Inferring Genetic Regulatory Network Structure: Integrative Analysis of Genome-Scale Data

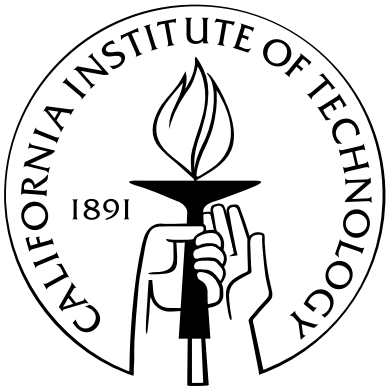
Thesis by

Christopher Edward Hart

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2005

(Submitted November, 22 2004)

Acknowledgements

Much of the work in this thesis came about through discussions with my advisor, Barbara Wold. She was always available, willing, and excited to discuss things with me. I learned a great deal during our our discussions, even when, as they often did, digress from our intended topic, or topics. My advisory committee complemented Barbara's mentoring very well, lending me support and ideas throughout my time at Caltech. The Wold lab provided good fellowship, good advice, and good ideas.

Specifically in writing this manuscript I received help from several friends and colleagues. I give much thanks to Titus Brown, Eliot Bush, Leslie Dunipace, Tracy Teal, Erich Schwartz, Brian Williams and even my grandfather Harry Hart for helping me proofread both the text, and ideas presented within.

I also need to thank my friends and family, both near and far, that kept me sane and happy throughout my time in graduate school. I also thank my parents for the curiosity they nurtured.

Abstract

With the aim of uncovering regulatory relationships that underly biological processes, we constructed a framework of computational tools and techniques to relate disparate genome-scale data within and across datasets. Using these tools we focus on the yeast cell cycle and the transcriptional network driving the transition into and out of G1. Through integrative analysis of genome-scale datasets we were able to recover many of the previously known transcriptional regulatory connections within the yeast cell cycle. We also found several novel hypothetical connections yet to be experimentally validated.

Much of the analysis of large-scale gene expression data has relied heavily on the application of clustering algorithms to identify sets of co-expressed genes (clusters). In chapter 2 we introduce several new techniques for comparing and evaluating microarray data, specifically focusing on clustering results. We discuss the need for quantitative methods for evaluating clustering methods, and discuss the application of comparative analysis of clustering results.

Remarkably, our analysis shows the results from any clustering algorithm are quite sensitive to slight perturbations to the data. Yet, the underlying structure revealed by most clustering algorithms remains fairly stable. These findings have a pragmatic impact on how clustering results should be interpreted and used. Chapter 3 uses the tools introduced in chapter 2 and performs a systematic comparison of the influence of noise on the stability and reliability of clustering results.

In chapter 4 we demonstrate the use of artificial neural networks (ANNs) to infer regulatory networks by combining expression data and protein:DNA binding data. We then compare these regulatory relationships to the presence of transcription factor binding sites. We also note evolutionary stability in some of the components of this network by compar-

ing results to other species of yeast.

Contents

List of Figures

Chapter 1

Introduction

The interactions of transcriptional regulators with target genes form regulatory networks that partially control many biological processes. The recent proliferation of sequenced genomes and the development of genome-scale functional assays provide a new context in which to better understand these regulatory networks. Using this data we can address questions regarding the evolution, structure and function of transcriptional regulatory networks more comprehensively across entire genomes. This complements and contrasts data from more conventional methods, which largely, are based on extrapolation from detailed studies of small numbers of genes.

1.1 The Yeast Cell Cycle: A Model of What's To Come

Throughout this work we use the yeast cell cycle as both a test case for evaluating new computational techniques and as an illustration of how these techniques can improve our understanding of the regulatory networks underlying a biological process. Much is understood regarding the genetics and molecular biology of the yeast cell cycle (for a recent review see [Breeden, 2003]). For evaluating the relative significance and abilities of new computational techniques this is quite valuable as the existing knowledge can be used as a type of internal control to compare results against. Further, because of the ease of experimentation, the well-established genetics, and a relatively small well-annotated genome, there has been an early adoption of high-throughput methods and genome-scale technologies in yeast. This has resulted in a vast amount of publically available high quality data

[Dolinski et al., 2004, Csank et al., 2002]. Complementing this, several additional yeast genomes have been fully sequenced: 7 related *Saccharomyces* species [Cliften et al., 2003, Kellis et al., 2003] and the more distantly related *Schizosaccharomyces pombe* [Wood et al., 2002] and Candida albicans [Jones et al., 2004].

Gene expression is carefully regulated within the yeast cell cycle and between 10-20% of genes have been observed to exhibit cell cycle-dependent gene expression [Lichtenberg et al., 2004]. Complex post-translational regulation, in particular a network of kinases (Cdks) and regulated proteolysis, is integral to regulating and maintaining the proper progression of the yeast cell cycle [Mendenhall and Hodge, 1998]. However, the downstream effect of many of these reactions is to change the transcription of genes which in turn play a pivotal role in the progression of the yeast cell cycle (reviewed [Breeden, 2000, Breeden, 2003]).

An overall conclusion from genome-wide RNA expression profiling throughout the cell cycle is that each phase of the cell cycle is characterized by very specific gene expression patterns [Cho et al., 1998, Spellman et al., 1998]. Transitions into and maintenance of cellular states, such as the cell cycle phases, are of particular interest in not only the cell cycle but also in development. The relationship between the structural features of the regulatory network and the functional mechanisms that underlie cell cycle progression may be similar in some developmental contexts as well.

The cell cycle is broken into four major phases that are recognized primarily based on their cellular activities. We focus on the transcriptional network underlying cell cycle progression. Specifically, G1 and the transition into it from M phase and the progression out of it into S phase. The G1 phase of the cell cycle is of particular interest as its duration, unlike other phases, is controlled depending on the growth conditions of the cell [Forsburg and Nurse, 1991]. In addition, based on genome-wide expression profiling the G1 phase of the cell cycle has the most pronounced gene expression pattern of any of the cell cycle phases (chapters 2 and 3).

Two Swi6-containing dimeric complexes, SBF (SCB binding factor) and MBF (MCB binding factor) have been identified and are considered to be the primary transcriptional regulators of G1 gene expression [Andrews and Herskowitz, 1989, Koch et al., 1993]. SBF binds DNA together with Swi4 and recognizes the SCB (Swi4 cell cycle box) binding site

while MBF binds DNA with Mbp1 and recognizes the MCB (Mlu I cell cycle box) binding site [Nasmyth, 1985, Breeden and Nasmyth, 1987, Koch et al., 1993]. Remarkably many genes that appear to be primarily regulated by MBF contain only a single instance of MCB, its consensus binding site (chapter 4). Yet a single site is not sufficient to drive expression *in vivo* when placed in a reporter construct [Lowndes et al., 1991]. We discuss this further in chapters 4 and 5.

Progression out of G1 and into S phase is termed *Start*. It marks the initiation of DNA replication and commits the cell to mitosis. This is a highly guarded process and checkpoints, such as the DNA damage checkpoint, are in place to ensure that cellular and hereditary integrity are maintained. Ultimately the activity of Cdc28, primarily modulated by the G1 cyclins Cln1, Cln2 and Cln3, controls the exit from G1 and the entry into S (reviewed [Mendenhall and Hodge, 1998]). Once past *Start* the transcription of the mitotic cyclins increases and they, in turn, repress the G1 cyclins [Amon et al., 1994]. Coupled with the inherent instability of the Cln proteins the cell is able to shut down the G1 processes [Schneider et al., 1998].

At the M/G1 transition many of the B-type cyclins (Clbs) are targeted for proteolysis by the anaphase promoting complex (APC). During this process Swi5 is dephosphorylated by Cdc14 because of the degradation of Clb5. Dephosphorylation of Swi5 allows it to translocate to the nucleus where it begins activating early G1 gene expression [Visintin et al., 1998]. Among the genes activated by Swi5 are Sic1 and Rme1. Expression of Sic1, a cdk inhibitor, further inactivates any remaining Clbs [Knapp et al., 1996]. Rme1 encodes a nuclear localized transcription factor which seems to be involved in the induction of CLN2 and also inhibits meiosis through repression of IME1 (Initiator of Meiosis 1) [Frenz et al., 2001].

Meanwhile, mediated by the ECB (Early cell cycle box) binding site, SWI4 and CLN3 transcription is also activated during the M/G1 transition [MacKay et al., 2001]. The specificity of the ECB site comes from its interplay with neighboring sites and transcriptional cofactors. The MADS box protein Mcm1 is bound *in vivo* to ECB elements throughout the cell cycle, so is unlikely to impart cell cycle phase specificity on its targets [Mai et al., 2002]. Neighboring sites for the transcriptional repressors Yox1 and/or Yhp1 have been identified

and shown to be functional in many of the M/G1 expressed genes (including SWI4 and CLN3) [Pramila et al., 2002]. Also adding specificity are flanking forkhead sites which recruit Fkh2-Ndd1 complexes [Koranda et al., 2000].

1.2 A Genomic Context

1.2.1 Genetic Regulatory Networks

As observed in the yeast cell cycle, one perspective is that the repertoire and dynamics of gene expression largely control the state and fate of a cell [Davidson, 2001, Ptashne and Gann, 2002]. Although the exact biochemical process of transcriptional regulation is complex and not fully understood, simplifying models can provide critical insights [Bolouri and Davidson, 2002]. For instance, genetic regulatory network models that relate transcriptional regulators to target genes and the cis-elements they bind have been shown to provide a scaffold of information on which to build new hypotheses and better understand development [Davidson et al., 2003]. Likewise, interaction models based on protein:protein [Uetz et al., 2000, Ho et al., 2002] and protein:DNA interaction measurements [Lee et al., 2002, Horak et al., 2002, Harbison et al., 2004] have been superimposed onto biochemical and genetic pathways to better understand the evolution and relationship of these pathways in different species [Kelley et al., 2004].

1.2.2 Transcriptional Regulatory Networks and Co-expression

Many genes show similar expression patterns and can be classified into co-expression groups which exhibit comparable expression patterns across a variety of conditions [DeRisi et al., 1997, Wodicka et al., 1997, Cho et al., 1998, Spellman et al., 1998]. Specific cis-regulatory interactions of transcriptional regulators play a central role in transcriptional control for most genes. As such, many co-expressed genes also appear to be regulated by the same transcriptional regulators [Spellman et al., 1998, Cho et al., 1998]. However, not all co-expressed genes are co-regulated, and not all co-regulated genes are co-expressed (chapters 2 and 4). These observations lead to many questions regarding the nature of co-expression.

Why are co-expressed genes co-expressed? Do the same set of transcriptional regula-

tors drive their expression? What features of the regulatory sequence of these genes are important in driving their expression patterns? How do regulatory features evolve? Answers to these questions will yield information on how genetic regulatory networks drive cellular physiology and development. Although only partially addressed within this thesis, these questions underlie much of the work presented.

1.3 Relating Genome-Scale Data to Transcriptional Regulatory Networks

1.3.1 Dealing With Data Quality

High throughput techniques create vast amounts of genome-scale data. Yet, a caveat of these new approaches is often increased noise and loss of experimental precision. Replication and careful experimental design can only alleviate some of these problems. As an example, when performing genome-wide RNA expression profiling using microarrays, noise is introduced at each step of the experiment. Tu et al. (2002) systematically surveyed the amount of noise introduced at the critical steps during a typical microarray experiment: RNA extraction, RNA-to-target synthesis, and hybridization. The RNA extraction and target synthesis were found to be the dominant source of noise within a particular experiment. However, the variance between complete biological replicates dwarfed these sources of noise. Chapter 3 focuses on gaining a better understanding of the implications of noise on our ability to identify groups of co-expressed genes, or clusters. By developing metrics to quantify and dissect the consistency of clustering results we gain an appreciation for which features of clustering results are likely artifact and which features are likely true.

Gaining a more detailed understanding of clustering results is clearly important as they often form the basis for downstream analysis, such as the network inference discussed in chapter 4. Further, cluster membership is often used as the basis for the functional annotation of genes in many databases, such as Saccharomyces Genome Database (SGD) [Dolinski et al., 2004]. Different clustering results can also arise from clustering data from different experiments, such as time-course RNA expression data collected from cells ex-

posed to different stimuli. Chapter 2 describes a computational framework built around the need to be able to better understand the differences and similarities between clustering results. We then show how these differences and similarities can have implications on uncovering functional relationships between co-expression and regulatory networks.

Although large-scale datasets have limitations, they are mostly accurate. As discussed in the following sections, different large-scale datasets impinge on regulatory networks differently. By leveraging multiple large-scale datasets together some of their limitations can be overcome. We use this idea in chapter 4 to infer cell cycle phase-specific transcriptional regulatory connections from genome-wide RNA expression data and genome-wide protein:DNA binding data. Gifford and co-workers also applied this idea when devising the GRAM (Genetic Regulatory Modules) algorithm. Similarly, working to discern regulatory modules from ChIP/chip data in yeast they demonstrate that when genes show a persistent pattern of co-expression across a wide variety of conditions they could include binding measurements that would have otherwise been considered marginal, without sacrificing false positive rates [Bar-Joseph et al., 2003].

1.3.2 Expression Profiling: DNA Microarrays

The ability to measure co-expression on a genomic scale has been largely driven by the development of microarray technologies during the past decade [Pease et al., 1994, Schena et al., 1995]. They have been used to study many model organisms. A common feature in each implementation of the technology is the linking of DNA molecules of known sequence to specific locations on a solid substrate. Each of these features, known as probes, are complementary to a target sequence that is produced from each gene's mRNA. Current state of the art technologies allow for the creation of arrays capable of measuring the expression levels of up to $\sim 40,000$ genes simultaneously. Although specifics vary between microarray platforms, in each case RNA is extracted from cells and labeled targets are produced representing each RNA molecule. The labeled targets are then hybridized to the microarray and the relative amount of label at each feature is converted to a relative expression level for each gene. These measurements comprise the gene expression profile, or expression signature of the

cells being assayed. Gene expression profiles for cells under varying conditions are typically assembled into a single dataset. These varying conditions can be: kinetic time-points across a biological process; dose-response measurements; different tissues; diseased tissues such as tumors; or different mutant strains or individuals. In each case, the expression level of each gene across each condition is referred to as an expression trajectory or expression vector.

Expression data provides a very powerful readout of the activity of genetic regulatory networks but cannot be used in isolation to understand the architecture of the network. The expression level of a gene is a result of both the rate of transcription and the rate of degradation of the mRNA. Although often the impact of post-transcriptional regulation on expression levels is ignored, these processes can affect the measured expression level for a gene. For instance, in yeast the half-life of the vast majority of poly-A mRNA has been shown to vary between 10 and 30 minutes, although in the extremes decay rates vary between 3 and 90 minutes [Wang et al., 2002]. Another complication is that the expression level of an mRNA molecule is only a surrogate for measuring the actual protein concentrations in the cell, and the protein product of a gene is often what invokes influence in cellular behavior. Further post-translational modification of proteins is often critical in regulating many processes.

1.3.3 Measuring *in vivo* Transcription Factor Interactions : Chromatin Immunoprecipitation / Microarray analysis (ChIP/chip)

Chromatin immunoprecipitation (ChIP) is an assay to measure the *in vivo* binding activity of transcriptional regulators [Orlando, 2000]. By exposing cells to a crosslinking agent such as formaldehyde, the protein:DNA interactions that are occurring at the time of crosslinking are captured. Using antibodies or an affinity tag directed specifically against a particular regulator, both the regulator and the DNA to which it is bound at the point of crosslinking can be retrieved. Using gene specific primers, a PCR assay can be used to show selective enrichment for DNA of a particular regulatory sequence. This method has been extended using microarray technology, where intergenic sequences are printed onto

slides. The bound DNA is then evaluated by ligation-mediated PCR amplification, labeling and hybridization to the intergenic microarray. By using microarrays as opposed to conventional methods, the *in vivo* binding activity of a transcriptional regulator can be measured across an entire small genome [Ren et al., 2000, Iyer et al., 2001] or one or two chromosomes in larger genomes [Martone et al., 2003, Cawley et al., 2004]. This technique is often referred to as ChIP/chip or ChIP/array. High throughput application of this methods has allowed Young and colleagues to collect the *in vivo* binding profiles for nearly all transcriptional regulators in yeast [Lee et al., 2002, Harbison et al., 2004]. These methods are especially powerful in gaining a better understanding regulatory networks, as they provide a direct measure of the transcriptional interactions which underlie many cellular processes.

Current experimental limitations, mostly cost and time constraints, in the large-scale application of ChIP/chip prevent resolving the cellular context of the measured interactions. Investigators can only survey the binding activity of all transcriptional regulators either in a single cellular state, or a heterogeneous population of cellular states. As a consequence many relevant regulatory connections are either missed, or the time and space domain in which observed connections have regulatory significance is blurred. Further, regulators may bind upstream to a gene *in vivo* with no effect. As discussed in chapter 5 in some cases, but likely not all cases, this may be because the regulator influences the transcription of a neighboring gene.

The kinetics of the transcriptional response of a cell can be well captured using existing microarray technology in some settings, such as yeast cultures. Changes in the gene expression signature of a cell are reflections of the underlying regulatory network that drives them. Coupling kinetic expression data to heterogeneous ChIP/chip data allows for the dynamics of one set of measurements to be imposed on another set of non-dynamic measurements (Chapter 4). More direct time resolved ChIP/chip data would also facilitate this, but this is currently impractical on a large scale.

1.3.4 Genomic Sequence

Expression data can be kinetic and ChIP/chip binding data is usually collected from physiologically heterogeneous cells and is not dynamic. The genome underlying all of it is static, with only a few exception such as mating type switching in yeast. However, the genome is mostly informationally complete since it contains the information necessary to regulate itself. Understanding how all the information encoded within a genome relates to the functional regulatory networks of the cell is a long term goal. Working towards this goal, in chapters 4 and 5 we first infer cell cycle phase-specific regulatory connections from expression data and chIP/chip data. We then relate some of the regulatory connections to conserved enrichment and disenrichment of specific transcription factor binding sites upstream of genes that show phase-specific expression patterns.

Mapping regulatory connections between transcriptional regulators and target genes is difficult in part because the regulatory sequences to which they bind are difficult to define either experimentally or computationally. Experimentally, SELEX methods can be used to define binding sequences. These methods use a particular transcription factor to enrich for a biased sub-population of randomized oligonucleotides [Klug and Famulok, 1994]. Other *in vitro* methods also exist, but none isolate the *in vivo* binding sequences for a transcription factor. This results in the loss of both the cellular and genomic context of a binding site which can influence its function. For example, in yeast Ndd1 interacts with DNA through Fkh2, and the binding activity of Fkh2 is likely modulated by Ndd1 as deletion of Fkh2 suppresses the lethality of Ndd1 depletion [Koranda et al., 2000]. Therefore, unless the assay surveyed the Ndd1-Fkh2 complex, the measured *in vitro* binding affinities may be different than what occurs *in vivo*.

Computationally even describing the sequences to which regulators bind to is a challenge. Qualitative descriptions, such as an IUPAC consensus, specify the allowed degeneracy for each position within the recognition site a transcriptional regulator binds to. These descriptions lack any ability to describe position specific base biases. More quantitative descriptions based on a weight matrix can describe position specific base biases, but both of these descriptions implicitly assume that each position in the sequences that a transcrip-

tional regulator binds to is independent from every other [Benos et al., 2002].

Beyond defining and describing the sequences to which transcriptional regulators bind, another challenge is identifying which occurrences of these sequences within the genome are functional. This is an especially extensive problem in larger genomes such as mouse and human because of the vast amounts of intergenic sequence. Although several databases have been constructed to maintain a compendium of binding sites [Galperin, 2004], there are large inconsistencies in the quality and source of the binding site descriptions. By using comparative genomics, regulatory sequences that are under selection can be isolated to help uncover functional elements (reviewed in [Miller et al., 2004]). Comparative analysis can also be employed to highlight other characteristics that have been conserved throughout evolution, such as the statistical enrichment of the presence of binding sites (Chapter 4). In addition, by using comparative genomics to map gene orthologues between different species, expression datasets as well as regulatory networks can be compared [Stuart et al., 2003, Kelley et al., 2004]. By leveraging the similarities and differences across different species, evolutionary relationships can be exposed onto the underlying regulatory networks.

In the following chapters, using several different large-scale datasets we explore the biology of the yeast cell cycle.

Part I

Comparing, Mining and Understanding Clustering Results

Chapter 2

Framework for Quantitative Comparison and Exploration of Microarray Clusterings

2.1 Introduction

A key step in analyzing most large-scale gene expression studies is clustering or otherwise grouping gene expression data vectors and conditions (individual RNA samples) into sets that contain members more similar to each other than to the remainder of the data. To do this biologists have at their disposal a wide range of techniques including supervised and unsupervised machine learning algorithms and various heuristics, such as k-means, phylogenic-like hierarchical clustering, Expectation Maximization of Mixture models, Self Organizing Maps, Support Vector Machines, statistical models, Fourier analysis, etc. [Cho et al., 1998, Eisen et al., 1998, Golub et al., 1999, Tamayo et al., 1999, Ross et al., 2000, Ihmels et al., 2002]. Their purpose is to detect underlying relationships in the data, but different algorithms applied to a given dataset typically deliver different and only partly concordant results. Sometimes the differences in cluster organization and content are so large as to cast doubt on any resulting gene or condition lists. Still more diverse outcomes arise from using different distance metrics, initialization conditions, and data preprocessing protocols. Is one clustering result objectively more correct than another? How important are the differences and what specific gene groups or samples within a clustering are most affected? Are specific intersecting subsets of genes robust and consistent from

one algorithm to another? Do the inconsistencies highlight overall dataset properties that are important or are they concentrated on differences of marginal biological significance? To answer such objectively we needed a way to make systematic, quantitative comparisons and we needed tools to effectively mine the resulting comparisons. In a similar manner, as data sources become more extensive, there is a growing need to quantitatively compare clustering results derived from different studies, and good comparative tools should also handle this class of problems.

To address these needs, we developed a mathematical and computational framework designed for comparative clustering analysis. Confusion matrices are the foundation for the comparisons. A confusion matrix effectively summarizes the pairwise intersections between clusters derived from two clustering results. These similarities can then be quantified by applying scoring functions to the confusion matrix. We use two different scoring functions for this purpose: 1) Normalized Mutual Information (NMI) which measures the amount of information shared between the two clustering results [Forbes, 1995]. 2) A linear assignment (LA) method which quantifies the similarity of two clusterings by finding the optimal pairing of clusters between two clustering results and measuring the degree of agreement across this pairing [Gusfield, 2002]; this work. Prior to this work, metrics for evaluating the total number of data point pairs grouped together between two different clusterings began to address the need for quantifying overall differences [Rand, 1971, Hubert and Arabie, 1985, Levine and Domany, 2001, Ben-Hur et al., 2002]. Ben-Hur et al. (2002) used this to help determine an optimal number of clusters (K) and to assess the overall validity of a clustering. These prior techniques did not, however, offer the capacity to isolate and inspect the similarities and differences between two different clusterings. We also introduce application of receiver operator characteristic (ROC) analysis to this class of problems [Peterson, 1954, Swets, 1988]. ROC enables us to quantify the distinctness of a given cluster relative to another cluster or relative to all non-cluster members.

