ parameters, which represent the non-dominant time points. Additionally, there is a shared parameter across the wave classes for the negative values at each time point, $\lambda_{t-}$. There are thus $T \times (M + 1)$ parameters, or specifically 24 in this case. The probability distribution functions for the parameters after EM has converged are shown in Figure 3-5. Despite the fact that $M = 5$ for both types of clustering, 16 more parameters are estimated in the Gaussian clustering.




(a) Aggregate score




(b) Random subject groups




(c) Normalized score

Figure 3-4: Log-likelihood scores as a function of the number of clusters for the Gaussian clustering

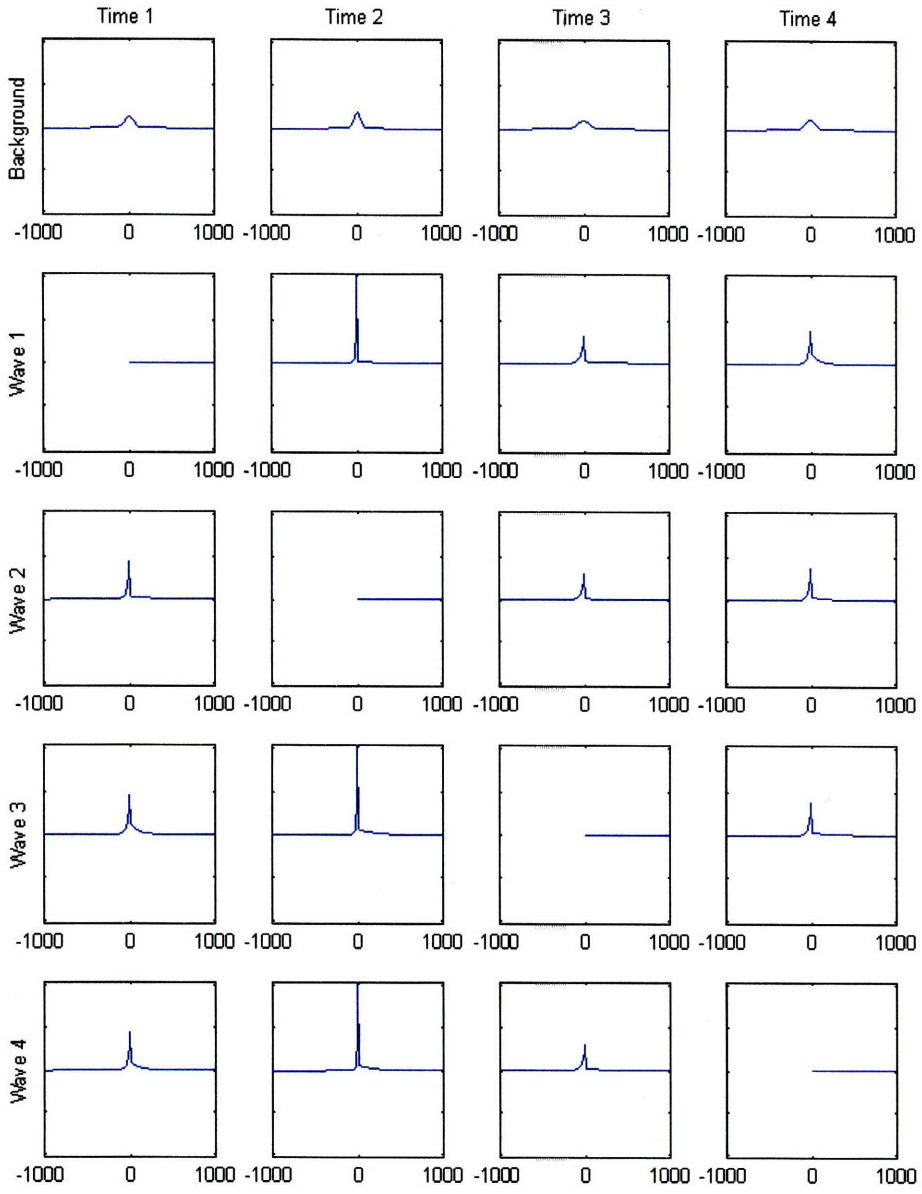



Figure 3-5: Probability distribution functions for the background class and four wave classes at each of the four time points, after the EM algorithm converges (healthy subjects)

### 3.3.2 Cluster labels

The purpose of the clustering was to identify the subset of genes that are differentially active and have common structure out of the original data, which is shown in Figure 3-6. In this section, we compare the structure of the clusters resulting from both the Gaussian clustering and the wave clustering.

For the wave clustering, we threshold the MAP values of cluster assignments to select 500 genes for each subject group, as was explained in Section 3.2.2. We reorder the genes for each subject group according to the cluster labels and then, within a cluster, according to the MAP values. There is visual structure evident in the data in Figure 3-7, as high values in yellow or red start at the top left (first wave, $t-1$) and move down the diagonal to the bottom right (fourth wave, $t_4$).

Because the Gaussian clusters have no *a priori* interpretation, an analogous visualization would include all 50,000 genes. Because such a large percentage of these genes are background, and in order to compare the results in a consistent way, we use the same 500 genes selected by the wave clustering, but instead reorder them by the Gaussian cluster labels and MAP values. This is shown in Figure 3-8, where for each of the groups, the 500 relevant genes are separated into either two or three clusters. It does not appear that the clusters correspond in any way to the patterns of expression we see for the wave clustering.

We show quantitatively in Table 3.1 how the cluster labels differ. We compare the cluster memberships for the Aggressive group given results from the wave clustering and from a random trial of Gaussian clustering. Background genes comprise three entire Gaussian clusters and a large proportion of the other two remaining clusters. Nearly all wave 1 and wave 2 genes were in cluster 1, but the other two waves were split between clusters 1 and 2. This result is similar to what was seen for other trials and other subject groups. Additionally, larger values of $M$ subdivided the background class further, but did not change the classification of non-background genes.

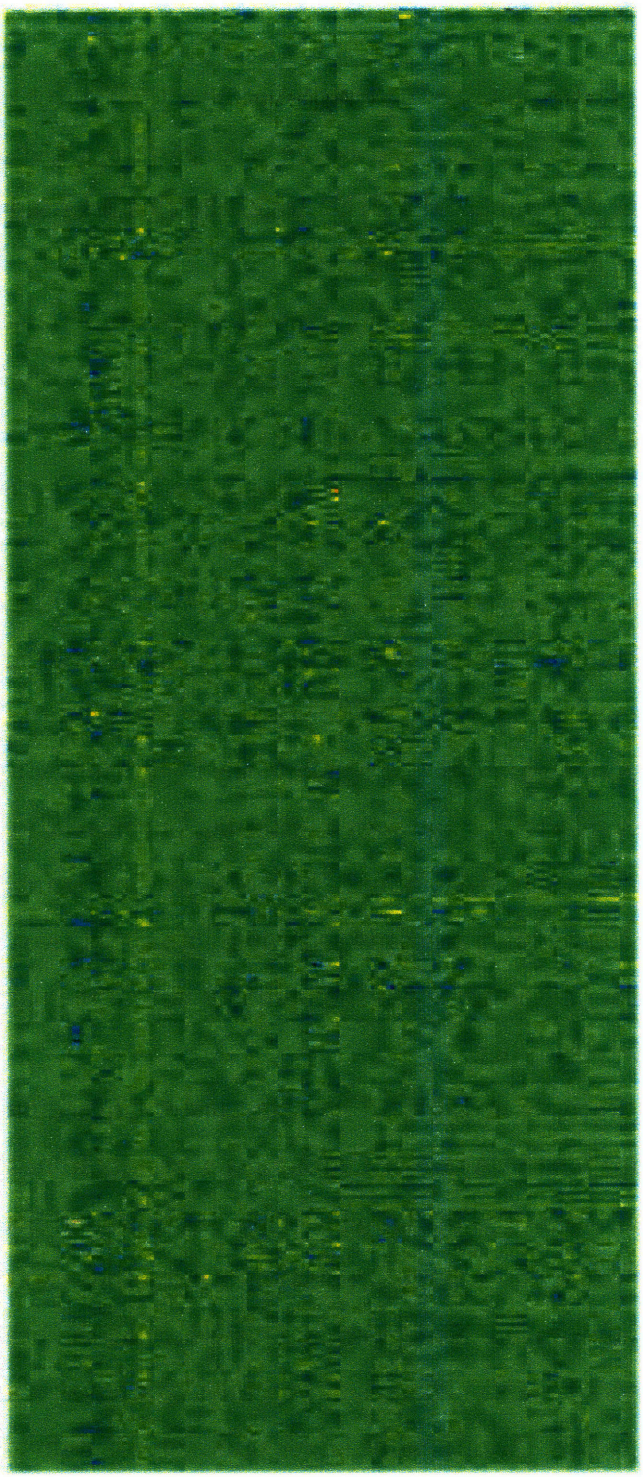


Figure 3-6: Log ratios of stimulated to unstimulated expression data. Each row of the image corresponds to a given gene. There are four columns, corresponding to time points, for each of the 17 subjects.

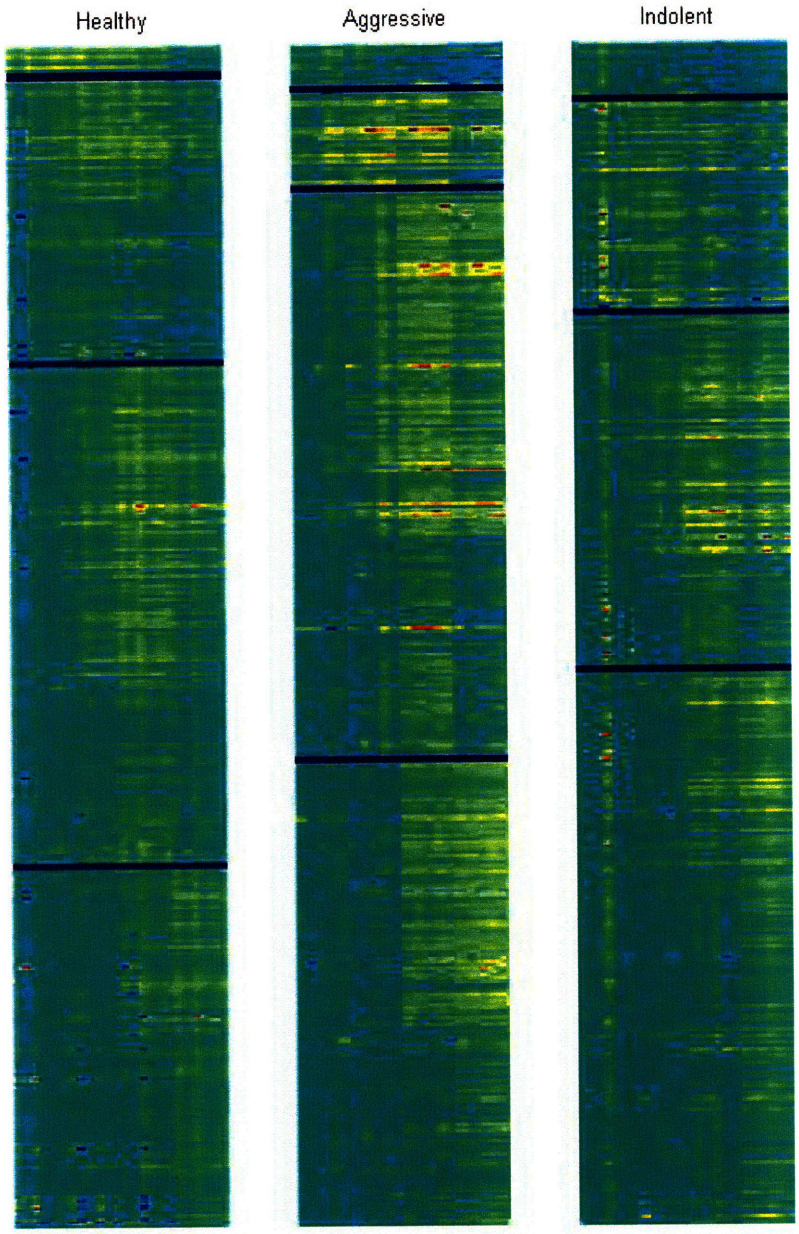



Figure 3-7: Log ratios of stimulated to unstimulated expression data, reordered according to wave clustering results. Only the 500 genes with the highest *a posteriori* values for cluster membership are included.

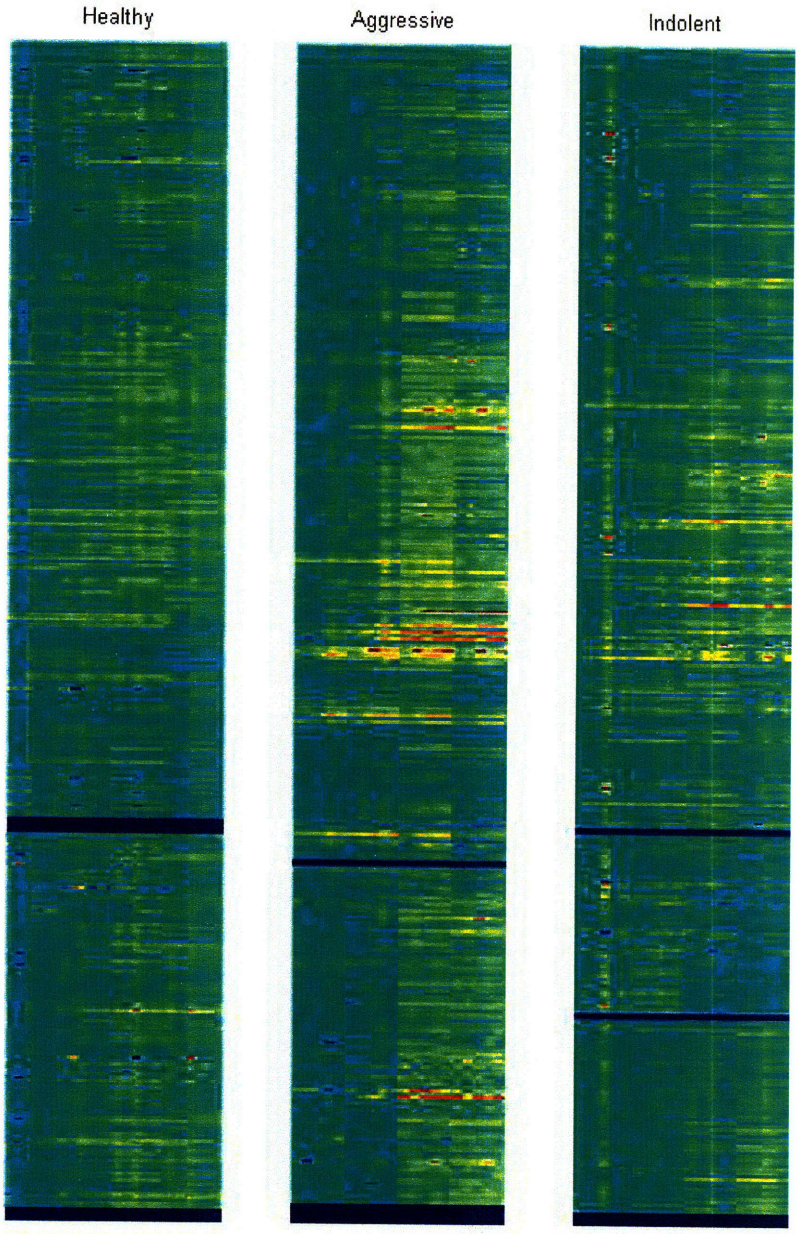



Figure 3-8: Log ratios of stimulated to unstimulated expression data, reordered according to Gaussian clustering results. Only the 500 genes selected by the wave clustering are shown.

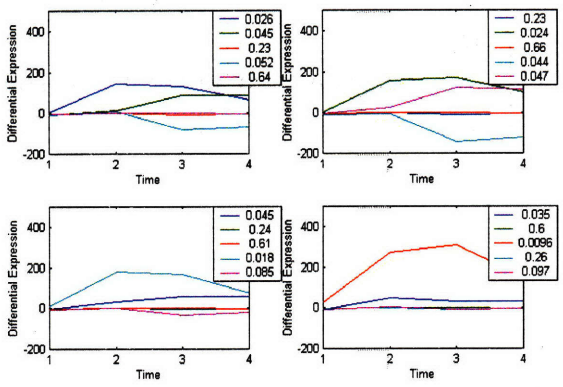
Table 3.1: Comparison of labels for Gaussian clusters and wave clusters

| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| Background | 1056 | 2374 | 3742 | 13876 | 33065 |
| Wave 1 | 16 | 1 | 0 | 0 | 0 |
| Wave 2 | 37 | 3 | 0 | 0 | 0 |
| Wave 3 | 194 | 48 | 0 | 0 | 0 |
| Wave 4 | 106 | 95 | 0 | 0 | 0 |

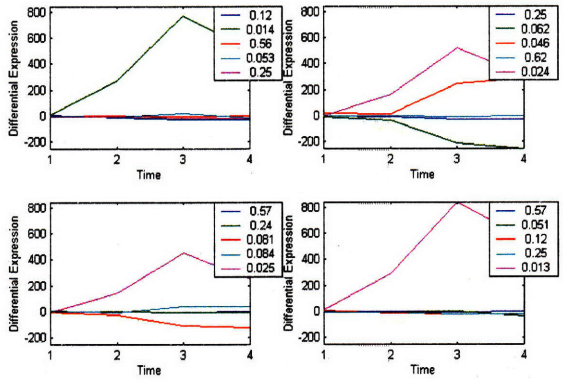
### 3.3.3 Cluster profiles

For the Gaussian clustering, we show in Figure 3-9 several sample results for each subject group because of the effects of the random initialization. The mixture proportion for each of the clusters is included in the legend. There are generally two to three clusters that are nearly zero-mean that comprise the vast majority of the genes, which would correspond largely to the background class in the wave clustering (see Table 3.1). The remaining clusters are a smaller proportion of the genes but have larger differential expression. The cluster means for the non-background classes are relatively unstable and though the majority of the background genes are grouped with one another, many are included in the other clusters as well, which is why the means have relatively small magnitude. Standard deviation error bars are not included in the figure because they are, at each time point, several times as large as the mean. We show this with the bar plots in Figure 3-11(a), where we compute a ratio of the standard deviation at each time point to the maximum absolute value of each cluster mean. We use the maximum value instead of the actual mean value at each time point because it better represents the scale of the cluster mean. For example, at an off-peak time point (usually $t_1$) the mean value may be very small and that makes the standard deviation ratio very large, even though what is most important is the ratio at the cluster's peak time point.

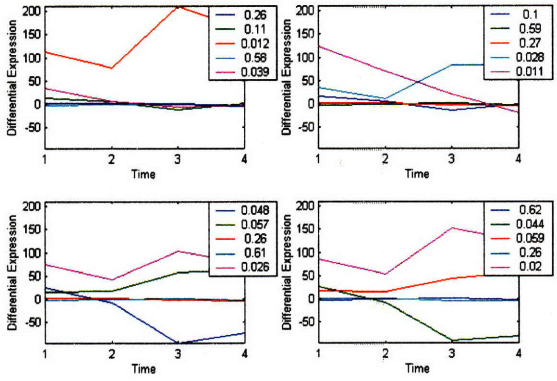
We compute the cluster means and standard deviations for wave clustering as




(a) Healthy




(b) Aggressive




(c) Indolent

Figure 3-9: Cluster means for four random trials of Gaussian clustering (M=5) for each subject group. Because the results greatly depend on the initialization, cluster means are not consistent across trials.

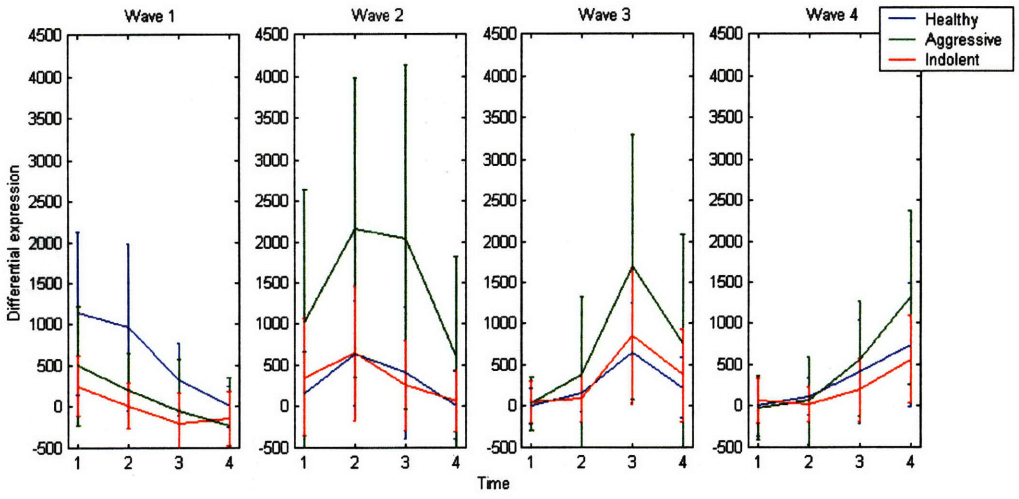
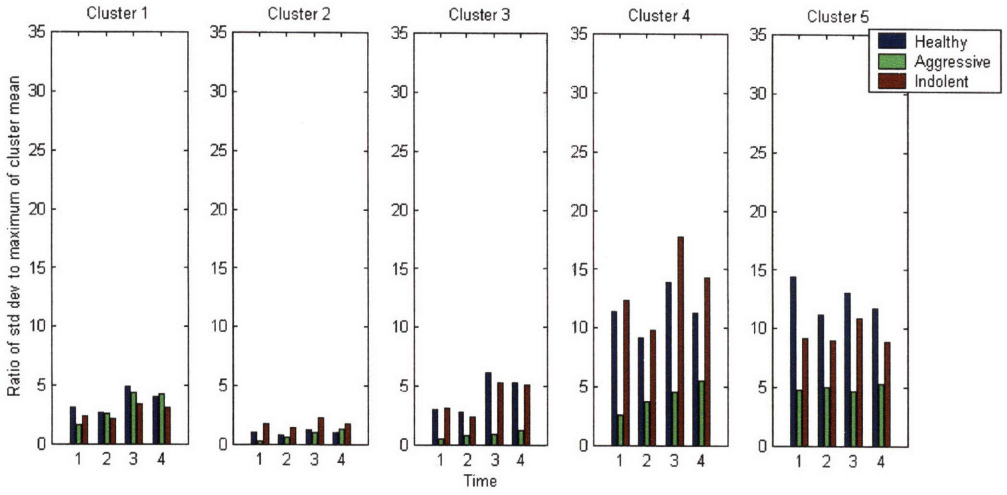


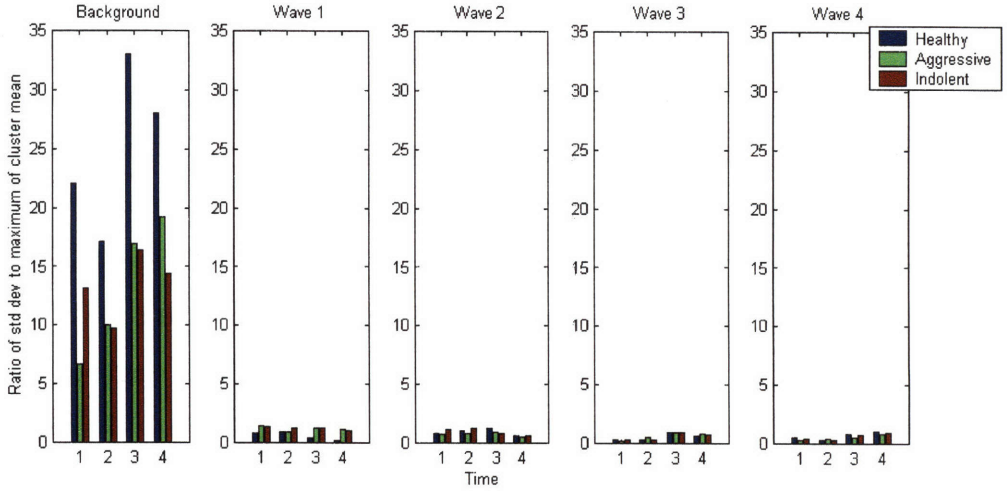

Figure 3-10: For wave clustering, the cluster means for each of the three subject groups.

well, showing the four non-background classes in Figure 3-10. The means have larger magnitude as compared to the Gaussian clustering because fewer of the background genes are included in these clusters. The wave clusters have consistent shapes for each subject group, with comparatively smaller intra-cluster variance relative to the means (see Figure 3-11(b)). For the background class, there is a very high ratio because the means are essentially zero, and all genes that do not fit the wave models are assigned to the background class. This includes genes that are not highly differentially expressed at all and those genes that are not highly differentially expressed consistently across subjects within a group. Forcing these two types of genes into a single cluster means that this type of clustering does not effectively model these genes. However, because the purpose of the background class is only to identify genes to be excluded from further processing, it is not essential for them to well-represented by the cluster parameters.

The standard deviation ratio results in Figure 3-11 are important because when there is high variance relative to the mean, misclassification of genes is more likely. To quantify the confidence in the cluster assignment, we show in Figure 3-12 histograms of the MAP assignment values for the 500 selected genes under both clustering meth-




(a) Gaussian clustering




(b) Wave clustering

Figure 3-11: Ratio of standard deviation for each cluster at each time point to the maximum of the cluster mean

Table 3.2: Log-likelihood scores

| | Wave clustering | Gaussian clustering |
|---|---|---|
| Healthy | $-7.56e+06$ | $-6.59e+06$ |
| Aggressive | $-6.83e+06$ | $-5.89e+06$ |
| Indolent | $-7.40e+06$ | $-6.40e+06$ |

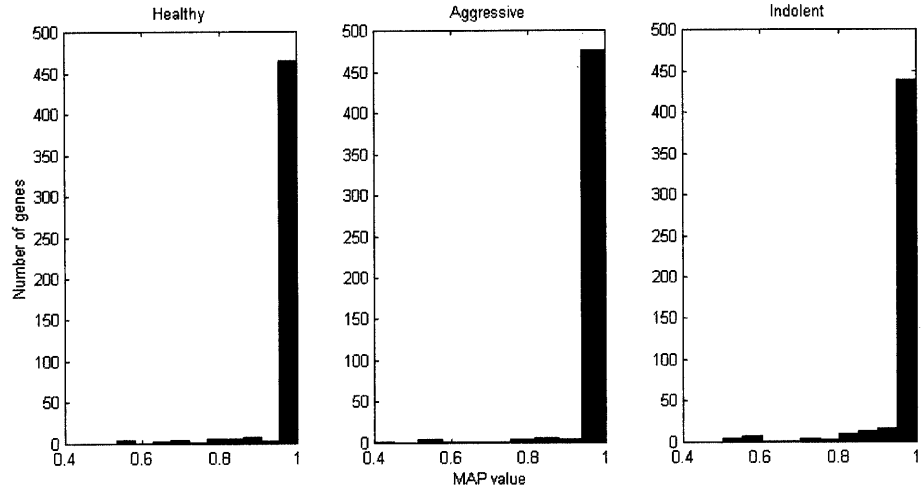
ods. The means are 0.9990, 0.9994, and 0.9993 for the wave clustering for each group, and 0.9822, 0.9864, and 0.9733 for the Gaussian clustering. While these are all high values, there is lower *a posteriori* probability for the Gaussian assignments.

### 3.3.4 Score comparisons

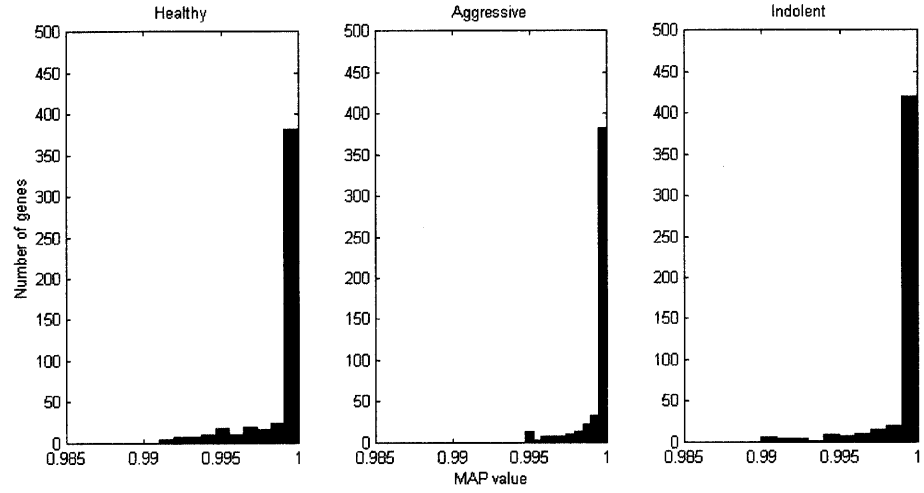
An obvious method to compare the performance of two EM-based clustering algorithms is to compute the log likelihood scores, as that is the score that the iterative EM algorithm is maximizing. This is the log of the likelihood of the data as a function of the parameters of the model. In Table 3.2, for every subject group, the score for the Gaussian clustering is superior to that of the wave clustering.

This result would imply that the Gaussian clustering does a better job modeling the structure of the data. Though there are 16 more parameters estimated in the Gaussian clustering for $M = 5$, this is not simply a case of a better fit because of more parameters. From Figure 3-4(a), we see that even for $M = 3$, the scores are still better when the number of parameters is the same for both clustering methods.

There are a few reasons for the difference in score. First, as explained in Section 3.3.3, background genes are not well modeled by the wave clustering. Because these genes make up an extremely high percentage of the total and the score is an aggregate value, it is greatly influenced by these genes. They could be better modeled if we were to subdivide the background class further, which is exactly what happens automatically in the Gaussian clustering, but that only adds unnecessary parameters to the modeling. A more important consideration is that we only allow for four types




(a) Gaussian clustering




(b) Wave clustering

Figure 3-12: MAP values for the 500 genes identified by the wave clustering

of non-background genes, when there obviously would be some genes that would be better represented by models that have high differential expression at multiple time points. This is a limitation, but we make choices based on the pilot study to use the smallest number of models (which is helpful for the predictive modeling) that capture most of the structure in the data.

### 3.3.5 Significance results

One method for estimating the statistical significance of a clustering result is the use of permutation statistics [56, 26]. Permutation methods for computing statistical significance exploit the exchangeability in order to generate valid samples from the null hypothesis. In our particular case, the null hypothesis states that the statistical properties of our sample do not vary with time. Consequently, to generate samples from this hypothesis, for each trial, we randomly permute the time points for each gene independently (but consistently within a subject group). The clustering algorithm is run on each of these permuted data sets and the log-likelihood score is stored. The estimated $p$-value is defined as the fraction of trials from the null hypothesis that result in a higher log-likelihood score than the original data. Comparisons of log-likelihood scores here are within the same model distributions, so we avoid the ambiguity that was an issue in 3.3.4 when we compared across different model shape distributions. In Figures 3-13 and 3-14, we show histograms of the likelihood scores for 100 trials. The distribution of the likelihood scores for the wave clustering over the Indolent group appears to be bimodal, but the reason for this is still an open question.

These distributions of permuted scores are compared to the scores for the actual data in Tables 3.3 and 3.4. For wave clustering, the actual data has a better score in each of the 100 trials, so our estimated $p$ value is $< 0.01$ for each subject group. For the Gaussian clustering, the actual data has a better score in each of the 100 trials as well. Though the statistical significance results of the two clustering methods are indistinguishable in terms of $p$-values, the distributions of the permuted scores are closer to the actual scores. We quantify this difference by computing an upper bound

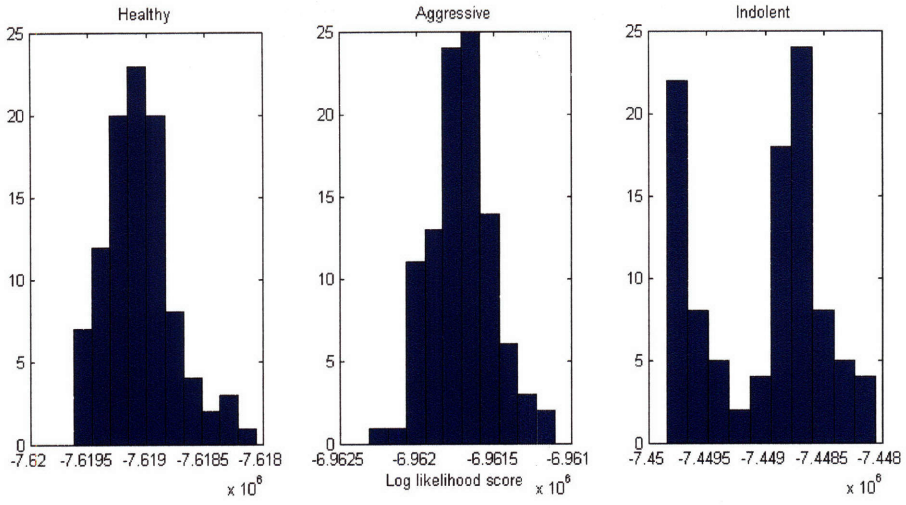



Figure 3-13: Histograms of log-likelihood scores for permuted data (Wave clustering)

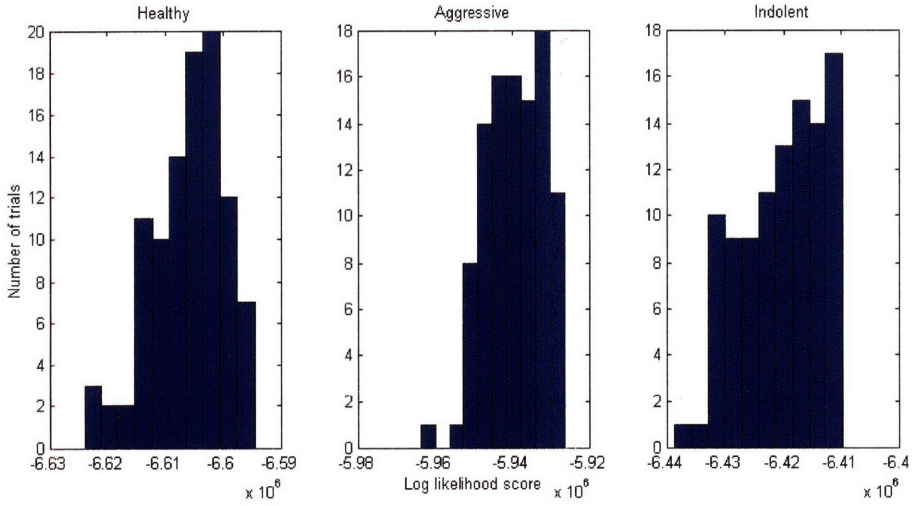



Figure 3-14: Histograms of log-likelihood scores for permuted data (Gaussian clustering)

Table 3.3: Significance for Wave clustering

| | Log-likelihood score | $p$-value | Chebyshev value |
|---|---|---|---|
| Healthy | $-7.56e+06$ | $< 0.01$ | $\leq 3e--05$ |
| Aggressive | $-6.83e+06$ | $< 0.01$ | $\leq 3e--06$ |
| Indolent | $-7.40e+06$ | $< 0.01$ | $\leq 1e--04$ |

Table 3.4: Significance for Gaussian clustering

| | Log-likelihood score | $p$-value | Chebyshev value |
|---|---|---|---|
| Healthy | $-6.59e+06$ | $< 0.01$ | $\leq 0.09$ |
| Aggressive | $-5.89e+06$ | $< 0.01$ | $\leq 0.03$ |
| Indolent | $-6.40e+06$ | $< 0.01$ | $\leq 0.06$ |

on the probability of a value higher than the actual score coming from the distribution of permuted scores, assuming we can characterize the distribution with a mean and variance. We use a one-tailed variant of the Chebyshev inequality, with $k > 0$, which is:

$$\Pr((X - \mu) \geq k\sigma) \leq \frac{1}{1 + k^2}. \tag{3.26}$$

These values are included Tables 3.3 and 3.4. We see a difference of two to four orders of magnitude, depending on the group, in favor of the wave clustering. Because the Chebyshev value is only an upper bound, it provides the worst case for the probability of the permuted distribution producing a higher value than the actual score, so it is not equivalent to actually running the corresponding number of permuted trials.

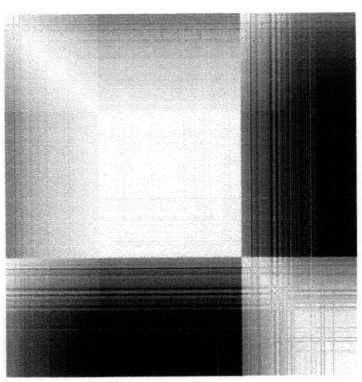
We interpret the statistical significance of both methods of clustering as the algorithms having captured the differences in the parameters of the distributions of the data at each time point. While this does not imply temporal coherence, it does show that the structure of the data is such that the order of the time-series is important.

### 3.3.6 Stability

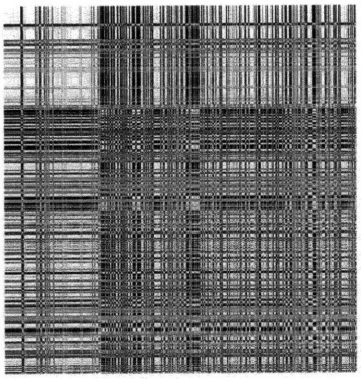
As we recognize that there is inconsistency even within relatively homogeneous subject groups and that microarray data is noisy, it is particularly that important the clustering algorithms be robust. Beginning with the Gaussian clustering, we investigate the effects of random initialization. Figure 3-15 is an image of how often two genes are clustered together over 25 trials of Gaussian clustering. Because of the difficulty in displaying a $50,000 \times 50,000$ matrix, only the 500 genes per subject group, selected by the wave clustering and re-ordered according to a random Gaussian trial (as in Figure 3-8), are shown in the left column. There is clear structure, showing that the 500 genes are subdivided into two or three fairly consistent groups. We see in the right column of Figure 3-15, however, as we reorder according to the wave clustering results, that these clusters do not correspond to the wave clusters at all. Because the initialization is set for the wave clustering, there is no dependence on a random initialization.

In addition to the random initialization, we know that microarray data is noisy. We model this noise by adding zero-mean Gaussian noise ($\sigma = 40$, about an order of magnitude smaller than the positive deflections of wave 1 genes) to the input data, in order to test the robustness of the clustering. We show in the left column of Figure 3-16 an image of how often genes are clustered together, reordered according to the Gaussian cluster labels in the same way as the left column of Figure 3-15. The same clusters appear, although they are now somewhat less stable. In the right column, as they are ordered by wave type, there is once again little structure.

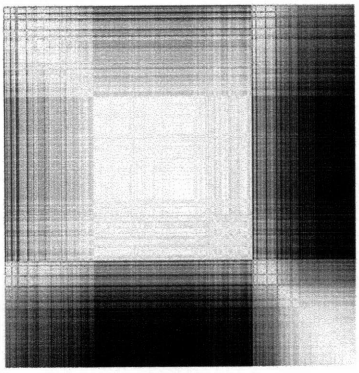
We compare the stability of the Gaussian clusters to the results of having perturbed the input data ($\sigma = 40$) for the wave clustering, shown in Figure 3-17. The genes that have the same label are generally clustered together in most of the trials. This comparison is shown quantitatively in Figure 3-18, where we compute the mean fraction of trials a gene is clustered with the genes in the same wave (excluding itself) and the mean fraction of trials a gene is clustered with other genes in the three remaining waves. For the Gaussian clustering, with the exception of wave 1 genes for



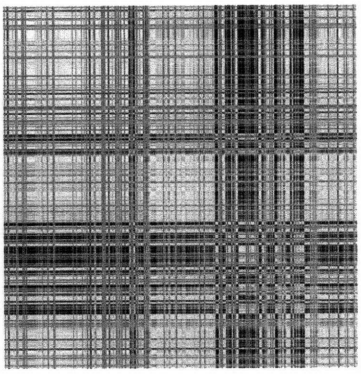
(a) Healthy (Gaussian ordering)



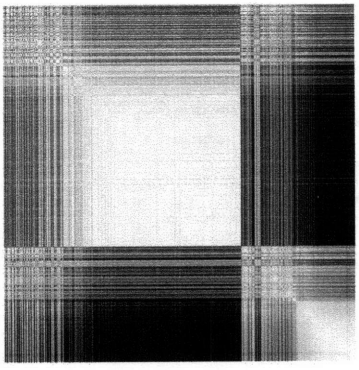
(b) Healthy (Wave ordering)



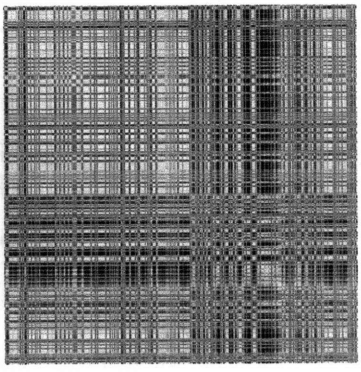
(c) Aggressive (Gaussian ordering)



(d) Aggressive (Wave ordering)

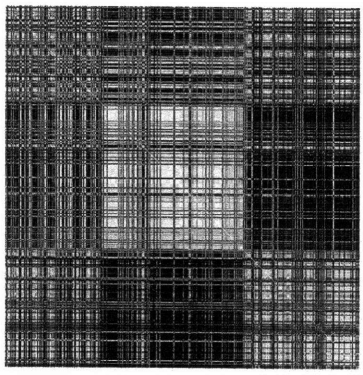


(e) Indolent (Gaussian ordering)

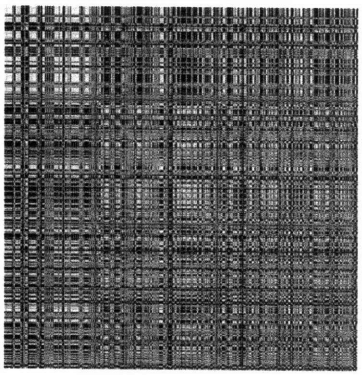


(f) Indolent (Wave ordering)

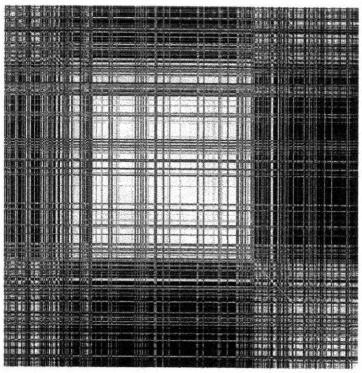
Figure 3-15: Cluster stability for random initializations for Gaussian clustering



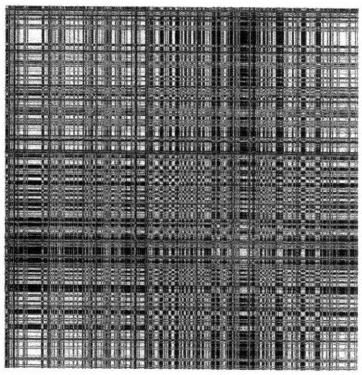
(a) Healthy (Gaussian ordering)



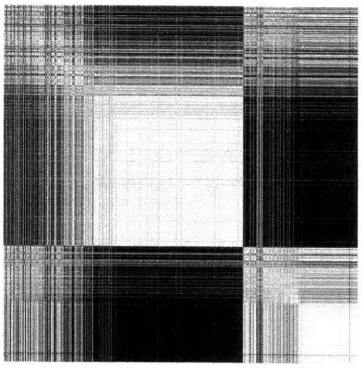
(b) Healthy (Wave ordering)



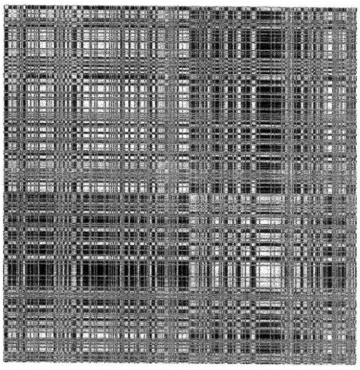
(c) Aggressive (Gaussian ordering)



(d) Aggressive (Wave ordering)

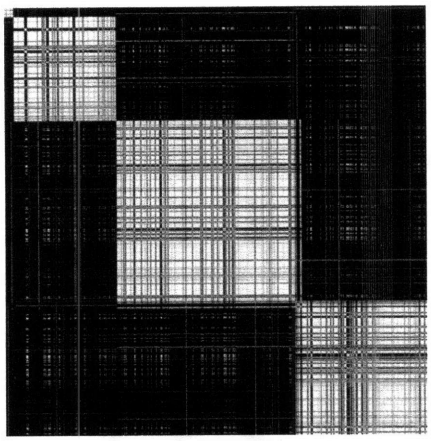


(e) Indolent(Gaussian ordering)

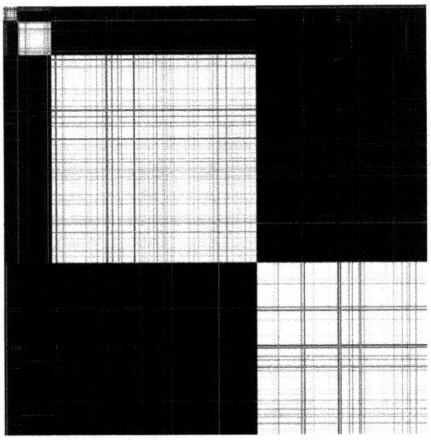


(f) Indolent (Wave ordering)

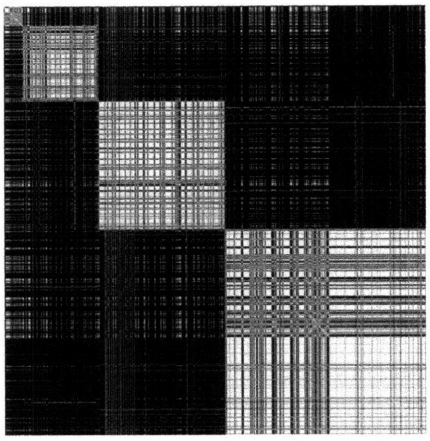
Figure 3-16: Cluster stability for perturbed input data for Gaussian clustering



(a) Healthy



(b) Aggressive



(c) Indolent

Figure 3-17: Cluster stability for perturbed input data for wave clustering

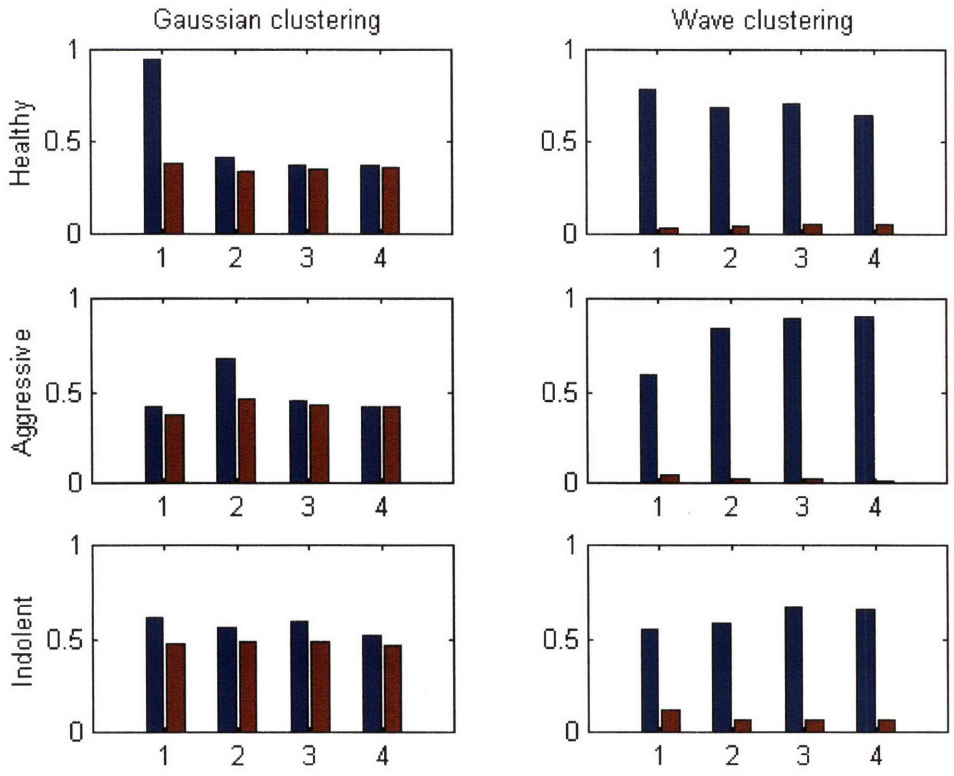



Figure 3-18: Blue is the mean fraction of trials a gene is clustered with other genes in the same wave, red is the mean fraction that a gene is clustered with other genes in the three remaining waves.

the healthy subjects, these values are nearly equal, which means there is no differentiation between waves of genes. The wave clustering, however, consistently clusters genes in the same waves together and in different waves separately.

## 3.4 Summary

We have shown with our clustering that we can represent the temporal structure of the BCR stimulation response with few models which correspond to waves of expression that peak at each of the four time points. We obtained classifications for all of the genes, identifying those that appear to be relevant to BCR and at what time point they peak. The effect of subject group on the labelings is discussed in

Section 6.5. Compared to the Gaussian clustering, the wave clusters have consistent shapes for each subject group, with smaller intra-cluster variance relative to the means. Genes are also consistently clustered in the same waves together and in different waves separately, despite perturbations of the input data with noise. Both clustering procedures show that there is statistically significant temporal structure that can be extracted from the data. We have shown for our application that the wave clustering produces stable, consistent and meaningful models of the differential expression profiles of genes related to BCR stimulation. In addition to the results presented in this chapter, wave clustering serves as a first step in order to make predictive modeling, as described in Chapter 5, tractable.

# Chapter 4

# Hierarchical Clustering

In much of the genetic research in B-CLL, the focus has been on classifying patients into subgroups to more effectively plan a course of treatment. These approaches generally consist of identifying a single gene or small number of genes that are useful in discriminating between the aggressive and indolent forms of the disease and do not explain why or how those genes are behaving differently and what effect they have on the resulting regulatory network. Without a more global picture of the differences in cellular behavior, designing targeted treatments is virtually impossible. On the other hand, combining information from all genes for a patient clustering approach is a challenge because there are thousands more features (genes) than samples (patients). These difficulties are well summarized by Xing and Karp [82]: samples may be clustered accurately in many different ways, most of which are not necessarily biologically meaningful; a large number of features are not relevant; and generalization is difficult because clustering on the large feature space tends to overfit the data.

This is exactly the case for our B-CLL data set as we have measurements from more than 50,000 genes but only 17 subjects. We use BCR stimulation to attempt to identify those genes that are relevant to the BCR signaling pathway, but this may not necessarily be the same subset of genes that will accurately discriminate between patient populations. Some differentially expressed genes will behave similarly across all patients, and some, which are useful for patient classification, will behave differently. Because we ultimately intend to infer models of temporal interaction that

show where similarities and differences occur, both of these types of genes should be included in further analysis.

In this chapter, we formulate an integrated approach for combining clustering of patients with classification of genes, such as the method in Chapter 3. The result will provide a breakdown of relatively homogeneous subject groups, which is necessary for the implicit assumption of consistent expression behavior in predictive modeling (see Chapter 5).

## 4.1 Background

The simultaneous clustering of both genes and samples is a common challenge in bioinformatics. Current thinking in molecular biology holds that only a small subset of genes participate in any cellular process of interest, and that a cellular process takes place only in a subset of samples. Therefore, biclustering applies to virtually any problem where data from any different experimental conditions (disease category, immune response, temperature, time, etc.) has been acquired. The first part of this section will address current work in biclustering, but because our samples are not simply a value for each gene for each patient, but an entire temporal expression profile, we can take advantage of the extra dimension of data in an interesting way. In the second part of this section, we discuss related problems in other fields which utilize hierarchical EM algorithms and actually better address the goal for our application.

### 4.1.1 Biclustering

Cheng and Church [15] were the first to apply biclustering to gene expression data, simultaneously clustering both genes and conditions in order to identify groups of genes that are co-regulated under a subset of conditions. Their algorithm identifies biclusters with low mean squared residue, defined as the variance of the set of all elements plus the mean row variance plus the mean column variance. Because this problem is NP-hard, an efficient node-deletion algorithm is presented. Similar methods have been proposed more recently which extend the original algorithm by using

a probabilistic algorithm to discover sets of possibly overlapping biclusters [83] or by iteratively refining the biclusters by adding more genes and/or conditions [87].

Coupled Two-Way Clustering (CWTC) [31] looks for pairs of a relatively small subset of features (either genes or samples) and of objects (samples or genes), such that when the objects are represented using only those features, clustering yields stable and significant partitions. Finding such pairs of subsets is also computationally hard; the CTWC method produces such pairs in an iterative process. Other algorithms use bipartite graph representation [73], hyperplanes in the data [29], preprocessing combined with linear algebra [74], probabilistic models[66], feature filtering [82], and block mixture modeling [32]. For a more complete summary, Madeira and Oliveira [47] survey recent work in biclustering. Time series data is also relatively rare, and when it is available, it is often treated as simply additional experimental conditions. This is addressed by Zhang et al. [86], where the typical mean residue score measure is used, but deletion from a bicluster is only allowed at the starting and ending points of the time interval.

For all of these techniques, as for clustering in general, evaluating the quality and validating the results is difficult. One common approach is to embed synthetic biclusters in the data and then retrieve them with the clustering algorithm. Additionally, there are data sets available for simple model organisms, such as yeast, where much of the transcriptional regulatory network is well-understood. Finally, data sets from diseases, where the patient subgroups and many of the key genes are known, are also available. There is, however, no other immune response time series data set similar to the BCR acquisition we analyze here.

### 4.1.2 Hierarchical EM

Our application is dissimilar to other biclustering approaches because we are not simply looking for genes that make up significant biclusters. We do not want to solely identify those genes that discriminate patient populations. As in Chapter 3, the clustering is not only intended to be a result in and of itself, but also as a first step in the predictive modeling discussed in Chapter 5, so a complete classification of genes

for each patient cluster is necessary. Also, because we are trying to model all subjects, we do not want to ignore those that do not have perfectly consistent behavior. In other words, instead of isolating subsets of the data that form biclusters, we intend to label the entire matrix of data. The problem is therefore better described as clustering a data set that is comprised of patients that can be divided into subgroups, which in turn are comprised of genes that can be divided into subgroups as well. This is closely related to probabilistic Latent Semantic Analysis [37], where a collection of documents (or subjects) contains topics (or waves) which are comprised of terms (or genes), though there are important differences. We assign subjects to subpopulations based on the wave to which each gene, which would correspond to classifying a document based not on which topics are present, but on which terms belong to which topics (i.e. how their meanings change depending on the context). Additionally, terms are often repeated in documents exactly, while microarray data is noisy and would require considerable pre-processing to attempt eliminate what would be equivalent to inconsistent spelling. Despite differences in the details, the formulation is similar. It is also analogous to work done in image classification [12, 76], word segmentation [57], and learning object maps for robot environments [1], in which EM is used to learn parameters at multiple layers of a hierarchy.

## 4.2 Patient clustering

We begin by temporarily ignoring the hierarchy and simply clustering the patients. We treat each patient as a sample and each gene as a feature, which is a natural alternative to the clustering approach presented in Chapter 3, where we treated each gene as a sample and each patient as a feature. To model this, we assume that subjects come from a finite mixture of probability distributions. The likelihood function, where $c$ represents the hidden patient class label, is defined as:

$$L(\Theta) = \sum_{c=1}^{C} p(X; c, \Theta) p(c|\Theta) \tag{4.1}$$

We use EM (see Chapter 3 and Appendix A) to cluster by combining the data for each patient into an $NT \times 1$ feature vector $\vec{x}_p$, where $N$ is the number of genes and $T$ is the number of time points in the expression profile. Each cluster has $N \times T$ independent Gaussian distributions, each with a mean $\vec{\mu}_c$ and variance $\vec{\sigma}_c^2$. $\Theta^k$ includes all parameters of all the clusters at iteration $k$. Because we believe there are three main subgroups (Healthy, Aggressive, and Indolent), we set $C = 3$. In the Expectation step, we compute the $P \times C$ matrix of weights, where each entry is defined as:

$$P(c|\vec{x}_p, \Theta^k) = \frac{P(\vec{x}_p|c, \Theta^k)P(c|\Theta^k)}{\sum_{c'=1}^{M} P(\vec{x}_p|c, \Theta^k)P(c|\Theta^k)} \tag{4.2}$$

$$= w_{pc} \tag{4.3}$$

where, for a mixture of Gaussians, $P(\vec{x}_p|c, \Theta^k)$ is:

$$P(\vec{x}_p|c, \Theta^k) = \prod_{i=1}^{NT} \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left(\frac{-(x_{ip} - \mu_{ic})^2}{2\sigma_{ic}^2}\right) \tag{4.4}$$

In the Maximization step, we solve for $\vec{\mu}_c$:

$$\vec{\mu}_c = \frac{\sum_{p=1}^{P} w_{pc}\vec{x}_p}{\sum_{p=1}^{P} w_{pc}} \tag{4.5}$$

This is equivalent to taking a weighted average of the data to compute the cluster mean. Similarly, the variance for a cluster is (using the value for $\vec{\mu}_c$ just computed):

$$\vec{\sigma}_c^2 = \frac{\sum_{p=1}^{P} w_{pc}(\vec{x}_p - \vec{\mu}_c)^2}{\sum_{p=1}^{P} w_{pc}} \tag{4.6}$$

Finally, the mixture proportion for a given class, $P(c|\Theta^k)$, is computed as:

$$P(c|\Theta^k) = \frac{1}{P}\sum_{n=1}^{N} w_{pc} \tag{4.7}$$

In clustering samples (patients), many of the genes are irrelevant, i.e. they do not contribute to a meaningful, consistent clustering [82]. To reduce the dimensionality of the feature space, we use the gene classification results in Section 6.5 to obtain a subset of genes believed to be relevant to the BCR signaling pathway. While not all of these are necessarily useful to differentiate across subject groups, it does

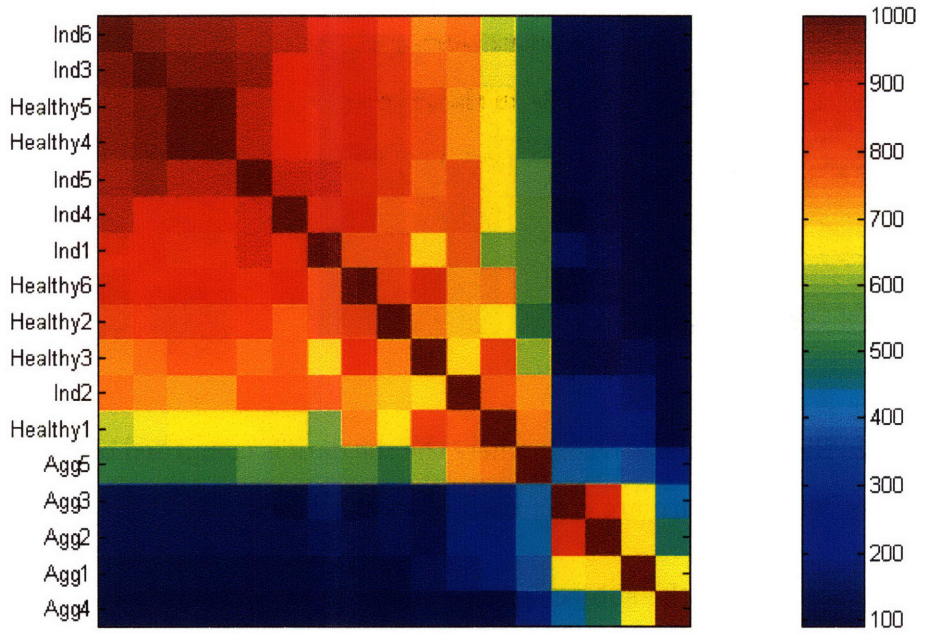



Figure 4-1: Affinity matrix for the 17 subjects, showing how often each subject pair was clustered together over the 1000 trials

provide a more reasonable size set of genes with which to cluster. There are a total of 1028 genes (see Figure 6-18) as compared to 54,613. Also, EM only guarantees convergence to a local maximum, so the final clusters depend on the initialization. Over the $k = 1000$ trials, we randomly assign each of the subjects to a cluster, compute the resulting means and variances, and then iterate between the Expectation and Maximization steps described above until the cluster memberships are stable. Given the cluster assignments for each trial, we compute an affinity matrix in which each entry represents the number of times any pair of subjects were clustered together. In Figure 4-1, this matrix has been reordered to best show the structure in the results.