The comparison algorithms are integrated into a set of interactive analysis tools collectively called CompClust. It enables a user to organize, interrogate and visualize the comparisons. In addition to comparative cluster analysis an important feature of this soft-

ware is that it establishes and maintains a link between the outputs of clustering analyses and the primary expression data, and, critically, with all other desired annotations. In the sense used here, "annotations" include other kinds of primary and metadata of diverse types. This gives important flexibility in data mining and permits expanded analyses that include results from other kinds of experiments such as global protein:DNA interactions (ChIP/Array), protein:protein interactions, comparative genome analysis, or information from gene ontologies.

CompClust methods and tools are agnostic about the kind of microarray data (ratiometric, Affymetrix, other) and the types of algorithms used. We used the tools to analyze two different sets of yeast cell cycle expression data that were clustered by four very different methods: A statistical clustering algorithm (Expectation Maximization of a Mixture of Diagonal Gaussian distributions (EM MoDG)) (this work); a human-driven heuristic [Cho et al., 1998]; a Fourier transform algorithm designed to take advantage of a periodic time course patterns [Spellman et al., 1998]; and an agglomerative version of the Xclust phylogenetic ordering algorithm [Eisen et al., 1998]; and this work. We then show that gene groups derived from these comparative analyses can be integrated with data on evolutionarily conserved transcription factor binding sites to identify regulatory modules. The results begin to illustrate how a more quantitative and nuanced understanding of both global and local features in the data can be achieved, and how these can be linked with diverse kinds of data types to infer connectivity between regulators and their target gene modules.

2.2 Results

2.2.1 Mathematical tools for organizing and quantifying microarray clusterings

Confusion matrices and comparative metrics. A confusion matrix summarizes all pairwise intersections between all clusters from any two clusterings of the same data. A confusion matrix is the matrix of cardinalities of all pairwise intersections between two different clusterings (see Methods). We then apply different scoring functions to the confusion ma-

trix to quantify similarity: 1) Normalized Mutual Information (NMI) measures the amount of information shared between two clusterings [Forbes, 1995], 2) Linear Assignment (LA) optimizes the number of data vectors in clusters that correspond to each other, thereby identifying the optimal pairing of clusters. LA also reports the percentage of data vectors contained within those clusters, and this can be used to assess similarity of results globally over the entire dataset and locally on a cluster pair by cluster pair basis ([Gusfield, 2002]; and this work). See methods for mathematical descriptions of confusion matrices, NMI and LA. The combined use of LA and NMI metrics can provide immediate insight into the nature of global differences between two microarray clusterings by capitalizing on the fact that NMI is asymmetric and LA is symmetric (see methods 2.4.5 and table 2.1). This readily discriminates instances in which one clustering is different from the other, but is essentially a refinement of the other verses a fundamentally different view of the data structure.

Confusion arrays organize and display comparative analyses Given two different clusterings of a dataset and a global evaluation of their similarity via NMI and LA, we next needed a way to systematically compare clusters derived from one algorithm with those from another in a way that is more effective and intuitive than inspection of gene lists. To do this we define the confusion array, which is a direct extension of a formal confusion matrix. Each cell of such a confusion array for two different clusterings contains the intersection set between the two parent clusters (as opposed to the cardinality of this set, as in a confusion matrix; see Methods 2.4.2). In the context of the CompClust system, the confusion array cells can then be interactively mined. Confusion arrays for two different clusterings, one an Affymetrix yeast cell cycle dataset [Cho et al., 1998] and the other a deposition ratiometric dataset [Spellman et al., 1998] are shown in Figures 2.1 and 2.5, and are analyzed further below.

Understanding cluster relatedness: Receiver Operator Characteristic (ROC) measures cluster overlap Whatever algorithm has been used to cluster data, it is useful to find out how distinct each cluster is from all the others and how distinct any particular

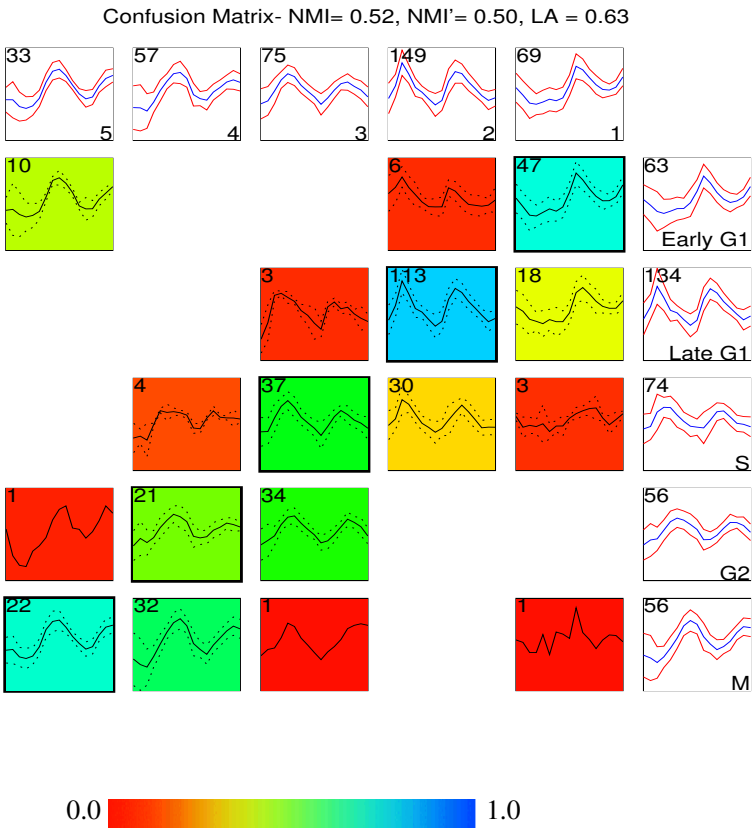


Figure 2.1: Comparing two clustering results using a confusion array. Shown in this comparison is a supervised clustering result published in the original study by Cho et al. (1998) and results from running an unsupervised clustering (EM MoDG, see methods) on the same Affymetrix microarray dataset profiling yeast gene expression through two cell cycles. The confusion array is composed of a grid of summary plots. Each summary plot displays the mean (blue or solid) expression level of a group of genes as well as the standard deviation (red or dashed). Summary plots with a white background represent clusters from either the the Cho et al. (1998) clustering result (along the right most column) or the EM MoDG clustering result (along the top row), cluster names are in the lower right corner, and the number of genes in each cluster is displayed in the upper left corner. Summary plots with a colored background represent cells within the confusion array (see methods) where each cell C_{ij} represents the intersection set of genes that are in common between the Cho et al. (1998) cluster i and the EM MoDG result cluster j. Again the upper left hand corner display the number of genes within a confusion matrix cell. The background of each plot is colored according to a heat-map (scale below) that registers the proportionate number of genes in the cell compared with the corresponding cluster in the EM MoDG result. Intersection cells with dark outlines indicate the optimal pairings between the two data partitions, as determined from the linear assignment calculation (eq. 2.2). Quantitative measures of overall similarity between the two clustering results using both Linear Assignment (LA) and Normalized Mutual Information (NMI), are displayed in the graph title (see methods).
.

cluster is from another specific cluster. This is especially pertinent when membership in a cluster will be translated into a gene list that ultimately becomes a functional annotation or defines which genes will be input into higher-order analyses. To address this issue we applied classical Receiver Operator Characteristic (ROC) analysis (Methods). In this context, cluster assignment is used as the "diagnosis" and the distance of each expression vector from the cluster mean vector is the "decision criterion". The corresponding ROC curve plots the proportion of cluster members versus the proportion of non-cluster members as the distance from the cluster centroid increases (figure 2.2). This can be interpreted geometrically as expansion of a hypersphere from the cluster centroid until all members of the cluster are enclosed. Thus, when one cluster is completely separate from all other data, all of its members are closer to the cluster center than all non-members and the area under the ROC curve is 1.0 (Figure 2.2B). When a cluster is not fully separable from the remainder of the data, the ROC curve rises more slowly and the area under the ROC curve < 1.0. In the limit, when the two classes are perfectly mixed, the ROC curve closely follows X=Y and the area under the curve drops to ~ 0.5 (figure 2.2D). The shape of the ROC curve also contains additional information about how cluster overlap is distributed, and this information can be used to choose useful data mining cut-offs that mark discontinuities and cluster substructure (see below in section 2.2.5 and figure 2.4). It can also be used interactively within CompClust to explore and select data vectors (genes) that are closer or more distant from the cluster center. Selection of vectors not assigned to the cluster yet positioned at overlapping distances from its center, is also possible and is often instructive (section 2.2.5 and figures 2.2, 2.4 below).

2.2.2 Comparing clusterings of yeast cell cycle microarray datasets

We next performed comparative analyses on clustering results from two different yeast microarray time course datasets (one Affymetrix and one ratiometric), each composed of genes that are differentially expressed over the cell cycle [Cho et al., 1998, Spellman et al., 1998]. These comparisons provide valuable perspective, since gene classification results from the original gene clusterings of these time courses have been mined in many subsequent stud-

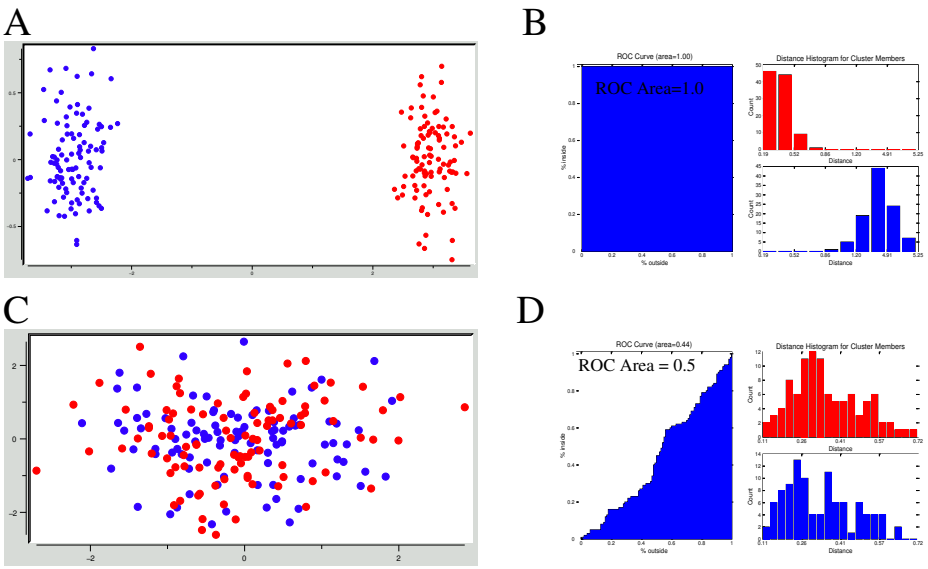


Figure 2.2: Example receiver operator characteristic (ROC) curves to assess cluster overlap. An ROC curve (panels B and D, left side) is drawn as a function of moving outward from a cluster center and counting the proportion of cluster members (blue points) encountered along the Y-axis vs the proportion of non-cluster members (red points) encountered along X-axis. The collection of distances from every point within a cluster and every point outside a cluster is binned and used to create the distance histograms (panels B and D, right side). Shown in red is the distance histogram for cluster members and cluster non-members are shown in blue. Two extreme cases are exemplified in this figure. *A)* Example expression data falling into two completely discrete clusters highlighted in red and blue. *B)* The corresponding ROC curve (left) and distance histograms (right) for the sample data shown in panel A. Notice since all cluster members are encountered before any non-cluster members the area under the ROC curve is 1.0. The distance histograms also show this perfect separation. *C)* Example expression data falling into two completely overlapping clusters highlighted in red and blue. *D)* The corresponding ROC curve (left) and distance histograms (right) for sample data shown in panel B. Notice since cluster members and non-cluster members are encountered at an equal rate as a function of distance from the cluster center the ROC curve approximates the line $X = Y$ and the area under the ROC curve ~ 0.5. This overlap is also highlighted in the distance histograms because the distributions of distances for cluster members completely overlaps with that of the distribution of distances for non-cluster members.

ies and have been introduced as gene annotations in widely used databases (Incytes' YPD [Csank et al., 2002], SGD (http://yeastgenome.org)). We generated a new clustering for each dataset, in each instance selecting an algorithm that differs substantially from the one used in the original publication but should also be entirely appropriate for the dataset. For the Cho et al. (1998) dataset we used EM MoDG (Expectation Maximization of a Mixture of Diagonal Gaussians [Dempster et al., 1977], which is an unsupervised method that searches for the best statistical fit to the data modeled as a mixture of Gaussian distributions. The heuristic used in the original report [Cho et al., 1998] is a supervised method based on biologist's knowledge of cell cycle phases. The heuristic focused on the time of peak expression for each gene trajectory to guide assignment of each gene to one of five time domains associated with Early G1, Late G1, S, G2, and M phases of the cell cycle. For the second dataset [Spellman et al., 1998], we performed agglomerative phylogenic hierarchical clustering of the tsCDC15-mutant synchronized data. This algorithm is based on the widely used Xclust phylogenetic ordering algorithm [Eisen et al., 1998], onto which we grafted an agglomeration step designed to establish objective boundaries in the tree (Methods). This result was compared to the result reported by [Spellman et al., 1998], in which they used a Fourier transform-based algorithm to assign expression vectors to phases of the cell cycle.

2.2.3 Global similarity measures

Comparison of the two clusterings of Affymetrix data from Cho et al. (1998) gave a global Linear Alignment (LA) score of 0.63 and Normalized Mutual Information (NMI) scores of 0.52 and 0.50, immediately indicating that EM MoDG and the heuristic classification have produced substantially different results. The LA value of 0.63 says that the optimal pairing of clusters still classifies 37% of the genes differently between the two algorithms. ROC curves and ROC areas were generated for each cluster (figure 2.3). Viewed in aggregate, this ROC analysis showed that clusters from EM MoDG are all better separated from each other than are any clusters from the original Cho et al. (1998) heuristic. Thus the ROC indices for EM MoDG are all 0.96 or above, and four of the five clusters are ≥ 0.98. In

contrast, the heuristic classification groups had ROC values as low as 0.82 for S phase and no better than 0.97 (M phase). By this criterion, we can say that EM clustering is a superior representation of the underlying data structure.

How are these differences between clustering results distributed over the dataset? We used PCA (Principle Component Analysis) to determine whether the two clusterings were globally similar or different in the way they partitioned the dataspace. PCA projects high dimensional gene expression vectors (each dimension here corresponding to a different RNA sample) into a different and lower dimensional space (usually two or three) [S. Raychaudhuri, 2000], in which the new PCA dimensions have each been selected to explain the maximal amount of variance in the data. A common feature of microarray datasets is that the first few principle components often capture most of the variation in the data (here 64%). Using CompClust to view the cluster means in PCA space allowed us to assess relationships between clusters from the two algorithms. Relative positions of cluster means in the PCA display the cell cycle progression in a counterclockwise pattern that is quite similar for the two algorithms. The absolute positions of the cluster centers in PCA space differ, but not extravagantly so, for most clusters. This is interesting because the coherence in overall structure would seem to contradict the rather high dissimilarities in cluster composition measured by the criteria LA and NMI, and shown graphically in the confusion array (figure 2.1). Considered together the results argue that the overall data structure, reflecting phases of the cell cycle, is robust and has been treated rather similarly by the two algorithms, even though 37% of individual gene expression vectors were assigned differently. This raises the question of which gene vectors have been differentially assigned and what biological meaning, if any, should be attached to the differences. These questions are addressed in sections 2.2.4, 2.2.6 and 2.3.2 by examining specific gene groups in the confusion array.

Using the ratiometric data of Spellman et. al. (1998) a comparison of the original Fourier-based algorithm versus agglomerated Xclust produced NMI and LA scores of 0.39, 0.41 and 0.60, respectively. These scores indicate the two clusterings are even more different in membership assignment, with 40% of genes falling outside the optimal linear assignment pairing. Since both NMI and LA scores are low, gene memberships for some

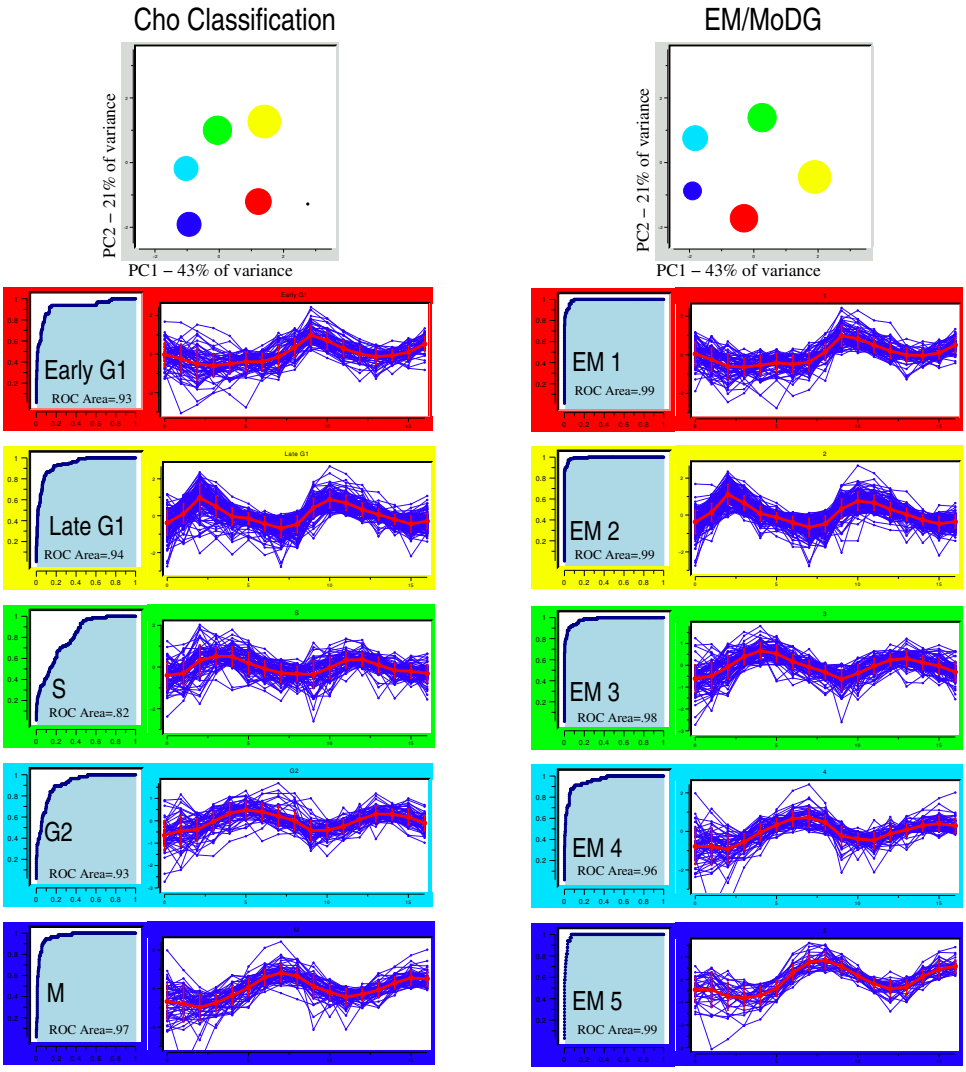


Figure 2.3: Principle component analysis, receiver operator characteristic (ROC) plots, and trajectory summary views of clusters from the Cho classification and an unsupervised clustering (EM MoDG) of an Affymetrix yeast cell cycle time course [Cho et al., 1998]. The top panel for each clustering results shows cluster means projected into the top two dimensions of the principle component space defined by the expression data (capturing 64% of the variance). The area of the marker size for each cluster is proportional to the number of genes in each cluster. Below are ROC curves (left) and trajectory summaries (right) for each cluster. The trajectory summaries display every gene's expression profile within a cluster as a blue line with time along the X-axis and expression along the Y-axis. The red line within each trajectory summary represents the mean expression level for the cluster. ROC area values are displayed within the ROC curve for each cluster. The background colors for the trajectory summaries and the PCA projection have been matched within each clustering result. In addition linear assignment was used to find the optimal mapping of clusters between the Cho classification and the EM MoDG result and the colors have been set accordingly.

clusters must be truly scrambled rather than being simple combinations of cluster unions and subsets (see Methods 2.4.5 and Table 2.1). PCA projection (figure 2.6) showed that some major cluster centers from the two algorithms are positioned very differently, both absolutely and relatively (note the yellow cluster corresponding to the Fourier S-phase group). The confusion array shows that XclustAgglom clusters often combine genes that are members of adjacent Fourier clusters, though in some cases it joins vectors from non-adjacent groups (the confusion around S phase is complex). ROC curves and scores also indicate that XclustAgglom has done a slightly better job of segregating data into discrete groups that reflect underlying data structure, while the Fourier analysis groups are less coherent and often seem to mix members of kinetically adjacent groups as detailed in the confusion array (Figures 2.5 and 2.6). This may be due, in part, to the use of a small number of then known genes to center landmark phases by the Fourier algorithm. The fact that this phase assignment was a "somewhat arbitrary" step in the original analysis was pointed out by Spellman et al. (1998).

2.2.4 High resolution cluster comparison

Confusion arrays can be used to explore issues raised by global analyses and to mine relationships between individual clusters in more detail. The latter activity can then be used to make refined and edited gene lists based on expert opinion or on computationally objective criteria. We applied linear assignment to the confusion matrix of the Cho heuristic and the EM MoDG results and produced the corresponding adjacency matrix (Methods, Equation 2.3). This delivered an objectively optimized pairing of EM cluster 1 with Cho "Early G1"; EM cluster 2 with Cho "Late G1", and so on, as shown in the array visualization (figure 2.1). Each cell in the confusion array contains the corresponding gene vectors and displays the calculated mean vector for each intersect cell in the array.

The confusion array highlighted relationships that were not clear from figures 2.3 or 2.6. For example, in the Affymetrix platform data [Cho et al., 1998], both algorithms identified two gene classes within G1, (red and yellow respectively in the PCA analysis of figure 2.3). However, the EM-1 cluster shares only 67% of its content with the Cho "Early G1," and

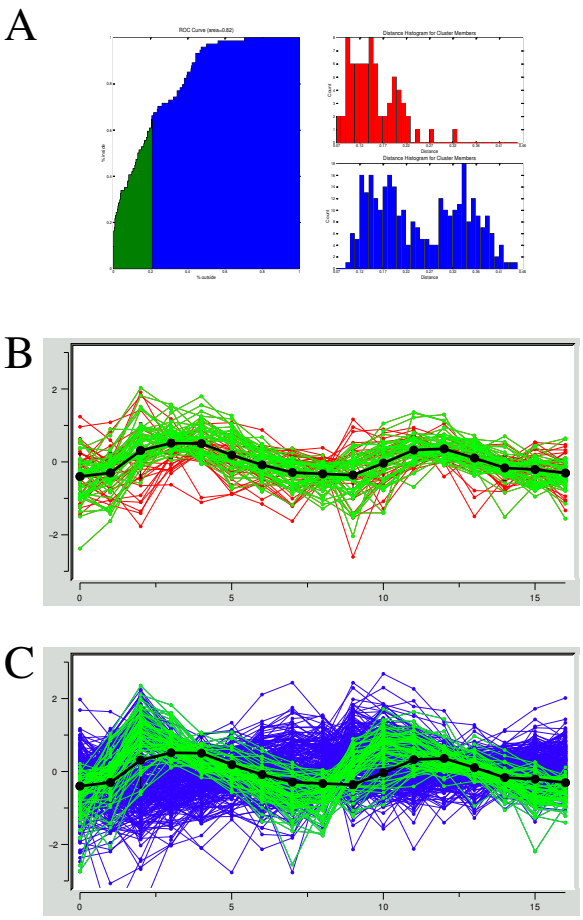


Figure 2.4: Receiver operator characteristic (ROC) analysis of the S phase cluster of Cho et al. 1998. *A)* ROC curve (left) shows the overlap between this cluster of 74 genes and genes from all other clusters in the time course analysis (383 genes in total, selected by inspection by Cho et al (1998) for cycling behavior). The area under the ROC curve is 0.82. The area under the curve highlighted in green demonstrates selection of genes from S-phase that overlap with other clusters least. At the shown distance threshold, 66% of genes from the Cho determined S-phase cluster are selected, and the overlap with only $\sim 20\%$ non-S-phase genes. (A Right) Correlation distance histograms illustrating the distribution of distances to the center of the S phase cluster for non-cluster members (bottom/blue) and for all S phase cluster members (top/red). B) Expression trajectories for the 74 genes in the S phase cluster, highlighting in green cluster members represented by the green highlight in panel A. C) Expression trajectories for all genes outside the S phase cluster of the Cho clustering highlighting in green non-cluster members represented by the green highlight in panel A.

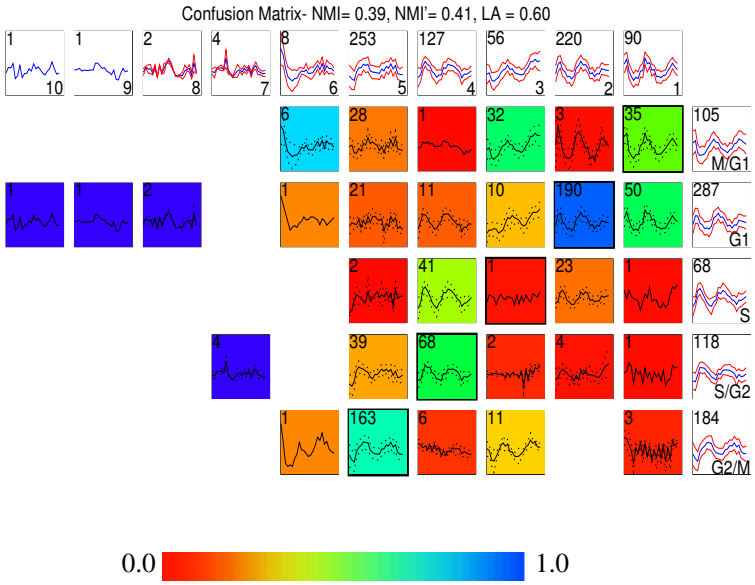


Figure 2.5: Comparing two clustering results on a ratiometric microarray dataset using a confusion array. Shown in this comparison is a Fourier clustering result published in the original study by Spellman et al. (1998) and results from running an unsupervised clustering (XclustAgglom, see methods) on the same ratiometric microarray dataset as the Fourier analysis was run on. Details of the figure layout are discussed in the legend of figure 2.1. Here the 5 Fourier clusters are shown along the rows, while the 10 XclustAgglom clusters are displayed across the columns.

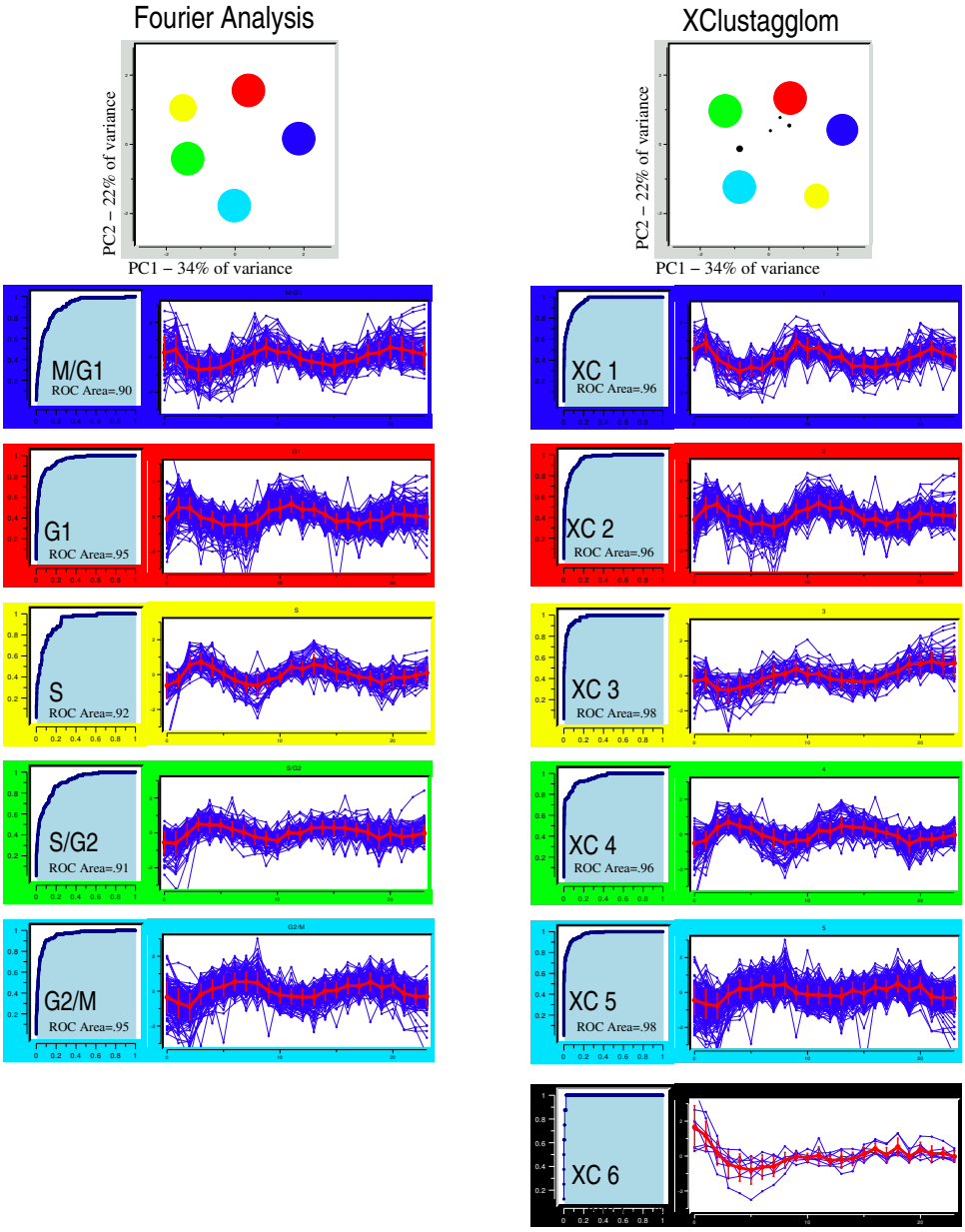


Figure 2.6: Principle component analysis, ROC plots, and trajectory summary views of clusters from the Fourier classification and an unsupervised clustering (XclustAgglom) results from the ratiometric yeast cell cycle time course [Spellman et al., 1998]. Details of the figure layout are the same as for figure 2.3. Only the 6 largest clusters are shown in the XclustAgglom. Clusters that do not have an optimal pairing by linear assignment with a Fourier cluster are colored black. Note that PCA summary calls attention to the low quality of the S/XC3 pairing places it between XC5 and XC1.

most remaining genes fall into the Cho "Late G1" cluster (Figure 2.7A). A straightforward hypothesis is that the statistical EM algorithm simply could not justify dividing G1 vectors into early and late G1 kinetic groups as the heuristic had. The confusion array, however, makes it clear at a glance that a different data feature is driving the G1 sub-groupings. EM1 genes are upregulated only in the second cycle, while EM2 genes are upregulated in both cycles. The array also shows that the Cho Early G1 group contains a set of 10 genes that appear much more consistent with a coherent M phase group that corresponds to EM5.

Because the focus of the heuristic classification was mainly on the second oscillation, it suppressed the distinction between single cycle and two cycle G1 patterns, while "paying more attention" to fine-structure kinetic differences of the second cycle. EM MoDG, on the other hand, treated all features with equal weight, and centered the clusters without prior guidance about their relationship to cell cycle phase. The confusion array intersect cells then parsed fine kinetic differences with EM1 by separating 47 vectors that more closely resemble the early G1 cluster versus 18 that are more like late G1 cluster. Thus the intersect cell captured the two distinct ways in which the algorithms segregate G1 genes and dissected parent clusters accordingly.

In the confusion array the distribution of members from Cho "S-phase" cluster are shown to overlap almost evenly between either EM2 (41%) or EM3 (49%) from the EM MoDG result. A simple biological interpretation is that the kinetic boundary between Late G1 and S-phase is not very crisp, regardless of whichever algorithm is used to try to define them. An alternative explanation is that this is an instance where one algorithm is frankly superior to the other in defining a coherent expression group. The latter explanation is supported by the ROC curves (Figure 2.3). The EM3 ROC area is a very high value at 0.98 versus 0.82 for the Cho heuristic S phase group, indicating that the Cho S-phase group overlaps much more with data outside its group than does the corresponding EM3 cluster.

2.2.5 Dissecting individual clusters using ROC

Further ROC-based analysis of the S phase cluster from the Cho et al. (1998) classification is shown in figure 2.4. The ROC curve shows how far from the cluster mean one needs

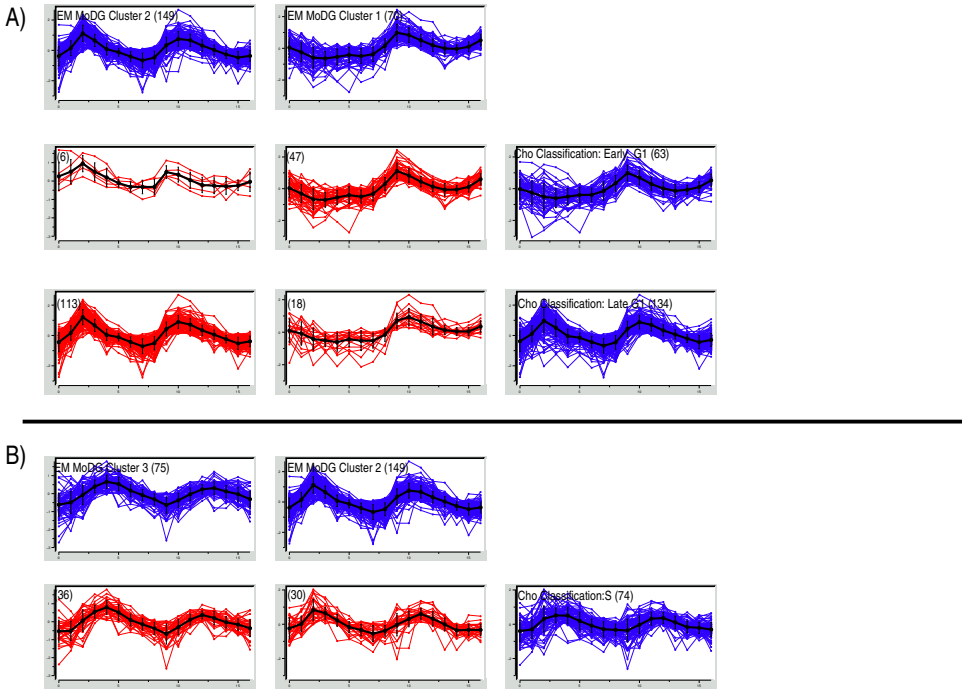


Figure 2.7: .

Selected confusion array cells from figure 2.1 highlighting cluster membership differences for genes with peak expression during the G1 and S phases of the cell cycle. The trajectory summaries display an expression profile for every gene with time along the X-axis and expression along the Y-axis. Blue trajectory summaries show parent clustering results (EM MoDG along the columns, the Cho classification along the rows). Intersection cells from the confusion array are shown in red. Mean vectors for each gene set are shown in black with error bars proportional to the standard deviation. The total number of gene expression vectors in each cell is shown in parentheses. *A)* G1 genes are subdivided differently by the two algorithms. EM MoDG separates genes upregulated only during the second phase of the cell cycle from those upregulated during both the first and second cycles. The Cho classification separates G1 based primarily on peak time in the second cycle. Figure 2.8 illustrates these observed kinetic distinctions being a result of these genes belonging to distinct regulatory modules. *B)* Detailed comparison from the confusion array of Figure 2.1 showing the S-phase cluster of the Cho classification is subdivided nearly equally among EM-2, EM-3 (optimal match by LA) clusters.

to expand a hypersphere to include a given fraction of vectors from the cluster. Inspection of the ROC curve and the corresponding histogram (Figure 2.4A) identified a natural discontinuity separating the first 66% of genes that are nearer the cluster center from the remainder. For additional data mining, we therefore set a boundary at 66% on the ROC curve and then inspected all gene vectors from the entire cell cycle dataset that fall within that boundary. $\sim$20% of gene vectors inside this dataspace threshold had been assigned to other clusters. Panels 2.4B and 2.4C allow inspection of gene trajectories that were either interior or exterior to the boundary. This tool is useful for reviewing and "pruning" lists of putatively co-expressed genes in an objective manner.

2.2.6 Comparative clustering integrated with transcription factor motifs to identify regulatory modules

CompClust is designed to integrate different kinds of data by linking each gene with other data, annotations, and results of meta-analyses. There are many ways to use other data sets to identify relationships between, for example, observed patterns of RNA co-expression and other data that help to answer the question: Are similarly expressed genes co-regulated? A group of genes that are co-expressed may also be co-regulated, but this is far from assured. Co-expressed genes can instead arrive at the same expression pattern by the action of two (or more) different regulators. Conversely, genes that are co-regulated by the same factor(s) at the transcriptional level may not display identical RNA expression patterns for a variety of reasons, including differential turnover rates. For these reasons other kinds of data are needed to help determine which co-expressed genes are, in fact, transcriptionally co-regulated and to provide evidence for the identity of factor(s) driving co-regulation. Here we show how the occurrence of evolutionarily persistent transcription factor binding sites can be mapped informatively onto gene expression clusters from a confusion array to predict the structure of transcription modules.

The observation of two distinct sets of genes, one that peaks during both the first and second cell cycles after release from arrest, and another restricted to only the second oscillation (figure 2.7), suggests that they might be regulated differently at the level

of transcription. Prior work has led to the view that MCB and SCB sequence motifs bind Mbp1/Swi6 (MBF) or Swi4/Swi6 (SBF) factor complexes to drive G1 specific transcription [Nasmyth, 1985, Breeden and Nasmyth, 1987, Koch et al., 1993]. Thus many genes are believed to be selectively and specifically expressed in G1 due to their membership in either MBF or SBF regulatory modules. The two modules are also thought to be partly distinct from each other, with some genes apparently being strongly governed by either Swi4 or Mbp1 [Horak et al., 2002, Iyer et al., 2001], and reviewed by [Breeden, 2003].

We therefore calculated a motif conservation score (MCS, see Methods) to quantify the conserved enrichment of a consensus site within 1kb of the start ATG in sequence data from the seven available yeast genomes [Cliften et al., 2003, Kellis et al., 2003]. We then asked if different intersect cells within the confusion array are differentially and significantly enriched for these known candidate motifs. The EM2/Late G1 intersect cell was highly enriched, above chance, for MCB and SCB. 79 of 113 genes (70%) were enriched for MCB compared with the expectation of 13 such genes for randomly selected samples of 113 yeast genes. 18% are enriched for SCB sites compared with an expectation of only 6 genes by chance (figure 2.8A). Also, the vast majority of genes with above threshold MCB or SCB MCS scores also have significant in vivo binding activity for either MBF or SBF as measured by Lee et al. (2002) (figure 2.8 C and D). In contrast, the EM1/Early G1 intersect cell, whose genes peak only once during the time course, showed no significant enrichment for either MCB or SCB (figure 2.8A).

Given that the genes in the EM1/Early G1 confusion array cell show low MCS scores for both MCB or SCB, what factor(s) could be responsible for the EM1/Early G1 intersect pattern? We searched for another binding motif that showed an enrichment for this group and found that the SWI5/ACE2 motif is enriched so that $\sim 30\%$ are above threshold, a value double that expected by chance (figure 2.8A). Figure 2.8B shows these genes ordered according to their Swi5 MCS scores, where the highest Swi5 MCS scores correlate with very intense expression in the second cycle. This, in turn, correlates strongly with in vivo factor binding by both Swi5 and Ace2 taken from the chromatin immunoprecipitation data of Lee et al., (2002).

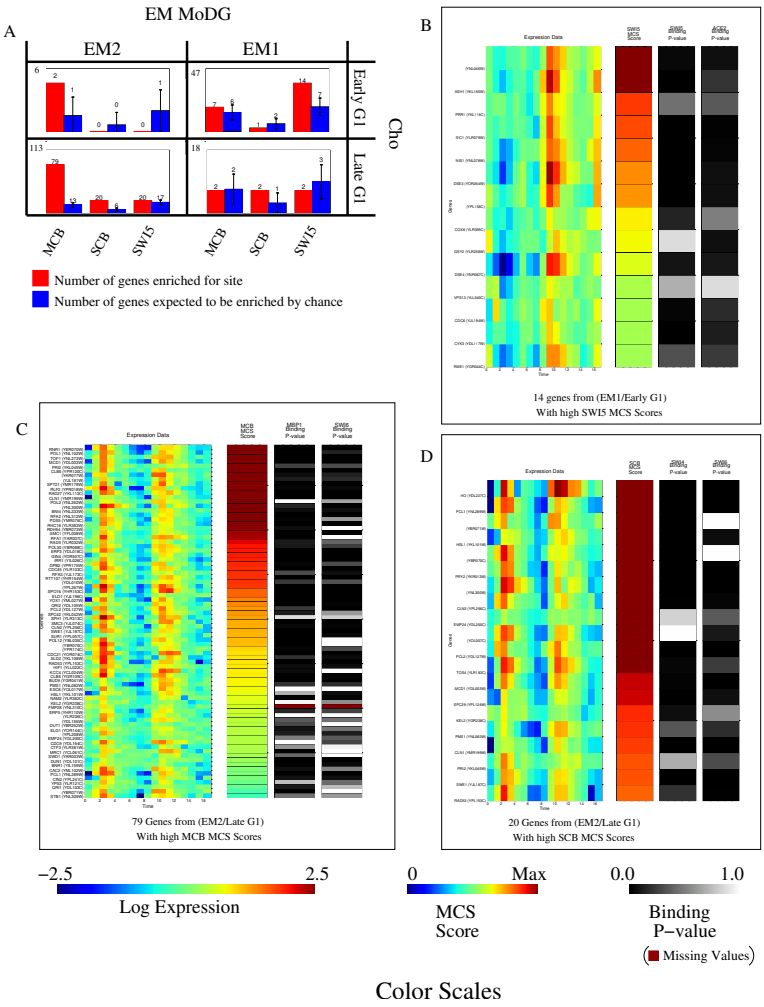


Figure 2.8: Integrating expression data, regulatory motif conservation, and protein-DNA binding information. *A)* Binding site enrichment in genes from the four confusion matrix cells of figure 2.7 that dissect genes in the G1 cell cycle phase. Shown in red are the observed number of genes with a MCS score above threshold for each motif. Shown is blue are the number of genes expected by chance, as computed by bootstrap simulations. The total number of genes each cell contains is in the upper left. *B, C and D)* Heatmap displays showing expression data on the left, followed by MCS scores for a specified motif, followed by in vivo protein:DNA binding data for transcription factors implicated in binding to the specified consensus. Color scales for each panel are at the bottom of the figure. For the MCS scores, the color map ranges for 0 to the 99th percentile to minimize the influence of extreme outliers on interpretation. *B)* Shown are 14 genes that fall within the EM1/Early G1 intersection cell and have a conserved enrichment in the presence of the SWI5 consensus as measured by MCS scores (see methods equation 2.4.9) *C)* Shown are 79 genes that fall within EM2/Late G1 intersection cell and have a high MCS score for MCB. *D)* Shown are 20 genes that fall within EM2/Late G1 intersection cell and have a high MCS score for SCB. In each heatmap genes are ordered by decreasing MCS score. Significant correlation can be seen between a high MCS score, protein:DNA binding and the expected expression pattern.

2.3 Discussion

As illustrated for yeast cell cycle data, differences among clustering algorithms, implementation parameters, and individual dataset structures make it difficult - even limiting - to simply select one clustering result and expect it to produce a fully informative data model. And without objective comparative criteria, it is difficult to tell by inspection whether one clustering is significantly "better" than another. The mathematical, computational and visualization tools that collectively comprise CompClust allow one to run diverse unsupervised and supervised algorithms, compare the results using unbiased quantitative tools, and dissect similarities and differences between clusters and clusterings. Specifically, we showed that linear assignment (LA) and normalized mutual information (NMI) metrics, receiver operator characteristic (ROC) analysis, principle component analysis (PCA) projections, and interactive confusion array analysis can effectively quantify and visualize similarities and differences. By coupling the resulting comparative analyses with a flexible visualization system within CompClust and, especially, by using confusion arrays to organize comparisons, it becomes relatively easy to identify global and local trends in expression patterns and to find out which features are fragile to algorithm choice or other variations. The tools are also useful for investigating substructure within individual gene clusters and for showing how a cluster from one analysis relates to a cluster from another analysis. CompClust (including source code) and associated tutorials are available at http://woldlab.caltech.edu/compClust/. The principle capabilities presented above can be used through a graphical user interface (GUI) that is introduced within the tutorials using examples presented above.