The Aggressive patients separate out from the other two groups, which confirms our intuition, given that their prognosis is so different from Healthy subjects and Indolent patients. Indolent and Healthy, which are expected to be more closely related, are often clustered together. This provides some basis for believing that there

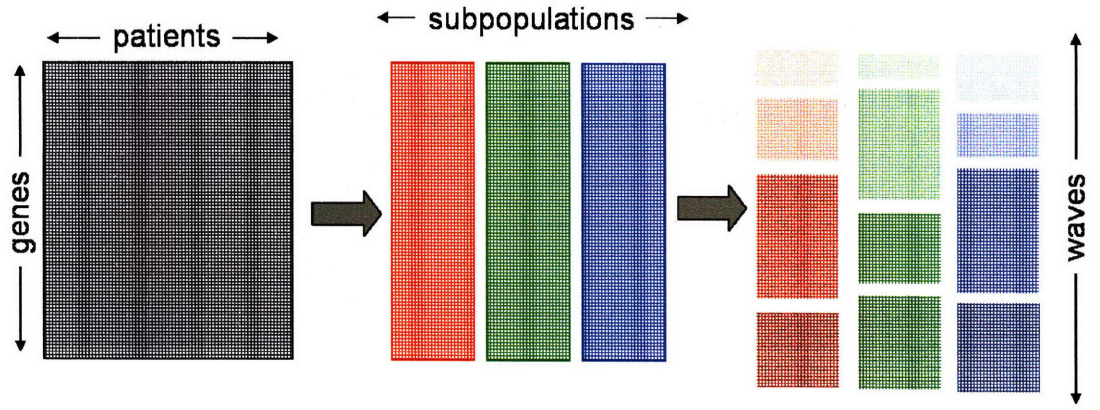



Figure 4-2: Hierarchical clustering, where the the data is comprised of subpopulations of subjects, which in turn are comprised of waves of genes

are differences among the subject groups that can be identified, but is not entirely satisfactory because the subset of 1028 genes used for clustering were selected based on consistent within-group behavior. Addressing this issue requires simultaneous clustering of genes and patients.

## 4.3 Hierarchical formulation

Combining patient clustering and gene classification into an integrated approach is accomplished by adding an additional level of hierarchy into an EM formulation, which is shown in the diagram in Figure 4-2. We assume that there is an underlying distribution of patients, which each have an underlying distribution of genes. The first layer of the hierarchy remains the same as the clustering of subjects in the previous section, but $p(X|c,\Theta)$ becomes:

$$p(X|c,\Theta) = \sum_{m=1}^{M}\prod_{n=1}^{N} p(\vec{x}_{np}|m,\Theta)p(m|n,c,\Theta), \tag{4.8}$$

where $n$ specifies the gene and $m$ specifies the gene model. Therefore, the complete data likelihood function is now:

$$L(\Theta) = \sum_{c=1}^{C}\prod_{p=1}^{P} p(c|p,\Theta)\sum_{m=1}^{M}\prod_{n=1}^{N} p(\vec{x}_{np}|m,\Theta)p(m|n,c,\Theta) \tag{4.9}$$

The probability for a given gene $\vec{x}_{np}$ in a given class is defined by:

$$p(\vec{x}_{np}|m,\Theta) = \prod_{t=1}^{T} p(x_{npt}|m,\Theta) \tag{4.10}$$

Unlike the formulation in 3.2.2, it is no longer necessary to take a product over patients because the patients are combined in the first level of the hierarchy. However, the rest of the formulation is almost identical. We also assume for now that the gene model parameters are shared across patient clusters, which means that the size and shape of the expression profiles are the same, and it is the wave labels that differ. For waves 1 through 4, the "dominant" time point is defined as being strictly positive, which corresponds to an increase in gene expression at that time point due to the initial stimulation. When $m$ (the class label) corresponds to $t$ (the time point), the equations are:

$$p(x_{npt}|m,\Theta) = \lambda_{m*} \exp\left(-\lambda_{m*} x_{npt}\right) \text{ for } x_{npt} > 0 \tag{4.11}$$

$$p(x_{npt}|m,\Theta) = 0 \qquad \text{for } x_{npt} \leq 0 \tag{4.12}$$

At the other time points, the two-sided Laplacians are defined by:

$$p(x_{npt}|m,\Theta) = \frac{\lambda_{mt+}}{2} \exp\left(-\lambda_{mt+} x_{npt}\right) \quad \text{for } x_{npt} > 0 \tag{4.13}$$

$$p(x_{npt}|m,\Theta) = \frac{\lambda_{t-}}{2} \exp\left(\lambda_{t-} x_{npt}\right) \qquad \text{for } x_{npt} \leq 0 \tag{4.14}$$

The parameters for the distributions at the non-dominant time points are $\lambda_{mt+}$, which are unique to each model and expected to be larger (modeling smaller deflections) than $\lambda_{m*}$, and $\lambda_{t-}$, which are shared across the four wave models. Because these parameters are shared, the negative deflections will be equally likely under each of the models and will not have an impact on the assignment of a gene to a particular class. The final class is the background, which is composed of genes that are not differentially activated with BCR stimulation and are modeled by a zero-mean Gaussian distribution at each of the four time points.

$$P(x_{npt}|m,\Theta) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(\frac{-x_{npt}^2}{2\sigma_t^2}\right) \tag{4.15}$$

The difference between this formulation and the one for the wave clustering presented in Chapter 3 is in how the information across subjects is combined. Instead of a fixed assignment of a patient to a group, where each patient is assumed to be an independent draw, the patient cluster assignments are iteratively adjusted depending on the consistency of the wave labels of their genes. In the Expectation and Maximization steps, we show how the parameters of the distributions and the class memberships of each gene and patient are estimated.

### 4.3.1 EM iterations

In the Expectation step, we define:

$$w_{nmpc} = p(c, m|\vec{x}_{np}, \Theta^k) \tag{4.16}$$

This is the joint posterior probability for a gene model $m$ and patient class $c$ for a given gene $n$ for a given patient $p$. It is computed by:

$$w_{nmpc} = \frac{p(\vec{x}_{np}|c, m, \Theta^k)p(c, m|n, p, \Theta^k)}{\sum_{c'=1}^{C}\sum_{m'=1}^{M} p(c', m'|n, p, \Theta^k)} \tag{4.17}$$

In the Maximization step, the update equations for parameters are very similar to those in Chapter 3.

$$\sigma_t^2 = \frac{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nmpc}x_{npt}^2}{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n=1}^{N} w_{nmpc}} \tag{4.18}$$

The remainder of the parameters are for the wave classes. For $\lambda_{m*}$, the peak time point, when $m = t$:

$$\lambda_{m*}^2 = \frac{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nmpc}x_{npt}}{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nmpc}} \tag{4.19}$$

Instead of summing over all $N$ genes, only those with $x_{npt} > 0$ are included. Currently, we impose a hard constraint on $\lambda_{m*}$, which guarantees large positive deflections at the cluster's peak time point. For the non-dominant time points, when $m \neq t$:

$$\lambda_{mt+}^2 = \frac{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nmpc}x_{npt}}{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{n\in x_{npt}>0} w_{nmpc}} \tag{4.20}$$

For the shared parameter $\lambda_{t-}$:

$$\lambda_{t-}^2 = \frac{\sum_{c=1}^{C}\sum_{m=1}^{M-1}\sum_{p=1}^{P}\sum_{n\in x_{npt}\leq 0} w_{nmpc}|x_{npt}|}{\sum_{c=1}^{C}\sum_{m=1}^{M-1}\sum_{p=1}^{P}\sum_{n\in x_{npt}\leq 0} w_{nmpc}} \tag{4.21}$$

It is important to note that because we sum over $C$, the model parameters $\lambda$ and $\sigma$ are shared across the patient classes. This implies that the shapes of the genes are expected to be relatively consistent across patient classes, which given the cluster means shown in Figure 3-10 is a reasonable assumption. With shared parameters the discrimination between patient classes occurs with *which* model the genes were assigned, not the shapes of the models. If the model parameters were different depending on the patient cluster, the only difference in the formulation would be that there would be no summations over $C$ and parameters would be computed independently for each class.

In computing the mixture proportions, we weight each gene's contribution to the model as before, but we also weight the contribution of the patient according to its patient cluster membership. For these updates, $p(c, m|n, p, \Theta^k)$ can be reduced to:

$$p(c, m|n, p, \Theta^k) = p(m|n, c, \Theta^k)p(c|p, \Theta^k). \tag{4.22}$$

Each of those is computed by:

$$p(c|p, \Theta^k) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} w_{nmpc}}{\sum_{c'=1}^{C}\sum_{n=1}^{N}\sum_{m=1}^{M} w_{nmpc}} \tag{4.23}$$

and

$$p(m|n, c, \Theta^k) = \frac{\sum_{p=1}^{P} w_{nmpc}}{\sum_{m'=1}^{M}\sum_{p=1}^{P} w_{nmpc}}. \tag{4.24}$$

These are two extremely useful values in understanding the clustering results. The first, $p(c|p, \Theta^k)$, provides the posterior probability for a cluster given a patient and the current parameter settings, which is used for the MAP assignment of a patient to a cluster. The second, $p(m|n, c, \Theta^k)$, is used for each gene's MAP assignment to a gene model within a patient cluster. It is represented by a $M \times 1$ vector that is the mixture proportion for each gene model. When there is only one non-zero value in that vector, that gene has the same label for every patient in the patient cluster. If

all the values were equal, the wave label would be completely inconsistent within the cluster. Therefore, the maximum of $p(m|n, c, \Theta)$ for a given gene and cluster provides measure of consistency that ranges from $\frac{1}{M}$ to 1. As we evaluate the quality of the clustering, measuring consistency of the labelings within the cluster is key.

### 4.3.2 Initialization

Cluster parameters are initialized in the same way as in 3.2.2, using an ML estimate of the parameters for an appropriate subset of the data at each time point. For the mixture proportions:

$$p(c|p, \Theta^0) = \frac{1}{C} \pm \epsilon_1, \tag{4.25}$$

where $\epsilon_1$ is a small random perturbation.

$$p(m|n, c, \Theta^0) = \epsilon_2 \qquad \text{for } m < M, \tag{4.26}$$

$$p(m|n, c, \Theta^0) = 1 - (M-1) \times \epsilon_2 \qquad \text{for } m = M \tag{4.27}$$

where $\epsilon_2$ is the approximate fraction of genes in each wave class.

## 4.4 Results on synthetic data

We first test the ability of the algorithm to label genes and patients simultaneously where the underlying structure of the data is known and is similar to that of the BCR data. In this toy example, we generate expression profiles over 4 time points for 100 genes and 8 patients. The distributions from which the expression profiles are drawn are parameterized by the variance, $\sigma^2_{mt}$ for zero-mean Gaussians, defined at each time point so the shapes roughly correspond to a background class and four wave classes.

To divide the data into two relatively consistent patient groups, the genes for half the subjects are randomly reordered. Now, for example, gene 1 could have been drawn from the wave 1 distribution for half the patients and drawn from the wave 2 distribution for the other half. The resulting data set is shown in Figure 4-3, and no visual structure is evident.

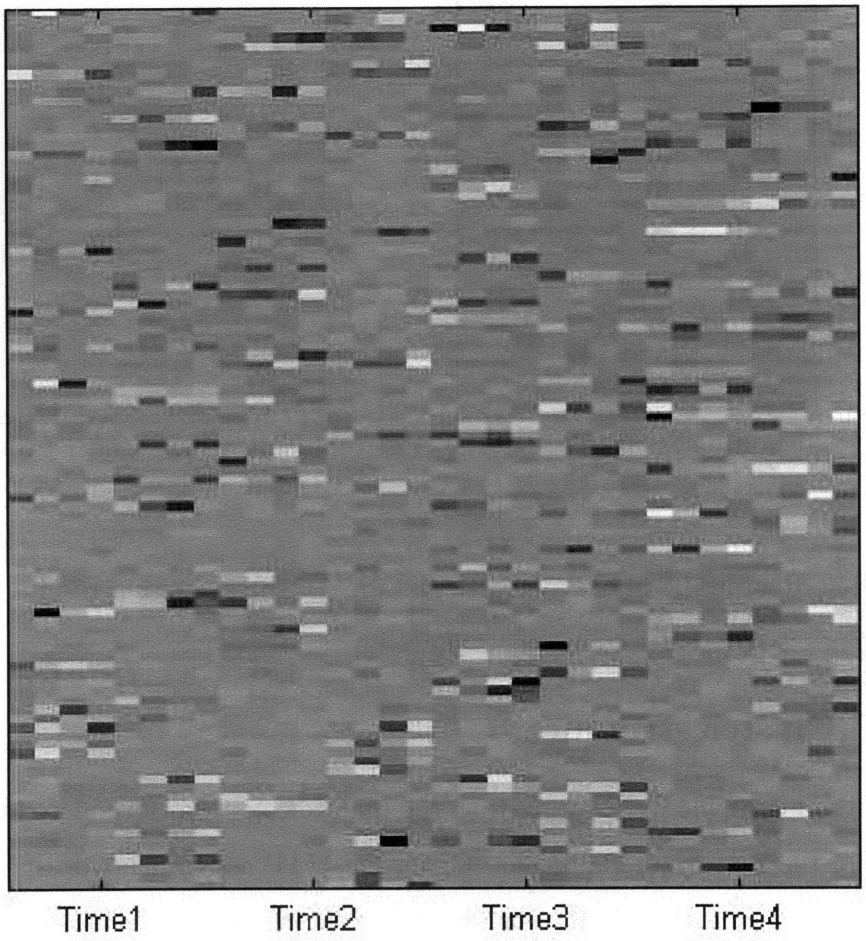



Figure 4-3: Synthetic data for 100 genes and 8 subjects, grouped by time point

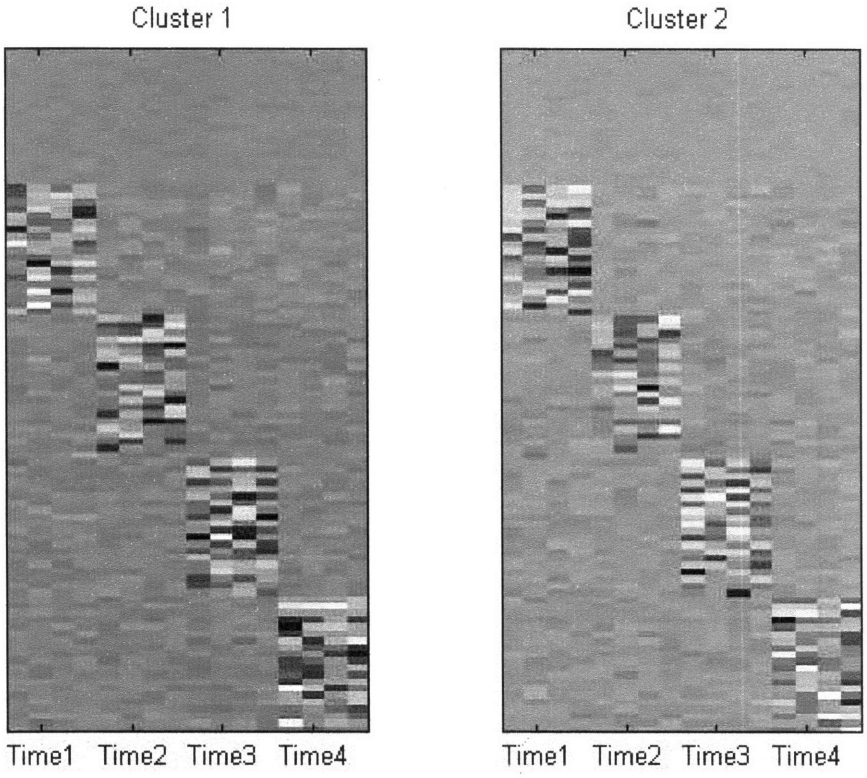



Figure 4-4: Reordering of the synthetic data given the clustering results. The order of the genes is different for clusters 1 and 2.

### 4.4.1 Cluster structure

We run the hierarchical clustering with the update equations formulated to estimate the $\sigma_{mt}^2$ of the zero-mean Gaussians. Given the MAP assignments of the patients to patient classes, and the MAP assignments of the genes to gene models, we reorder the data to show the cluster structure in Figure 4-4.

This shows, for a small example, that hierarchical clustering successfully recovers the structure of the data. Visually, it appears the the patients have been divided into two clusters in which all the genes have been labeled in a meaningful way. In order to confirm the intracluster consistency quantitatively, we compute the maximum $p(m|n, c, \Theta^k)$ for each gene in each cluster. As previously discussed, this ranges from $\frac{1}{M}$ (uniform distribution over all possible labels) to 1 (completely consistently

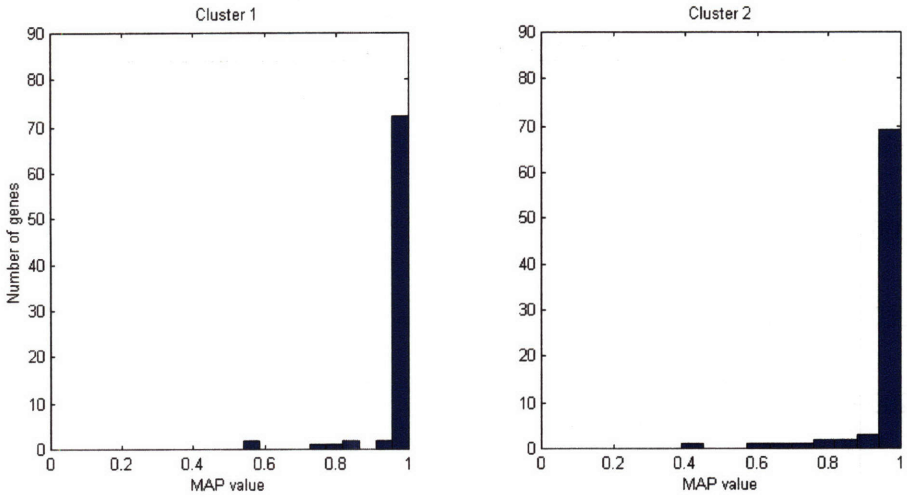



Figure 4-5: Histograms of the maximum $p(m|n, c, \Theta)$ values for each gene in each cluster

labeled). Histograms for the two clusters are shown in Figure 4-5. For nearly all the genes, the labeling is completely consistent within a cluster. The mean values for the distributions of each of the two clusters are 0.9761 and 0.9663, respectively.

### 4.4.2 Cluster separability

In the actual data, we expect some number of genes to be at least as consistent across groups as they are within groups. Not all genes would be expected to behave abnormally in the B-CLL patients, and microarray data is noisy. If the number of genes important to B-CLL is small as compared to the number responding to BCR stimulation, then one would not expect a hierarchical approach to necessarily correspond to clinically relevant subpopulations. In order to determine how separated two groups have to be in order for the clustering to converge to a consistent, accurate result, we vary the fraction of genes that are reordered for the second patient cluster. That is, when we only reorder 10% of the genes, 90% are drawn from the same distribution and should be labeled in the same way in both patient groups. This is obviously a more difficult clustering problem than the previous one in which all the genes were reordered. To measure the success of the clustering, we record the number

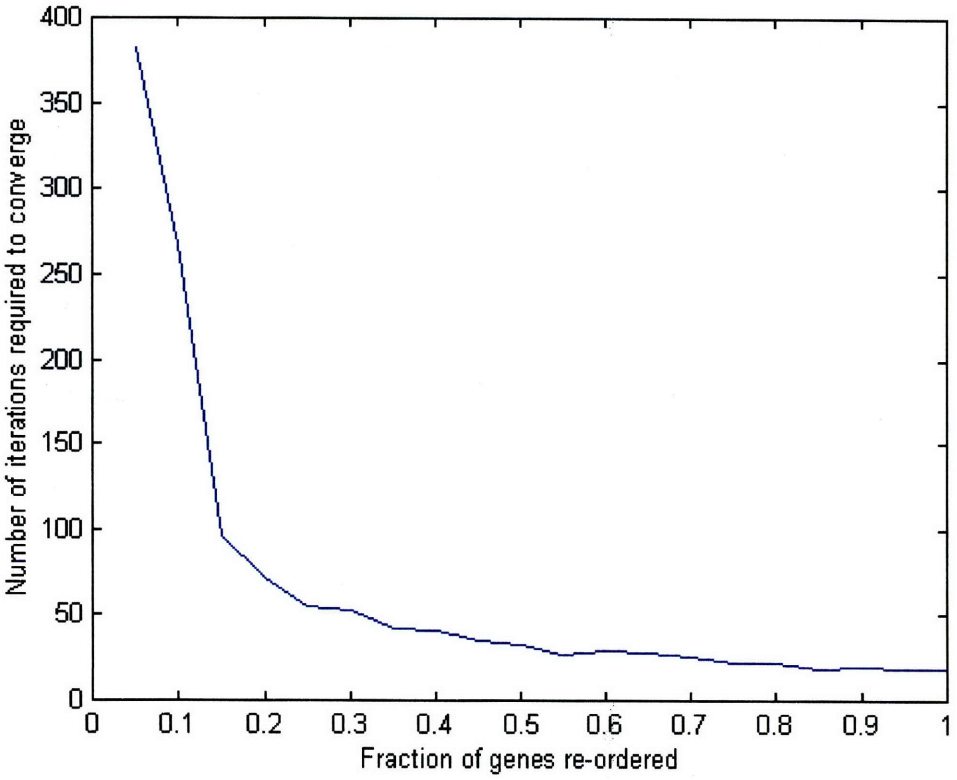



Figure 4-6: Iterations required for hierarchical clustering to converge as a function of the fraction of genes reordered to separate the two clusters

of iterations required to converge to a set criterion. At each fraction, we re-run the algorithm multiple times to average out the effects of initialization.

Figure 4-6 shows that as the fraction of genes reordered increases, the number of iterations required decreases. As expected, as clusters are more separated by increased amount of reordering, the algorithm is able to converge more quickly to the correct result. It is interesting to note, however, that even for as few as 5-10% of the genes differing across clusters, the patient groups are still accurately identified. This means that the algorithm is able to ignore the irrelevant information, and while the actual data is still much more challenging, it does show that this layered approach works when we know that our initial modeling assumptions were correct.

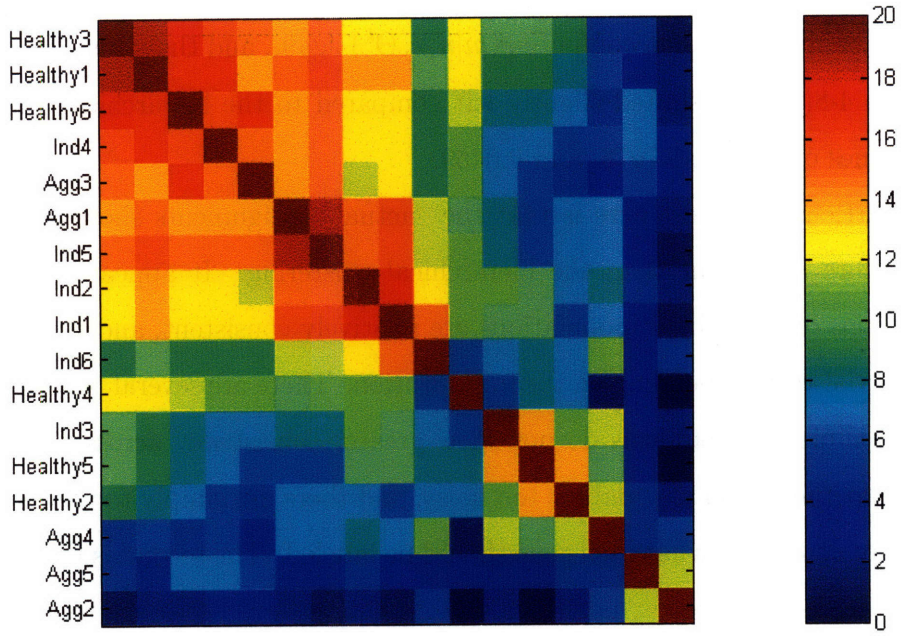



Figure 4-7: Affinity matrix for 17 subjects showing how often each patient pair was clustered together over the 20 trials of hierarchical clustering

## 4.5 Results on BCR data

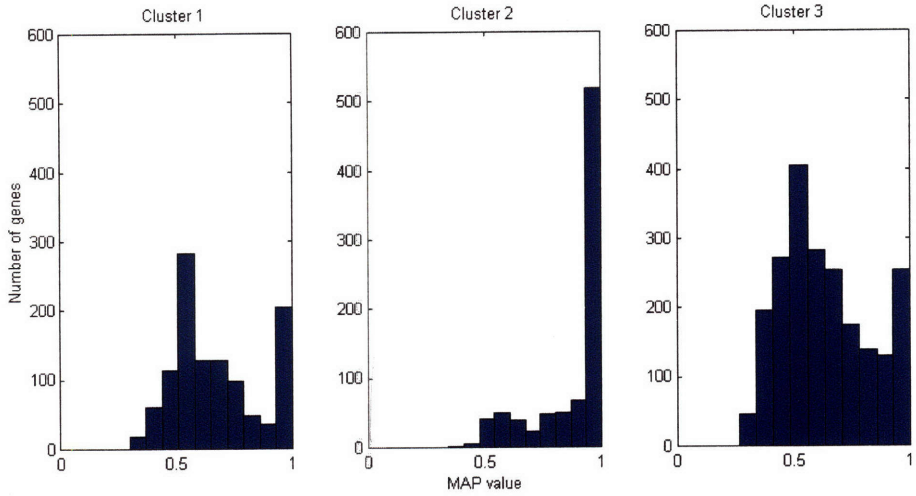
To cluster the data from the BCR data set of 17 subjects, we set $C = 3$, expecting to recover clusters similar to those in 4.2. With shared model parameters and random initialization over 20 trials, we compute how often pairs of patients were clustered together. This is shown in Figure 4-7, where the affinity matrix has been reordered according to pairwise similarity. There is structure; however, as was pointed out in [82], a "successful" clustering result is not necessarily biologically meaningful. In the image, we see a mix of subjects from each of the patient groups in each block of structure. This suggests that the hierarchical clustering found structure in the data that corresponds to these unexpected groups. By looking at the histograms in Figure 4-8(a), we see that there is a reasonable amount of consistency in the gene labels within a cluster. The means are values are 0.6742, 0.8866, and 0.6356 for each of the three clusters. To have a basis for comparison, we re-ran the hierarchical clustering

algorithm with the patient assignments fixed to the Healthy, Aggressive, and Indolent clinical labelings. We then compute $p(m|n, c, \Theta^k)$ again, and these histograms, shown in Figure 4-8(b), are shifted slightly left compared to the hierarchical results, with mean values of 0.6430, 0.7076, and 0.6469.

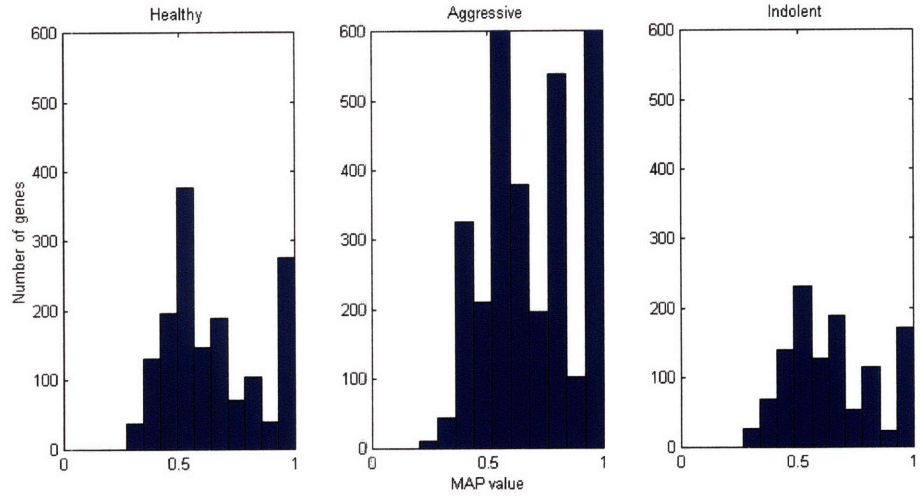
Of particular interest here is that the "actual" assignments show labelings that are no more consistent than the hierarchical clustering. It appears that multiple subdivisions of the subject populations are relatively consistent, and that the results for each trial vary, depending on the initialization. There are several potential reasons that the clinical breakdowns do not show greater consistency across subjects. The first is that the microarray data is too noisy, and there are simply too few subjects to accurately determine what each gene should be labeled. The second is that the groups actually are heterogeneous, i.e. more patient clusters are necessary to represent the data well. Finally, perhaps group differences are not captured by the differences in gene labeling but in difference in the wave model parameters. Unfortunately, again given the very limited number of subjects, when the model parameters are not shared across patient clusters, there is not enough data to support the number of free parameters and clustering fails entirely. For the purposes of this work, this is the only data set of its kind available, and until more subject data is collected, it is impossible to identify which of these potential issues is the reason for the unexpected clustering results. Therefore, in the analysis of the clustering in Chapter 3 and the inference of the predictive models in Chapter 5, we continue to use the clinical labelings.

## 4.6 Summary

In this chapter, we have presented a hierarchical EM approach to a biclustering problem, in this case the simultaneous assignment of genes to wave classes and subjects to patient clusters. While clusters embedded in a synthetic data set are successfully recovered, even with a small fraction of genes that differentiate between the two patient groups, it is less successful for the BCR data set. It does classify subjects into groups that show consistency wave labelings across genes, but these do not correspond to




(a) Hierarchical Clusters




(b) Actual Subject Groups

Figure 4-8: Histograms of the maximum $p(m|n, c, \Theta^k)$ values for each gene in each cluster

the Healthy, Aggressive, and Indolent subgroups. For that reason, we do not use an integrated approach in the clustering of the genes or the inference of the predictive models described in Chapter 5.

# Chapter 5

# Predictive Modeling

A key goal in understanding the behavior of cells and the effects of genetic differences is the reconstruction of networks of interaction. This has become more practical as recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Treating microarray data as a large collection of traditional experiments that simply need to be processed in an automated way ignores one of the key advantages. Global measurements allow us to search out patterns of expression and relationships between genes. By combining confirmatory results, which simply show that microarray analysis is consistent with expected behavior from the literature, and new results, we begin to form a more complete picture of what is occurring in the cell. Due to the complex procedures of microarray experiments, gene expression data often contains a large amount of noise. Additionally, while DNA microarrays enable measurement of the transcripts of every gene, we cannot measure protein expression levels or biochemical activity of gene products. These are, generally, what are actually physically interacting with one another and functioning in the cell. A comprehensive picture of the transcriptional regulatory network would require direct measurements. There is, however, a sensible link between the expression pattern and function of the gene product [9]. Therefore, while predictive models may show relationships between mRNA expression levels, these do not necessarily correspond to actual, physical interactions of RNA sequences or proteins.