2.3.1 Is one data model quantitatively better than the other?

By using LA and NMI metrics, we found that four different algorithms applied to two yeast cell cycle datasets produced clusterings that differed substantially. It is noteworthy that these differences were in the classification of gene expression vectors that all exhibit cycling behavior, since the data had been prefiltered to include only genes that cycle by the authors of the original papers. This means that differences in clustering results are not

attributable to the manner with which each algorithm handles background "noise" from genes that do not cycle or are not expressed significantly.

The magnitude of difference revealed by LA and NMI quantification raises the question of whether one data model is objectively superior to the other. ROC curves and area calculations delivered some useful guidance. Clusters from the original Cho heuristic and Fourier algorithms shared more internal overlap among clusters than do clusters from the EM MoDG or XclustAgglom algorithm. In the case of the Affymetrix dataset, the EM MoDG algorithm produced an objectively superior data partition partly because all parts of expression trajectories were weighted equally. This is not surprising since the statistically based EM MoDG algorithm should locate optimal natural groupings and define the best-justified boundaries between them based on the entire data vector for each gene. In contrast, the expert-supervised heuristic was tuned to a model of cell cycle expression that emphasized peak expression values and focused selectively on the second oscillation in the timecourse. This expert-imposed emphasis produced different cluster memberships, and it is left to the biologist to determine which is most appropriate for a given use and to decide how to use information on intersection and mutual exclusion of individual genes. In other studies using entirely different datasets, we have seen that one clustering is sometimes broadly and irrefutable superior to another, and in such cases poor ROC values across an entire clustering are diagnostic of poor quality. As such, we believe ROC can be useful for quickly screening out clusterings that are frankly inferior.

More detailed comparative analysis showed that for the Affymetrix dataset [Cho et al., 1998], EM MoDG and the Cho heuristic found a basic data structure dominated by, and consistent with, the major phases of the cell cycle (figure 2.3). This presented an apparent paradox, since the overall cell cycle phase structure was highlighted similarly by both algorithms, while the assignment of specific gene vectors to individual clusters was quite different, as shown by the LA and NMI scores (figure 2.1). Further investigation of cluster relationships in the context of confusion arrays, local ROC analysis, and ROC curve structure helped to resolve the paradox and, in doing so pointed to both technical quality issues between algorithms and to biologically pertinent substructure in the data. Thus, in some instances, assignment of gene vectors differed because the boundaries between clusters, such as G1

and S, were not as well chosen by one algorithm as by the other. Inspection of intersections in the confusion arrays also identified specific expression vectors and groups of vectors that were the most “ambiguous”. Sometimes the ambiguity appeared to be a data quality issue for a given gene, and highlighting these affords a user the opportunity to trim gene lists accordingly. In other cases differences between algorithms in cluster assignment portrays correctly the notion that phases of the cell cycle are not entirely separate with respect to mRNA synthesis and decay.

2.3.2 Inference of transcriptional modules

By analyzing gene groups corresponding to confusion array intersect cells we showed that G1 gene expression that peaks during both first and second cycles (confusion array intersect cell EM2/Late G1 of figures 2.7A and 2.8A) is prominently associated with a conserved enrichment of either MCB or SCB, the known “classical” G1 cis-acting regulatory sequence motifs [Nasmyth, 1985, Breeden and Nasmyth, 1987, Koch et al., 1993]. CompClust linking capabilities were then used to visualize correlations with in vivo protein:DNA binding data for Swi4 and Mbp1 [Lee et al., 2002], the factors that are expected to bind active SCB and MCB sites. In contrast we found that genes expressed solely during the second cell cycle oscillation (EM1/Early G1 intersection cell) are not enriched for MCB or SCB sites. We then identified an enrichment for the Swi5/Ace2 binding motif for this group (figure 2.8A); about 30% of the genes in the EM1/Early G1 intersect cell. Remarkably only $\sim 4\%$ of the EM2/late G1 group (which has the MCB/SCB two-oscillation pattern) was enriched for Swi5/Ace2 binding sites. Further independent support for a Swi5/Ace2 G1 regulatory node that includes these target genes comes from two sources. A prior study had identified Swi5 as the primary regulator conferring Early G1 gene expression to the EGT2 gene [Kovacech et al., 1996]. And another study [Doolin et al., 2001] identified 15 genes regulated by either Swi5 or Ace2. Of these, 11 are present in the input cycling group of 383 genes [Cho et al., 1998]. Seven of these are in the EM1/Early G1 group and have high Swi5 MCS scores. EGT2 is notably absent because Cho et al. (1998) had not included it in their original analysis. However, its expression pattern is consistent with this group

and has a very high Swi5 MCS score. Second, as with the MCB/SCB regulatory module, independent supporting data comes from in vivo protein:DNA interaction data from global chromatin immunoprecipitation [Lee et al., 2002]. Thus, $\sim 60\%$ of this EM1/Early G1 Swi5/Ace2 group have p-values below 0.05 for Swi5 or Ace2 in the global chromatin immunoprecipitation study of Lee et. al. (2002), and still more are relatively strong binders as indicated by the p-values shown in figure 2.8B.

Our observations raise questions about the structure of the G1 regulatory network and the regulatory modules that comprise it. The top scoring members of each candidate co-expression/co-regulation module displayed strong positive signals for transcription factor motif enrichment, good p-values for putative in vivo binding by the corresponding factor (or heteromeric complexes, in the case of Mbp1-Swi6 (MBF) and Swi4-Swi6 (SBF)), and typically they were among the most robust examples of their subcluster RNA expression pattern. These are reminiscent of classical molecular genetic studies performed on a "model gene". Systematically defining these genes as members of G1 regulatory modules (MCB, SCB, or Swi5/Ace2) was relatively straightforward. However, these robust transcriptional connections to a given expression pattern capture and account for only a *part* of each expression group. We think that the high quality connections supported by all data types generate, in effect a sparse network scaffold, rather than a complete and comprehensive network model, even though the input data are from comprehensive genome-scale assays. Thus, in the major confusion array sub-clusters we studied in detail, a substantial proportion of genes display the group-defining RNA expression pattern, yet they lack convincing motif enrichment scores and/or high level ChIP/chip binding. An interesting biological explanation is that there are additional regulatory factors that can also drive these G1 patterns of expression. Since G1 regulation in yeast already uses at least three regulatory modules, it seems entirely plausible that others might exist, and new datasets are likely to uncover theses. In addition, combinatoric regulatory modules likely account for expression of some G1 genes, and especially strong output from a single binding site based on its position, orientation or in vivo affinity, might account for others. Such combinatoric and single site configurations are likely to have marginal MCS scores, but other algorithms that explicitly address these alternatives can be added to CompClust to help identify them.

Other explanations for connections of uncertain quality include assay sensitivity and the incompleteness of current datasets. For example, reliable in vivo binding data (from microarray based ChromatinIPs) in the current state of the art may be biased in favor of genes having multiple binding motif instances for a factor rather than just one or two sites. With respect to completeness more comprehensive ChIP-chip data will add results for factors not included in the currently available data (R. Young, personal communication), making possible identification of new candidate regulators associated with particular gene sets, if they exist.

2.4 Methods

2.4.1 CompClust

As shown above the maturation of additional large-scale data types (global chromatin immunoprecipitation assays, more complete and highly articulated protein:protein interaction maps, GO ontology categories, evolutionarily conserved sequence features, other covariates) shifts the emphasis from analyzing and mining expression data alone to integrating disparate data types. A key feature of any system designed for integration is the ability to provide a many-to-many mapping of labels to data features and data features to other data features in a global way. CompClust provides these capabilities by maintaining and tracking linkages of multiple arbitrary annotations and co-variates with data features through almost any data transformation, merger, selection, or aggregation. In addition, many supervised and unsupervised machine learning algorithms are easily accessible within CompClust.

CompClust is primarily accessible through an application programming interface (API) and, with the use of Python's exposed interpretor, this provides a very rich command line interface (CLI). The major capabilities illustrated in this paper are accessible through a set of simple graphical user interfaces (GUIs) to offer a convenient starting point without learning Python commands. These GUIs will permit users to perform the major classes of analyses shown, though we note that these comprise only a fraction of CompClust capabil-

ities, as the flexibility and diversity of analysis paths is too great to reduce all of them to GUI form. This limitation can be overcome by using the Python command line environment. Python commands can be learned at the level needed in a relatively short time (a few weeks of part time effort) by users who do not have prior programming experience. The benefit is access to remarkable flexibility in interrogating datasets. This is a much closer match to the diversity of questions and comparisons that biologists usually want to make and to the spectrum of specific needs that arise in different studies.

The choice to implement CompClust in Python over other languages was made for several reasons which, considered in aggregate, argue it is the best available language to support the capabilities and analysis goals of CompClust: 1) Using Python's exposed interpretor, our API becomes immediately useful for analysis without the construction of a complex GUI. The exposed interpretor also speeds the development time. 2) Python's syntax is fairly straightforward and easy to learn for even non-programmers. 3) It is freely available and distributable under an open-source license. 4) Python has an extensive and standard library and in addition 3rd party extensions, including the Numeric package which provides powerful numeric analysis routines. 5) Python is also platform neutral and runs on the majority of systems including unix/linux, Microsoft Windows and the Mac OS.

2.4.2 Pairwise comparison of clusterings (partitions) using confusion arrays and matrices

Confusion arrays and matrices were used to make pairwise comparisons between different clusterings (mathematical partitions). A set of metrics were then applied to the confusion matrix to measure the nature and degree of similarity between two dataset partitions. Briefly, a confusion matrix M is the matrix of the cardinalities of all pairwise intersections between two partitions, A and B (eq. 2.1), where a partition of a dataset D is defined as a set of disjoint subsets of D whose union contains all elements of D. We define a confusion array simply as an array of all pairwise intersections between two partitions A and B of a dataset D. The cardinalities of these intersection sets form the confusion matrix C, whose elements are given by equation 2.1. C is a confusion matrix where:

$$C_{i,j} = |A_i \bigcap Bj| \tag{2.1}$$

and,

A_i: The data members of class i in A

B_j: The data members of class j in B

2.4.3 Linear Assignment

The Linear Assignment (LA) value for a confusion matrix is calculated between two partitions (clusterings) A and B by generating an optimal pairing so that there is, at most, a one-to-one pairing between every class in partitions A and B. This pairing is calculated by optimizing the objective function in equation 2.2, using the constraints given in equation 2.3 thus defining a linear assignment problem. Next, the maximum-cardinality bipartite matching of maximum weights algorithm [Gabow, 1973] was implemented for the optimization. After finding the optimal pairing, the LA score is simply the proportion of vectors (e.g. gene expression trajectories or conditions) included in the optimally paired clusters (eq 2.4). It is important to note that LA, unlike NMI, is a symmetric score so that $LA(A, B) = LA(B, A)$. In addition to quantifying the degree of similarity or difference between two partitions, the adjacency matrix M (eq 2.3) also provides a way to identify pairs of clusters that are globally most similar to each other between two partitions of the data. As illustrated for clusterings of yeast cell cycle regulated genes, this is especially useful for interactive examination of two clusterings.

$$E = -\sum_{ab} M_{ab} C_{ab} \tag{2.2}$$

where,

$$M_{ab} \in \{0, 1\} \ \wedge \ \sum_{a} M_{ab} \leq 1 \ \wedge \ \sum_{b} M_{ab} \leq 1 \tag{2.3}$$

Now,

$$LA = \frac{\sum_{a,b} M_{ab} C_{ab}}{\sum_{a,b} C_{ab}} \tag{2.4}$$

where,

M: adjacency matrix describing the pairing between A and B

C: the confusion matrix (eq. 1)

2.4.4 Normalized Mutual Information

The NMI (normalized mutual information) index [Forbes, 1995] quantifies how much information is lost, on average, when one clustering is regenerated from a second classification (Equation 2.5). A noteworthy difference from LA is that NMI is asymmetric.

$$NMI(A,B) = \frac{I(A,B)}{H(A)} = \frac{H(A) - H(B) - H(A,B)}{H(A)} = 1 - \frac{H(A,B) - H(B)}{H(A)} \tag{2.5}$$

where $I(A,B)$ is the shared information between the two partitions and it is normalized by the entropy of partition A, $H(A)$ defined as:

$$H(A) = \sum_{i \in partitions} p_i \cdot \log \cdot p_i \tag{2.6}$$

and,

$$p_i = \frac{\sum_j C_{i,j}}{n} \tag{2.7}$$

and the joint-information is:

$$H(A,B) = H(C) = \sum_j \sum_i \frac{C_{i,j}}{n} \log(\frac{C_{i,j}}{n}) \tag{2.8}$$

and

$$n = \sum_{i,j} C_{i,j} \tag{2.9}$$

2.4.5 Combined Use of Normalized Mutual Information (NMI) and Linear Assignment (LA)

NMI(A,B)	NMI(B,A)	LA	Implies
Low	Low	Low	Poor Similarity
Low	High	Low	B refines A
High	Low	Low	A refines B
High	High	High	Good Similarity

Table 2.1: Interpretations of commonly observed combinations of LA and NMI scores.

Given two clustering results A and B for which both NMI(A,B), NMI(B,A) and LA(A,B) values are high (nearing the maximum value of 1.0) the two clusterings are very similar, and when all three are significantly lower, they are very different. But when NMI(A,B) is high, NMI(B,A) is low and LA is low, then it is likely that A is a refinement of B. In this case, many clusters in B have been broken into two or more clusters in A (possible combinations summarized in here) (2.1). The magnitude of dissimilarity that is important is defined by the user and may vary considerably with the dataset, although values below 0.7 for both LA and NMI are usually viewed as quite different. Additional interpretation of differences measured by LA and NMI depends on more detailed analysis of the dissimilarities and their distribution over the dataset, as outlined above.

2.4.6 EM MoDG clustering of yeast cell cycle data

Expectation Maximization of a Mixture model of Gaussians was implemented with a diagonal covariance matrix model because the number of samples in the [Cho et al., 1998] cell cycle dataset was too small to fit a statistically valid full covariance matrix to each cluster [Dempster et al., 1977]. In order to ensure a near optimal initialization, each EM MoDG results was a result of selection of the best of 30 runs each initialized by placing the initial cluster centroids on K randomly selected datapoints. The run with best fit to the data

(i.e. had the lowest log-likelihood score) was used for the final clustering. Multiple best-of-30 runs were performed to verify that the quantitative measures and gene lists results reported here did not vary significantly. The EM MoDG code used here was developed by the NASA/JPL Machine Learning Systems Group.

2.4.7 XclustAgglom

We agglomerate the hierarchical tree returned by Xclust [Sherlock, 2000] based on a maximal cluster size threshold. Starting from the root, any subtree with in the tree with less than the maximal cluster size threshold is agglomerated into a cluster. In order to work with the familiar parameter K (number of clusters) we iteratively find the size threshold that will return as close to K clusters as possible. In practice this simple heuristic works best when K is over specified by roughly 2-4 times the expected number of clusters because it will generate several very small (often singleton) clusters that are outliers to core major clusters in the data (figure 2.5)

2.4.8 Data Preprocessing

Each microarray dataset was obtained from the cited authors. For the Cho et. al. (1998) data we removed any gene that did not show a sustained absolute expression level of at least 8 for 30 consecutive minutes. We then for each gene vector divided each timepoint measurement by the median expression value for the gene. For the Spellman et al. (1998) we linearly interpolated missing values using the available adjacent time points. For both datasets we log_2 transformed the resulting gene expression matrices. The datasets where then annotated with the original clustering results as published.

2.4.9 Motif Conserved Enrichment Score (MCS)

For each motif we translated the IUPAC consensus (Swi5/Ace2: KGCTGR, MCB: ACGCGT, SCB: CACGAAA) into a position weight matrix (PWM) where the probabilities or frequencies in the PWM is determined by the degeneracy of the IUPAC symbol. We calculate a log-odds ratio as described in equation 2.4.9 for the PWM occurring at every position in

the 1KB upstream of each ORF for each species available. We then sum the log-odds ratio over all possible positions where the log-odds ratio is greater than 7. The summed log-odds ratios for each species is then averaged together to generate an ORF specific motif enrichment score. In equation 2.4.9 below N is the total number of species compared, W is the length of the motif, p_{ni} is the probability from the PWM of position i being the nucleotide n, $n \in A, C, T, G$ and bg represents the probability of the window being generated from a background sequence model based on a second order hidden Markov model.

$$MCS = \frac{1}{N} \sum_{\forall windows} \frac{\prod_{i=0}^{W} p_{ni}}{bg}$$

Chapter 3

Influences of Measurement Noise, Data Preprocessing, and Algorithm Choice on Microarray Clustering Results

3.1 Introduction

Large scale gene expression data provides a global perspective on the diversity of expression patterns that occur through biological processes both with and without perturbations. Many clustering techniques have been applied to these data and have been shown to uncover natural groupings of co-expressed genes [Cho et al., 1998, Eisen et al., 1998, Golub et al., 1999, Tamayo et al., 1999, Ross et al., 2000, Ihmels et al., 2002]. Further analysis of groups of co-expression genes have led to the isolation of functional regulatory modules ([M. A. Beer, 2004, Tavazoie et al., 1999] and section 2) and correlations with functional annotations such as those found in the gene ontology [GOConsortium, 2001] resource have been utilized to gain a basic understanding of the types of processes that are being modulated [Doniger et al., 2003]. Clustering results also give insights into the general trends of expression behaviors that are utilized to respond to varying conditions.

Although the underlying goal of most clustering algorithms is to uncover clusters of genes with behaviors more similar to each other than to non-cluster members, the results can be strikingly different (Table 3.1). The results of a given algorithm can be confounded by selection of varying algorithm parameters such as initialization, seeding, distance metric or the number of clusters to find. It is important to understand the influences of these

parameters in the resilience of clustering results for biological interpretations.

	EM MoDG	KMeans	XClust
EM MoDG	1.0	0.67	0.77
KMeans	-	1.0	0.72
XClust	-	-	1.0

a) Yeast Cell Cycle Data

	EM MoDG	KMeans	XClust
EM MoDG	1.0	.45	.46
KMeans	-	1.0	.42
XClust	-	-	1.0

b) Tumor Data

Figure 3.1: Cluster similarity as measured by linear assignment (LA) scores (section 2.4.3) comparing clustering results of three different clustering algorithms when used with comparable parameters on exactly the same dataset. a) Clustering comparisons from a microarray dataset measuring the expression levels of all genes in yeast through two cell cycles [Cho et al., 1998]. b) Comparisons from a microarray dataset surveying the expression profiles of several different human breast tumors [Perou et al., 2000]

.

Microarray data is also often confused by measurement and biological noise [Tu et al., 2002, Novak et al., 2002, Yang et al., 2002]. When drawing on clustering results as a foundation for building new hypotheses and understanding a biological process, the stability of both the cluster memberships and also the cluster behaviors are important. Another source of variance in clustering comes from preprocessing and the decisions of which genes to cluster. An understanding of these influences on clustering results can seriously alter downstream interpretations of the results. Particularly sensitive to the reliability and the inclusion of genes that may or may not be warranted in clustering results are those methods that search for statistical biases within cluster members, such as those looking for enrichment of binding sites or GO categories.

Discerning an “optimal” or “best” clustering algorithm to use is dependent on far too many factors ranging from data quality, dataset size (ie. number of genes and number of conditions), and most importantly the underlying questions that need to be addressed. Using the CompClust framework we can gain an understanding of how the reproducibility of clustering results are effected by features of microarray data, such as noise, and the typical ways in which we process them.

Here we present a framework to understand the sensitivities and influences of measure-

ment noise, data preprocessing, and algorithm choice. We compare several clusterings of a time-resolved microarray experiment measuring the expression levels of every gene in yeast through two cell cycles. We also demonstrate the applicability of these methods on a microarray dataset measuring expression profiles of different breast tumor samples. We compare three different widely used decisive clustering algorithms: an expectation maximization (EM) based algorithm fitting data to a mixture of Gaussian [Dempster et al., 1977, Ghosh and Chinnaiyan, 2002] (and chapter 2), KMeans [Tavazoie et al., 1999], and an agglomerated phylogenetic clustering algorithm (Eisen et al. 1998, and Chapter 2) . Although there are many other clustering techniques available, these provide a baseline of sensitivities on a set of prototypical algorithms. We do not include self organizing maps (SOMs) [Tamayo et al., 1999] even though it is an often used technique for microarray analysis. Although an end result of the SOMs is a decisive clustering, the projection of the data onto a lower dimensional space is its primary advantage. Further, when using SOMs in a map agnostic fashion it is algorithmically quite similar to KMeans.

3.2 Methods

3.2.1 Dataset Preprocessing:

Each dataset was obtained from the original authors and loaded into CompClust. The yeast cell cycle data [Cho et al., 1998] was passed through a filter to ensure that all genes showed significant expression (an absolute intensity of 8) for at least 30 consecutive minutes. The "cycling" genes as identified by the original authors were primarily used for our clustering analysis. For the breast tumor dataset [Perou et al., 2000] we loaded the same normalized intensity data as the authors and used their selection of 1753 genes which exhibited a minimal of 4-fold change in at least one tumor sample from the mean expression level measured across all tumors. For both datasets noise was always added to the intensity data directly before we computed the log ratios that are used for visualization and clustering.

3.2.2 Comparing Clustering Results:

Every clustering result was compared pairwise using the CompClust framework as described in chapter 2. Specifically we constructed confusion matrices and quantified the similarity between any two clustering results using either linear assignment (LA) or normalized mutual information (NMI) (section 2.4).

3.2.3 Clustering Methods:

Each of the clustering algorithms were used as part of the CompClust framework. The EM MoDG and XClust with agglomeration are described in more detail in section 2.4. Because KMeans and EM MoDG are both initialized by randomly seeding the algorithms with starting cluster means we selected an optimal seeding. The optimal seed was selected in both cases for every clustering result reported. In order to select the optimal seed, each algorithm was run on the dataset 30 times with each repeated run starting from a different random seeding. The seeding that resulting in the best-fit model to the data was then reported as the final clustering result.