There are three major challenges in inferring the predictive networks. The first is that even given the subset of the original data clustered in Chapter 3, there are still many times more samples (genes) than there are observations (time points and subjects). We address this by imposing sparsity on the networks in several ways. The second is that it is not clear how to take into account temporal interactions, which we are able to do because of BCR stimulation and the acquisition of time series data. Finally, evaluating the quality of the resulting models is difficult. In addition to simply visualizing the networks, we perform leave-one-out cross-validation, analysis of network structure, and tests of robustness and statistical significance. Once the models are inferred, we identify genes for literature comparisons and biological validation experiments, which are discussed in Chapter 6.

## 5.1 Background

The most popular method for inferring the relationships between genes is to build coexpression networks. A variety of experimental conditions, gene deletions, or even species profiles [68] are used to determine which genes behave in similar ways. Probabilistic graphical models and Bayesian methods [28, 35, 38, 41, 60] are common. One drawback for much of this work is that while it is effective for small numbers of genes [4, 88, 89] or when the transcription factors are already known [63], it becomes computationally impractical as the number of genes increases. The complete transcription regulatory network has been built for yeast [3, 75], but even for a relatively simple organism like *Xenopus*, networks are still largely unknown [43].

More importantly, a coexpression network as typically viewed is not our goal. Instead of linking genes that behave similarly under set experimental conditions, we attempt to link genes that influence one another across time. Relationships between transcription factors and targets are more complex than coexpression. These relationships may be inverted or time-shifted [85]. Bickel et al.[6] introduce the concept of lag in time to the co-expression network, and more complicated temporal relationships are explored by Guthke et al.[34] and Tabus et al.[71].

Relevance networks inferred by Reis et al. [58] utilize dynamics in gene expression data in addition to static coexpression in order to take into account effector (enhancer/suppressor) type relationships. Darvish et al. [21] relate future expression levels of prototypes of gene clusters to past expression levels of each of the prototypes of each of the clusters to each other. Both of these methods use a yeast gene expression data set with a relatively high sampling rate in time. Yeast is often used as a model organism for studying transcriptional networks because it is easy to manipulate experimentally and has a relatively small number of genes, most of which are functionally annotated. Our BCR data set, however, includes an order of magnitude more genes and subjects from a variety of genetic backgrounds.

## 5.2 Inferring models of interaction

Ideally, we would like to infer relationships between every possible gene pair as:

$$\vec{x}_j = \sum_{i=1}^{N} \mathcal{F}_{ij}(\vec{x}_i) + \eta_j \tag{5.1}$$

where $\vec{x}_j$ is the vector of differential expression levels (each element corresponding to a time point) for the $j$th gene, $\mathcal{F}_{ij}()$ is a function which models the influence of gene $i$ on gene $j$, and $\eta_j$ is an additive noise term. Given limited data, care must be taken in the parameterization of such a model. Even for linear models, allowing independent relationships between every possible gene pairing would require inference of $N^2T^2$ parameters from $N \times T \times P$ data points. We therefore specify a number of simplifying constraints resulting in a tractable inference procedure. While it is reasonable to question any of these choices, they are supported by the subsequent empirical results.

In our model, we divide interactions, represented above by $\mathcal{F}_{ij}()$, into two pieces. The first is the mode of interaction $F_{k(i)k(j)}$, which models relationships of excitation, inhibition, and differences in scale between input and output genes. Because the wave assignments (as described in Chapter 3) classify genes according to temporal structure, we use those labels to limit the number of modes of interaction. We describe

the additional structure of the $F_{k(i)k(j)}$ matrices in detail in Section 5.2.1. Second, $w_{ij}$ represents the relative weight (between 0 and 1) of a prediction for a particular predictor on an output gene.

$$\vec{x}_j = \sum_{i=1}^{N} w_{ij} F_{k(i)k(j)}(\vec{x}_i) + \eta_j \tag{5.2}$$

The weight $w_{ij}$ is re-parameterized as $\alpha_{ij}$ by:

$$w_{ij} = \frac{\alpha_{ij}}{\sum_{i=1}^{N} \alpha_{ij}} \tag{5.3}$$

in order to define the global optimization:

$$\vec{\alpha}_j^*, \{F\}^* = \underset{\vec{\alpha}_j, \{F\}}{\arg\min} \left\| \vec{x}_j - \frac{\sum_{i=1}^{N} \alpha_{ij} F_{k(i)k(j)} \vec{x}_i}{\sum_{i=1}^{N} \alpha_{ij}} \right\|_2 \tag{5.4}$$

subject to the constraints:

$$\alpha_{ij} \geq 0 \tag{5.5}$$

$$\sum_{i=1}^{N} \alpha_{ij} \leq 1 \tag{5.6}$$

The reason for the constraints on $\alpha$ is that we expect that there are a limited number of "input" genes that influence the expression of any one "output" gene, consistent with biological literature. As the $F_{k(i)k(j)}$ matrices already describe the type of relationships over time between genes, e.g. inhibition, excitation, relative scaling, $\alpha$ values are only used to define how strong are those relationships in comparison to all the other possible input genes, so a negative connection would not be defined. The second constraint serves as an $L_1$ penalty imposing sparsity on the number of incoming edges. The most important advantages of $L_1$ penalization schemes are their convexity, and the strong sparsity of the results, in that most indices of the result are set exactly to zero [49]. An equivalent form of these constraints on $\alpha$ is provided in Section 5.2.2.

Global optimization of $\alpha$ and $\{F\}$ simultaneously is non-convex; however, optimizing the two sets of parameters is conditionally convex. If all the $\alpha_{ij}$ are known, solving for $F_{k(i)k(j)}$ requires a straightforward least squares calculation, and when the set of $F$ matrices is known, a solution for $\alpha$ is found using quadratic programming,

as implemented in MATLAB with the SeDuMi toolbox[69]. We therefore use coordinate descent, a commonly used optimization technique, in order to iteratively solve for $\{F\}$ and $\alpha$. Coordinate descent guarantees improvement of the objective criterion on each iteration, and is guaranteed to converge to a stationary point [5]. In Sections 5.2.1 and 5.2.2, we further explain the two steps of the iterative procedure. For this particular application, we found that for all three subject groups, this occurred after approximately 6 iterations, which is what we used for all the analysis described here. Upon completion, we have as output a connectivity matrix $[\vec{\alpha}_1\vec{\alpha}_2\cdots\vec{\alpha}_N]$, which gives us the strength of connection between genes, most of which are zero.

### 5.2.1 Wave interaction matrices

Using the labels assigned to each gene via wave clustering (see Section 3.2.2), we limit the types of interactions between genes by computing common linear predictor models indexed by wave type rather than allowing unique models for each gene-gene pair. The intuition is that temporal structure as defined in the previous section is a predictor of the way in which genes might interact. Consequently, wave 1 genes are not predicted by anything, but are able to influence waves 2, 3, and 4. Similarly, wave 4 genes do not predict any other waves, but are influenced by all other previous waves. These relationships are characterized by six $4 \times 4$ matrices, $F_{12}$, $F_{13}$, $F_{14}$, $F_{23}$, $F_{24}$, and $F_{34}$, where the first index represents the input wave type and the second represents the corresponding output wave type.

Regardless of which wave the gene belongs to, all $F$ matrices are lower triangular, enforcing temporal causality. Dependencies that span the same number of time steps have the same coefficient. This is equivalent to saying that the effects of a given gene will be stationary, influencing other genes in a consistent way over time. For example, if the gene excites the expression of the output at the immediately following time point, this will be true for $t_1$-$t_2$, $t_2$-$t_3$, and $t_3$-$t_4$. The gene may, however, have a longer term inhibitory effect for other time point pairs. Because the time points for this data set were not evenly sampled, these are not quite stationary relationships, but these requirements balance simplicity of the model and our reasonable expectations of

protein and gene interactions with the flexibility required to represent more complex relationships. For a single input-output pair of genes $\vec{x}$ and $\vec{y}$, which are each a $4 \times 1$ vector, the $F$ matrix is the linear transformation:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ a & 0 & 0 & 0 \\ b & a & 0 & 0 \\ c & b & a & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \eta \tag{5.7}$$

As an initial estimate of the models, we group genes by wave type. Given $N_{input}$ genes of the input type and $N_{output}$ genes of the output type, there are $N_{input} \times N_{output}$ possible combinations of these genes that need to be accounted for enumerated as:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N_{output}} \\ y_{21} & y_{22} & \cdots & y_{2N_{output}} \\ y_{31} & y_{32} & \cdots & y_{3N_{output}} \\ y_{41} & y_{42} & \cdots & y_{4N_{output}} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ a & 0 & 0 & 0 \\ b & a & 0 & 0 \\ c & b & a & 0 \end{bmatrix}}_{F_{k(input)k(output)}} \begin{bmatrix} x_{11} & x_{11} & \cdots & x_{1N_{input}} \\ x_{21} & x_{21} & \cdots & x_{2N_{input}} \\ x_{31} & x_{31} & \cdots & x_{3N_{input}} \\ x_{41} & x_{41} & \cdots & x_{4N_{input}} \end{bmatrix} + \eta \tag{5.8}$$

where $X, Y \in \Re^{4 \times (N_{input} \times N_{output})}$ with the genes ordered in such a way as to enumerate every possible pairing. The parameters $a$, $b$, and $c$ are estimated by minimizing the following least squares criterion:

$$\{a, b, c\} = \arg\min_{a',b',c'} \left\| Y - F_{k(input)k(ouput)}(a', b', c')X \right\|_2 \tag{5.9}$$

after reshaping the data into vectors and computing each element of $F_{k(input)k(output)}$ independently. This is equivalent to having all connection strengths $\alpha_{ij}$ set equal to one another, which is used only for the initialization. After the first iteration, using the $\alpha$ vectors resulting from the optimization in Section 5.2.2, we re-weight the matrices according to the strength of the connections, then re-estimate the $F_{k(input)k(output)}$ matrices as:

$$\{a, b, c\} = \arg\min_{a',b',c'} \left\| Y' - F_{k(input)k(output)}(a', b', c')X' \right\|_2 \tag{5.10}$$

where $Y'$ and $X'$ include all possible input and output pairings, scaled by the corresponding $\alpha_{ij}$.

### 5.2.2 Combining predictions

For a given output gene $x_j$, we use a weighted combination of the predictions, represented as:

$$\vec{\alpha}_j^* = \arg\min_{\vec{\alpha}_j} \left\| \vec{x}_j - \frac{\sum_{i=1}^{N} \alpha_{ij} F_{k(i)k(j)} \vec{x}_i}{\sum_{i=1}^{N} \alpha_{ij}} \right\|_2 \tag{5.11}$$

subject to the constraints:

$$\alpha_{ij} \geq 0 \tag{5.12}$$

$$\sum_{i=1}^{N} \alpha_{ij} \leq 1 \tag{5.13}$$

where $k(i), k(j)$ are the wave labels of gene $i$ and gene $j$, respectively, and $\alpha_{ij} \geq 0$ is their connection strength. An equivalent form would include the $L_1$ penalty on $\vec{\alpha}_j$ in the minimization as:

$$\vec{\alpha}_j^* = \arg\min_{\vec{\alpha}_j} \left\| \vec{x}_j - \frac{\sum_{i=1}^{N} \alpha_{ij} F_{k(i)k(j)} \vec{x}_i}{\sum_{i=1}^{N} \alpha_{ij}} \right\|_2 + \lambda \left\| \vec{\alpha}_j \right\|_1 \tag{5.14}$$

still subject to the constraint:

$$\alpha_{ij} \geq 0 \tag{5.15}$$

The relative weighting $\lambda$ of the L1 norm and the L2 norm of the prediction error can also be varied to adjust sparsity as the two are minimized simultaneously. This would be similar to adjusting the threshold on the $L_1$ constraint in the equivalent form in Equation 5.13. For either of these forms, we solve for $\alpha$ using quadratic programming, as implemented in MATLAB with the SeDuMi toolbox[69].

## 5.3 Visualization

Visualization of inferred networks is difficult because of the number of genes and the dynamic nature of the expression profiles and the interactions. We use Cytoscape software [65], and in Figures 5-1, 5-2, and 5-3, we show a default layout in which edges of the graph correspond to springs, creating an attractive force between nodes that are far apart, and a repulsive force between nodes that are close together. Each

node actually represents the entire temporal expression profile of the gene, and each edge represents a $4 \times 4$ matrix $F_{k(input)k(output)}$ and the connection strength $\alpha_{ij}$ between the two genes. As the nodes in the network are arranged in this way, we see that there a small number of genes in each network that have a very large number of outgoing edges. The network structure is described in more detail in Section 5.4.1.

In order to represent the dynamic networks statically, we show in Figure 5.3 the progression through waves of expression, which is related to time. Though a single node still represents the expression profile over 4 time points for that gene, we use the cluster labels of the genes, which corresponds to the time point at which they peak, to show which genes and which interactions dominate at a given time. In the leftmost column Figure 5.3, the wave 2 genes are highlighted in yellow. They are influenced by wave 1 genes only, via $F_{12}$. These connections are shown in red. In the center column of Figure 5.3, wave 3 genes are highlighted, and inputs that are direct (those from wave 2, which is only one time step removed) are shown in red, while inputs that are from wave 1 to wave 3 are shown in blue. Finally, in the rightmost column of 5.3, wave 4 genes are highlighted, with direct connections from wave 3 genes in red, inputs from wave 2 in blue, and inputs from wave 1 in green. Red connections appear to dominate the network, consistent with the assumption that genes are most likely to effect genes in the immediately following wave, even though this was not an explicit constraint. This is shown qualitatively here, but is confirmed quantitatively in Section 5.4.1.

## 5.4 Analysis

As explained in Section 5.1, inferring the networks, while challenging, is merely a starting point. From a mathematical standpoint, the network results can be evaluated in several ways. The first is the structure of the network, which includes the number and types of edges, as well as the parameter values. In addition, we evaluate the robustness of the edges of the network to input noise. Finally, using accuracy with respect to prediction error, we compute statistical significance and leave-one-out cross

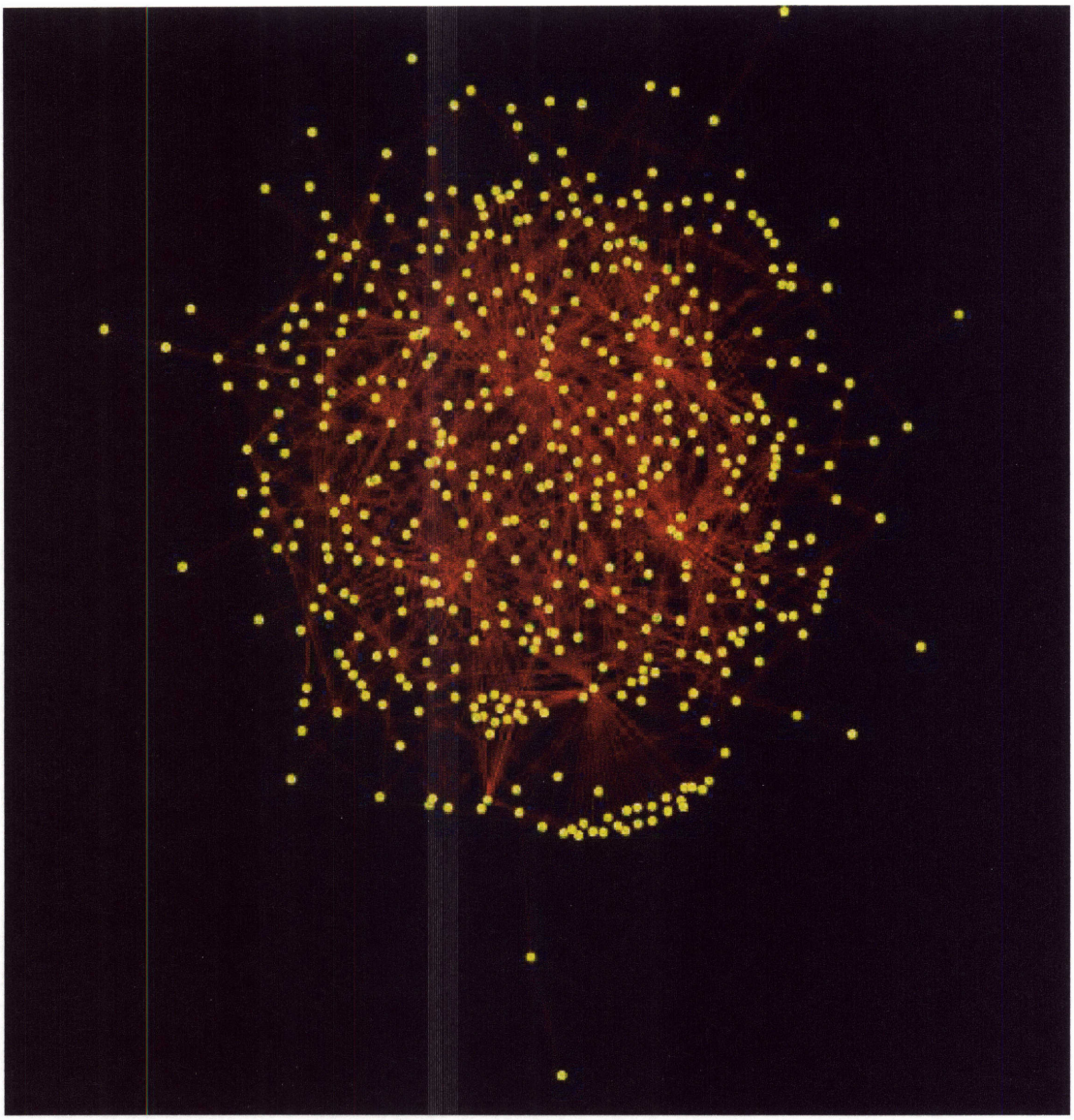


Figure 5-1: Network inferred for healthy subject group

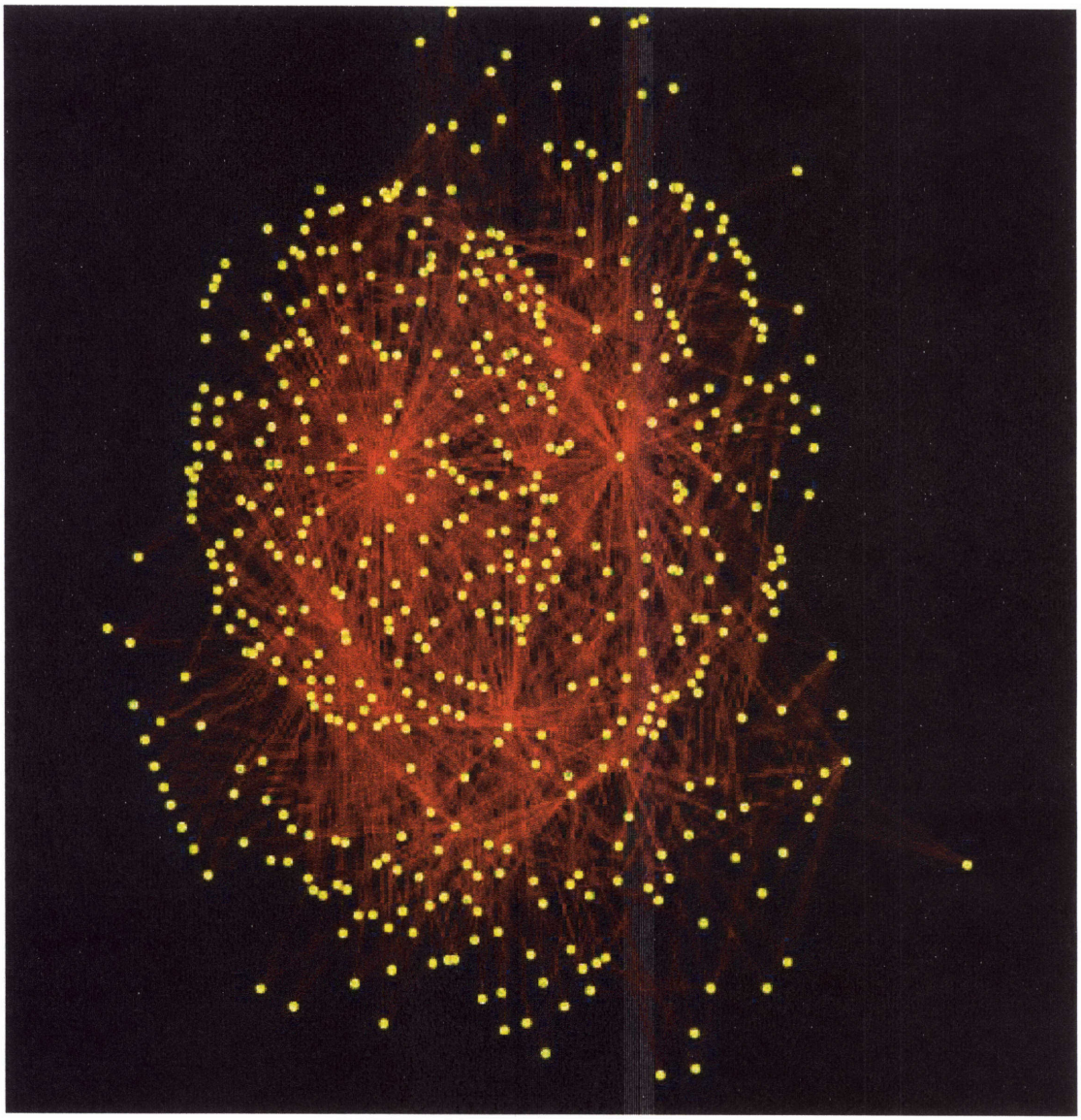


Figure 5-2: Network inferred for aggressive subject group

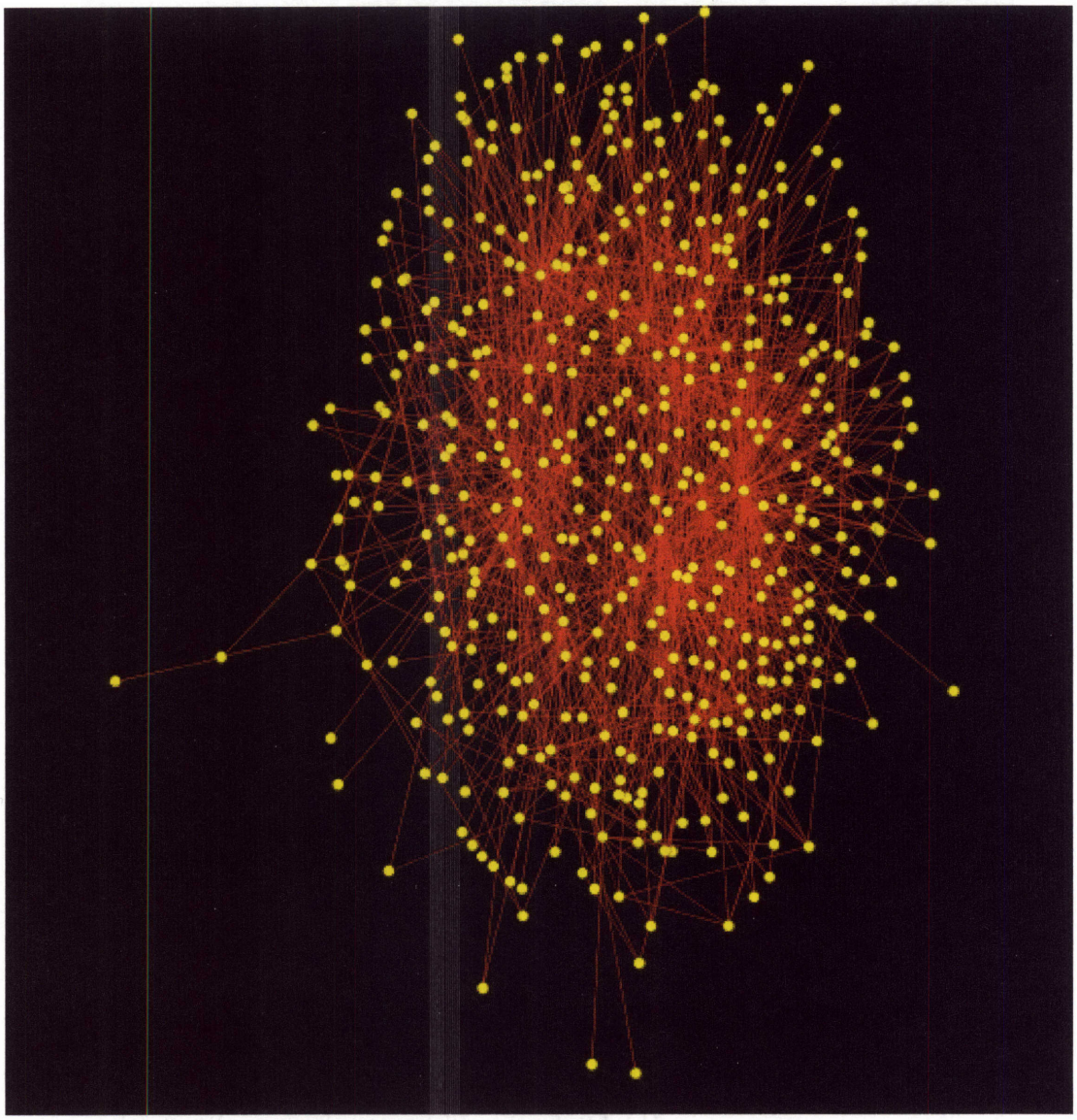


Figure 5-3: Network inferred for the indolent subject group

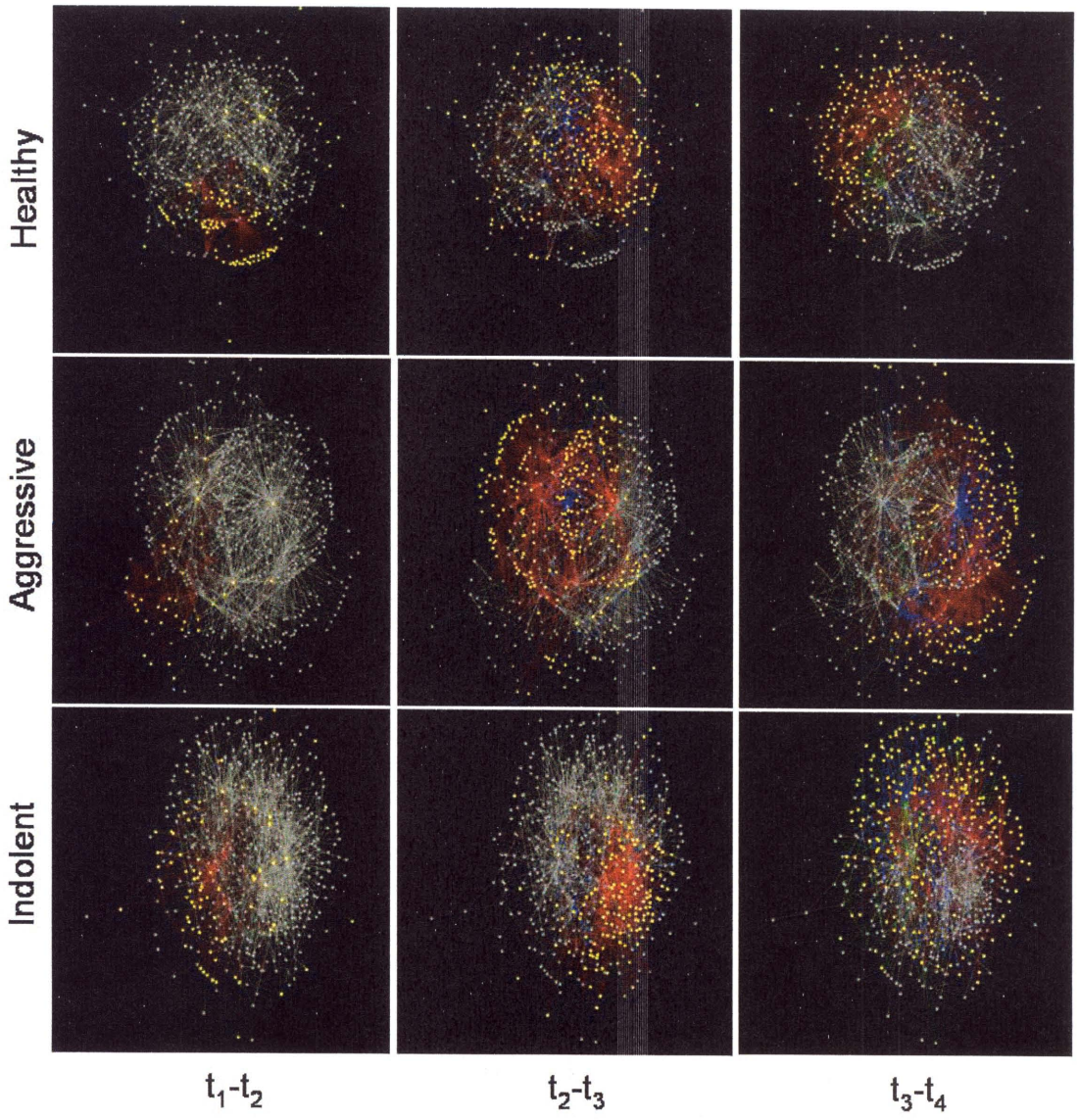



Figure 5-4: In the leftmost column, wave 2 genes are highlighted in yellow. They are influenced by wave 1 genes only, via $F_{12}$. These connections are shown in red. In the center column, wave 3 genes are highlighted, and inputs that are direct (those from wave 2, which is only one time step removed) are shown in red, while inputs that are from wave 1 to wave 3 are shown in blue. Finally, in the rightmost column, wave 4 genes are highlighted, with direct connections from wave 3 genes in red, inputs from wave 2 in blue, and inputs from wave 1 in green.

validation. None of these depend on local validation of the networks from biological experiments, which will be discussed in Chapter 6.

### 5.4.1 Structure

Each edge represents the link between two genes: there is the $F_{k(input)k(output)}$ matrix, which depends on the wave labels of the input and output genes, defining the predictive relationship over time, and there is the $\alpha$ weighting that defines the strength of the connection. We first describe the effects of the iterative method for adjusting the interaction matrices, then quantify the types of interaction matrices that make up the networks, and finally discuss the effects of varying the sparsity constraint on $\vec{\alpha}_j$, the number of edges present.