3.2.4 Perturbing Datasets With Synthetic Noise and Mix-in Data Vectors:

We perturbed our expression datasets to asses the stability of clustering results as a function of noise. For our synthetic noise experiments we add simple Gaussian noise to each intensity measurement before calculating ratio values. Since we are focusing on genes that have significant expression, this simple model approximates more complex functions that indicate low intensity values are subject to more noise [Tu et al., 2002]. The variance of the noise added to each dataset was proportional to the mean intensity value for the dataset. We added Gaussian noise that was pulled from a distribution of mean zero and a variance that was: 0.5%, 1.0%, 2.0%, 4.0%, 8.0% 16.0%, 32.0%, and 64.0% of the mean intensity value for the dataset. Estimations of noise values that occur during typical microarray experiments are dependent on several factors but for technical replicates the variance expected

is between 4.0% - 8.0% of the measured intensity value and biological repeats between 16%-32.0% [Tu et al., 2002, Novak et al., 2002, Yang et al., 2002] [1]

3.3 Results

Although largely reproducible microarray measurements are subject to a significant amount of noise [Tu et al., 2002, Yang et al., 2002, Novak et al., 2002]. Further, isolation of genes that are "of interest" which form the input to clustering algorithms often rely on methodologies that do not have easily definable boundaries, such as p-value cut-offs. We demonstrate here that these uncertainties have a significant impact on clustering results, especially with regard to the determination of cluster membership boundaries.

3.3.1 The Influence of Measurement Noise of Clustering Results

We compared the influence of noise on clustering results using two different datasets, one a time-course experiment measuring yeast gene expression through two cell cycles collected using Affymetrix gene chips [Cho et al., 1998], the other a comparison of different human breast tumor surgical samples collected using deposition microarrays [Perou et al., 2000]. We compared the clustering results before and after the addition of increasing amounts of noise (see methods) using three different clustering algorithms that are commonly used in microarray data analysis; Expectation Maximization fitting a mixture of diagonal Gaussians (MoDG) (chapter 2), K-Means [Tavazoie et al., 1999], and an agglomerated phylogenetic clustering algorithm (XClust) (chapter 2).

3.3.1.1 Global Influences of Noise on Cluster Membership:

The overall performance of each algorithm in the presence of increasing amounts of noise is summarized in figure 3.2 for both the yeast cell cycle time course data and the tumor sample dataset. Comparisons show that EM MoDG is slightly more tolerable to noise than

[1] Although these studies don't explicitly provide these data, our analysis of both published and unpublished datasets from a variety of platforms and experimental systems are largely in agreement with these estimates based on cited published reports.

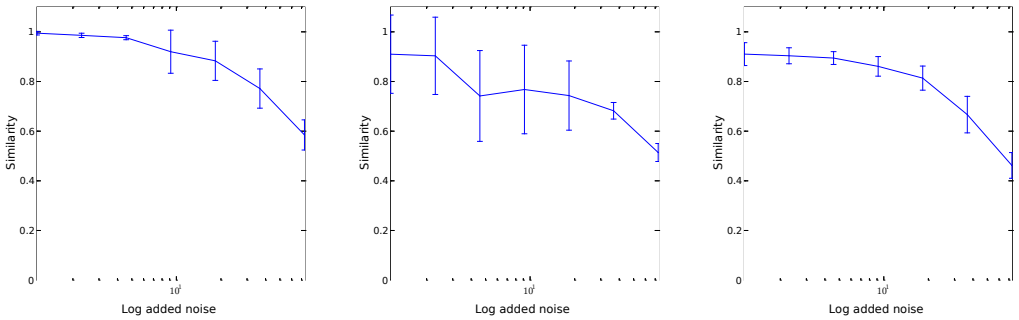
are KMeans or XClust, especially at modest noise levels. XClust and KMeans both show very large standard deviation of similarity scores even at a noise level that introduces a perturbation of only 0.5%, much less than the typical observed noise which are greater that 16%. However, all algorithms show dramatic sensitivity when noise level climb above 16%.

3.3.1.2 Differential Influence of noise on Clustering Results:

Figure 3.2 demonstrates that clustering results show substantial variations in the presence of noise. Naturally an understanding of how these differences are distributed through a dataset is critical. Are some genes consistently clustered the same while others are consistently clustered differently in each run of the algorithms? Figure 3.3 illustrates that the fluctuations in cluster assignments are not evenly distributed across the datasets and that some genes are almost always clustered similarly, but other genes are more often differentially clustered. Inspection of the results from the yeast cell cycle microarray data show that at low to moderate noise levels EM MoDG performs especially well and 90% of genes are classified the same in every run of the algorithm. As the noise levels increases EM MoDG clusters 75% of the genes the same way in roughly 75% of the runs. Comparatively, just over 50% of genes are classified the same way when using KMeans or XClust even when the datasets are permuted with very modest levels of noise (0.5%). XClust performance seems to be intermediate between EM MoDG and KMeans at moderate to high levels of noise where 75% of genes are classified the same in roughly 50% of runs. These observations are similar when clustering the breast tumor dataset. Fundamentally these results strongly indicate that cluster membership for a large proportion of data vectors is not a robust property.

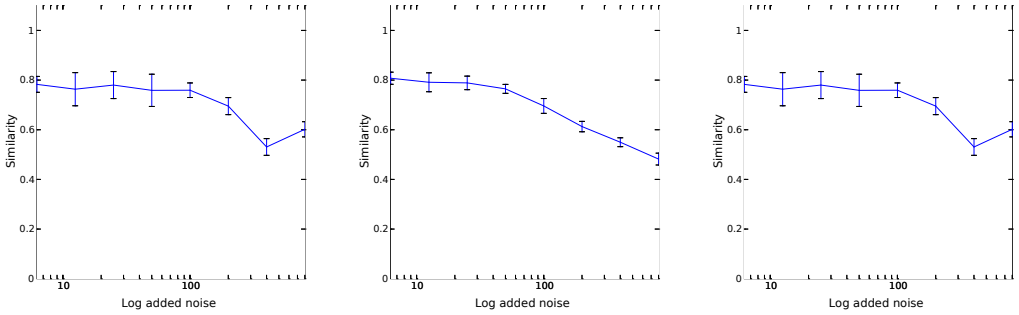
Figure 3.3 illustrates that cluster membership is unstable for a large proportion of genes but does not address how these genes are distributed within the dataset. We find that some clusters have a higher proportion of similarly clustered genes and less differentially clustered genes then other clusters. In particular the G1 cluster found in the yeast cell cycle expression dataset [Cho et al., 1998] contains genes that are clustered identically in almost every clustering run even at moderate noise levels when the other clusters are starting to

Cell Cycle Microarray Data [Cho et al., 1998]



(a) EM MoDG (b) KMeans (c) XClust

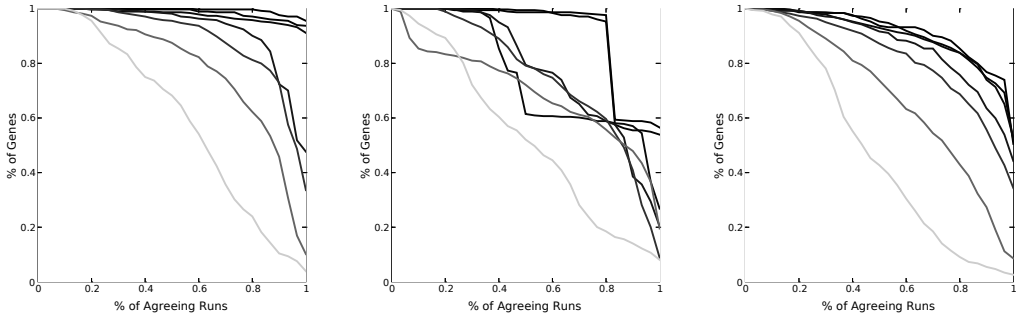
Breast Tumor Microarray Data [Perou et al., 2000]



(d) EM MoDG (e) KMeans (f) XClust

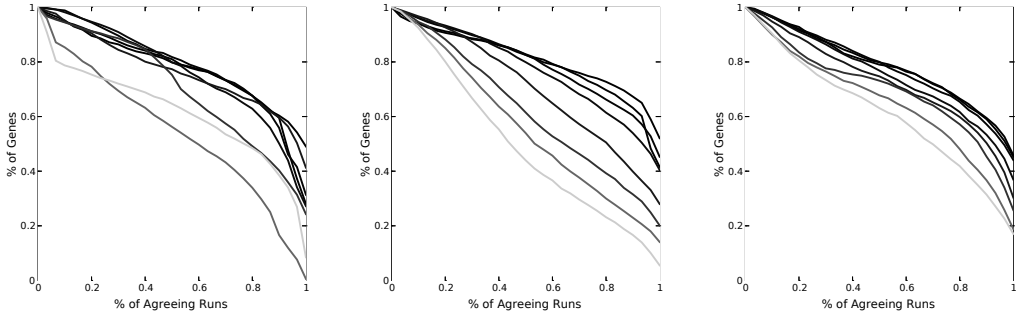
Figure 3.2: Performance vs Noise plots for each algorithm. In each plot the reference dataset was permuted by adding noise 30 times for each noise level. Each noise level adds Gaussian noise centered with mean zero and variance proportional to 0.5%, 1.0%, 2.0%, 4.0%, 8.0% 16.0%, 32.0%, and 64.0% of the mean intensity value for the dataset. The clustering result for each run was then compared with the clustering result before any noise was added to the dataset. a-c) EM MoDG, KMeans and XClust run on the "cycling" genes from yeast cell cycling microarray dataset [Cho et al., 1998]. d-e) The same algorithms run on the breast tumor microarray dataset [Perou et al., 2000]. All algorithm parameters were unaltered between the noise added and the original datasets. Similarity was quantified using linear assignment (LA) (chapter 2). Drawn is the mean similarity line for each algorithm, and the error bars show the standard deviation in the LA scores

Cell Cycle Microarray Data [Cho et al., 1998]



(a) EM MoDG (b) KMeans (c) XClust

Breast Tumor Microarray Data [Perou et al., 2000]



(d) EM MoDG (e) KMeans (f) XClust

Figure 3.3: Cluster Consistency Curves Across Increasing Noise. Each line draws the proportion of genes being clustered identically (y-axis) vs the proportion of clustering runs that classify them identically (x-axis) as computed by adjacency during the linear assignment computation (section 2.4.3). If every run of the algorithm returned the same clustering result a horizontal line at y=1.0 would be drawn. As points within the line pull towards the origin of the graph the clustering results are becoming less consistent with each other. Each line on these plots represents the consistency curve for the designated algorithm on the designated dataset at a specific noise level (see methods 3.2.4). The lighter the line the higher the noise levels. a-c) show consistency plots for EM MoDG, KMeans and XClust run on the yeast cell cycle dataset [Cho et al., 1998] respectively. d-e) show consistency plots for the algorithms on the breast tumor dataset [Perou et al., 2000]

contain many more genes that differentially clustered (figure 3.4). Similarly we find clusters in the breast tumor dataset [Perou et al., 2000] that are also more stable in the presence of noise (data not shown).

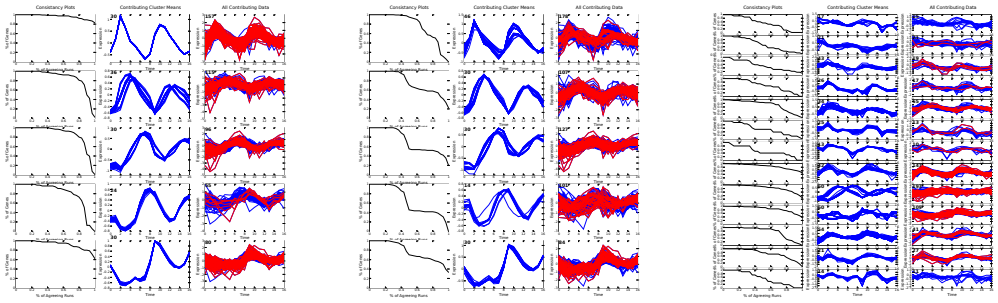
In addition to the cluster memberships of some clusters showing more fluctuations than other clusters the properties of the cluster means are also less stable. Cluster means provide a reduction of a complex microarray dataset into a small set of core prototype expression behaviors. In general these core behaviors are fairly stable and are recognizably similar to the means found in the original unperturbed dataset at modest noise levels. The PCA projections in figure 3.5 show again that the G1 (positioned at three-o'clock in the PCA projection) and M (positioned at six-o'clock in the PCA projection) clusters in the yeast cell cycle microarray dataset, in addition to having a more stable cluster membership, also exhibit high stability in the dataspace.

3.4 The Influence of Gene Filtering on Clustering Results

A major step in preprocessing microarray data before clustering revolves around selection of the interesting genes to cluster. Many techniques have been utilized for this and selection of the proper technique ultimately depends on the underlying biological questions that are being asked and the nature of the data. In the case of the cell cycle data, clustering and further analysis was performed on genes that exhibited cycling behaviors. With the breast tumor data, genes that show differential expression across the tumor samples were selected for clustering. Roughly 400 genes, or 6% of the data, was selected in the case of the yeast cell cycle, and ~ 1800 genes, 20% of the breast tumor data. As in most cases, these two analysis relied on hand tuning the input set of genes for further analysis or creating an arbitrary threshold on significance based on a particular measure (ie. p-value, fold-change). We sought to understand how diluting the selected genes with unselected, putatively "uninteresting" genes, would effect the clustering results of each of these algorithms.

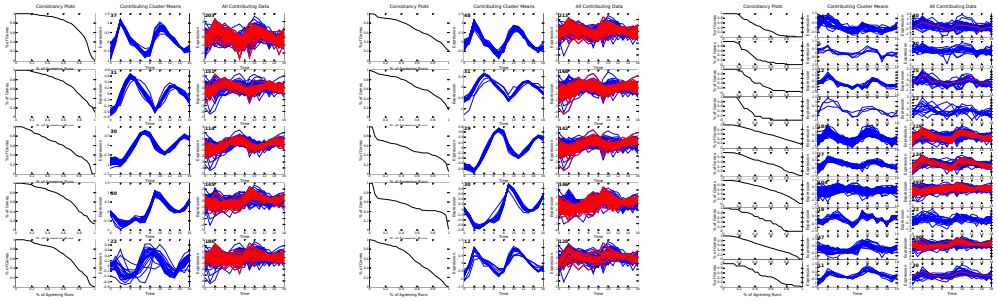
Figure 3.6 shows the influence of dilution on clustering results when we dilute selected "interesting" genes with additional genes from the remainder of the data. These additional genes we refer to as "background" genes. Both the cell cycle and the breast tumor

Perturbed with 4.0% noise



(a) EM MoDG (b) KMeans (c) XClust

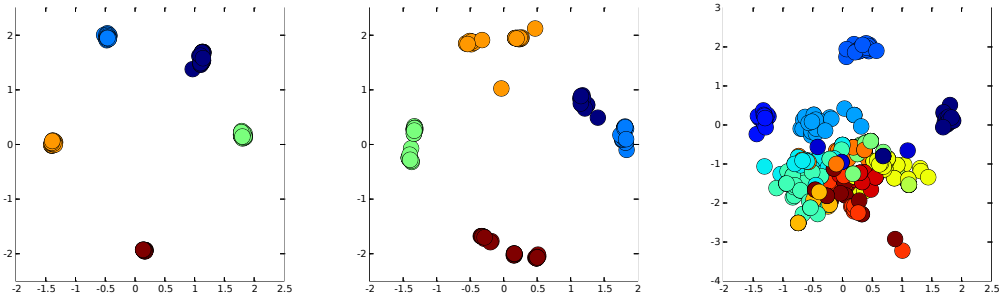
Perturbed with 16.% noise



(d) EM MoDG (e) KMeans (f) XClust

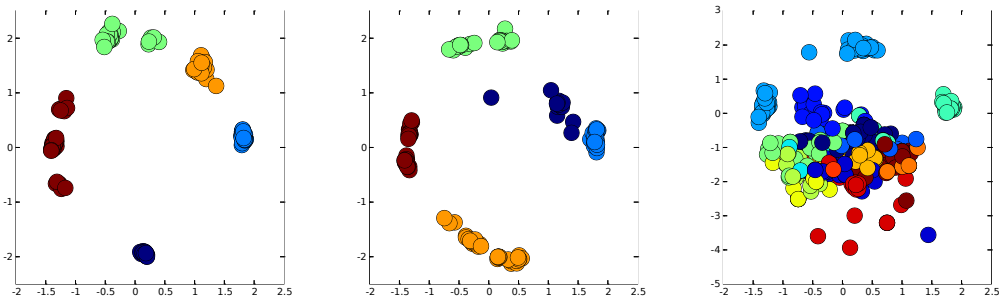
Figure 3.4: Cluster by cluster consistency analysis for clustering runs on the yeast microarray dataset [Cho et al., 1998] at a low noise level of 4.0% and a higher level of noise 16.0%. For each noise level each algorithm (EM MoDG (a,d), KMeans (b,e), and XClust (c,f)) was run 30 times after the dataset was perturbed by adding noise at the indicated level 3.2.4. The means from each clustering result were clustered into super-clusters using the algorithm that generated the clusters. Cluster consistency curves for all the genes that are contained within any cluster within the super-cluster are drawn on the left most column of each subfigure. The cluster means that compose each super-cluster is shown in the center column of each subfigure. The right most column of each subfigure shows all gene vectors contained within any of the clusters from the super-cluster. The gene vectors highlighted in red are clustered the same in at least 90% of the 30 clustering runs.

Perturbed with 2.0% noise



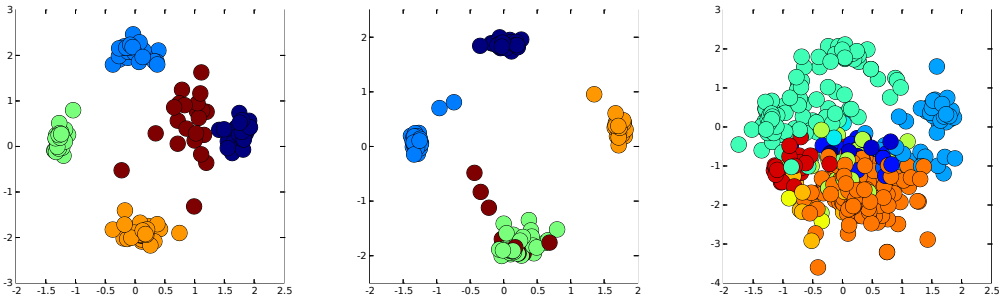
(a) EM MoDG (b) KMeans (c) XClust

Perturbed with 4.0% noise



(d) EM MoDG (e) KMeans (f) XClust

Perturbed with 16.% noise



(g) EM MoDG (h) KMeans (i) XClust

Figure 3.5: PCA cluster consistency for clustering runs on the yeast microarray dataset [Cho et al., 1998] at three noise levels, 0.2%, 4.0% and 16.0%. Clusterings were performed as described in figure 3.4. Each cluster mean was then projected into the top two dimensions of the PCA space defined by original data. Roughly as found in 2.3 the PCA space maps expression trajectories such that progression through the cell cycle is mapped into a counter-clockwise movement in the PCA space starting with G1 expression at three-o'clock and ending with M-phase at six-o'clock. Each mean is colored accordingly to its membership in super-clusters (as generated as in figure 3.4)

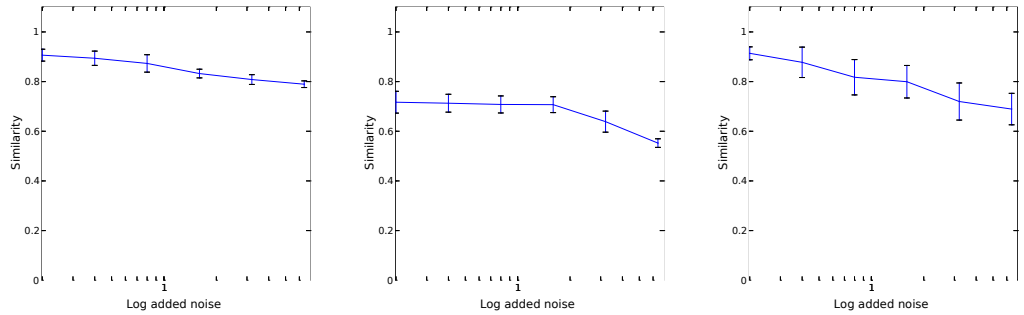
dataset were augmented with increasing numbers of "background" gene vectors from the unselected set of genes. We created augmented datasets from the original datasets that contain 20%, 40%, 80%, 160%, 320% and 640% "background" genes as compared with the original number of selected "interesting" genes. As with the permuted datasets in section 3.3.1 these augmented datasets were clustered and compared with clustering results from the original unaugmented dataset. Since each of the algorithms requires the specification of the number of clusters to identify (K) and the larger augmented datasets may require a higher or lower K value, we clustered each augmented dataset with varying K values allowing it to fluctuate +/- 2 from the K specified in the original clustering. We then chose the clustering result from augmented dataset that was most similar to the original clustering for comparisons.

Although the results from all of the algorithms are effected by dilution of the selected genes with "background" gene vectors, KMeans is most severely effected. Even at modest levels of noise nearly 30% of the genes are differentially classified. EM MoDG and XClust are similarly effected at low and modest levels of dilution, but EM MoDG shows more resilience in its classification at severe dilution. XClust also shows a much larger variability in its clustering results than does EM MoDG or KMeans as indicated by the error bars in figure 3.6.

As with the differential results observed between clustering results before and after addition of noise (figure 3.3), dilution results in some genes having more highly variable cluster membership while other genes are quite stable (figure 3.7). The cluster consistency curves in figure 3.7 again show KMeans performance being the most severely effected by diluting "interesting" genes with "background" genes. In the cell cycle data only $\sim 65\%$ of the genes retain there cluster memberships across just $\sim 75\%$ of the clustering runs. EM MoDG and XClust both showed better performance although XClust tends, to in few of the clustering runs, misclassify large proportions of the genes in the dataset. This behavior is exaggerated by increasing the dilution level.

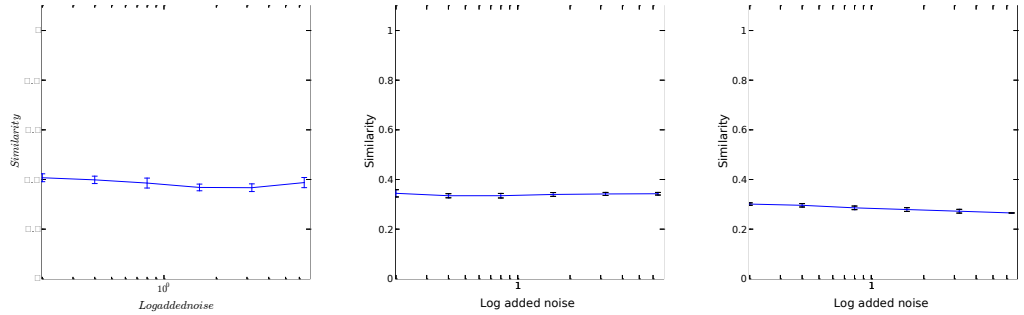
We attempt to isolate the genes that are systematically differentially clustered after augmenting the dataset with "background" vectors in figure 3.8. As in figure 3.4 cluster consistency is visualized for each of the prototype behaviors in the yeast microarray dataset.

Cell Cycle Microarray Data [Cho et al., 1998]



(a) EM MoDG (b) KMeans (c) XClust

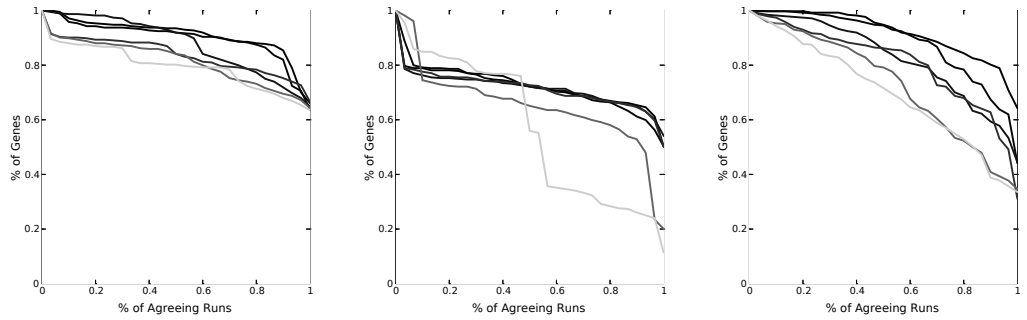
Breast Tumor Microarray Data [Perou et al., 2000]



(d) EM MoDG (e) KMeans (f) XClust

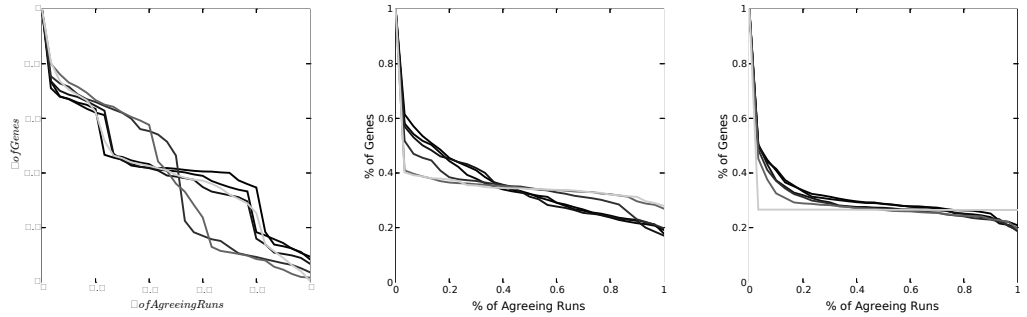
Figure 3.6: Performance vs "background" gene dilution for each algorithm. In each plot the reference dataset was permuted by adding "background" gene vectors. For each dataset the reference dataset was based on the selection of interesting genes by the original authors. In the yeast cell cycle data, these were genes that were identified to be cycling, for the breast tumor dataset these were genes that showed large amounts of differential expression 3.2.4. We created mixed datasets containing all the selected interesting genes and increasing numbers of gene vectors from the remainder of the dataset. Shown are results from clustering datasets containing 20%, 40%, 80%, 160%, 320% and 640% "background" vectors as compared to with the dataset. The clustering result for each run was then compared with the clustering result before any background vectors were added to the dataset, and only the forementioned selected interesting genes were used in scoring. a-c) EM MoDG, KMeans and XClust run on the yeast cell cycling microarray dataset [Cho et al., 1998]. d-e) The same algorithms run on the breast tumor microarray dataset [Perou et al., 2000]. All algorithm parameters were unaltered between the mixed dataset added and the original datasets. Similarity was quantified using linear assignment (LA) (Chapter 2). Drawn is the mean similarity line for each algorithm, and the error bars show the standard deviation in the LA scores across 30 separate mix-in clustering experiments

Cell Cycle Microarray Data [Cho et al., 1998]



(a) EM MoDG (b) KMeans (c) XClust

Breast Tumor Microarray Data [Perou et al., 2000]



(d) EM MoDG (e) KMeans (f) XClust

Figure 3.7: Cluster Consistency Curves Across Increasing Amounts of Dilution. As in figure 3.3 plot shows the degree of consistency between clusterings of an permuted dataset and in this case a dataset with "background" gene vectors augmented to it. Each line on these plots represents the consistency curve for the designated algorithm on the designated dataset with specific degree of dilution (see methods 3.2.4). The lighter the line the higher the dilution. a-c) show consistency plots for EM MoDG, KMeans and XClust run on the yeast cell cycle dataset [Cho et al., 1998] respectively. d-e) show consistency plots for the algorithms on the breast tumor dataset [Perou et al., 2000]

Interestingly as opposed to the consistency results after addition of Gaussian noise, dilution effects the clusters more evenly. The G1 expression clusters tend to be slightly more stable, but not appreciably so. The PCA projection of cluster means in figure 3.9 illustrates that the "background" effect these data because they exist in the the middle of the dataspace thereby imposing a balanced effect on cluster membership.

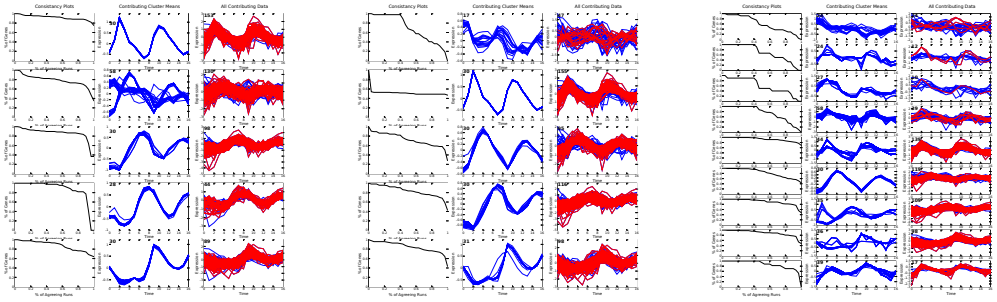
3.5 Discussion

Clustering techniques provide biologists with a statistically based method for reduction of the inherent complexity in expression datasets often needed to relate existing knowledge to the data and more importantly to extract new relationships from the data. Although clustering techniques are very powerful techniques caution needs to be utilized when leveraging their results. We demonstrate that the inherent noise in microarray datasets, even at the level of technical noise and certainly at the level of biological noise can have significant influences on clustering results (figure 3.2). Further we show that variations in preprocessing in the form of data selection also effects the clustering results (figure 3.6).

By isolating the stable features within clustering results and highlighting those features that are less stable to perturbations we gain a better perspective on global features of the data that are real and interesting as opposed to potentially artifact. For instance in the yeast cell cycle microarray experiment the core phases of the cell cycle are easily discernible by every algorithm we ran. Remarkably even when the majority genes are being differentially clustered in the majority of clustering runs this fundamental structure remains clear, especially in the case of EM MoDG. We can also use this analysis to identify genes that are particularly stable in cluster analysis and clustered identically even through multiple clusterings as highlighted in figure 3.4.

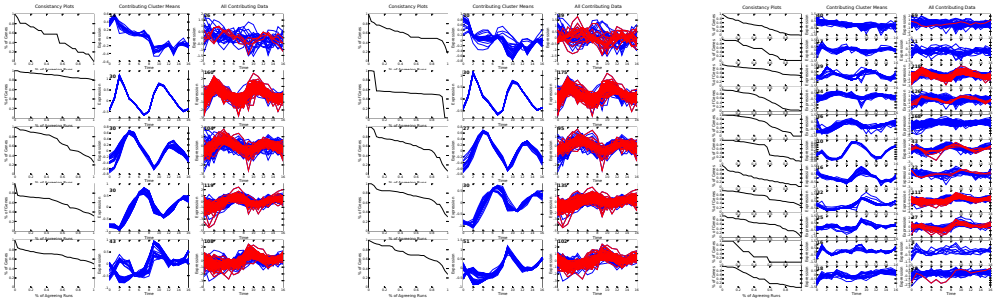
Efficient mining of large-scale expression data provided by microarrays requires careful selection of statistical tools to aid biologists in addressing hypotheses both conceived before the experiment and derived from the experimental data themselves. Often clustering provides an intermediate reduction of this data before secondary and tertiary analysis is applied to uncover functional relationships from the correlative relationships suggested by

Additional 20% more “background” genes



(a) EM MoDG (b) KMeans (c) XClust

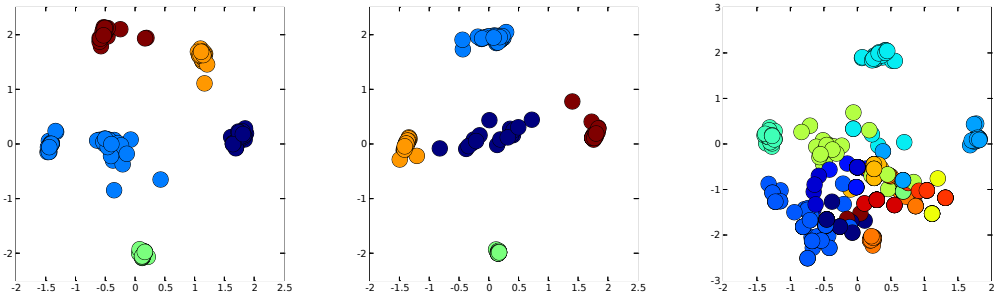
Addition 320% more “background genes”



(d) EM MoDG (e) KMeans (f) XClust

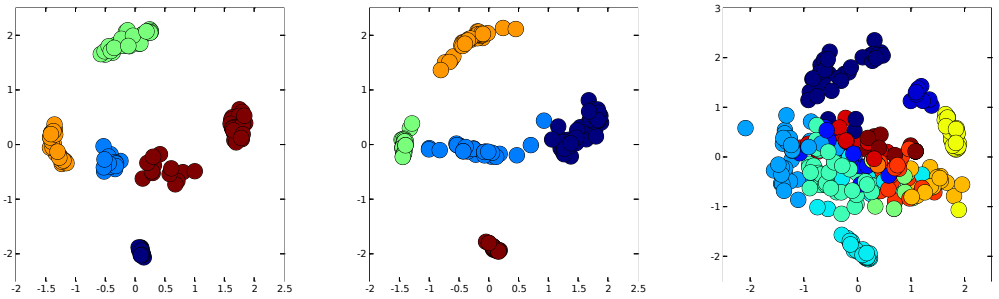
Figure 3.8: Cluster by cluster consistency analysis for clustering runs on the yeast microarray dataset [Cho et al., 1998] at both a modest and large dilution of genes. As described in section 3.2.4 data was clustered after varying the dilution of “interesting genes” with “background” genes. Sub-figures display cluster consistency curves on a cluster by cluster basis as in figure 3.4 for dilution level and for each algorithm (EM MoDG (a,d), KMeans (b,e), and XClust (c,f))

Additional 20% more “background” genes



(a) EM MoDG (b) KMeans (c) XClust

Addition 320% more “background genes



(d) EM MoDG (e) KMeans (f) XClust

Figure 3.9: PCA cluster consistency for clustering runs on the yeast microarray dataset [Cho et al., 1998] at two dilution levels; 20% and 320%. Clusterings were performed as described in figure 3.8. As in figure 3.5 each cluster mean was then projected into the top two dimensions of the PCA space defined by original data.

the clustering results. These techniques provide a foundation to evaluate and interrogate the nature of the clustering results to ensure that meaningful relationships are being pursued rather than artifacts.

Part II

Integrative Analysis of Gene Expression Data and Protein:DNA Interaction Data

Chapter 4

Inference of Cell Cycle Phase Specific Regulator-To-Gene Connections Using Artificial Neural Networks

Abstract

The yeast cell cycle is often divided into four distinct phases. Transcriptional regulation plays an important role in both the maintenance of, and the transitions between, these phases. Each phase has been shown to correspond with a group, or cluster, of genes that have phase-specific peak RNA expression, and are concordantly expressed throughout the cell cycle. We trained artificial neural networks to predict gene RNA expression patterns based solely on which transcriptional regulators have been measured to bind upstream. As the artificial neural networks share some structural properties with the transcriptional regulatory network that we are seeking to understand, we were able to derive a map of cell cycle phase-specific regulator-to-gene connections. Many of the inferred connections corresponded with previously identified regulatory connections. In addition, these connections often correspond to the conserved presence of known transcriptional binding sites across multiple yeast species.

4.1 Introduction

Cellular or developmental states are often characterized by sets of genes that are expressed consistently together. The yeast cell cycle is broken into at least four major cellular states. These states are referred to as the cell cycle phases. Genome-wide RNA expression profiling has associated each cell cycle phase with sets of genes, or clusters, that are observed to have phase-specific expression [Cho et al., 1998, Spellman et al., 1998]. We focus here on dissecting the transcriptional regulatory connections associated with each cell cycle phase. These regulatory connections are likely to underlie, in part, the establishment and execution of each phase, and the transitions between them.

One major goal of the work described in this chapter is to infer the temporal, or cell cycle phase, association of transcriptional regulatory interactions. Chromatin immunoprecipitation followed by microarray analysis (ChIP/chip) can assay the binding activity of a transcriptional regulator upstream of nearly every predicted gene in yeast. Through high-throughput adoption of ChIP/chip techniques [Ren et al., 2000, Iyer et al., 2001], Young and colleagues collected data for 204 of the currently 275 annotated yeast transcriptional regulators[1] for their respective binding activity upstream of nearly every gene in yeast [Lee et al., 2002, Harbison et al., 2004]. However, these data lack time or cell cycle phase resolution. In ChIP/chip experiments, for each measured transcriptional regulator a yeast strain is engineered such that the wildtype transcriptional regulator is replaced with an epitope-tagged version. Each strain was then grown and exposed to a cross-linker. After crosslinking the tagged transcriptional regulator, among other things, is covalently attached to the DNA in which it was bound to *in vivo* at the time of crosslinking. Using the epitope tag, the bound DNA is retrieved, amplified, labeled and hybridized to a DNA microarray with spots representing the intergenic, presumably regulatory, regions of the yeast genome. Mostly due to time and cost constraints these measurements are not performed time-resolved. As such, these data were collected from cells that are freely cycling. Therefore, the measured interactions between a transcriptional regulator and its target genes are the summation of all interactions that occur during any phase of the cell cycle that was suf-

[1] This number is derived from gene ontology annotations obtained from http://www.yeastgenome.org at the time of writing.

ficiently represented in the unsynchronized culture. Thus, the temporal, or phase-specific, aspects of these interactions are not explicit within the dataset and need to be reestablished from other data.

Using artificial neural networks (ANNs) we demonstrate that by coupling genome-wide time-resolved RNA expression data from microarrays [Cho et al., 1998] with large-scale measurements of genome-wide protein:DNA interactions from ChIP/chip experiments [Lee et al., 2002, Harbison et al., 2004] many of the known previously discovered regulatory connections associated with the cell cycle can be identified. In addition, several novel hypothetical regulatory associations are also found. With each of these regulatory relationships we also capture the cell cycle phase in which these regulatory associations are likely to be pertinent. From this, we begin to build a map of cell cycle phase specific regulator-to-gene connections.

Transcriptional regulation often relies on the aggregate effect of the interactions of multiple transcriptional regulators on a given target gene. These interactions of transcriptional regulators can have dramatic influences on their *in vivo* activity. For example, for many of the genes that show M-phase specific accumulation of mRNA during the yeast cell cycle, Mcm1 is bound upstream and capable of driving transcription constitutively throughout the cell cycle. For many Mcm1 dependent genes, expression is restricted to M-phase through adjacent binding of the transcriptional repressors Yox1 and Yhp1 [Pramila et al., 2002]. An important aspect of gaining an understanding of the regulatory networks that define and allow for the transitions between cell cycle phases, and cellular states in general, involves capturing these diverse kinds of regulatory interactions.

ANNs are structural models that have a long history in pattern recognition [Bishop, 1995]. They are often used in machine learning as “black boxes” to perform classification tasks. However, in this application the ANN model is of interest because the underlying structure of the network reflects structural properties underlying the regulatory networks we are seeking to understand, such as non-linear sparse interactions between transcriptional regulators and target genes [Mjolsness et al., 1991, Weaver et al., 1999, Vohradsky, 2001]. In this study we construct a simple ANN classifier that can be used to predict the expression behavior of a gene given only information regarding transcription factor binding activity

upstream of that gene. The ability of the ANN to predict expression behavior solely based on the binding activity of transcriptional regulators is a strong indication that the primary regulation of that gene's mRNA concentration is determined by the measured factors as opposed to other complicated regulatory processes such as highly controlled RNA processing, including degradation, or chromatin modifications.

The structural nature of the ANN model can be interrogated to build a map of regulator-to-target gene associations. It also provides a ranking of these associations for each regulator on regulating a particular RNA expression behavior. These relationships in effect cast the kinetic context of expression data onto connections implied from non-kinetic ChIP/chip studies. We constructed a simple single layer network (Figure 4.1) partially because we can more easily gain an intuitive understanding of the ANN. Further, addition of multiple layers into the network provided little gain in prediction accuracy. It is important to appreciate that, although the weights in this network do not have a direct biological analogy, they do provide a clue in understanding the importance of a regulator's binding or absence of binding in producing a particular gene expression pattern.

4.1.1 Methods

4.1.1.1 Data Preprocessing:

The microarray dataset measuring expression levels of nearly every gene in yeast throughout two cell cycles was obtained from the cited authors of Cho et al. (1998). These data were collected by Cho and colleagues from yeast cells synchronized using a cdc28TS arrest. RNA was extracted from the cells every 10 minutes for 170 minutes. Labeled target was synthesized from the extracted RNA and then hybridized to Affymetrix arrays. The resulting data processing was the same as in section 3.2.1. Briefly, any gene that did not show a sustained absolute expression level of at least 8 for 30 consecutive minutes was removed from the analysis. For each of the remaining 6174 gene vectors we divided each timepoint measurement by the median expression value across all time points for the gene. The log_2 of each ratio was then used to create the expression matrices that we used. Much of our analysis focuses on a set of 384 "cycling" genes in which Cho et al., 1998 identified

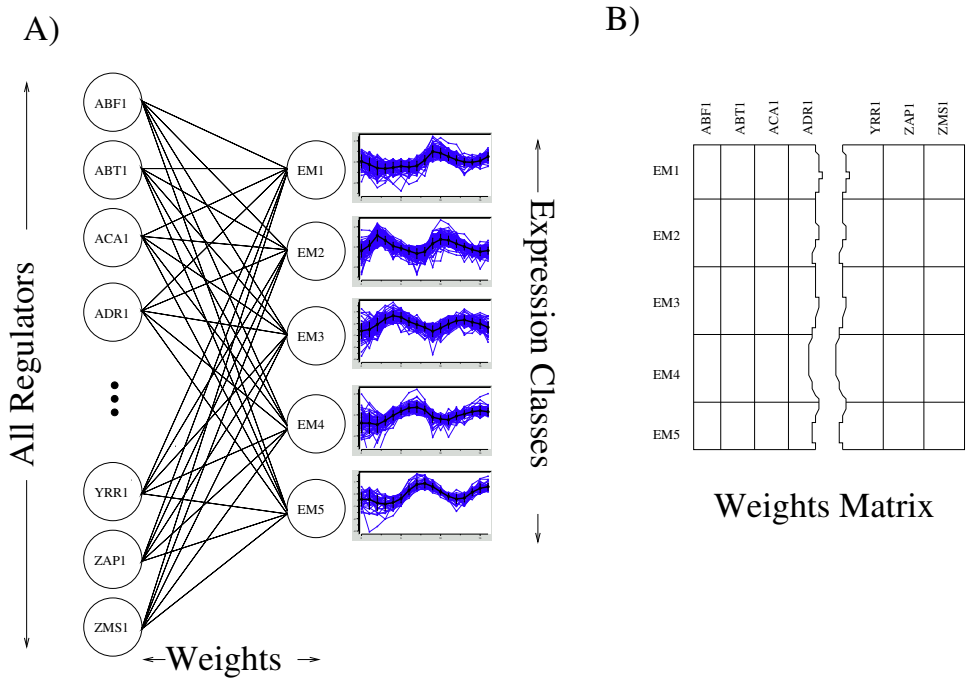


Figure 4.1: **The Artificial Neural Network Architecture (ANN)** A) Shown is the simple single layer network we trained to predict expression behavior based on the *in vivo* binding activity of $\sim 75\%$ of the transcription regulators in yeast. A 204 dimension vector containing the measured binding data from [Harbison et al., 2004] is used as the input vector. Given this binding vector the ANN was trained to predict during which of the five canonical cell cycle expression groups it is likely to be expressed. These expression classes were determined using EM MoDG (section 2.4.6) B) Matrix representation of the ANN. Each matrix cell, $W_{c,r}$, represents the real-valued connection strength, or weight, between a regulator (r) and an expression class (c) and is shown in A as an edge between a regulator and an expression class. These weights represent the importance of a regulator's binding activity or inactivity in the associated expression class

to show cell cycle dependent expression and which also passed our thresholds for being significantly expressed.

The protein:DNA interaction dataset (ChIP/chip) was collected from the cited authors of Harbison et al. (2004). No further processing was necessary with these data and the reported p-values were used for all of our analyses. Briefly, for each of the 204 assayed transcriptional regulators, Harbison and colleagues labeled targets synthesized from DNA that were enriched through chromatin immunoprecipitation (ChIP) using an affinity tag directed against the specific transcriptional regulator being measured. The targets synthesized from the ChIP enriched DNA was then co-hybridized along with targets synthesized and differently labeled from control DNA. Nearly every intergenic sequence in yeast was represented as a single feature on the microarrays. A binding ratio was then calculated based on the relative hybridization signal for targets synthesized from ChIP enrichment vs control DNA. Three biological replicates, starting from fresh yeast cultures each time, were performed. Based on an error model first described in [Hughes et al., 2000] and the three replicate binding ratios for each intergenic sequence, a p-value was calculated for each intergenic sequence. This p-value estimates the probability that a given transcription factor was bound to it.

4.1.1.2 Neural Network Implementation and Training:

Figure 4.1 illustrates the overall structure of the artificial neural networks (ANN) that we trained. We used backpropagation implemented by the UWBP package [Maclin et al., 1992] to train a simple single layer network with no hidden units. The “cycling” genes from the yeast microarray dataset were clustered using an expectation maximization algorithm fitting the data to a mixture of diagonal covariance Gaussians probability distributions (EM MoDG, section 2.4.6). We then trained artificial neural networks to predict the cluster membership of each gene based on the input vector of the binding probabilities for the 204 measured regulators. A best average network was created by iteratively splitting the data into testing and training datasets in which the training dataset contained 80% of the data and the testing dataset contained the remaining 20%. For each dataset split, ten neural networks were trained using different random seeds for each network. The network with the

best prediction accuracy on the testing dataset was then selected. This process was then repeated 40 times splitting the dataset into different testing and training datasets. The network weights from the resulting 40 selected "best" networks were then averaged together to create the average-of-bests neural network. We focused on this network for subsequent biological interpretation. The main goal was to identify regulatory connections between transcription factors and their target genes.