#### Wave interaction matrices

As described in Section 5.2, there are two methods for computing the $F_{k(input)k(output)}$ matrices, which are wave-specific modes of interaction. The first computes the least squares estimate of the matrix elements weighting all possible input/output pairs equally. The second iteratively adjusts $F_{k(input)k(output)}$ and $\alpha$ according to a coordinate descent method that decreases the prediction error. The prediction error for the network converges with the re-weighting of the input/output pairs according to $\alpha$ in the computation of $F_{k(input)k(output)}$ at each iteration. For all three subject groups, the error stops decreasing after approximately 4 to 6 iterations. There is some variation in when this happens, depending on the subject group, but because of computation time considerations and in order to have a set number of iterations for all further analysis, we use the adjusted $F_{k(input)k(output)}$ matrices and $\alpha$ values after 6 iterations.

We compare the error in Table 5.1. As the $F_{k(input)k(output)}$ matrices are optimized, there is guaranteed to be an improvement in the prediction error scores. For each network, we see about a 15% to 20% decrease in the prediction error. In Figure 5-5(a) and 5-5(b), we compare the values in the $F_{k(input)k(output)}$ matrices before and after the optimization. In the adjusted matrices, $a$, $b$,and $c$ are generally higher than

Table 5.1: Prediction Error with adjusted and unadjusted F matrices

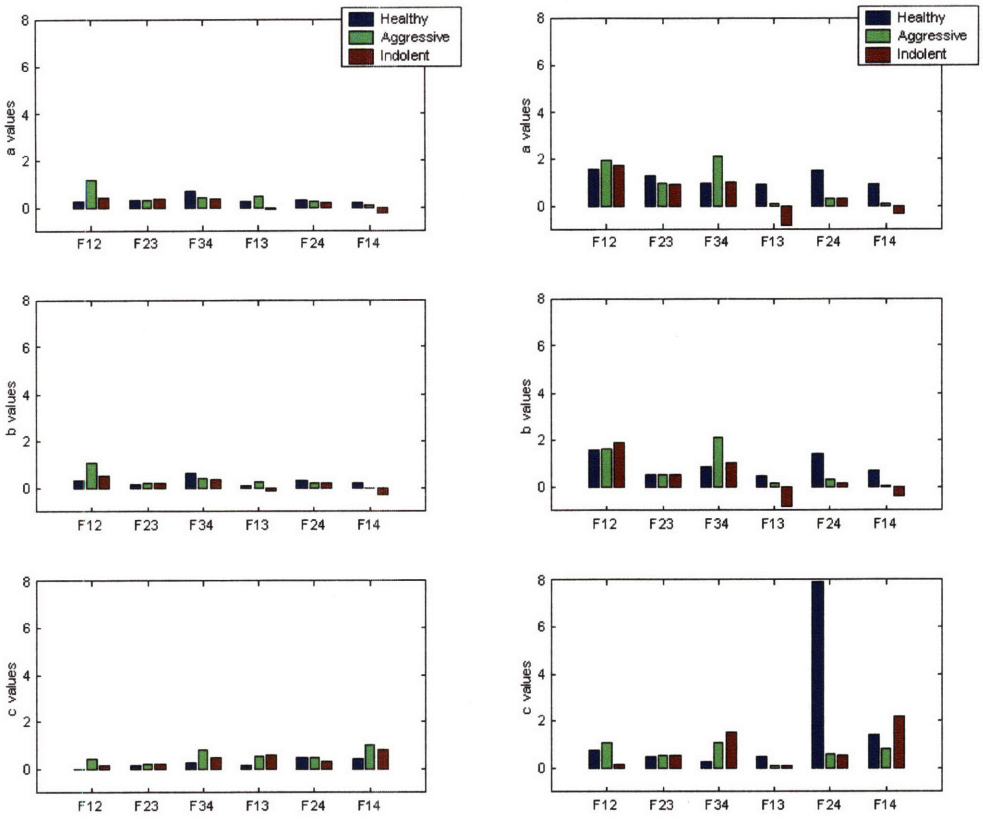
| | Unadjusted F | Adjusted F |
|---|---|---|
| Healthy | 167 | 139 |
| Aggressive | 140 | 108 |
| Indolent | 170 | 135 |

their unadjusted counterparts because the unadjusted values are influenced by input-output pairs of genes that are not related to one another. These pairs will, in general, average one another out, leading to no clear activation or inhibition relationships. During optimization, these gene pairs have less of an impact in the estimation of $F_{k(input)k(output)}$ because of their low connection strengths. Therefore, the larger values of $a$, $b$, and $c$ are from the consistent relationships between the waves. The reason for the very high value of $c$ in $F_{24}$ for the Healthy subject group is unknown.

**Edges**

We see qualitatively in Figure 5.3 that there appear to be many "red" edges. These edges correspond to $F_{12}$, $F_{23}$, $F_{34}$, which represent connections between adjacent waves. We present this numerically as edge counts in Table 5.2, which shows that the connections between adjacent waves (the first three columns of the table) do in fact dominate the network. They comprise 80%, 78%, and 69% of the edges in the Healthy, Aggressive, and Indolent networks, respectively. Interactions are temporally local, which provides evidence of global waves of transcription in the context of synchronized, isolated pathway stimulation.

A second element that was visually evident was the appearance of a small number of genes with a large number of outgoing edges. In Figure 5-6, we show histograms of the number of outgoing edges on a per gene basis. We define these as "hub" genes. While the sparsity constraints on $\alpha$ are imposed to limit the number of *incoming* edges to any one gene, the number of outgoing edges was not constrained. A total of 18 genes (5 for Healthy, 7 for Aggressive, and 6 for Indolent) have more than 50




(a) Unadjusted F matrix parameters

(b) Adjusted F matrix parameters

Figure 5-5: Comparison of the adjusted and unadjusted $F$ matrix parameters

Table 5.2: Edge counts

| | F12 | F23 | F34 | F13 | F24 | F14 |
|---|---|---|---|---|---|---|
| Healthy | 144 | 469 | 380 | 126 | 52 | 69 |
| Aggressive | 117 | 647 | 416 | 88 | 228 | 24 |
| Indolent | 174 | 478 | 410 | 55 | 289 | 138 |

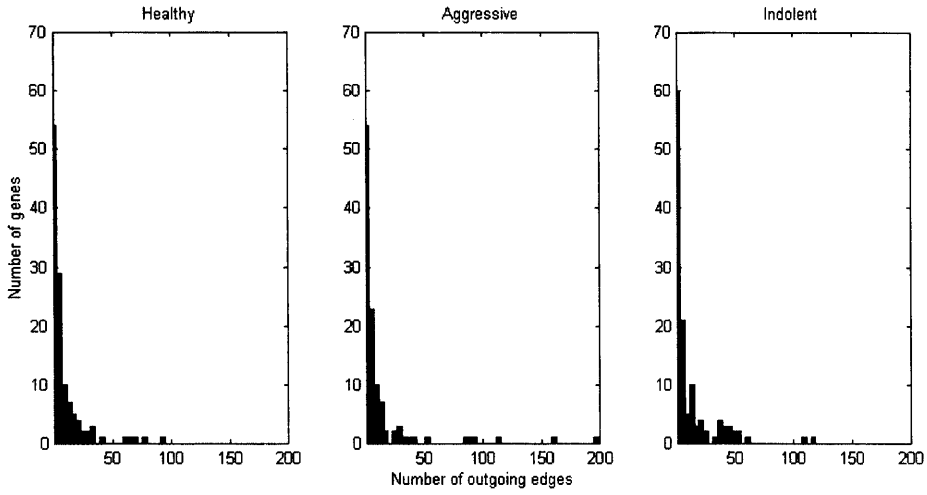



Figure 5-6: Histograms of the number of outgoing edges on a per gene basis

outgoing edges. More than 90% of the genes have fewer than 10 outgoing edges.

In Chapter 6, we discuss what is known biologically about these genes and their targets, as well as further experiments to validate the related parts of the networks. This scale-free structure, however, is consistent with has been shown previously for B-cell networks [4] and for transcriptional regulatory networks in general [8].

### Varying sparsity constraints on alpha

We set the sparsity constraint on $\vec{\alpha}_j$ such that the $L_1$ norm was $\leq 1$ in Section 5.2. We made this choice because we assume that all inhibition, excitation, and scaling effects between input and output pairs are modeled with the $F_{k(input)k(output)}$ matrices, and $\alpha_i j$ values simply provide connection strengths. These connection strengths can also be thought of as weighting the average contributions of the input genes. For example, if one gene $i$ is able to perfectly predict another gene $j$ through the $F_{k(i)k(j)}$ matrix, without any contribution from any other input genes, then $\alpha_{ij} = 1$ and all other values in the $\vec{\alpha}_j$ are zero. In this case, a stronger constraint on the $L_1$ norm (a lower threshold) would erroneously favor input genes that have larger predictions, instead of those predictions which are on the same scale as the output gene.

We have, however, effectively adjusted this sparsity constraint up and down using

Table 5.3: Significance

| | Prediction Error | Min error | $p$-value |
|---|---|---|---|
| Healthy | 139 | 163 | $< 0.01$ |
| Aggressive | 108 | 129 | $< 0.01$ |
| Indolent | 135 | 167 | $< 0.01$ |

the equivalent form in the optimization. As expected, as a more severe constraint is imposed, the number of edges decreases. It would allow for greater control over the sparsity of edges in the network if that is what is desired for another application.

### 5.4.2 Significance results

We test the statistical significance of the models that have been inferred using permutation tests [56] in a method similar to the one in [6], where the false detection rate of a connection between two genes in the model is calculated. We first define a null hypothesis: there is no temporal structure in that data that can be represented with predictive models. To generate samples from this hypothesis, for each trial, we randomly permute the time points for each gene independently (but consistently within a patient group). Accurate $p$-values necessitate re-running the full optimization (including wave clustering) because wave labels depend on the temporal structure as well. While estimating $\alpha$ for the predictive models, the $L_2$ norm of the prediction error is minimized for each output gene. We sum over all the genes to get an aggregate score for the network and the current data set for the patient group. The statistical significance of our hypothesis that there is temporal structure in the data is defined as the fraction of trials in which the permuted data has a better score than the actual data.

In Table 5.3 and Figure 5-7, for all three subject groups, the error is lower in all 100 trials given the original data for a $p$-value of $< 0.01$. This demonstrates that labeling, model constraints, and connectivity constraints are not arbitrary. We

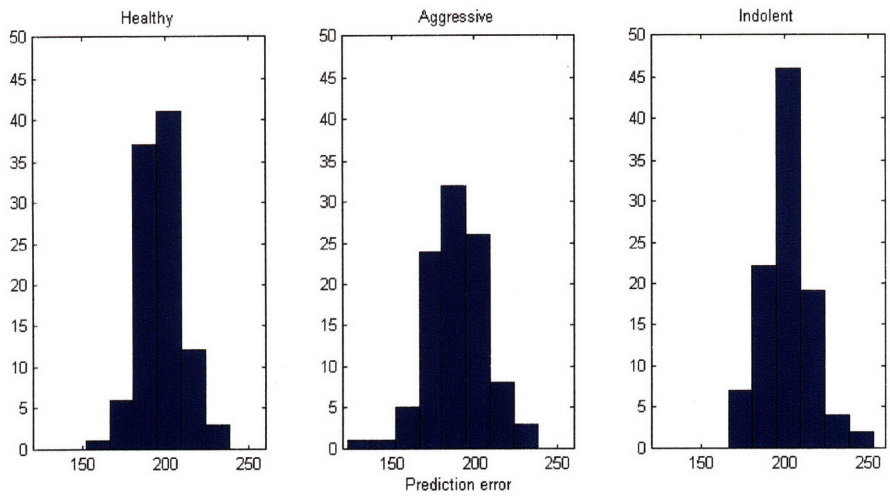



Figure 5-7: Permutation histograms of prediction error for networks

conclude therefore that temporal structure exists and that it can be captured to some degree by a small number of temporal structures and interaction types.

### 5.4.3 Stability

We test the stability of the edges in the network in order to show the robustness of the inference procedure to noise for the networks in general and for specific subregions. We model the effects of noisy microarray data by adding zero-mean Gaussian perturbations (ranging from $\sigma$=10 to $\sigma = 40$, about an order of magnitude smaller than the positive deflections of wave 1 genes) to the input data. For 25 trials, the networks are re-estimated using the same subset of genes that had been previously identified and labeled by the wave clustering. In Figure 5-8, we show "survival curves" for consistent edges. As a function of the number of trials, these curves are computed from the fraction of edges in the original network that are present. Virtually 100% of the edges each of the original networks are present in at least 1 trial. However, anywhere from 15% (Healthy, $\sigma = 40$) to 80% (Aggressive, $\sigma = 10$) are present in all 25 trials. The Aggressive subject group seems to be more consistent than the other two groups, likely because there are only five patients in the group as opposed to six for the other two groups.

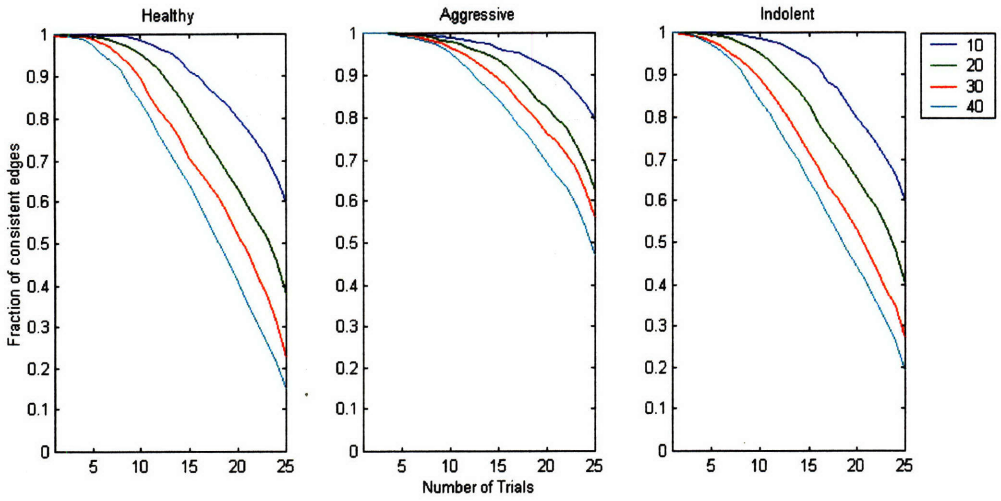



Figure 5-8: Stability of the edges for all three networks with differing levels of noise added to the input data.

We show these figures to give a general context for how the entire networks behave, so we can compare these results to those for subnetworks. We identified in Section 5.4.1 a small number of "hub" genes that have a large number of outgoing edges, which we expect to play an important role in the networks. In Figure 5-9, we display survival curves for the outgoing edges of the hub genes compared to the average curves for the networks ($\sigma = 20$, which is a representative result for the other values of $\sigma$). In nearly all cases, the outgoing edges of the hub genes are more consistent than the averages. Because the outgoing edges associated with the hub genes are more stable than the networks as a whole, they are less likely to have been produced at random, and therefore provide a starting point for biological validation of the networks.

### 5.4.4 Cross validation

In this section, we perform leave-one-out cross validation to test the consistency of the networks within and across subject groups. We use a single subject from the original data set as the testing data, and the remaining observations as the training data. We repeat this such that each subject is used once as the testing data, held out from the inference of its own type of network (healthy, aggressive, or indolent). For

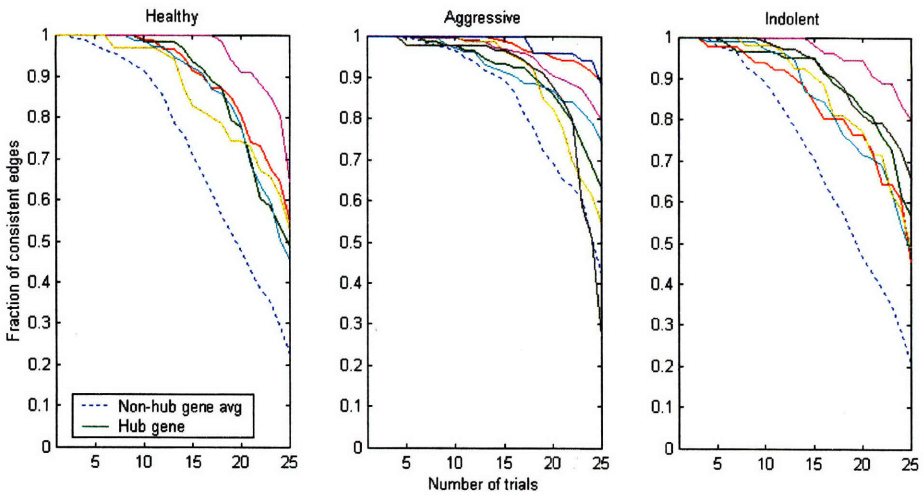



Figure 5-9: Stability of the edges for the hub genes as compared to the non-hub genes.

example, for subject Healthy1, the Healthy network was inferred using the data from subjects Healthy2 through Healthy6. Using the input data for Healthy1, we compute the aggregate prediction error for that network as well as the prediction error for the network inferred from all the Aggressive subjects and the network inferred from all the Indolent subjects. The results are shown in a bar plot in Figure 5-10.

For four of the six healthy subjects, the healthy network has the lowest prediction error. For three of the five aggressive subjects, the aggressive network has the lowest prediction error. For five of the six indolent patients, the indolent network has the lowest prediction error. If we were to use the cross validation results to classify subjects into groups, it would be marginally successful. This implies that even within groups, there is inconsistency across patients. Some of the inconsistency can be attributed to the noisiness of the microarray data, but we also expect that genetic differences play a role. Particularly for the B-CLL patients, as was explained in Section 2.1.2, groupings based on disease outcome or on various genetic mutations yield different results. Because our data set includes such a small number of patients, outliers are not only not well-classified but also bias the networks in which they are included.

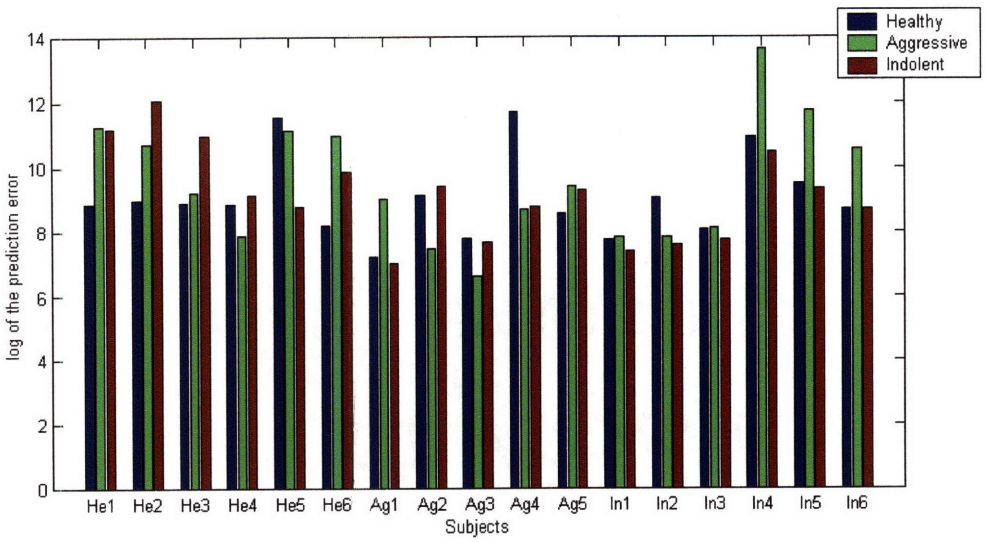



Figure 5-10: Error measures for leave one out cross-validation, as well as for the error across the other two models

## 5.5 Summary

In this chapter, we have introduced a method for inferring sparse models of temporal interaction for genes relevant to the BCR signaling pathway using the wave cluster labels from Chapter 3, which have a consistent biological interpretation, to define a small number of modes of interaction between the genes and by effectively limiting the number of genes influencing each target gene. In order to represent the dynamic networks statically, we show in Figure 5.3 the progression through waves of expression, which is related to time. We first describe the effects of the iterative method for adjusting the interaction matrices, then quantify the types of interaction matrices that make up the networks, and finally discuss the effects of varying the sparsity constraint on $\vec{\alpha}_j$ the number of edges present. We test the statistical significance of the models that have been inferred using permutation tests. This demonstrates that labeling, model constraints, and connectivity constraints are not arbitrary. We conclude therefore that temporal structure exists and that it can be captured to some degree by a small number of temporal structures and interaction types. Also, in nearly all cases, the outgoing edges of the hub genes are more stable than the

averages of the networks as a whole. We perform leave-one-out cross validation to test the consistency of the networks within and across subject groups.

# Chapter 6

# Biological Implications

In Chapters 3 and 5, we presented the results of clustering and predictive model inference from a mathematical perspective. We have shown that the temporal structure of the data has been successfully captured, but the purpose of the modeling was also to provide insight into the genetic program of the B cell receptor signaling pathway. The ultimate objective is to use the network to design different approaches for affecting the evolution of the gene activity profile of the network [25]. In a disease-specific sense, analysis of the differences in the networks between healthy subjects and B-CLL patients could eventually guide potential treatments.

Therefore, from a biological standpoint, we first validate the networks that were inferred and then use the networks to design future experiments. Validation is done in three ways: 1) comparisons of the genes that play important roles in the network to those that have been previously determined to be relevant to BCR activation or B-CLL, 2) comparisons of the links in the network to known physical interactions, and 3) most importantly, intervention experiments that test network predictions by overexpressing or silencing a gene. This third method distinguishes between causality and correlation in local regions of the network. For example, if two genes are regulated by a common factor, with its effects on one lagging in time as compared to the other, the former may appear in the network as having been influenced by the latter (either in addition to or instead of the common factor). Intervention experiments tease out these differences, as only genes that are actually influenced by the manipulated gene,

either directly or indirectly, will show substantial and predictable changes as compared to the control. Because of the relatively large number of genes in the networks (500 per subject group), one important issue is still gene selection and the question of how we can take advantage of the structure of the networks and the cluster labels to identify genes to target.

We also compare the clustering results across the three subject groups. As was explained in Chapter 4, there is not adequate consistency within the current subject groups to combine clustering of genes and subjects in an integrated approach, which was the reason we used the known clinical labelings (Healthy, Aggressive, Indolent). We do, however, analyze aspects of the clustering results to determine how cluster labels relate to subject group and to combine the networks into a single visualization.

## 6.1 Network structure

While sparsity constraints imposed during the predictive model inference procedure (see Section 5.2.2) limited the number of *incoming* edges to any one gene, the number of *outgoing* edges was not constrained. In Figure 6-1, histograms of the number of outgoing edges per gene show that the inferred networks have a scale-free structure. This is consistent with what has been found in protein interaction networks [8], where a small number of proteins serve as hubs for very large numbers of interactions. A transcriptional regulatory network, which was inferred from a large set of expression profiles that represented perturbations of B cell phenotypes, shows similar structure [4].

For each of the three subject groups, more than 90% of the genes have fewer than 10 outgoing edges, and 18 genes (5 for Healthy, 7 for Aggressive, and 6 for Indolent) have more than 50. We define these as "hubs". From Section 5.4.3, when perturbing the input data with noise, the edges related to hub genes are more consistent than those in other regions of the networks. These hubs appear to not only be important to the network predictions, but their connections to their targets are relatively stable as well. They therefore provide a starting point as candidates for literature comparisons

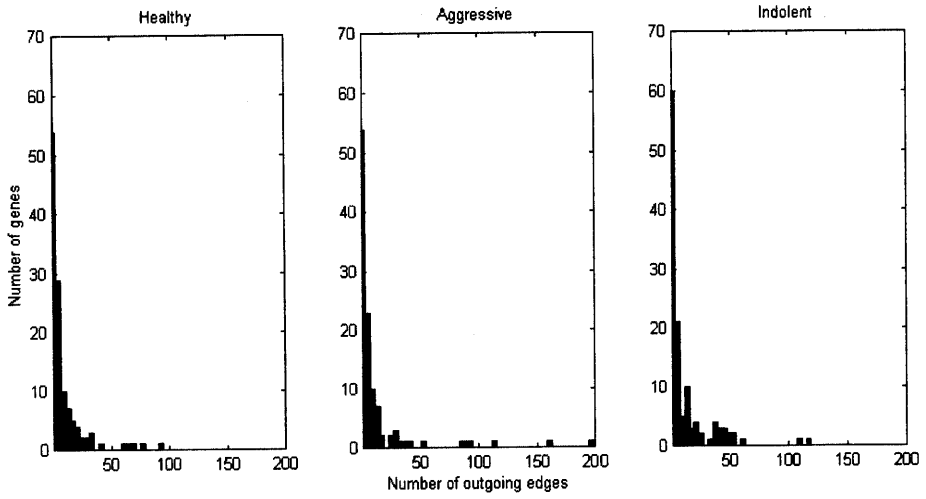



Figure 6-1: Histograms of the number of outgoing edges on a per gene basis

and for further experiments. In the following section, we investigate these highly connected genes.

## 6.2 Hub genes

The probe set labels for the 18 network hubs represent 15 known genes. Of these, *EGR1*, *NR4A1*, and *ZFP36* are already known to be expressed after BCR stimulation [64, 52]. Those three genes, in addition to the hubs *BTG2*, *CXXC5*, and *NR4A3*, are transcription factors [51], which by definition are genes that influence others. Each transcription factor stimulates the transcription of target genes through the binding to specific DNA regulatory element located at the beginning of each target gene. The number of possible binding sites through the genome is different for each transcription factor, ranging anywhere from fewer than 200 to more than 10,000, depending on the network [79]. It is important to note that because the networks are only *inferred* relationships and not necessarily actual physical interactions, identification of hubs in the network that are known transcription factors is a confirmatory result.

Table 6.1 shows the complete list of the probe set labels for the hubs identified in the networks, with the threshold on the minimum number of outgoing edges set at 50, for each of the three subject groups. We also include the gene name [80], if

it is known. In addition to the transcription factors (in bold), these genes code for motor proteins (*DNAH3*), ribosomes (*RPL10A, RPL15, RPL27A, RPL37A, RPS27*), transport proteins (*SLC2A3*), and enzymes (*SAT*) [48].

As we have established that a number of the network hubs are transcription factors which could potentially have influence on target genes, we compare individual links in the network to potential physical interactions. For other BCR-related subnetworks, Basso et al. [4] validate around 40% of inferred connections using chromatin precipitation (chIP-chip) experiments, so we do not expect perfect correspondence. In our case, the links in the network could be a result of the modeling decisions made to make inference tractable, artifacts of noise in the data, or could appear to be direct connections when the gene actually acts on the target indirectly through an intermediate transcription factor. As there is not yet ground truth for all possible interactions, regions of the network are analyzed manually. We begin with the subnetwork associated with *EGR1*, a hub in the Healthy network with 93 outgoing edges. The probe sets corresponding to these edges represent 76 different known genes. We compare the inferred links in the network to actual potential physical interactions by searching for the *EGR1* binding site in the promoter region of these putative target genes using the TESS database [62]. Of the 93 potential target probe sets, 37 have a binding motif for *EGR1*. This confirms that *EGR1* may in fact be acting directly on some of the targets to which it is linked in the network, despite having inferred the relationships based on only the mRNA expression profiles.

As this analysis is completed for the remaining hub genes, we will lower the threshold for outgoing edges in order to identify more genes for further research and experimentation. Figure 6-2 shows how the number of genes that would be selected changes as a function of the threshold. Gene selection is an important problem in microarray analysis, given that it is simply not possible to test every possible interaction experimentally, so our results guide experimental design based on which genes appear to be related statistically. For the hubs in particular, we have greater confidence in the stability of those local regions of the networks and the potential for physical interactions, which we can then provide to the biologists to investigate in a systematic way.

Table 6.1: Hub probe set labels and corresponding gene names (transcription factors in bold)

| Probe Set | Name |
|---|---|
| 220725_x_at | DNAH3 |
| 201693_s_at | **EGR1** |
| 216236_s_at | SLC2A3 |
| 229563_s_at | RPL10A |
| 230333_at | SAT |
| 200741_s_at | RPS27 |
| 201236_s_at | **BTG2** |
| 203034_s_at | RPL27A |
| 214041_x_at | RPL37A |
| 222145_at | unknown |
| 224516_s_at | **CXXC5** |
| 227404_s_at | **EGR1** |
| 216979_at | **NR4A3** |
| 227224_at | RALGPS2 |
| 221475_s_at | RPL15 |
| 201531_at | **ZFP36** |
| 202340_x_at | **NR4A1** |
| 238276_at | unknown |

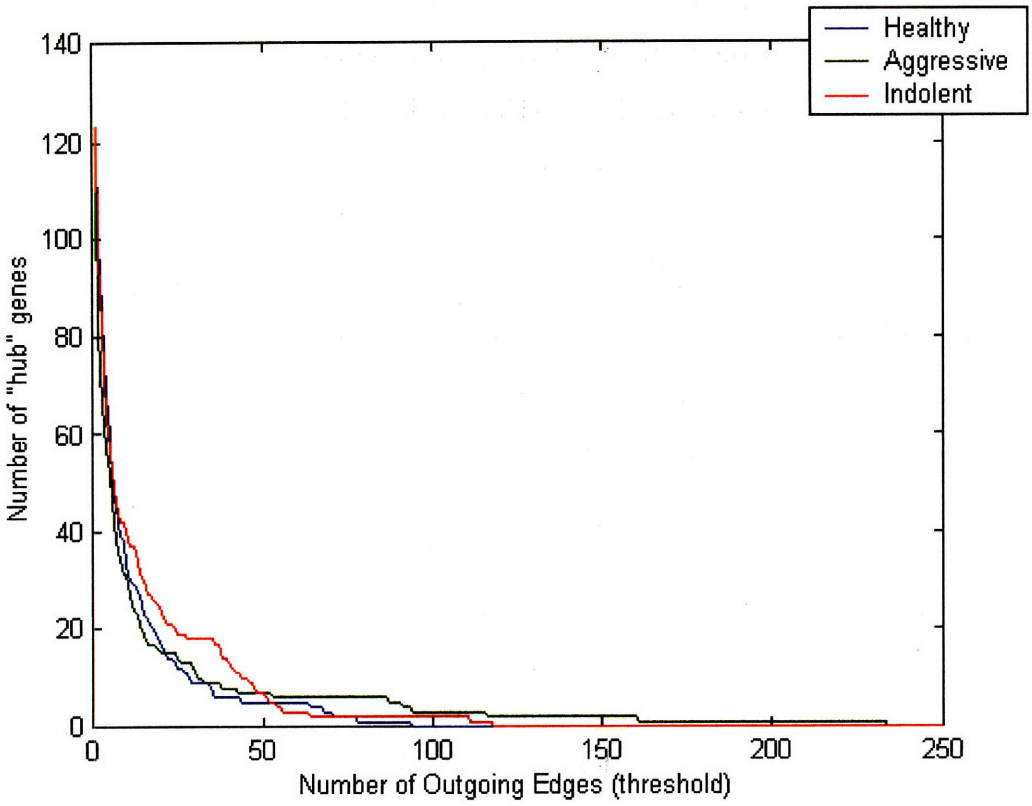



Figure 6-2: Number of hub genes selected given the threshold on the number of outgoing edges

## 6.3 Sensitivity analysis

Once the networks have been inferred, testing the predictions is the obvious next step. One way to do this is to manipulate the expression levels of a single gene, either by silencing or overexpression, in an intervention experiment. These intervention experiments are expensive and time-consuming to perform because as we test temporal interactions, time series acquisitions are necessary. Therefore, much thought must go into selecting the target gene. It should peak early in the BCR activation response, so that it will have downstream effects in the cascade of reactions that can be measured over the time course of the experiment. It also should be expected to influence many

other genes, so model predictions can be compared to actual measurements. Identification of genes with these two properties are precisely what wave cluster labeling and the inference of the predictive models provide. First wave genes (as labeled in Chapter 3) are, by definition, those that peak early in the response. Hub genes in the networks (as inferred in Chapter 5) have many direct connections, presumably to target genes they influence. However, in order to be more general, we take into account both the direct and indirect connections in this section.

In order to identify potential target genes, we propagate the effects of manipulation of each of the first wave genes for all three of the networks. This is more complex than simply computing all the predictions, as we also include the indirect effects. For example, a first wave gene may influence second, third, and fourth wave genes directly, but the second and third wave genes that it influences may also have connections to other third and fourth wave genes. More precisely, for a first wave gene $\vec{x}_i$, second wave genes $\vec{x}_j$ are connected only directly:

$$\vec{x}_j = \alpha_{ij} F_{12} \vec{x}_i \tag{6.1}$$

Third wave genes $\vec{x}_k$ are connected directly to $\vec{x}_i$ via the interaction matrix $F_{13}$ and indirectly through second wave genes:

$$\vec{x}_k = \alpha_{ik} F_{13} \vec{x}_i \tag{6.2}$$

$$+ \sum_{j \in w2} \alpha_{jk} F_{23} \alpha_{ij} F_{12} \vec{x}_i \tag{6.3}$$

Fourth wave genes $\vec{x}_l$ are connected directly to $\vec{x}_i$ via $F_{14}$ and indirectly through second wave genes, third wave genes, or both:

$$\vec{x}_l = \alpha_{il} F_{14} \vec{x}_i \tag{6.4}$$

$$+ \sum_{j \in w2} \alpha_{jl} F_{24} \alpha_{ij} F_{12} \vec{x}_i \tag{6.5}$$

$$+ \sum_{k \in w3} \alpha_{kl} F_{34} \alpha_{ik} F_{13} \vec{x}_i \tag{6.6}$$

$$+ \sum_{k \in w3} \alpha_{kl} F_{34} \sum_{j \in w2} \alpha_{jk} F_{23} \alpha_{ij} F_{12} \vec{x}_i \tag{6.7}$$

Given these equations, we predict the effects on all $\vec{x}_j$, $\vec{x}_k$, and $\vec{x}_l$ as $\vec{x}_i$ is manipulated. In the intervention experiments, there is little control over how the target is manipulated beyond overexpression (an increase) or silencing (a decrease) for the entire time course, so we model this simply as unit change $\vec{x}_i = [1111]^T$. The results of the changes in all the genes in the network are calculated, and we display the predicted change in the expression profiles over time in Figures 6-3, 6-4, and 6-5.

In a qualitative way, we select those genes to which the network is most sensitive to changes. These are the probe sets corresponding to the columns in Figures 6-3 through 6-5 which have the largest effects, either in terms of numbers of genes influenced or magnitude of change. These are listed in the Table 6.2. While there is some overlap with the previously identified hubs (identified with asterisks), not all hubs are first wave genes and not all genes to which the network is sensitive have $> 50$ outgoing edges. Three of these genes (*FOS, IER2, JUN*) were found previously to be expressed by inducing early response genes with BCR activation [52], which shows that the genes that play an important role in the interaction networks also are relevant to the BCR signaling pathway. The results of the sensitivity analysis, which identify those genes that peak early in the time course and affect many downstream genes, guide the design of intervention experiments to test the predictive ability of the networks. We provide a visualization of a small set of genes for the biologist to evaluate and then to select for the time-consuming and expensive gene silencing or overexpression. Much like the testing of binding sites and potential physical interactions related to hub genes, the selection of targets for intervention is a key step in the process. Though these results are specific to BCR stimulation and B-CLL, this type of analysis could be used for other applications in order to systematically analyze the effects of manipulating regions of a network. In the following subsection, we evaluate the results of one intervention experiment, the silencing of *DUSP1*.

### 6.3.1 *DUSP1* Silencing

After BCR stimulation, we identified increased expression of *DUSP1* (*MKP1*), known to be involved in cell cycle regulation [45] and the inhibition of which potentiates *JNK*-

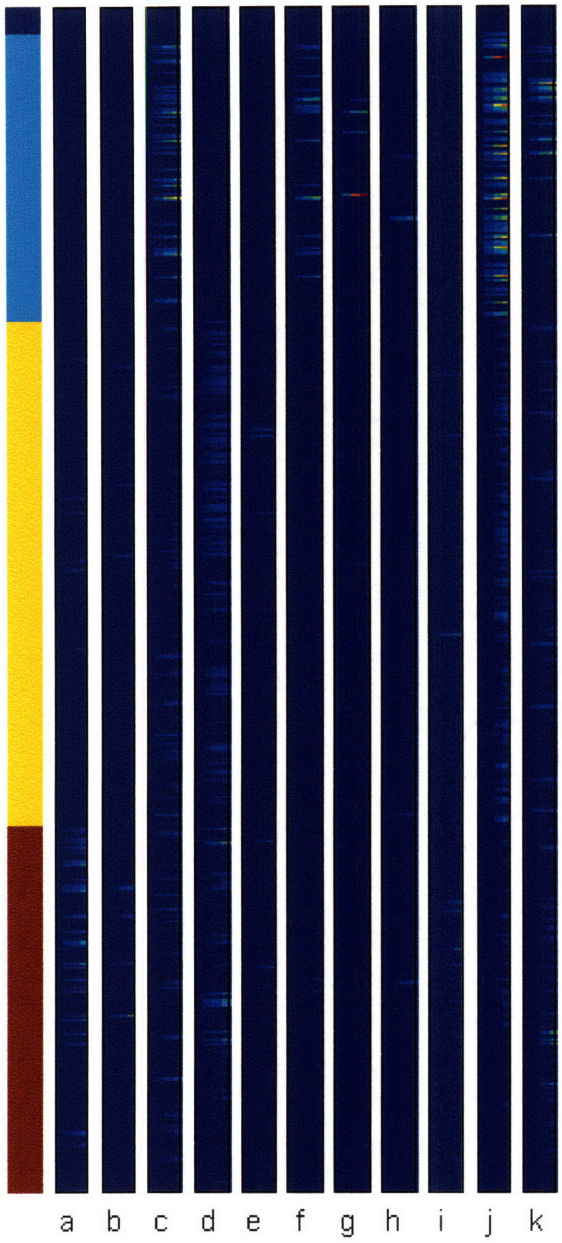



Figure 6-3: For the healthy network, the effects of uniformly perturbing each of the first wave genes on all other genes in the network over time. The leftmost column corresponds to the wave label.

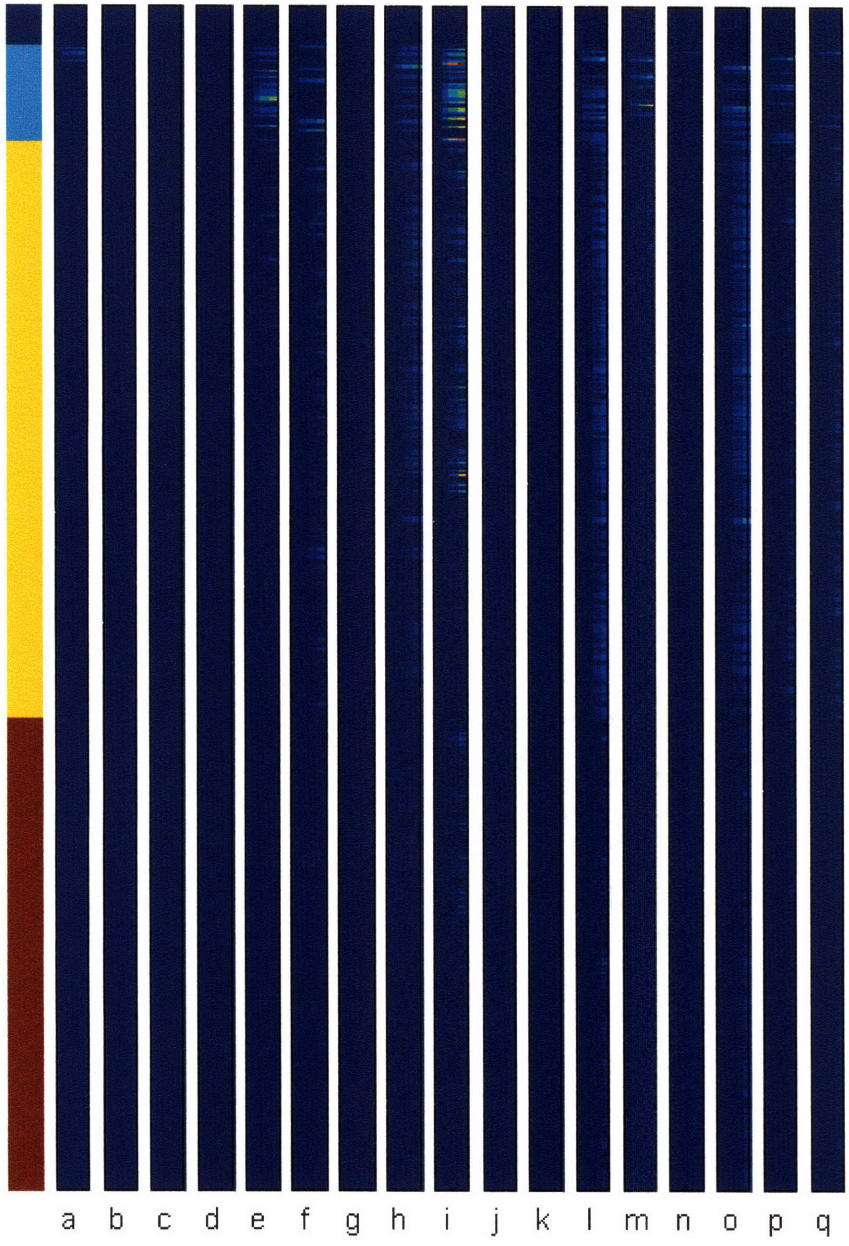



Figure 6-4: For the aggressive network, the effects of uniformly perturbing each of the first wave genes on all other genes in the network over time. The leftmost column corresponds to the wave label.

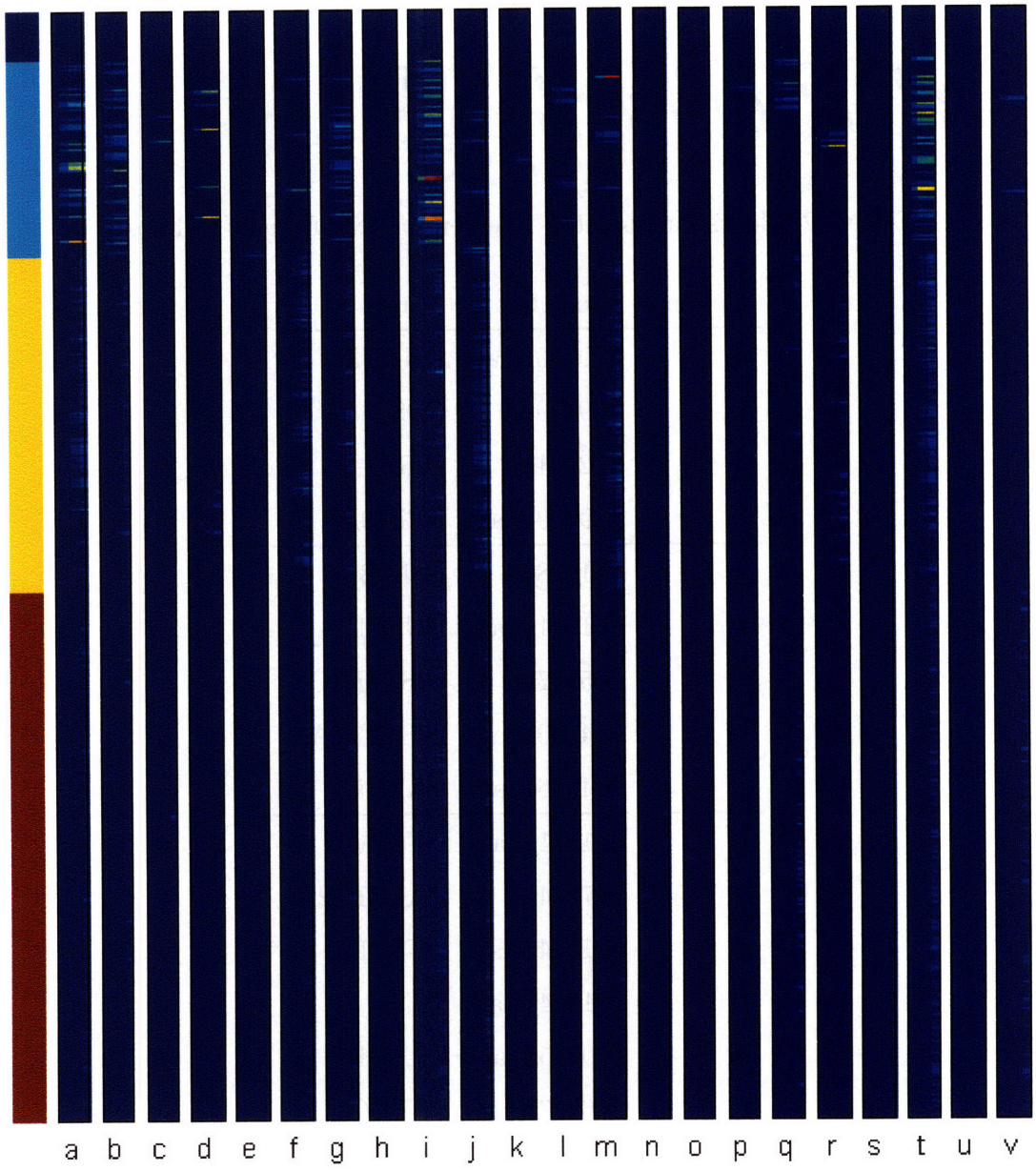



Figure 6-5: For the indolent network, the effects of uniformly perturbing each of the first wave genes on all other genes in the network over time. The leftmost column corresponds to the wave label.

Table 6.2: Genes for potential intervention experiments (* hubs)

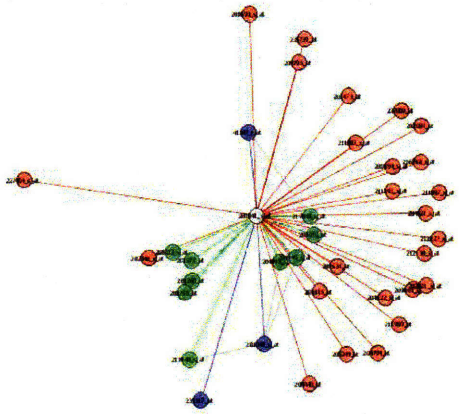
| Column | Probe Set | Name |
|---|---|---|
| c | 220725_x_at | DNAH3* |
| f | 209189_at | FOS |
| g | 202081_at | IER2 |
| j | 229563_s_at | RPL10A* |
| e | 211998_at | H3F3B |
| i | 201041_s_at | DUSP1 |
| l | 205967_at | HIST1H4C |
| m | 214290_s_at | HIST2H2AA |
| o | 224516_s_at | CXXC5* |
| a | 201041_s_at | DUSP1 |
| b | 227224_at | RALGPS2* |
| d | 201044_x_at | DUSP1 |
| i | 201464_x_at | JUN |
| m | 209685_s_at | PRKCB1 |
| t | 226227_x_at | TALDO1 |

related apoptosis [53]. We examine the consequences of down-regulation of *DUSP1* using small interfering RNA after BCR activation in a patient with the aggressive subtype of B-CLL in experiments described in detail by Vallat et al. [77]. By silencing *DUSP1*, a first wave gene, the effect of the intervention is more local to the peak at $t_1$ than overexpression, which would yield expression that is uniformly high, and the shape of *DUSP1* profile would be lost. Though the shape is not key to the analysis here where we test the predictive ability of the network, it is important for future experiments where more control over the expression profile is necessary.

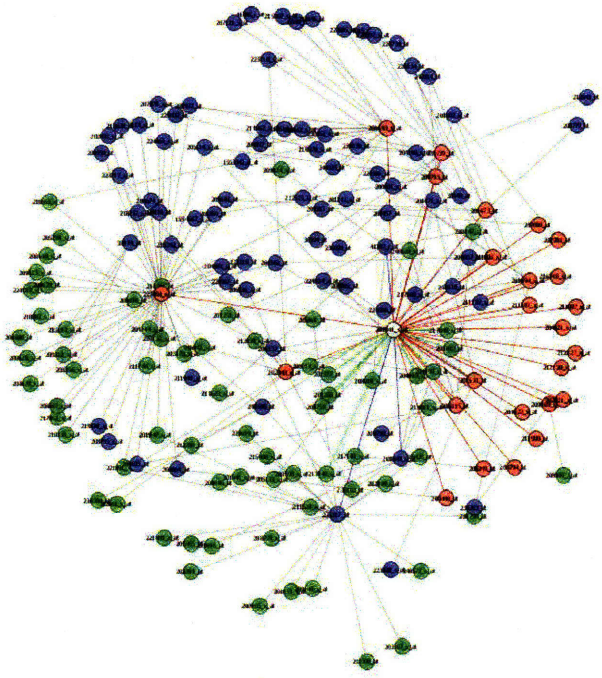
In the Aggressive network, *DUSP1* has direct connections to 37 genes. This subnetwork is shown in Figure 6-6(a). It is important to remember that these do not necessarily correspond to actual physical interactions, and microarray experiments cannot provide this information because they only measure mRNA expression levels. The intervention experiment described in this section, however, will help to identify which genes are actually influenced as compared to ones that are simply correlated with *DUSP1*. In the subnetwork, 25 of the 37 connections are to second wave genes via $F_{12}$ (in red), which confirms what was evident globally in the networks in Chapter 5. Most of the interactions are temporally local, despite not having imposed such a constraint on the inference procedure. One of these connections is to a hub in the Aggressive network, *EGR1*, a second wave gene. There are also 136 indirect connections, many of them through *EGR1*, which explains in part why *DUSP1* was identified as one of the potential target genes to manipulate in the previous sensitivity analysis. The combination of direct and indirect connections is shown in Figure 6-6(b).

### Experimental Procedure

For the data set used for the *DUSP1* experiment, details of the experimental procedure are found in Vallat et al. [77]. While these experiments allow for greater control over changes in expression, they are difficult to perform. In addition to identifying the gene targeted for silencing, the RNAi with the correct action must be obtained, and then it must be delivered inside the primary leukemia cells. It is extremely rare



(a) Direct connections



(b) Direct and indirect connections

Figure 6-6: *DUSP1* subnetwork (Aggressive network). *DUSP1* is white, second wave genes are red, third wave genes are blue, and fourth wave genes are green.

to combine silencing, receptor stimulation, and collection of microarray data over time in a systematic way. Therefore, we have a unique opportunity to validate the temporal interaction networks by testing their predictive ability of changes due to silencing.

In the experiment itself, B cells were isolated from one UM-CLL patient (corresponding to the Aggressive subtype) previously included in the microarray experiment as described in Section 2.2. Cells were transfected with *DUSP1* siRNA or siCONTROL non-targeting siRNA. Half of the cells were BCR stimulated as described previously, and RNA was collected at the same four time points. After reverse transcription, real time quantitative PCR was performed to assess the silencing effect observed on *DUSP1* to confirm that its expression level had been decreased as compared to the control. *DUSP1* mRNA expression was analyzed over four time points and the six experimental conditions (*DUSP1* silenced (US/S), siCONTROL non targeting siRNA transfected (US/S) and control non transfected B cells (US/S)). cRNA was then hybridized to the HG-HU133 plus2.0 microarray as described previously.

### Results

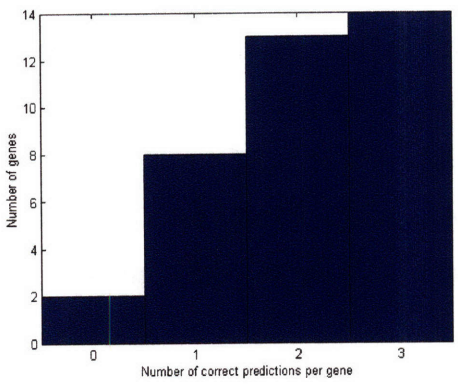
For each gene, we compute differential expression (stimulated minus unstimulated) for the original patient data, the control, and the silencing. Because of the noisiness of microarray data, the limitations of the models, and the other patients that were included in the clustering of the genes and the inference of the predictive models, we do not expect perfect quantitative predictions. For that reason, we evaluate the quality of the predictions by whether or not for a given gene, the models correctly predicted an increase or decrease due to the silencing relative to the control. We do not predict the value at the first time point because of the temporal causality constraints, i.e. changes at $t_1$ do not propagate to the rest of the genes until $t_2$. This means that for every gene in the subnetwork, there is a maximum possible score of three, which corresponds to correct predictions in direction of change at $t_2$, $t_3$, and $t_4$. Figure 6-7 shows a histograms of these prediction scores on a per gene basis, where Figure 6-7(a) shows the scores for the 37 genes directly connected to *DUSP1* in the

network and Figure 6-7(b) shows the scores for the 136 indirectly connected genes. The average prediction scores are 2.05 and 2.15, respectively. Thus for genes that are expected to be influenced by *DUSP1*, the model predictions are correct nearly 70% of the time, considerably better than chance, which would be 50%. Also, in addition to noise effects, some of the genes would be expected to be simply correlated with, not influenced by, *DUSP1* and would not change predictably.

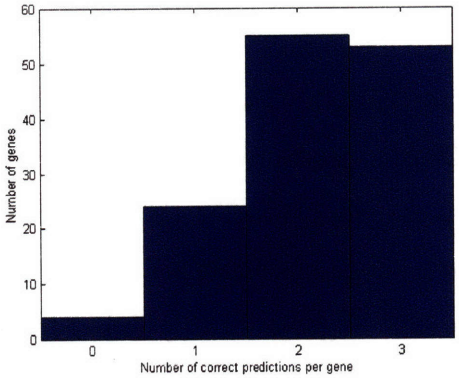
In Figure 6-7(c), the scores for genes that are linked to *DUSP1* neither directly nor indirectly are shown. These have an average score of 1.21, or correct predictions approximately 40% of the time. These genes that are unrelated to *DUSP1* should, in theory, have the same expression levels in both the *DUSP1* siRNA and the siCONTROL trials. Any differences that are seen in the data would be due to noise. The model is obviously unable to predict these noise effects, so it should do no better than chance, which is exactly what occurs. This strengthens the results with respect to those genes that are related to *DUSP1*, as the model is uniquely able to predict changes due to gene silencing.

In addition to the aggregate scoring results, we show in Figure 6-8 the actual data, the predictions, and the input genes for *EGR1* (a hub in the Aggressive network). In the upper left, the differential expression profiles for *EGR1* in the original data, the *DUSP1* silencing and the silencing control are shown. We display the network predictions for each of those three cases in the upper right, and the model predicts the direction of change correctly at time points $t_2$ and $t_3$. The three smaller plots consist of the three input genes in the Aggressive network for *EGR1*, with *DUSP1* on the left. The effect of the silencing on the *DUSP1* profile is most evident at $t_1$, where the differential expression decreases approximately five-fold.

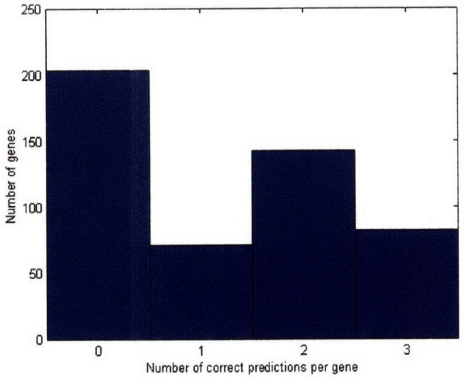
The dominant input to *EGR1* is *DUSP1*, with a connection strength of $\alpha = 0.77$. While there is no confirmed interaction between *EGR1* and *DUSP1*, these two genes belong to the same functional group, according to the graph of the *MAPK* signaling pathway presented in the KEGG PATHWAY database [40]. *DUSP1* is also known as *MKP*, and it regulates *JUN* and *FOS* transcriptional activity, and thus the transcription of *EGR1*. We also show example results for specific genes in Figures 6-9




(a) Direct Links




(b) Indirect Links




(c) No Links

Figure 6-7: Histogram of the number of correct predictions per gene (maximum of 3) for *DUSP1* silencing.

through 6-13, with a mix of successful and unsuccessful predictions and genes from waves 2, 3, and 4. These are the same style as Figure 6-8, where the actual data is in the upper left, the prediction in the upper right, and the expression profiles of the input genes are shown below. For those genes with prediction scores of 3 (Figures 6-9 and 6-10), the connection strengths between the targets and *DUSP1* are $\alpha = 0.55$ and $\alpha = 0.49$, respectively. Because *DUSP1* is the gene with the largest connection strength for each of the targets, it is reasonable to expect the network to successfully predict changes. Two genes with lower connection strengths, or less confidence in the *DUSP1*-target relationship, are shown in Figures 6-11 and 6-12. Their prediction scores are zero. Finally, the fourth wave gene shown in Figure 6-13 has a score of 2 and a connection strength of $\alpha = 0.4$. These results are typical of those from the remaining 30 genes directly linked to *DUSP1*.

## 6.4 Literature comparisons

In the previous two sections, we identified genes that appear to play important roles in the genetic program of the BCR signaling pathway by using the structure of the networks and sensitivity analysis. From previous work in B cells, other genes have been identified as well. In the following section, we discuss two well-characterized examples of these, *NF-kB* and *MYC*.

### 6.4.1 *NF-kB*

The transcription factor *NF-kB* is believed to play a role in lymphoid tissue development, immune, inflammatory, and environmental stress responses, and neuronal signaling [61, 81]. As the experiments in this work are based on immune response (activation of the B cell receptor), differences in the signaling response across subject groups could be related to changes in the *NF-kB* expression profile or in the genes that it influences. We investigate *NF-kB* in the temporal interaction networks and compare the outgoing edges to known biological interactions.