4.1.1.3 Consensus Site Enrichment Calculations:

In order to determine whether an expression cluster showed an enrichment in genes that contain a particular consensus site we calculated the likelihood of the observed enrichment, or depletion, being a chance occurrence according to a binomial model of occurrence probabilities. We count the observed number of genes that have at least one instance of a consensus sequence within the 1KB directly upstream of the coding sequence for all genes in an expression cluster versus the number of genes that would be expected by chance. As no known background sequence model is completely provably correct, for each consensus sequence we calculate the expected background frequency ($\hat{f}$) using a bootstrapping method. We randomly selected 1000 different sets of genes the same size as the cluster being compared (n). These randomly selected background sets are drawn from either the entire genome or from only the "cycling" genes which were used in training the ANNs. The number of genes that contain at least a single instance of the consensus is counted for each randomly selected set. The average count across the 1000 samples is normalized and used as our estimate of the expected number of genes within a cluster that have a single occurrence within 1KB upstream (E_c). Since the chances of any given gene within a cluster having a given consensus sequence within the 1KB upstream can be assumed to be independent, we can estimate the probability of finding the observed number of counts (O_c) using a standard binomial distribution (4.1). If the site is enriched we estimate the p-value for the likelihood of finding at least the observed count, but if the site is depleted we calculate likelihood of finding at most the observed count (equation 4.2).

$$P(i|c,n) = \binom{n}{i} \left(\frac{c}{n}\right)^{i} \left(1 - \frac{c}{n}\right)^{n-i} \tag{4.1}$$

$$p = \begin{cases} \sum_{i=0_c}^{n} P(i|E_c,n) & \text{if } O_c > E_c \\ 1 - \sum_{i=0_c}^{n} P(i|E_c,n) & \text{if } O_c \leq E_c \end{cases} \tag{4.2}$$

4.1.2 Results

4.1.2.1 Predictability of Expression Patterns

Using only the binding activity upstream of a gene across 204 of the 275 annotated transcriptional regulators in yeast, as measured by the global binding data reported by Harbison et al. (2004), our neural network classifier was able to place 86% of cell cycle regulated genes into their proper phase specific expression pattern (Figure 4.2). Although we were able to construct an average-of-bests network that showed this high degree of reliability in predicting gene expression behavior, individual ANNs trained on only 80% of the data and tested on the remaining 20% had an average predication accuracy of $\sim 50\%$, a minimal prediction accuracy of $\sim 40\%$ and a maximal prediction accuracy of $\sim 65\%$. Interestingly 125 genes in the dataset were predicted correctly by every ANN we trained, and 108 genes in the dataset were incorrectly classified by every ANN trained. The remaining genes in the dataset were predicted correctly only by a fraction of the individual ANN runs. Shown in figure 4.3 is the relative reproducibility of the rank order of regulators when we compare a best-of-average ANN built on the first 20 ANN runs with a best-of-average ANN built on the second 20 ANN runs. The ranking of regulators was based on the sum-of-squared weights taken across all expression classes for each regulator. This ranking paradigm focuses on the regulators that have the most significant weights, both positive and negative, in the ANNs computation of expression class predictions. In general, the ranks of regulators are stable across multiple training runs. Figure 4.4 shows the distribution of predictability across the EM clusters. There is an enrichment among genes that show the highest predictability to be in EM2; cluster that corresponds best to late G1.

In contrast to the predictability biases found from inspection of individual ANN runs,

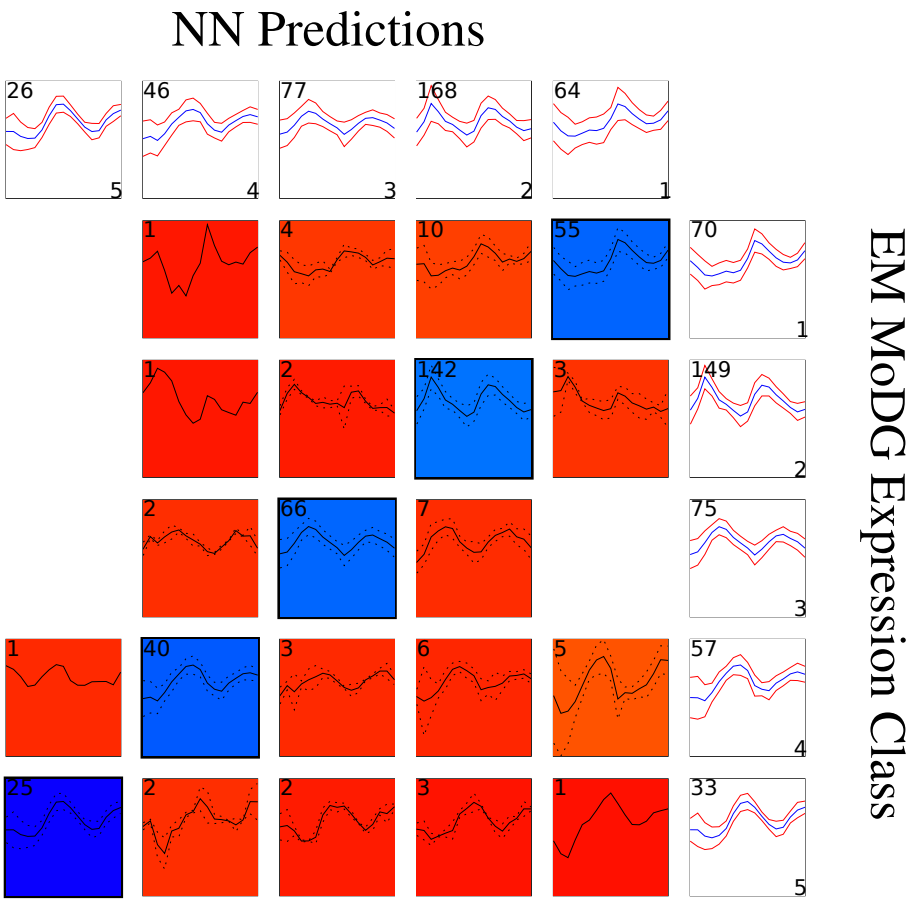
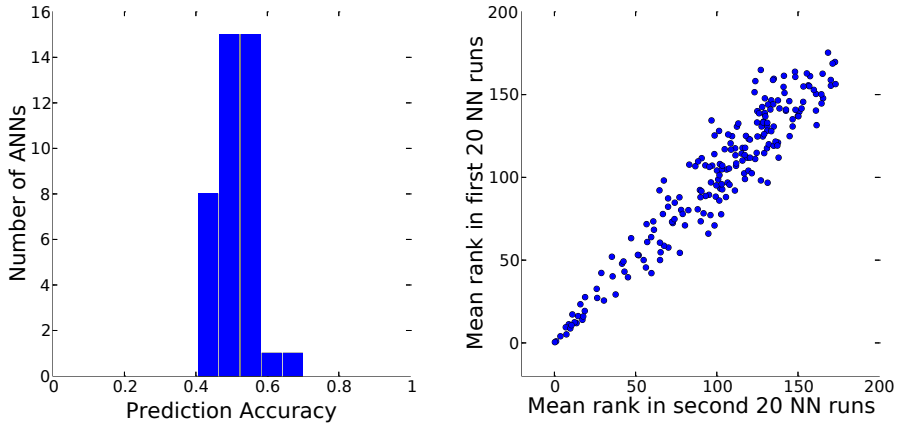
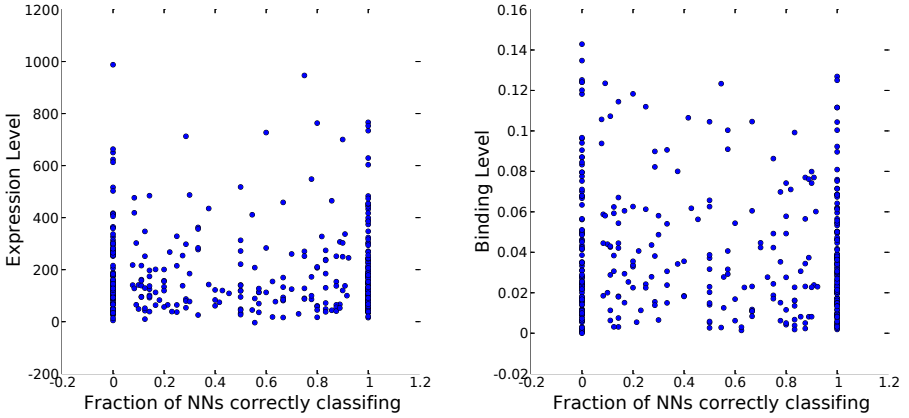


Figure 4.2: **Confusion Array showing the average-of-best ANN vs EM MoDG expression classes** (see methods 4.1.1.2). Here we compare the expression class prediction of the average-of-bests ANN which was created by averaging 40 ANNs trained to predict expression behavior from the binding data available for a gene. Each of the 40 ANNs were trained on 80% of the data and tested on the remaining 20% and they were selected as the best performing network out 10 networks trained on the same data split but initialized with differing seeds. These two classifications have a similarity of .86 by linear assignment



(a) ANN Accuracy

(b) ANN Reproducibility



(c) Predictability Vs Expression Level

(d) Predictability Vs Binding Level

Figure 4.3: **ANN Prediction Accuracy Histogram and correlations with binding and expression levels**. We trained 40 ANNs (see methods 4.1.1.2) to predict a gene expression behavior from only the regulator binding activity upstream to its start of transcription. For each network we trained on 80% of the data and tested on the remaining 20%. a) The distribution of ANN accuracy across the 40 trained ANNs. Along the x-axis are bins of accuracy ranges, the y-axis counts the number of ANNs that showed the designated prediction accuracy. b) Displays the relative reproducibility of the ANN rankings. Each regulator was ranked by its net influence in the ANN using the sum of squared weights across the classes in the weight matrix ($\sum_c w_{c,r}^2$). Shown is a scatter plot of the regulator ranks from the first 20 ANNs vs the second 20 ANNs trained. c) Scatter plot of the predictability (fraction of ANNs correctly classifying a gene correctly) vs mean absolute expression level of the 4 highest measured time points for each gene. d) Predictability vs mean binding level for the 10 highest bound regulators.

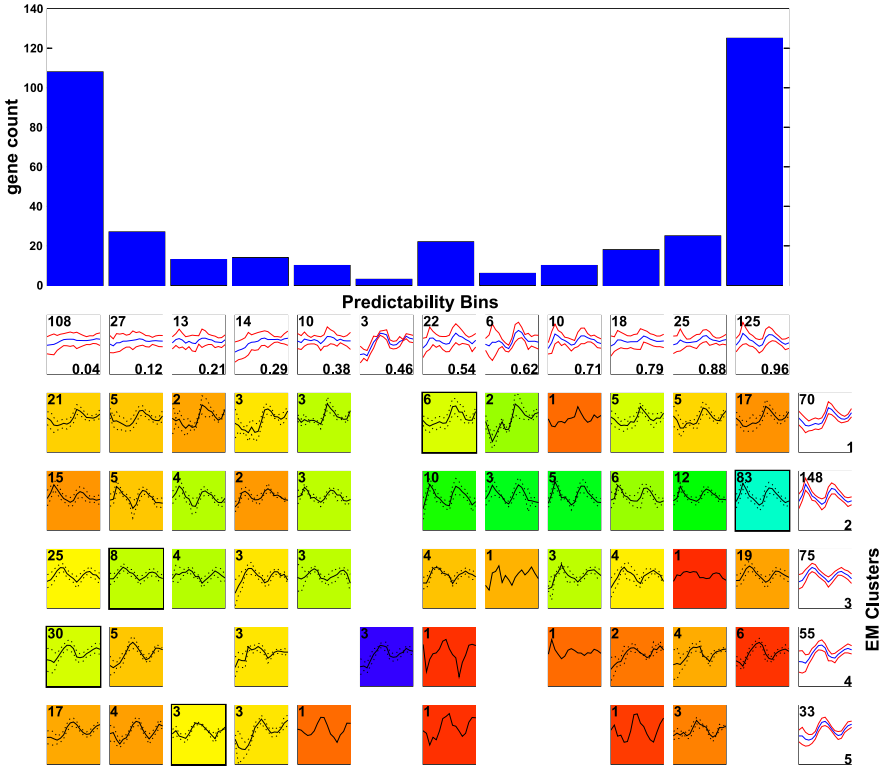


Figure 4.4: **Distribution of Neural Network Prediction Accuracy across EM MoDG Clusters**. The y-axis on the top panel measures the number of genes correctly classified by the indicated fraction of the trained ANNs (x-axis, bin range specified in the lower right corner of corresponding confusion array cells). Each bin is then broken up across the 5 EM MoDG clusters using a confusion array. The color map within the confusion array in the lower panel is shown as in figure 2.1

the average-of-bests ANN does not show a bias in expression class prediction. The confusion array shown in figure 4.2 illustrates that expression class prediction errors are evenly distributed throughout the dataset and is not specific to a single phase of the cell cycle for the average-of-bests ANN. Further, in the case of predicting the canonical G1 expression behavior which is represented by EM2, 65% of the misclassifications are slight errors whereby a gene is predicted to be expressed in either of the neighboring classes (ie. EM1 or EM3). However, in the case of EM1 56% of the errors are by misclassification into one non-adjacent kinetic expression cluster EM4.

4.1.2.2 Parsing the ANN Weight Matrix and Relating Inferred Regulatory Presence to Binding Site Presence

We next interrogated the weight matrix from the average of best network to find out which regulator's binding, or absence of binding, is important for predicting a gene's expression class. By sorting the regulators by their sum-of-squares ranks over of the expression classes, many previously known associations found by the network are highlighted. Figure 4.5 shows the weight matrix from the average-of-best network after sorting the regulators by importance. Shown are the top scoring 20% measured by using the sum of squared weights across the different expression classes. The very top regulators Swi6, Ndd1, Stb1, Fkh2, Mbp1 are all known regulators of the cell cycle and show appropriate associations with the expression patterns for which they are well known. For instance, Swi6 and Mbp1 are the first and sixth ranked regulators. They are known to function together as a heterodimer [Koch et al., 1993] and are well established positive activators of G1 gene expression. This is exactly where the weight matrix shows the strongest positive associations. The absence of swi6 binding relative to other clusters is also used by the networks as a strong indicator of gene expression during G2/M as represented by the low weights for Swi6 in EM4 and EM5. Ndd1 and Fkh2, the second and fourth ranked regulators, are another set of transcription factors which are functional together in a complex [Koranda et al., 2000]. They are found to have associations with S/G2 behavior in the neural networks, again recapitulating the activity of its known target genes.

Inspection of regulator weights, sorted on an expression class by expression class basis,

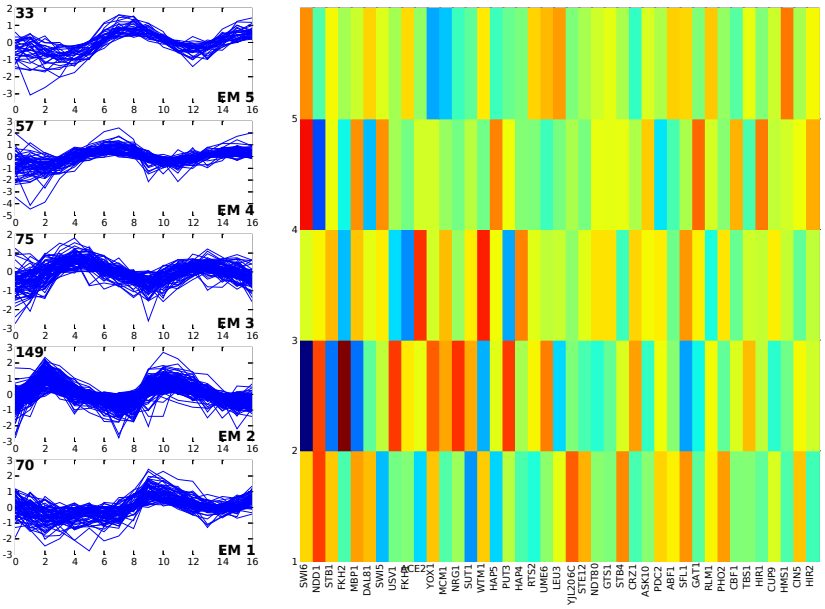


Figure 4.5: **Weight matrix for the Average-of-bests ANN**. Shown are the top 20 regulators after sorting each regulator by importance in predicting expression behavior using a sum-of-squared weights measure. The left hand column shows a trajectory summary for each expression cluster as classified by EM MoDG. The right hand color map represents the weight matrix where expression classes are displayed along the rows corresponding to the drawn trajectory summaries. Regulators are sorted along the columns in rank order. Each cell is colored proportional to its value in the weight matrix.

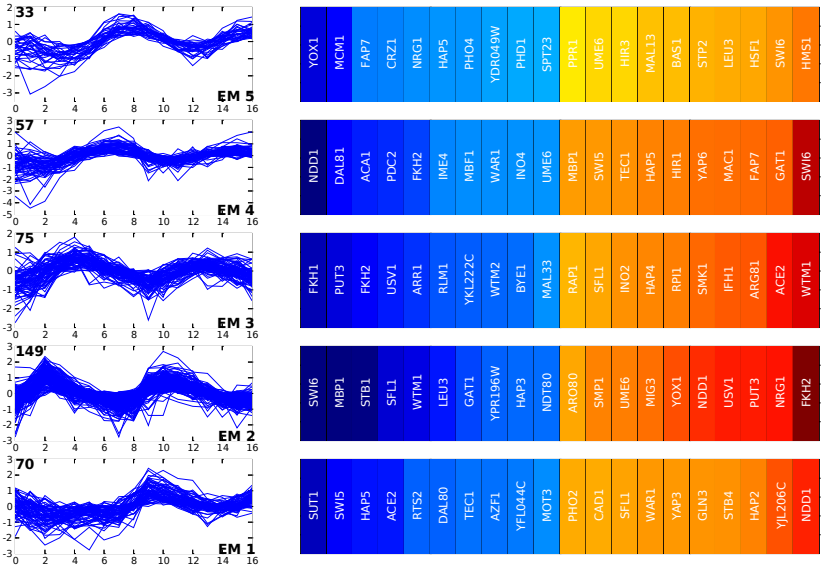
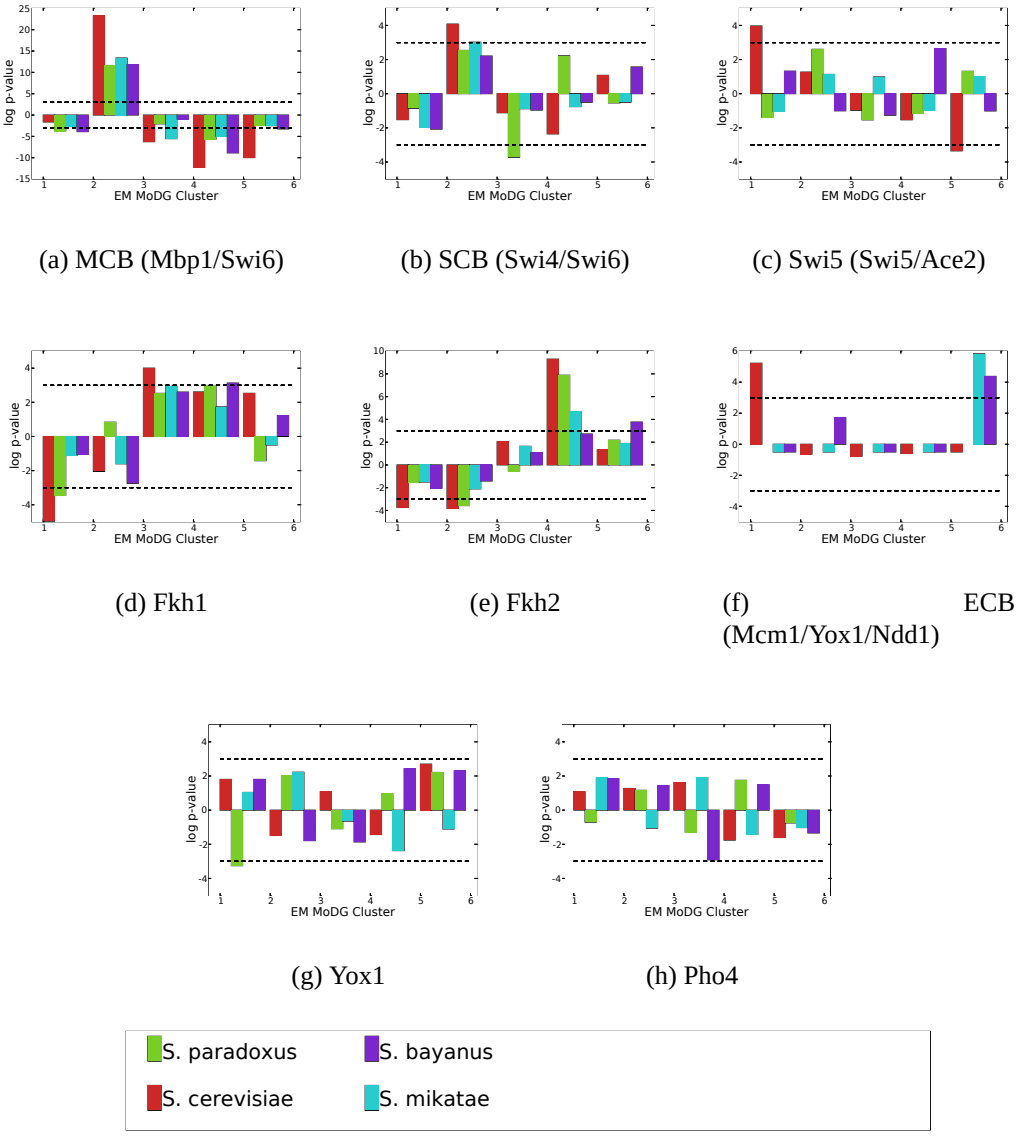


Figure 4.6: **ANN weights sorted on an expression class basis**. Shown are the ANN weights from the average-of-best network as in figure 4.5 with the exception that the top and bottom regulators for each class are displayed. The regulator ranking for each class is simply based on its weight in the weight matrix for each expression class. Detailed annotations for these regulators are listed in table A

reveals even more details of the connections inferred by the ANNs. Figure 4.6 and Table A show for each expression class the top ten positive and negatively associated regulators. The associations discovered by the neural networks imply a domain of activity or inactivity during the cell cycle that would be otherwise difficult to reveal using simple statistics, such as mean binding activity across an expression class. This is true even though the network weights are not directly interpretable as functional interactions. We observed that sometimes when a regulator is implicated with a particular expression class there is a corresponding enrichment of genes whose upstream genomic sequence contain the regulator's binding site (Figure 4.7). Site enrichment patterns for genes containing at least one binding site for a given factor are further supported by the conservation of this phenomenon across several of the sequenced species of Saccharomyces.