In the subset of genes selected by the clustering, probe sets corresponding to *NF-*

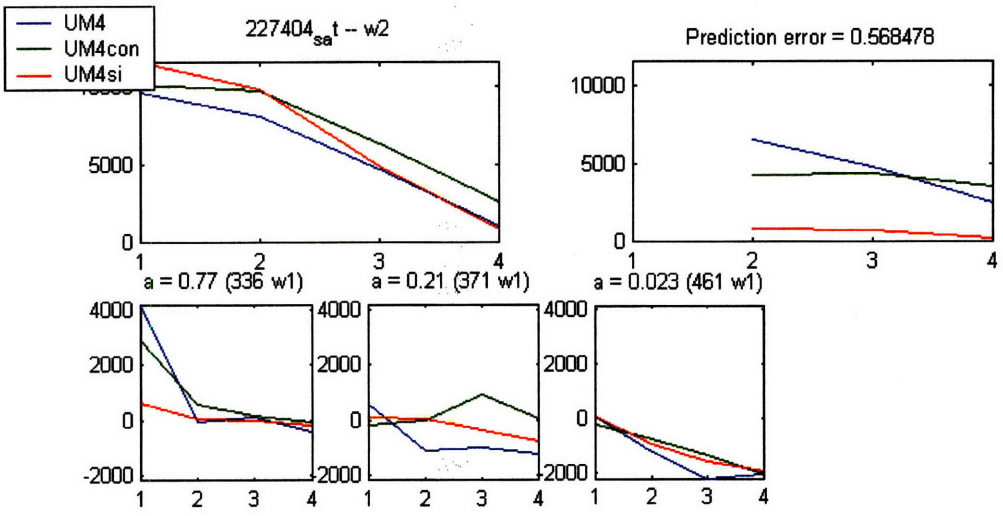



Figure 6-8: *EGR1*, a network target of *DUSP1*, prediction score is 2. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

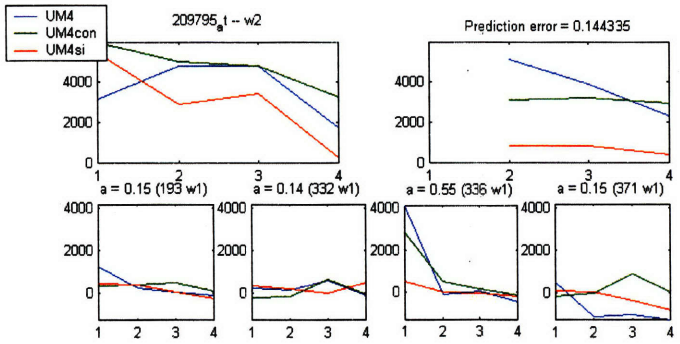



Figure 6-9: Network target of *DUSP1*, prediction score is 3. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

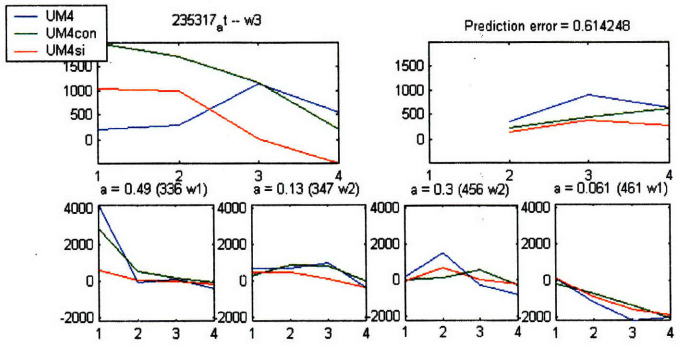



Figure 6-10: Network target of *DUSP1*, prediction score is 3. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

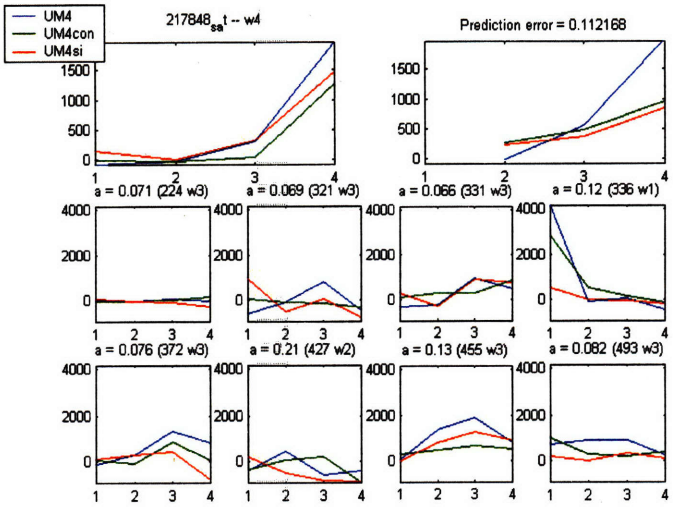



Figure 6-11: Network target of *DUSP1*, prediction score is zero. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

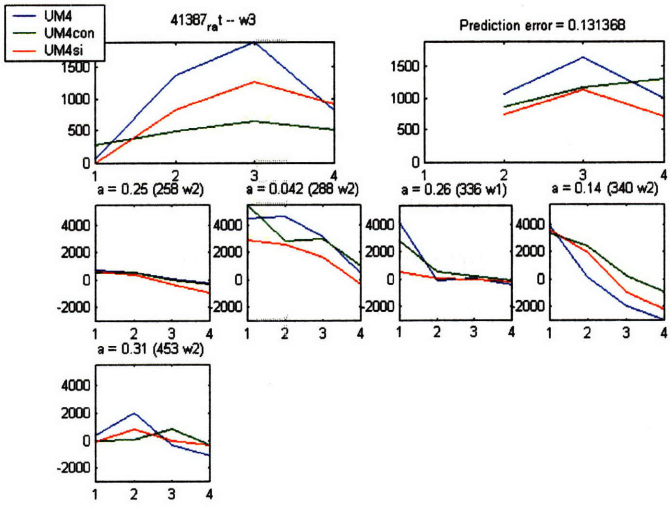



Figure 6-12: Network target of *DUSP1*, prediction score is zero. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

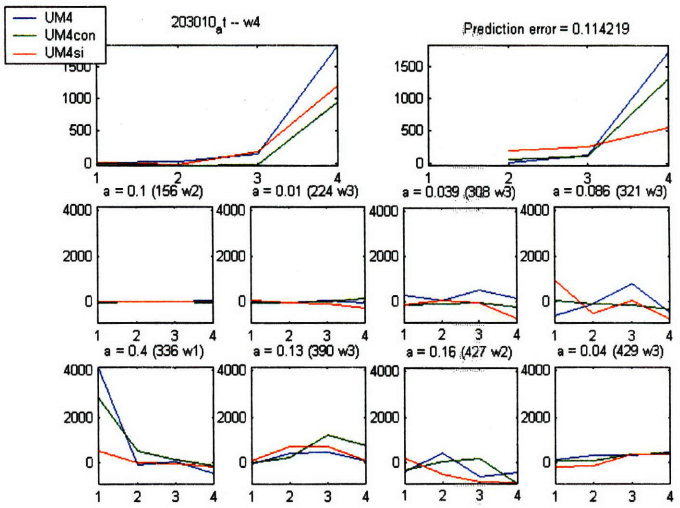



Figure 6-13: Network target of *DUSP1*, prediction score is 2. Top left is the actual expression profile, top right is the prediction, and the remaining plots are the input genes.

k*B* are present for the Aggressive and Indolent groups. For the Aggressive network, *NF*-k*B* has no outgoing edges. This does not necessarily mean that it does not influence any other genes, only that it dos not provide the best predictions, either due to subject inconsistency, noise, or modeling choices, for those genes that it is believed to target. For the Indolent network, *NF*-k*B* has 3 outgoing edges which correspond to relationships with *PIM1*, *TNIP1*, and *HLA-F*. At least one of these connections, the edge to *HLA-F*, has been confirmed biologically as *HLA-F* is a direct gene target of *NF*-k*B* [61]. *TNIP1* [36] and *PIM1* [90] are related to *NF*-k*B*, but there is not yet evidence of a direct interaction. Because many of the genes are not annotated and all interactions are certainly not yet know, verification of local regions of the network is difficult, but these results show that the inferred links may correspond to physical interactions in this case.

### 6.4.2 *MYC*

The *MYC* proto-oncogene is believed to encode a transcription factor and emerged as one of the top 5% of largest cellular hubs in a transcriptional regulatory network inferred by Basso et al. [4]. Because *MYC* is extensively characterized as a transcription factor, the inferred interactions were compared to those previously identified by biochemical methods. In the networks that we inferred in Chapter 5, *MYC* was present in all three subject groups (labeled as wave 2 in Healthy and wave 3 for both Aggressive and Indolent). The expression profiles for each of the subjects are shown in the top row of Figure 6-14. However, it had no outgoing edges in any of the networks, which was unexpected. The first possible reason for this was that it is much larger in magnitude than the cluster means for the waves to which it belongs, which are shown in the bottom row of Figure 6-14. Because the interaction matrices ($F_{k(input)k(ouput)}$) are based on the overall mode of interaction between pairs of waves, the predictions of outlier genes such as *MYC* are not necessarily well represented. This modeling choice did not appear to have affected known transcription factors in the earlier waves but is an issue here.

A second reason for the lack of outgoing edges for *MYC* is that other genes better predict the expression profiles of its known targets. We show examples from each of the three subject groups in Figures 6-15, 6-16, and 6-17. The network inputs for these genes actually appear to be less consistent than *MYC*, so either there are inter-subject differences or noise effects that cause these genes to more accurately predict the targets.

## 6.5 Comparisons across subject groups

From Chapter 2, we expect that some genes behave similarly across subjects. Many of these will be non-BCR pathway related, and there will be no expression difference in stimulated and unstimulated results. These were intentionally filtered out as part of the background. Some genes that are be expected to be active in the BCR pathway may still behave in similar ways across groups, while others will behave differently

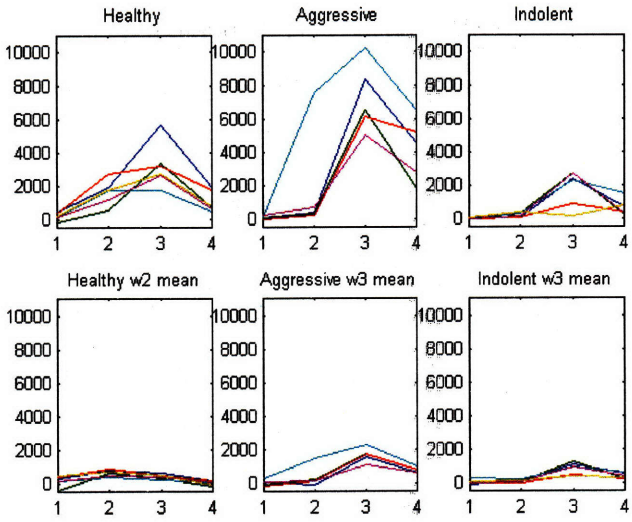



Figure 6-14: The top row has the expression profiles for *MYC* for each of the subjects, divided by subject group. The bottom row has the cluster means of the wave class for *MYC*.

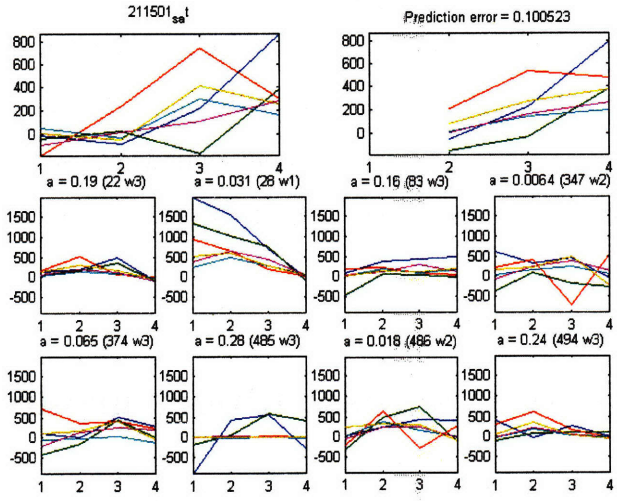



Figure 6-15: Prediction and network inputs of an example known target of *MYC* for the Healthy group.

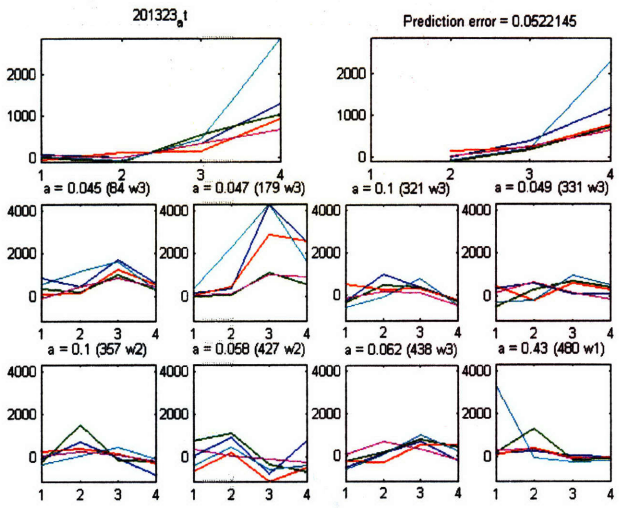



Figure 6-16: Prediction and network inputs of an example known target of *MYC* for the Aggressive group.

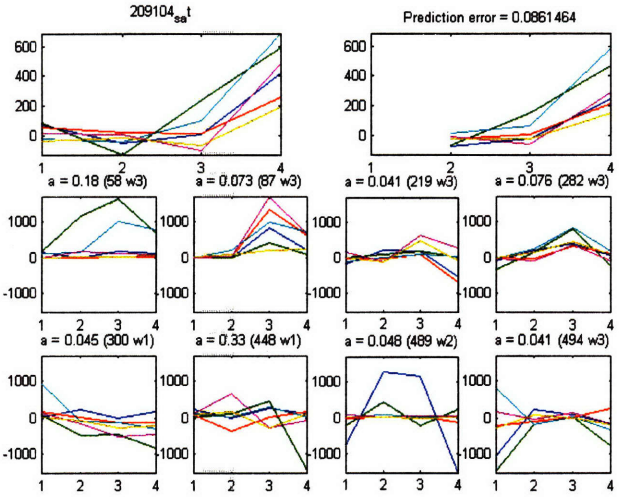



Figure 6-17: Prediction and network inputs of a known target of *MYC* for the Indolent group.

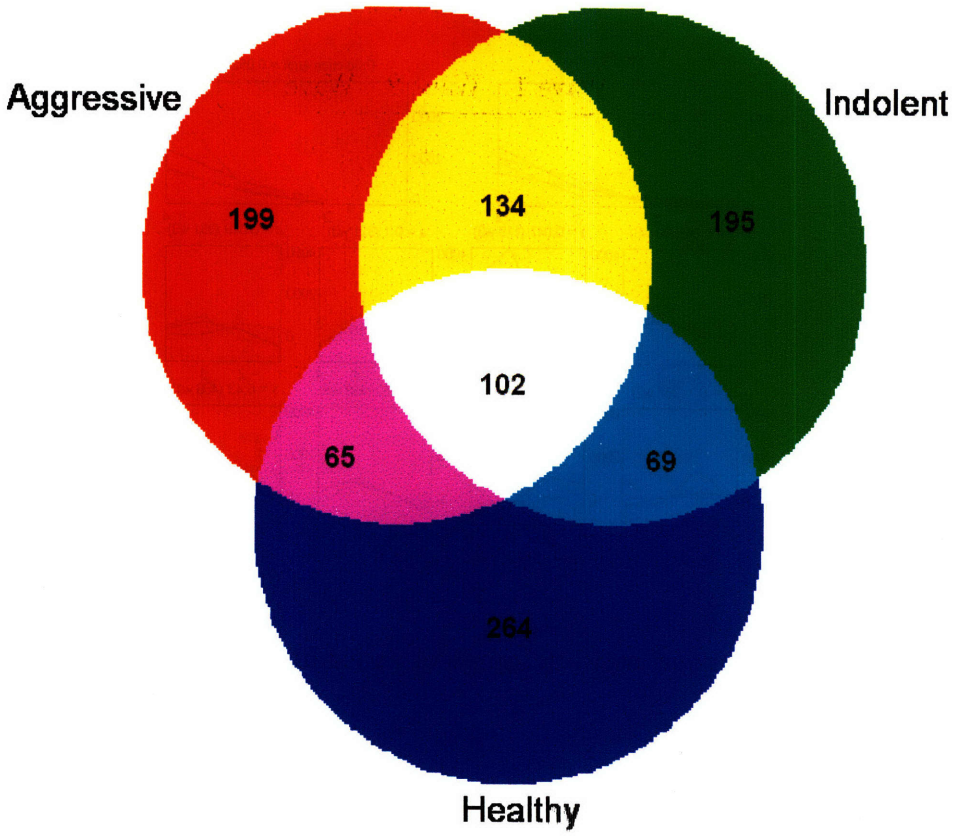



Figure 6-18: Overlap of genes selected by wave clustering (500 per group)

either because of genetic differences, which may be related to the disease.

While the clustering is currently done independently within each subject group, we compare the results across groups. In Figure 6-18, a Venn diagram displays a total of 1028 genes that are included when all 3 groups are combined. Table 6.3 shows the consistency of the labeling of the genes in this group. Interestingly, the Aggressive and Indolent groups are the most similar in terms of the wave labels while in general, it is the Aggressive group that is most dissimilar from the other two (see Chapter 4). We can speculate that perhaps the differences between the two B-CLL groups that cause the drastically differing prognoses is more due to magnitude and not shape of the relevant gene expression profiles.

Table 6.3: Wave label comparison across subject groups

| | Wave 1 | Wave 2 | Wave 3 | Wave 4 | Total |
|---|---|---|---|---|---|
| Healthy-Indolent | 0 | 31 | 40 | 53 | 124 |
| Healthy-Aggressive | 0 | 20 | 61 | 40 | 121 |
| Aggressive-Indolent | 3 | 31 | 87 | 93 | 211 |
| Healthy-Agg-Ind | 0 | 19 | 27 | 28 | 74 |

### 6.5.1 Cross validation

We perform leave-one-out cross validation to test the consistency of the clustering within and across subject groups. As in Section 5.4.4, we use a single subject from the original data set as the testing data, and the remaining observations as the training data. We repeat this such that each subject is used once as the testing data, held out from the clustering of its own type of network (Healthy, Aggressive, or Indolent). For example, for subject Healthy1, the cluster parameters, memberships, and log-likelihood scores were computed using the data from subjects Healthy2 through Healthy6. Using the input data for Healthy1, we compute the log-likelihood score for the data under the cluster parameters estimated from all the aggressive subjects and the parameters estimated from all the indolent subjects. The results are shown in a bar plot in Figure 6-19. If we were to use this method to classify subjects into groups based on the maximum log-likelihood score, only 1 of 6 Healthy subjects are correctly labeled as Healthy, 4 of 5 Aggressive subjects as Aggressive, and 6 of 6 Indolent subjects as Indolent. The log-likelihood scores corresponding to evaluation under the cluster parameters for the Healthy and Indolent groups are close to one another in most cases, which is expected because of the the similarities in cellular behavior for those two subject groups. Still, this does not separate the subjects successfully. As we compute the log-likelihood score only based on the 1028 genes (those represented in Figure 6-18), the classification results are even less successful.

For every subject, evaluation under the Aggressive parameters yields the best log-

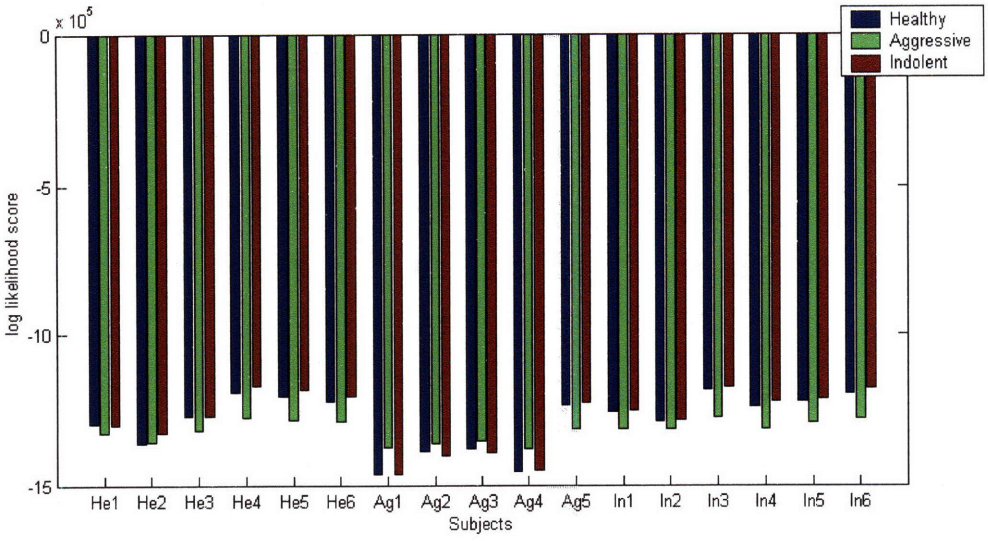



Figure 6-19: Log likelihood scores for leave one out cross-validation, as well as the scores under the other two sets of cluster parameters

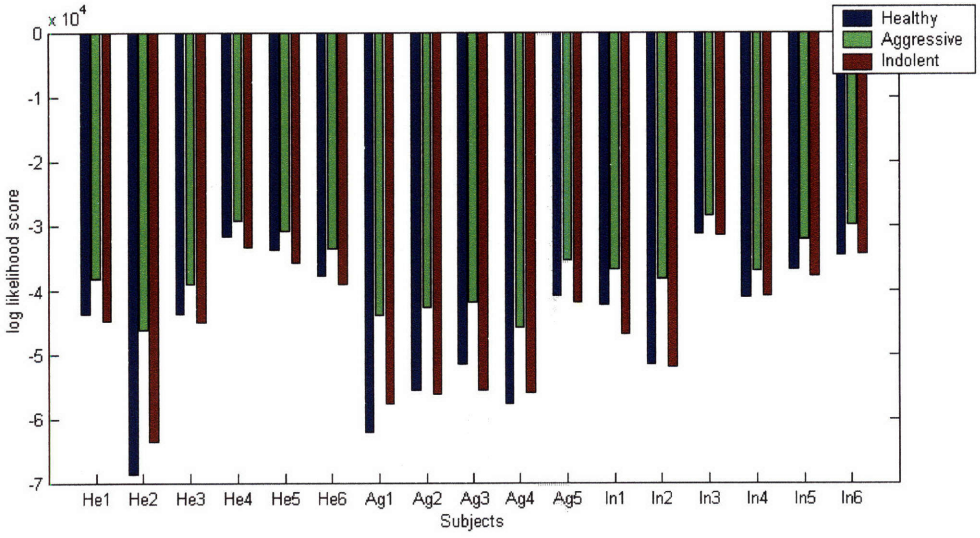



Figure 6-20: Log likelihood scores for leave one out cross-validation, as well as the scores under the other two sets of cluster parameters, for the 1028 selected genes

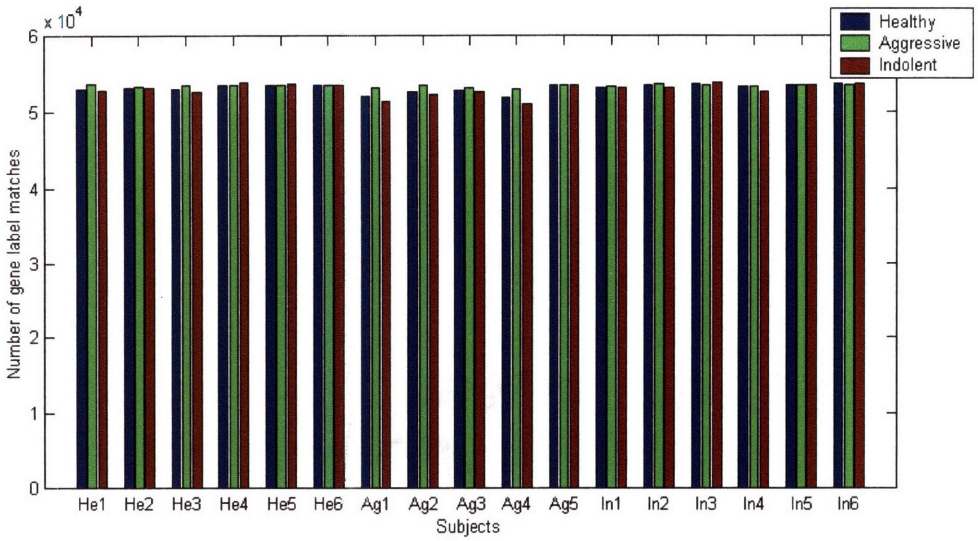



Figure 6-21: Match scores for leave one out cross-validation, as well as the scores for the other two subject groups.

likelihood scores. The reason for this is, in general, that the Aggressive group has the largest magnitude differential expression, and the cluster parameters reflect these large deflections. As we select only the top 500 genes (based on maximum *a posteriori* values) from each group, these tend to have larger deflections compared to the remaining genes in the same waves, so they fit better under the Aggressive parameters. Similar to the results in Chapter 3, log-likelihood scores do not necessarily capture the quality of the clustering. The only thing that it provides here for discrimination between subject groups is the shape of the expression profiles as opposed to the label of the gene. For example, a gene may be classified as a second wave gene in the Healthy group and as long as it fits the second wave cluster parameters for the Aggressive group, the score will be high, even if the same gene is classified as third wave for the Aggressive group. Therefore, instead of log-likelihood scores, we compute a "match score," where we count the number of cluster label matches for the left-out subject and each subject group. For all genes, the results are shown in Figure 6-21.

Under this criterion, 4 of 6 Healthy subjects are classified as Aggressive, and the other 2 are classified as Indolent. The 5 Aggressive patients are classified correctly. 3

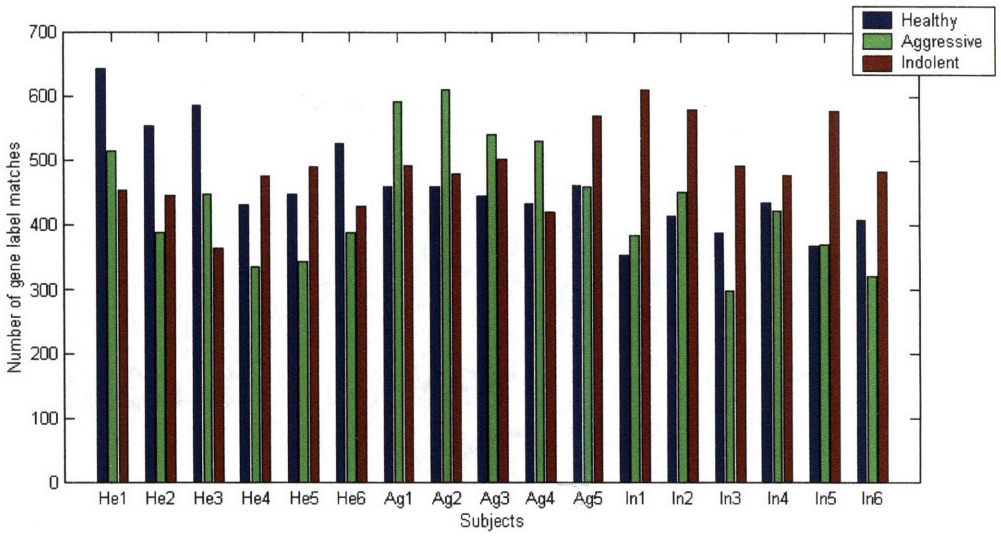



Figure 6-22: Match scores for leave one out cross-validation, as well as the scores for the other two subject groups for the 1028 selected genes.

of 6 Indolent patients are classified correctly, as 1 is classified as Healthy, and 2 are classified as Aggressive. We have already established that there are a considerable number of genes unrelated to the BCR pathway and likely unrelated to B-CLL that would be included in the computation. Therefore, we again compute the match score based only the 1028 genes previously selected (see Figure 6-22).

This correctly classifies 4 of 6 Healthy, 4 of 5 Aggressive, and 6 of 6 Indolent. One interesting note is that Aggressive5 was actually originally classified in [78] as a Mutated/Indolent patient, and that is the one that is misclassified here as well. Regardless, this method correctly classifies 14 of 17 subjects, and while this is not perfect, it does serve to show that some number of the 1028 genes behave differently across subject groups.

### 6.5.2 Network overlap

The network clusters were computed and the predictive models inferred independently for each subject group, and part of the long term goal of this work is to determine how to target and treat local differences that occur in a disease-specific sense. A logical

Table 6.4: Edges in common in the interaction networks

| | Number of Edges |
|---|---|
| Healthy-Indolent | 9 |
| Healthy-Aggressive | 11 |
| Aggressive-Indolent | 37 |
| Healthy-Agg-Ind | 1 |

step would be to visualize the networks simultaneously. First, this requires placing 1028 genes and more than 4000 edges in the image, which is shown in Figure 6-23. We use the colors from the Venn diagram in Figure 6-18 to show which genes are common across which subject groups. For example, red corresponds to Aggressive only, green to Indolent only, and yellow to both Aggressive and Indolent. The genes in white are those that are common to all three subject groups. The wave cluster labels are also not consistent across subject groups, as was shown in Table 6.3. Therefore, not only is there a limited subset of genes that are present in more than one subject group, but those that are may have different labels, so the networks have a very small amount of overlap with respect to their edges. The number of edges that are in common are shown in Table 6.4. As before, there is the most overlap between the Aggressive and Indolent groups, likely because of the greater similarity in wave labels. Because these common edges comprise only about 1% of the total edges in the combined network in Figure 6-23, the visualizations that best show the network structure are those that separate the subject groups completely, as in Figure 5.3.

Ideally, we would infer a combined network by adjusting the formulation described in Chapter 5 by combining information for all subjects in common parts of the network, those that correspond to general B cell response. It would take advantage of the additional subject data available for those regions, averaging out non-disease-related inconsistency and noise. Population or disease-specific differences would then allow analysis of the relatively small number of genes that discriminate between groups. A combined network would take us one step closer to determining where exactly the

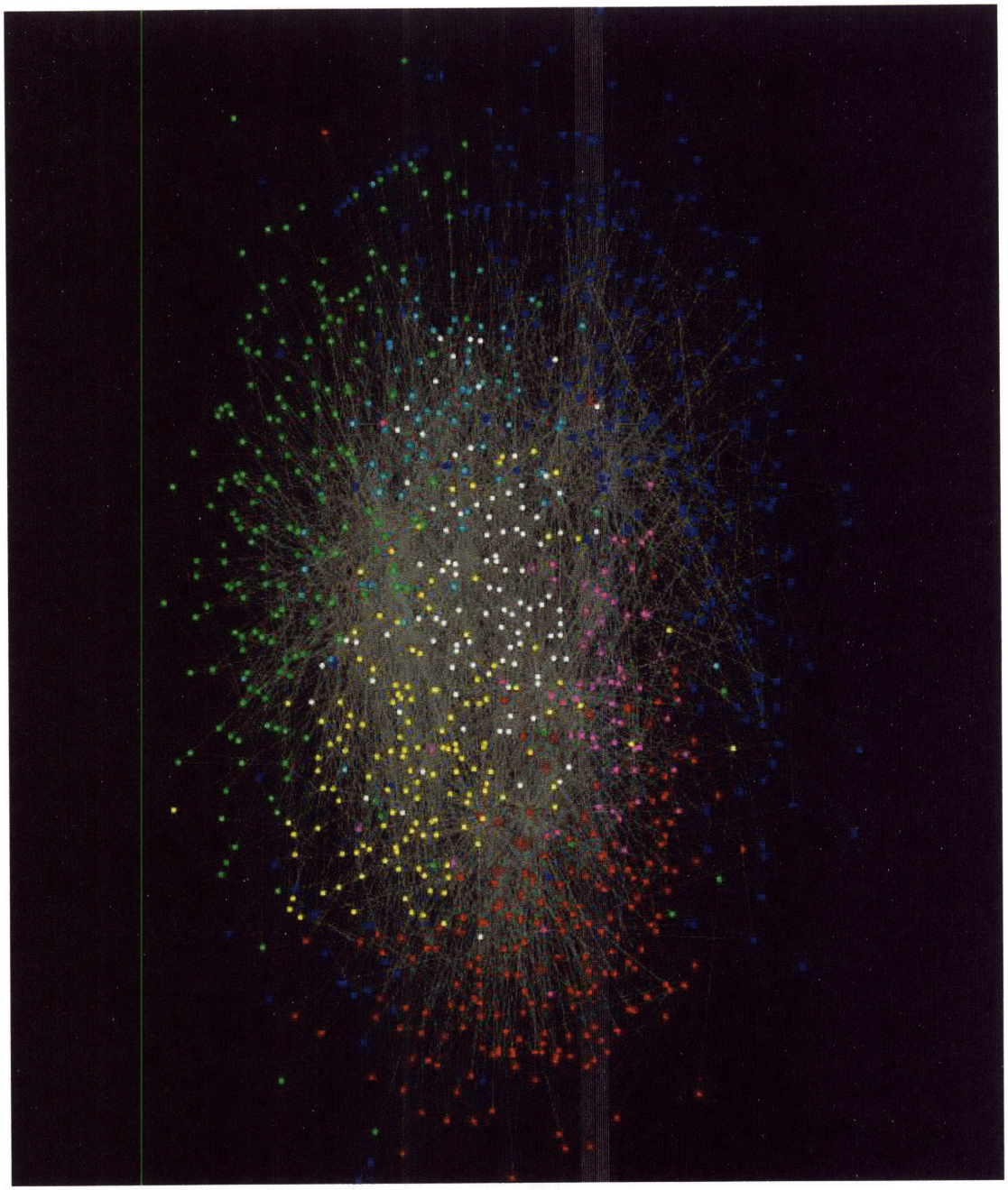


Figure 6-23: Combined network, where the node colors correspond to the colors in Figure 6-18.

differences between B-CLL patients and healthy subjects occur, as well as how to manipulate the cell to repair those differences.

One part of inferring a combined network is determining how to combine subjects across groups in the assignment of cluster labels. As was discussed in Chapter 4, given our extremely limited number of subjects, a consistent, meaningful breakdown of subject groups and genes is not yet possible in our data set. An additional opportunity for future work is in not only combining the subject groups but also in combining the clustering of the genes with the how they predict their targets. Because of the simplifying assumptions made in the predictive model inference step, how genes interact is largely determined by which wave class they have been assigned. It is likely that not all wave 1 genes interact with all wave 2 genes in the same way, for example. While this could be addressed by allowing predictions to be drawn from a distribution around the limited modes of interaction, it could also be solved by integrating the clustering and predictive model inference.

## 6.6 Summary

In this chapter, we have presented the biological implications of the clustering and the predictive modeling, providing both confirmatory results and insight into the design of future experiments. Given the scale-free structure of the temporal interaction networks, we identified a small number of genes with a large number of outgoing edges as network hubs. Also, with sensitivity analysis, we propagated the effects of manipulation of first wave genes to the remaining genes in the network. These techniques allowed us to select genes that play important roles in the network and served as starting points for literature comparisons in terms of gene function (e.g. known transcriptions factors that had been labeled as hubs) and physical interactions. Guided by previous biological results, we also investigated the behavior of *NF-kB* and *MYC* and their targets. In addition to comparisons to previous work, the results of an intervention experiment, the silencing of *DUSP1* show the predictive ability of the network. In order to test the network further, and to differentiate between causal

and correlative relations, similar intervention experiments could be performed based on the results of the sensitivity analysis.

For the clustering, which has been performed independently within each subject group, we compared wave labels across subject groups. Leave-one-out cross validation showed that label match scores based on the subset of genes identified as differentially active as a result of BCR stimulation are successful in classifying 14 of 17 subjects. Finally, we combine the genes identified into a single visualization of the interaction network, address the reasons for presenting the networks separately, and suggest potential paths for future work.

# Chapter 7

# Conclusions

In this work, we recognized that for B cell chronic lymphocytic leukemia (B-CLL), understanding the underlying disease process is a vital component in making progress toward effective diagnosis and targeted treatment. Little is known thus far, except that the response to B cell receptor stimulation appears to be related to key differences between the B-CLL patients and healthy subjects, as well as the aggressive and indolent forms within the B-CLL patient population. We therefore inferred predictive models of temporal gene interaction for the B cell receptor (BCR) signaling pathway.

To that end, we took advantage of the advances in microarray technology to collect differential expression data of more than 50,000 genes for 17 subjects (both healthy and suffering from B-CLL), coarsely sampled with 4 time points over 390 minutes. This large amount of data led us to consider clustering techniques to identify which genes were most relevant to the BCR signaling pathway and at what times they were most active. We designed simple statistical models in order to capture the temporal behavior seen in the pilot study, and estimated the parameters in an Expectation-Maximization framework, resulting in clusters with a biological interpretation. Compared to a naïve approach that did not require any prior knowledge of the cluster structure, the wave clusters have similar shapes for each patient group, with smaller intra-cluster variance relative to the means. Both clustering procedures show that there is statistically significant temporal structure that can be extracted from the data, but the wave clustering also consistently grouped genes in the same

waves together and in different waves separately, despite perturbations of the input data with noise. We have shown for our application that the wave clustering procedure produces stable, consistent and meaningful models of the differential expression profiles of genes related to BCR stimulation. In addition, wave clustering serves as a first step in order to make predictive modeling, as described in Chapter 5, tractable.

Because of the extremely limited number of subjects, the inference problem for temporal gene interaction is severely ill-posed. Therefore, we used the cluster labels, which have a consistent biological interpretation, to define a small number of modes of interaction between the genes. We also imposed sparsity to limit the number of genes influencing each target gene. Both of these choices were made in order to make inference tractable, but they are also biologically supported. We tested the statistical significance of the models, demonstrating that labeling, model constraints, and connectivity constraints are not arbitrary. We conclude, therefore, that temporal structure exists and that it can be captured to some degree by a small number of temporal structures and interaction types.

We showed that the scale-free network structure is consistent with what has been found in protein interaction networks [8], where a small number of proteins serve as hubs for very large numbers of interactions. We addressed the issue of distinguishing relationships of causality from those simply of correlation with intervention experiments. Because intervention experiments are expensive and time-consuming to perform, selection of potential target genes by combining wave cluster labeling and sensitivity analysis of the networks is an important contribution of this work. The results of one of these intervention experiments, the silencing of *DUSP1* in an patient with aggressive subtype of B-CLL, showed the models of gene interaction are uniquely able to predict changes due to gene silencing. This type of biological experiment provides a confirmatory result, but identification of other potential targets of genetic intervention also guides future experiments.

Though our application here is specific to B-CLL and BCR stimulation, similar techniques could be applied to virtually any type of experimental conditions because one of the fundamental problems in biology is understanding genomic function. In

addition to the purely scientific value, this knowledge can be applied to investigating the mechanisms of human disease. Predictive models will permit us to investigate the underlying cause of diseases and help us to develop targeted therapies. While DNA microarray technology has made it possible to monitor the expression levels for tens of thousands of genes in parallel, the large number of genes and the complexity of biological networks greatly increases the challenges of comprehending and interpreting the data for this purpose. The scale of the experimental data available has increased so rapidly that it is simply not possible to reconstruct networks manually, and therefore, interpretation of gene expression data requires sophisticated computational methods to gain insight into the inner workings of cells. This means that the collaboration in this work is but one example of how machine learning can provide insights into the global patterns and relationships of expression in cellular behavior.

# Appendix A

# Derivation of the EM Algorithm

In Chapters 3 and 4, we use the Expectation-Maximization algorithm to cluster genes, subjects, or both. This appendix shows the derivation of the algorithm originally presented by Dempster, Laird and Rubin [22], with the aid of details in intermediate steps from a tutorial by Borman [7]. We also provide a simple example of a Gaussian mixture model and derive the parameter update equations.

## A.1 Derivation

The data $X$ is a random vector which results from a parameterized family, and the goal is to find the values of the parameters $\Theta$, such that $P(X|\Theta)$ is a maximum, in the presence of missing or hidden data. The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate by maximizing the log-likelihood function, which is defined as:

$$L(\Theta) = \ln P(X|\Theta) \tag{A.1}$$

At each iteration, we intend to compute an updated estimate of $\Theta$ such that:

$$L(\Theta) > L(\Theta^k) \tag{A.2}$$

where the superscript $k$ denotes the value at the current iteration. We therefore maximize the difference:

$$L(\Theta) - L(\Theta^k) = \ln(P(X|\Theta) - \ln(P(X|\Theta^k) \tag{A.3}$$

We now incorporate the hidden variable, which, in the case of clustering, is the class label. If the hidden data (labels) were known, we would easily be able to compute the parameters for each of the classes.

$$P(X|\Theta) = \sum_{m=1}^{M} P(X|m,\Theta)P(m|\Theta) \tag{A.4}$$

Substituting back in to Equation A.3:

$$L(\Theta) - L(\Theta^k) = \ln \sum_{m=1}^{M} P(X|m,\Theta)P(m|\Theta) - \ln P(X|\Theta^k) \tag{A.5}$$

At this point, we introduce Jensen's Inequality, which states:

$$\ln \sum_{n=1}^{N} \lambda_n x_n \geq \sum_{n=1}^{N} \lambda_n \ln x_n \tag{A.6}$$

where:

$$\lambda_n \geq 0 \tag{A.7}$$

$$\sum_{n=1}^{N} \lambda_n = 1 \tag{A.8}$$

In this case, $P(m|X,\Theta^k)$, the posterior probability of a given class, fulfills both of these conditions for $\lambda_n$. We substitute back into A.3 again, this time using Jensen's Inequality to move the summation outside the log:

$$L(\Theta) - L(\Theta^k) = \ln \sum_{m=1}^{M} P(X|m,\Theta)P(m|\Theta)\frac{P(m|X,\Theta^k)}{P(m|X,\Theta^k)} - \ln P(X|\Theta^k) \tag{A.9}$$

$$\geq \sum_{m=1}^{M} P(m|X,\Theta^k) \ln \frac{P(X|m,\Theta)P(m|\Theta)}{P(m|X,\Theta^k)P(X|\Theta^k)} \tag{A.10}$$

Our objective is to choose values of $\Theta$ so that $L(\Theta)$ is maximized, so we define the function $Q(\Theta|\Theta^k)$ that is guaranteed to be less than or equal to $L(\Theta)$:

$$Q(\Theta|\Theta^k) = L(\Theta^k) + \sum_{m=1}^{M} P(m|X,\Theta^k) \ln \frac{P(X|m,\Theta)P(m|\Theta)}{P(m|X,\Theta^k)P(X|\Theta^k)} \tag{A.11}$$

In the Expectation Step, we compute $P(m|X,\Theta^k)$ based on the current parameter values, and then in the Maximization Step, we use that compute $\Theta^{k+1}$ which maximizes

$Q(\Theta|\Theta^k)$:

$$\Theta^{k+1} = \arg\max_{\Theta} Q(\Theta|\Theta^k) \tag{A.12}$$

$$= \arg\max_{\Theta} L(\Theta^k) + \sum_{m=1}^{M} P(m|X,\Theta^k) \ln \frac{P(X|m,\Theta)P(m|\Theta)}{P(m|X,\Theta^k)P(X|\Theta^k)}) \tag{A.13}$$

$$= \arg\max_{\Theta} \sum_{m=1}^{M} P(m|X,\Theta^k) \ln \left(P(X|m,\Theta)P(m|\Theta)\right) \tag{A.14}$$

## A.2 EM Example

We provide a very simple example of how the EM algorithm can be used for clustering. The EM equations are derived for the case of a mixture of Gaussians, in which the class label is the hidden data. The derivations are similar for the more complicated clustering in Chapters 3 and 4.

### A.2.1 Generative Model

The data $\vec{x}$ is an $N$x1 vector, which we assume has been generated from a mixture of $M$ Gaussians, each with a mean $\mu$ and variance $\sigma^2$. $\Theta^k$ includes all parameters of all the clusters at iteration $k$.

$$P(\vec{x}|\Theta) = \sum_{m=1}^{M} \prod_{n=1}^{N} P(x_n|m,\Theta) \tag{A.15}$$

### A.2.2 Expectation Step

$$P(m|x_n,\Theta^k) = \frac{P(x_n|m,\Theta^k)P(m|\Theta^k)}{\sum_{m'=1}^{M} P(x_n|m,\Theta^k)P(m|\Theta^k)} \tag{A.16}$$

$$= w_{nm} \tag{A.17}$$

This provides an $N$x$M$ matrix of weights that will be used to compute the parameters in the Maximization Step. Specifically for the mixture of Gaussians:

$$P(x_n|m,\Theta^k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right) \tag{A.18}$$

where $P(m|\Theta^k)$ is the current mixture proportion for class $m$ under the parameters $\Theta^k$.

### A.2.3 Maximization Step

We first assume a product model, in which each element of $\vec{x}$ is an independent draw:

$$P(\vec{x}|m,\Theta) = \prod_{n=1}^{N} P(x_n|m,\Theta) \tag{A.19}$$

and by taking the log of both sides:

$$\ln P(\vec{x}|m,\Theta) = \ln \prod_{n=1}^{N} P(x_n|m,\Theta) \tag{A.20}$$

$$= \sum_{n=1}^{N} \ln P(x_n|m,\Theta) \tag{A.21}$$

The expectation over all possible models is then:

$$\sum_{m=1}^{M} P(m|\vec{x},\Theta^k) = \sum_{m=1}^{M}\sum_{n=1}^{N} P(m|x_n,\Theta^k) \tag{A.22}$$

$$= \sum_{m=1}^{M}\sum_{n=1}^{N} w_{nm} \tag{A.23}$$

Therefore, the computation of $Q(\Theta|\Theta^k)$ reduces to:

$$Q(\Theta|\Theta^k) = L(\Theta^k) + \sum_{m=1}^{M}\sum_{n=1}^{N} w_{nm} \ln(P(x_n|m,\Theta)P(m|\Theta)) \tag{A.24}$$

We begin with $\mu$ for a given class $m$ and drop all the terms that do not depend on $\mu$ as we simplify:

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm} \ln(P(x_n|m,\Theta)P(m|\Theta)) \tag{A.25}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}(\ln P(x_n|m,\Theta) + \ln P(m|\Theta)) \tag{A.26}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}(\ln P(x_n|m,\Theta)) \tag{A.27}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}(\ln \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right)) \tag{A.28}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}\left(\frac{-(x_n-\mu)^2}{2\sigma^2}\right) \tag{A.29}$$

$$\tag{A.30}$$

Taking the derivative with respect to $\mu$ and setting it equal to zero gives us:

$$\sum_{n=1}^{N} w_{nm}(x_n - \mu) = 0 \tag{A.31}$$

Solving for $\mu$:

$$\mu = \frac{\sum_{n=1}^{N} w_{nm} x_n}{\sum_{n=1}^{N} w_{nm}} \tag{A.32}$$

This is equivalent to taking a weighted average of the data to compute the cluster mean. In a similar way, the variance for a cluster is (using the value for $\mu$ just computed):

$$\sigma^2 = \frac{\sum_{n=1}^{N} w_{nm}(x_n - \mu)^2}{\sum_{n=1}^{N} w_{nm}} \tag{A.33}$$

Finally, the mixture proportions require the addition of an additional constraint because they must sum to one:

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm} \ln(P(x_n|m,\Theta)P(m|\Theta)) - \lambda(\sum_{m=1}^{M} P(m|\Theta) - 1) \tag{A.34}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}(\ln P(x_n|m,\Theta) - \ln P(m|\Theta)) - \lambda(\sum_{m=1}^{M} P(m|\Theta) - 1) \tag{A.35}$$

$$\arg\max_{\Theta} \sum_{n=1}^{N} w_{nm}(\ln P(m|\Theta)) - \lambda(\sum_{m=1}^{M} P(m|\Theta) - 1) \tag{A.36}$$

Taking the derivative with respect to $P(m|\Theta)$ and setting it equal to zero gives us:

$$\frac{\sum_{n=1}^{N} w_{nm}}{P(m|\Theta)} - \lambda = 0 \tag{A.37}$$

Rearranging to solve for $P(m|\Theta)$ yields:

$$P(m|\Theta) = \frac{1}{\lambda} \sum_{n=1}^{N} w_{nm} \tag{A.38}$$

To solve for $\lambda$, we sum both sides over $M$:

$$\sum_{m=1}^{M} P(m|\Theta) = \frac{1}{\lambda} \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nm} \tag{A.39}$$

By definition, the left side is equal to 1, so:

$$\lambda = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nm} \tag{A.40}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} \tag{A.41}$$

$$= \sum_{n=1}^{N} 1 \tag{A.42}$$

$$= N \tag{A.43}$$

Finally, substituting in the value for $\lambda$:

$$P(m|\Theta) = \frac{1}{N} \sum_{n=1}^{N} w_{nm} \tag{A.44}$$

# Bibliography


[1] D. Anguelov, R. Biswas, D. Koller, B. Limketkai, S. Sanner, and S. Thrun. Learning hierarchical object maps of non-stationary environments with mobile robots. In *In Proceedings of the 17th Annual Conference on Uncertainty in AI (UAI)*, 2002.

[2] Ziv Bar-Joseph, Georg Gerber, David K. Gifford, and Tommi S. Jaakola. A new approach to analyzing gene expression time series data. In *International Conference on Research in Computational Molecular Biology*, 2002.

[3] Ziv Bar-Joseph, Georg K. Gerber, Tong Ihn Lee, Nicola J. Rinaldi, Jane Y. Yoo, Francois Robert, D Benjamin Gordon, Ernest Fraenkel, Tommi S. Jaakkola, Richard A. Young, and David K. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 2003.

[4] Katia Basso, Adam A. Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 2005.

[5] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[6] David R. Bickel. Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics*, 21(7), 2005.

[7] Sean Borman. The expectation maximization algorithm – a short tutorial. URL: http://www.seanborman.com/publications/EM_algorithm.pdf, 2004.

[8] Dennis Bray. Molecular networks: The top-down view. *Science*, 301, 2003.

[9] Patrick O. Brown and David Botstein. Exploring the new world of the genome with dna microarrays. *Nature Genetics*, 1999.

[10] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12), 2004.

[11] Stefano Casola, Kevin L. Otipoby, Sibille Humme, Nathalie Uyttersprot, Jeffery L. Kutok, Michael C. Carroll, and Klaus Rajewsky. B cell receptor signal strength determines B cell fate. *Nat Immunol*, 2004.

[12] Annabelle Chardin and Patric Perez. Unsupervised image classification with a hierarchical EM algorithm. In *Proc. Int. Conf. on Computer Vision*, 1999.

[13] Liguang Chen, John Apgar, Lang Huynh, Frank Dicker, Teresa Giago-McGahan, Laura Rassenti, Arthur Weiss, and Thomas J. Kipps. ZAP-70 directly enhances IgM signaling in chronic lymphocytic leukemia. *Blood*, 105(5), 2005.

[14] Liguang Chen, George Widhopf, Lang Huynh, Laura Rassenti, Kanti R. Rai, Arthur Weiss, and Thomas J. Kipps. Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. *Blood*, 2002.

[15] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000.

[16] Nicholas Chiorazzi, Katerina Hatzi, and Emilia Albesiano. B-cell Chronic Lymphocytic Leukemia, a clonal disease of B lymphocytes with receptors that vary in specificity for (auto)antigens. *Ann NY Acad Sci*, 2005.

[17] Iole Cordone, Serena Masi, Francesca Romana Mauro, Silvia Soddu, Ornella Morsilli, Tiziana Valentini, Maria Luce Vegna, Cesare Guglielmi, Francesca Mancini, Sonia Giuliacci, Ada Sacchi, Franco Mandelli, and Robert Foa. p53 expression in B cell chronic lymphocytic leukemia: A marker of disease progression and poor prognosis. *Blood*, 91(11), 1998.

[18] F. Crick. Central dogma of molecular biology. *Nature*, 1970.

[19] Peter E. Crossen. Genes and chromosomes in chronic B-cell leukemia. *Genes and Chromosomes in Chronic B-cell Leukemia*, 94, 1997.

[20] Rajendra N. Damle, Tarun Wasil, Franco Fais, Fabio Ghiotto, Angelo Valetto, Steven L. Allen, Aby Buchbinder, Daniel Budman, Klaus Dittman, Jonathan Kolitz, Stuart M. Lichtman, Philip Schulman, Vincent P. Vinciguerra, Kanti R. Rai, Manlio Ferrarini, and Nicholas Chiorazzi. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood*, 1999.

[21] Alireza Darvish, E. Bak, K. Gopalakrishnan, R. H. Zadeh, and Kayvan Najarian. A new hierarchical method for identification of dynamic regulatory pathways from time series DNA microarray data. In *IEEE Computational Systems Bioinformatics Conference*, 2004.

[22] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1977.

[23] Qian Diao, We Hu, Hao Zhong, Juntao Li, Feng Xue, Tao Wang, and Yimin Zhang. Disease gene explorer: Display disease gene dependency by combining Bayesian networks with clustering. In *2004 IEEE Computational Systems Bioinformatics Conference*, 2004.

[24] Judith Dierlamm, Lucienne Michaux, Arnold Criel, Iwona Wlodarska, Herman Van Den Berghe, and Dieter Kurt Hossfeld. Genetic abnormalities in chronic lymphocytic leukemia and their clinical and prognostic implications. *Cancer Genet Cytogenet*, 94, 1997.

[25] Edward R. Dougherty, Airuddha Datta, and Chao Sima. Research issues in genomic signal processing. *IEEE Signal Processing Magazine*, 2005.

[26] Jason Ernst, Gerard J. Nau, and Ziv Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 2005.

[27] G. Fleury, A. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop. Clustering gene expression signals from retinal microarray data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[28] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303, 2004.

[29] Xiang-Chao Gan, Alan Wee-Chung Liew, and Hong Yan. Biclulstering gene expression data based on a high dimensional geometric method. In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, 2005.

[30] Stephen B. Gauld, Joseph M. Dal Porto, and John C. Cambier. B cell antigen receptor signaling: Roles in cell development and disease. *Science*, 2002.

[31] Gad Getz, Erel Levine, and Eytan Domany. Coupled two-way clustering analysis of gene microarray data. *Proc Nat Acad Sci U S A*, 22(97), 2000.

[32] Gerard Govaert and Mohamed Nadif. An EM algorithm for the block mixture model. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27(4), 2005.

[33] O. Guipaud, L. Deriano, H. Salin, L. Vallat, L. Sabatier, H. Merle-Beral, and J. Delic. B-cell chronic lymphocytic leukaemia: a polymorphic family unified by genomic features. *Lancet Oncol*, 2003.

[34] Reinhard Guthke, Ulrich Mller, Martin Hoffmann, Frank Thies, and Susanne Tpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8), 2005.

[35] Alexander J. Hartemink, David K. Gifford, Tommi S. Jaakkola, and Richard A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 2002.

[36] Kai-Li He and Adrian T. Ting. A20 inhibits Tumor Necrosis Factor (TNF) alpha-induced apoptosis by disrupting recruitment of TRADD and RIP to the TNF receptor 1 complex in Jurkat T cells. *Molecular Cell Biology*, 2002.

[37] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainity in Artificial Intelligence*, 1999.

[38] Seiya Imoto, Tomoyuki Higuchi, Takao Goto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Combining microarrays aand biological knowledge for estimating gene networks via bayesian networks. In *Proceedings of the 2003 IEEE Bioinformatics Conference*, 2003.

[39] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11), 2004.

[40] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 2006. URL: http://www.genome.jp/kegg/pathway/hsa/hsa04010.html.

[41] Rishi Khan, Yujing Zeng, Javier Garcia-Frias, and Guang Gao. A Bayesian modeling framework for genetic regulation. In *Proceedings of the IEEE Computer Society Bioinformatics Conference*, 2002.

[42] Ulf Klein, Yuhai Tu, Gustavo A. Stolovitzky, Michela Mattioli, Giorgio Cattoretti, Hervé Husson, Arnold Freedman, Giorgio Inghirami, Lilla Cro, Luca Baldini, Antonino Neri, Andrea Califano, and Riccardo Dalla-Favera. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *Journal of Experimental Medicine*, 2001.

[43] Tetsuya Koide, Tadayoshi Hayata, and Ken W. Y. Cho. Xenopus as a model system to study transcriptional regulatory networks. *Proc Natl Acad Sci U S A*, 102(14), 2005.

[44] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci*, 2001.

[45] Maoxiang Li, Jun-Ying Zhou, Yubin Ge, Larry H. Matherly, and Gen Sheng Wu. The phosphatase MKP1 is a transcriptional target of p53 involved in cell cycle regulation. *J Biol Chem*, 2003.

[46] David J. Lockhart, Helin Dong, Michael C. Byrne, Maximillian T. Follettie, Michael V. Gallo, Mark S. Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Horton, and Eugene L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 1996.

[47] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational biology and Bioinformatics*, 2004.

[48] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, 2005. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene.

[49] Dmitry M. Malioutov. A sparse signal reconstruction perspective for source localization with sensor arrays. Master's thesis, Northeastern University, 2003.

[50] Linda Matsuuchi and Michael R Gold. New views of BCR structure and organization. *Curr Opin Immunol*, 2001.

[51] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. Kel, O. Kel-Margoulis, D. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 2003.

[52] Paul L. Mittelstadt and Anthony L. DeFranco. Induction of early response genes by cross-linking membrane Ig on B lymphocytes. *Journal of Immunology*, 1993.

[53] R. Mizuno, M. Oya, T. Shiomi, K. Marumo Y. Okada, and M. Murai. Inhibition of MKP-1 expression potentiates JNK related apoptosis in renal cancer cells. *Journal of Urology*, 2004.

[54] Mark Naylor and J. Donald Capra. Mutational status of Ig VH genes provides clinically valuable information in B-cell chronic lymphocytic leukemia. *Blood*, 1999.

[55] Wei Pan, Jizhen Lin, and Chap T. Le. Model-based cluster analysis of microarray gene-expression data. *Genome Biology*, 3(2), 2002.

[56] Jr. Paul W. Mielke and Kenneth J. Berry. *Permutation Methods: A Distance Function Approach.* Spriger-Verlag New York, Inc, 2001.

[57] F. Peng and D. Schuurmans. A hierarchical EM approach to word segmentation. In *In Proceedings of the 6th Natural Language Processing Pacic Rim Symposium*, 2001.

[58] Ben Y. Reis, Atul S. Butte, and Isaac S. Kohane. Extracting knowledge from dynamics in gene expression. *Journal of Biomedical Informatics*, 34, 2001.

[59] Andreas Rosenwald, Ash A. Alizadeh, George Widhopf, Richard Simon, R. Eric Davis, Xin Yu, Liming Yang, Oxana K. Pickeral, Laura Z. Rassenti, John Powell, David Botstein, John C. Byrd, Michael R. Grever, Bruce D. Cheson, Nicholas Chiorazzi, Wyndham H. Wilson, Thomas J. Kipps, Patrick O. Brown, and Louis M. Staudt. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *Journal of Experimental Medicine*, 2001.

[60] Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6), 2005.

[61] Joerg Schreiber, Richard G. Jenner, Heather L. Murray, Georg K. Gerber, David K. Gifford, and Richard A. Young. Coordinated binding of NF-$\kappa$B family

members in the response of human cells to lipopolysaccharide. *Proc Natl Acad Sci*, 2006.

[62] Jonathan Schug and G. Christian Overton. TESS: Transcription element search software on the WWW. Technical Report CBIL-TR-1997-1001-v0.0, Computational Biology and Informatics Laboratory, University of Pennsylvania School of Medicine, 1998. URL: http://www.cbil.upenn.edu/tess.

[63] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2), 2003.

[64] Vicki L. Seyfert, Vikas P. Sukhatme, and John G. Monroe. Differential expression of a zinc finger-encoding gene in response to positive versus negative signaling through receptor immunoglobulin in murine B lymphocytes. *Molecular and Cell Biology*, 1989.

[65] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Biomolecular interaction networks cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 2003.

[66] Aizheng Sheng, Yves Moreau, and Bart De Moor. Biclustering microarray data by Gibbs sampling. *Bioinformatics*, 19, 2003.

[67] Freda K. Stevenson and Federico Caligaris-Cappio. Chronic lymphocytic leukemia: revelations from the B-cell receptor. *Blood*, 103(12), 2004.

[68] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302, 2003.

[69] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 1999.

[70] Alexander Sturn, John Quackenbush, and Alatko Trajanoski. Genesis: Cluster analysis of microarray data. *Bioinformatics*, 18(1), 2002.

[71] Ioan Tabus, Ciprian Doru Giurcaneanu, and Jaakko Astola. Genetic networks inferred from time series of gene expression data. In *First International Symposium on Control, Communications and Signal Processing*, 2004.

[72] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander, and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci*, 1999.

[73] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18, 2002.

[74] A.H. Tewfik and Alain B. Tchagang. Biclustering of genes with coherent evolutions. In *Proc. 2005 IEEE International Workshop On Machine Learning For Signal Processing*, 2005.

[75] Amy Hin Yan Tong, Guillaume Lesage, Gary D. Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F. Berriz, Renee L. Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S. Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan Krogan, Zhigian Li, Joshua N. Levinson, Hong Lu, Patrice Menard, Christella Munyana, Ainslie B. Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L. Wong, Lan V. Zhang, Hongwei Zhu, Christopher G. Burd, Sean Munro, Chris Sander, Jasper Rine, Jack Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P. Roth, Grant W. Brown, Brenda Andrews, Howard Bussey, and Charles Boone. Global mapping of the yeast genetic interaction network. *Science*, 303, 2004.

[76] Rudolph Triebel, Wolfram Burgard, and Frank Dellaert. Using hierarchical EM to extract planes from 3D range scans. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2005.

[77] Laurent D. Vallat, Corey A. Kemper, F. Davi, John W. Fisher, and John G. Gribben. Temporal dependence model of BCR gene expression. *Science (In preparation)*, 2006.

[78] Laurent D. Vallat, Yuhyun Park, Cheng Li, and John G. Gribben. A temporal genetic program specific for aggressive chronic lymphocytic leukemia cells after B-cell receptor cross-linking. *Journal of Clinical Investigation (In preparation)*, 2006.

[79] Bas van Steensel. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nature Genetics*, 2005.

[80] Hester M. Wain, Michael Lush, Fabrice Ducluzeau, and Sue Povey. Genew: the human gene nomenclature database. *Nucleic Acids Research*, pages 169–171, 2002. URL: http://www.gene.ucl.ac.uk/nomenclature/.

[81] Shannon L. Werner, Derren Barken, and Alexander Hoffman. Stimulus specificity of gene expression programs determined by temporal control of IKK activity. *Science*, 2005.

[82] Eric P. Xing and Richard M. Karp. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, 17, 2001.

[83] Jiong Yang, Haixun wang, Wei Wang, and Philip Yu. Enhanced biclustering on expression data. In *Proceedings of the Third IEEE Symposium on bioinformatics and bioengineering*, 2003.

[84] Ryo Yoshida, Tomoyuki Higuchi, and Seiya Imoto. A mixed factors model for dimension reduction and extraction of a group structure in gene expression data. In *Computational Systems Bioinformatics Conference*, 2004.

[85] Haiyuan Yu, Nicholas M. Luscombe, Jiang Qian, and Mark Gerstein. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends in Genetics*, 19(8), 2003.

[86] Ya Zhang, Hongyuan Zha, and Chao-Hisen Chu. A time-series biclustering algorithm for revealing co-regulated genes. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, 2005.

[87] Zonghong Zhang, Alvin Teo, Beng Chin Ooi, and Kian-Lee Tan. Mining deterministic biclusters in gene expression data. In *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering*, 2004.

[88] Dongxiao Zhu and Alfred O. Hero. Gene co-expression network discovery with controlled statistical and biological significance. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[89] Dongxiao Zhu and Alfred O. Hero. Network constrained clustering for gene microarray data. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[90] Nongliao Zhu, Luis M. Ramirez, Rosaline L. Lee, Nancy S. Magnuson, Gail A. Bishop, and Michael R. Gold. CD40 signaling in B cells regulates the expression of the Pim-1 kinase via the NF-KB pathway. *Journal of Immunology*, 2002.