Interpretation of the ANN revealed a strong negative association of Swi6 and Mbp1,

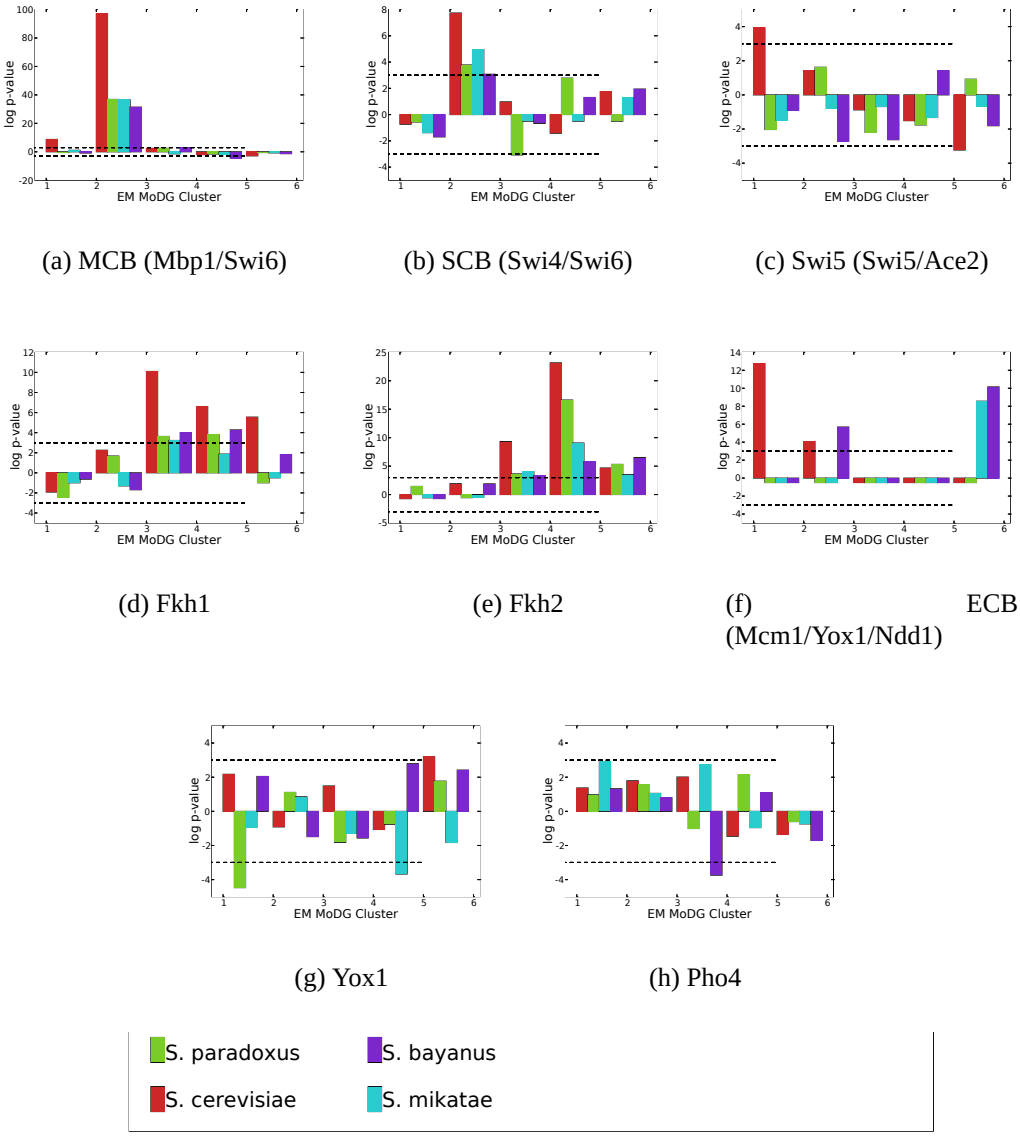


(a) MCB (Mbp1/Swi6) (b) SCB (Swi4/Swi6) (c) Swi5 (Swi5/Ace2)

(d) Fkh1 (e) Fkh2 (f) ECB (Mcm1/Yox1/Ndd1)

(g) Yox1 (h) Pho4

Figure 4.7: **Binding Site Enrichment and Depletion**. For several of the regulators highlighted by strong positive or negative association with particular expression classes (figure 4.6) we calculated site enrichment p-values for each EM MoDG cluster across each of seven sequenced Saccharomyces species (see methods 4.1.1.3). Each p-value was calculated using only the cell cycle identified genes that were also used as input genes to the ANN. Each block of bars along the x-axis represent log p-values (y-axis) for a EM MoDG clusters. Each bar within these blocks are log p-value measurements for a different Saccharomyces species as indicated by the color legend. Enrichment is shown as positive values (-log p-values) and depletion is shown as negative values(log p-values). The species have been arranged by evolutionary distance from S. cerevisiae. From left to right: S. cerevisiae, S. paradoxus, S. mikatae, S. bayanus. A dashed line along the graphs at p-value = .05 has been drawn to help visualize the scale difference between the plots. a-h) enrichment bar charts for the specified binding sites, if the binding site is referred to by a name other than the regulator that binds to it, the regulators that bind are parenthetically shown. i) A displaying the color map used for each bar.

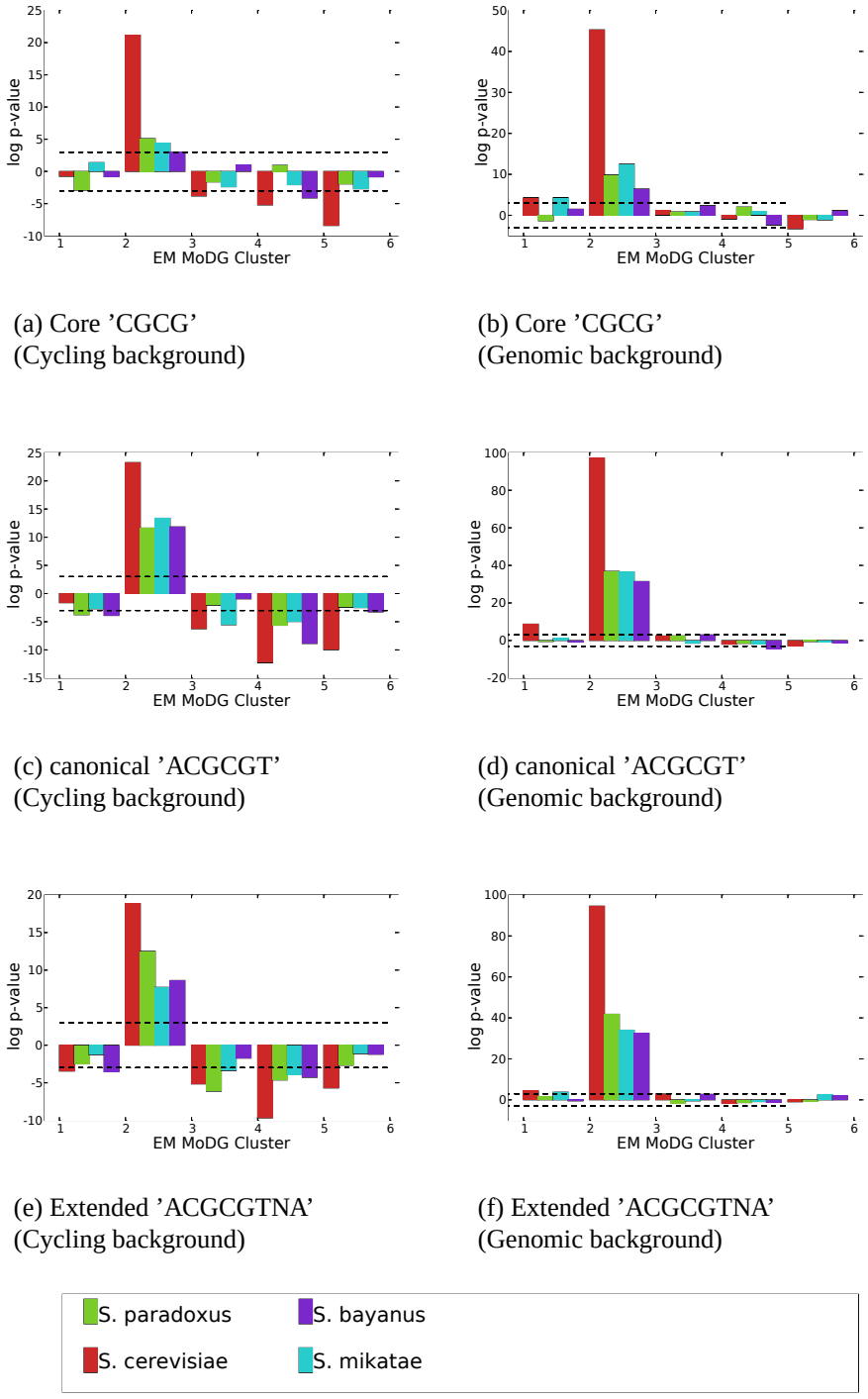


(a) MCB (Mbp1/Swi6) (b) SCB (Swi4/Swi6) (c) Swi5 (Swi5/Ace2)

(d) Fkh1 (e) Fkh2 (f) ECB (Mcm1/Yox1/Ndd1)

(g) Yox1 (h) Pho4

Figure 4.8: **Binding Site Enrichment and Depletion using whole genomes for backgrounds**. As shown in figure 4.7 enrichment p-value across each of the sequenced Saccharomyces species is shown. In this case, cluster enrichment p-values are calculated using the whole genome as the background set rather than just the cell cycle group.



(a) Core 'CGCG'
(Cycling background)

(b) Core 'CGCG'
(Genomic background)

(c) canonical 'ACGCGT'
(Cycling background)

(d) canonical 'ACGCGT'
(Genomic background)

(e) Extended 'ACGCGTNA'
(Cycling background)

(f) Extended 'ACGCGTNA'
(Genomic background)

Figure 4.9: **Influence of MCB site specification on Enrichment and Depletion Statistics** Enrichment and Depletion p-values for increasingly stringent definitions of the MCB binding site. As shown in figures 4.7 and 4.8 site enrichment statistics for each site definition are shown using only "cycling" genes (a,c,e) or all genes (b,d,f) for the background expectation calculations.

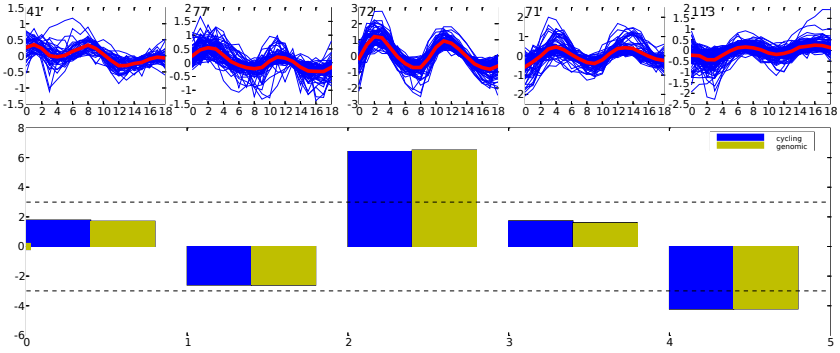
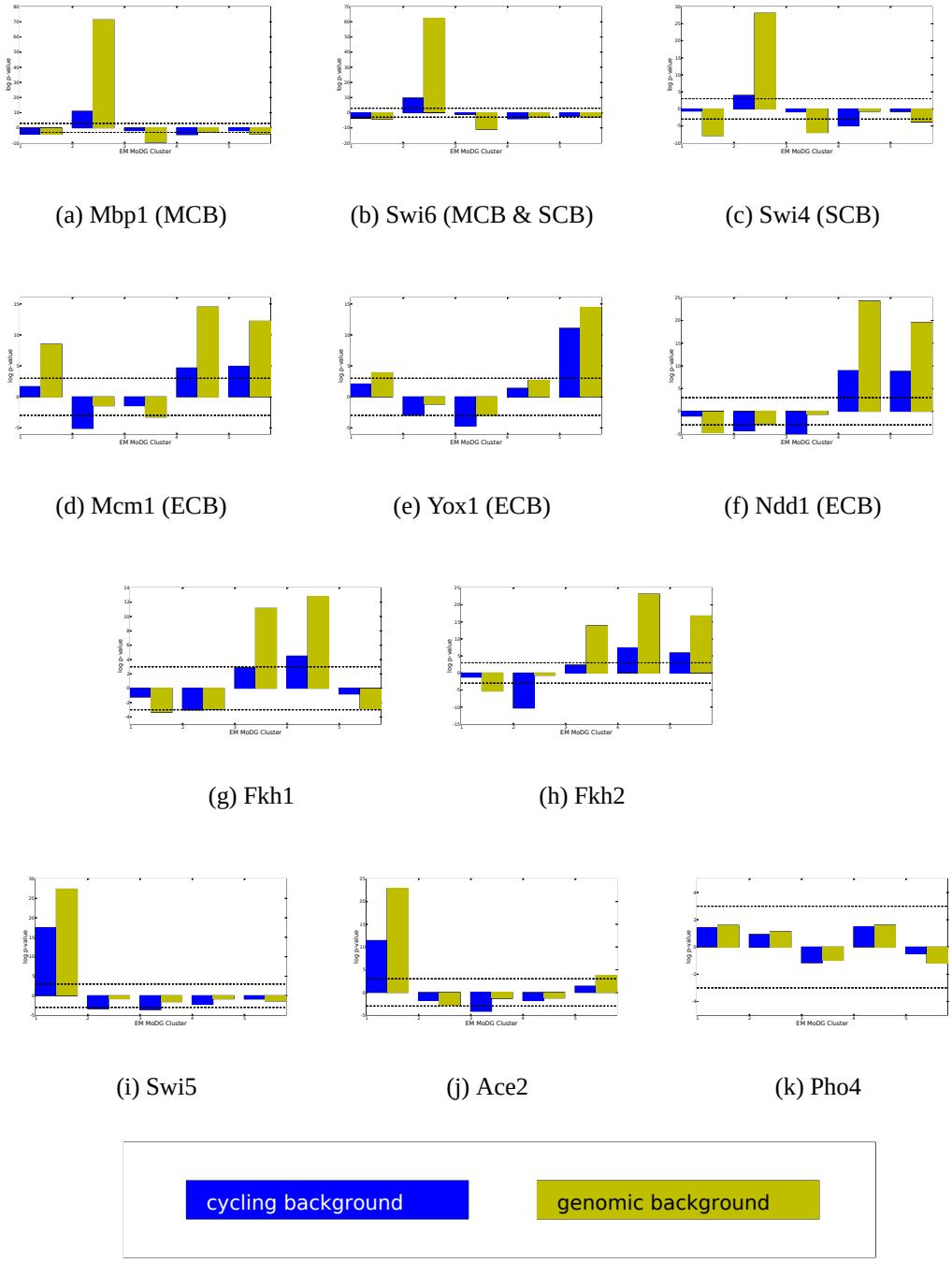


Figure 4.10: Binding Site Enrichment and Depletion for S. Pombe. Shown are the MCB enrichment p-values for S. pombe based an EM MoDG clustering of the expression data from Rustici et al., 2004. Cluster summaries for each of the expression clusters are shown along the top panels, red lines are the mean expression trajectory and cluster sizes are in the upper left corner. Below is a bar chart of p-values. Shown are the p-values normalized against only the cycling genes (blue) and p-values normalized against the whole genome (red).

with EM4 and EM5. This lead us to ask if the canonical hexameric binding site for MBF, the complex of Mbp1 and Swi6 which binds to the MCB site, was statistically depleted in those clusters. We did not find a statistically conserved depletion of MCB sites in either EM4 and EM5. Although, we do find a statistically significant enrichment of MCB sites within EM2, the cluster most strongly associated with Mbp1 and Swi6. This enrichment and depletion pattern remains conserved even in S. pombe (Figure 4.10). We investigated the dependency of this phenomenon on the definition of the binding site by repeating the analysis for 3 differing site definitions; the core 'CGCG', the canonical 'ACGCGT' and the extended site 'ACGCGTNA'. Even the core 4-mer, although to a lesser degree, showed significant enrichment in EM2. The most stringent site revealed very similar patterns of occurrence to the canonical site.

To directly compare the enrichment of sites to the observed binding data we performed an analogous computational experiment but instead of counting the number of genes in each cluster with at least one binding site, we counted the number of genes that were measured to have significant binding by ChIP/chip. As in the site presence experiments we calculated the probability of the observed gene count for each EM MoDG cluster (Fig-



(a) Mbp1 (MCB) (b) Swi6 (MCB & SCB) (c) Swi4 (SCB)

(d) Mcm1 (ECB) (e) Yox1 (ECB) (f) Ndd1 (ECB)

(g) Fkh1 (h) Fkh2

(i) Swi5 (j) Ace2 (k) Pho4

Figure 4.11: **Enrichment and depletion of regulator binding within each EM cluster**. We calculated an enrichment p-value measuring the degree in which each EM MoDG cluster is enriched or depleted for genes that are bound *in vivo* by the shown regulators. A gene was considered bound if it had been measured to have a p-value $\leq .05$ by Harbison et al., 2004. Shown with blue bars are the enrichment values when each cluster's enrichment is estimated using only genes within the "cycling" genes, yellow bars using all genes in the genome. These calculations are analogous to those discussed in section 4.1.1.3. Dashed lines at p$\leq$.05 are shown to visualize the scale difference.

ure 4.11). The enrichment for binding closely parallels the enrichment we observe for the site presence data for many of the regulators. However, the binding pattern of Swi5 and Ace2, which both bind to the same binding site, have *in vivo* binding patterns that are much more consistent with the known biological roles of Swi5 and Ace2 in early G1 regulation [Doolin et al., 2001]. The binding pattern of Yox1, whose sequence presence showed virtually no enrichment or depletion, shows striking agreement with the known biology of its involvement as transcriptional repressor whose action restricts gene expression to M phase [Pramila et al., 2002]. It should be noted that these simple statistics do not reveal the same level of structure as did the ANN weight matrix. For example, the ANNs provide a relative ranking of each regulator association within each RNA expression class.

4.1.3 Discussion

We show that the majority of cell cycle regulated genes can be classified into expression classes based solely on which transcriptional regulators have been measured to bind *in vivo* to a gene's upstream activating sequence. By using artificial neural networks (ANNs) to construct the classifier we also gain a structural model that relates both the presence and the absence of regulator binding activity to phase specific cell cycle gene expression. By ranking both positive and negative weights from ANNs we were able to assign priorities to both known and and previously unknown associations that regulate the cell cycle.

4.1.4 Examining the Connections Inferred by the ANN

The sum-of-squared weights sort order used in figure 4.5 highlights the most important cell cycle regulators as found by the ANNs. The top 12 regulators, with the exception of Dal81 and Usv1, all have well established roles in cell cycle regulation. Although neither Dal81 or Usv1 have been implicated in cell cycle regulation, because they were given high phase specific weights by the ANNs (figure 4.6), it suggest they may either directly or indirectly play a role in the cell cycle.

The weight matrix of the ANN as sorted and displayed in figure 4.6 more directly associates regulators with particular expression classes. We find that Swi5 and Ace2 are

among the top four genes positively associated with EM1. As discussed in more detail in chapters 1 and 2, Swi5 activates early G1 genes after being dephosphorylated by cdc14 and entering the nucleus [Visintin et al., 1998].

Swi6 and Mbp1 are represented in EM2, the canonical G1 cluster. These are the two canonical components of MBF, the heterodimeric complex that binds the MCB element and drives G1 expression [Koch et al., 1993]. The next most important regulator of the G1 group, according to the results from the ANN, was Stb1, which has recently been shown to act in vivo at MBF-regulated genes through interaction with Swi6 [Costanzo et al., 2003]. Missing from what would be expected from previous experiments is Swi4. This is interesting as it indicates the Swi4 binding has little predictive power in associating genes with G1. This is further supported in figure 4.11 where, although there is a large enrichment for Swi4 binding in EM2 when compared to the entire genome, there is only a marginally significant enrichment when we compared the enrichment for Swi4 binding in EM2 with only cell cycle genes. This could be an artifact of poor ChIP/chip efficiency or rather it might be a reflection of a more diffuse role for Swi4 in the cell cycle, although we know of no experimental support for such a function. Additional specific measurements of Swi6 activity and binding could be made to answer this question.

Fkh1 and Fkh2 are both found associated with EM3, the S/G2 expression cluster, which parallels the inferences made regarding their function in double knockout experiments [Zhu et al., 2000]. Ndd1 has a large positive weight in the weight matrix for EM4, indicating that as the cell progresses closer towards M-phase the role of Ndd1 in regulating this pattern increases, as is thought to be the case from more conventional experiments involving single gene chromatin immunoprecipitation experiments [Koranda et al., 2000]. Lastly during M-phase, the ANN implicates Yox1 and Mcm1 as the two largest transcriptional contributors to this expression pattern. Pramila et al. (2002) show that Yox1 and/or Yhp1 act as a transcriptional repressor to restrict transcription of genes containing the early cell cycle box (ECB) from being driven by Mcm1. Yox1 is shown by the weight matrix to have a positive association with M-phase expression, even though the factor is a repressor of transcription. Although perhaps counter-intuitive, this is actually expected and understandable because the net effect of having Yox1 binding is indeed to restrict expression to

M-phase dynamics.

The weight matrix from the ANN also associates many transcriptional regulators with phase-specific activity in the cell cycle that may no be appreciated yet. Many of the top ranking regulators appear to be involved in metabolic or related processes (table A). Given the central role of the cell cycle, finding hypothetical associations is not surprising. It is encouraging, however, many of the previously known regulatory connections are revealed. Also, several of the newly suggested connections have supporting data in the literature.

For instance SUT1 is strongly implicated in early G1 by the ANNs. Little is known about SUT1 other than that it is involved in regulating hypoxic genes and has been shown to work through physical interaction with the cyc8-Tup1 complex [Regnacq et al., 2001]. This is interesting as cyc8-Tup1, like Swi5, is a general transcription factor that recruits the chromatin modifying complexes SWI/SNF and SAGA complexes to derepress genes. Additionally the kinase activity of the cyclin dependent kinase (cdk) Srb10 (also known as Ssn3) contributes to the repression of roughly 15% of Tup1 repressed genes [Green and Johnson, 2004]. SUT1 may play a role in the regulation of early G1 gene expression. Because of the nature of cdc28 arrest it is likely that SUT1 activity is involved in regulating genes during the M/G1 transition before cdc28 become active.

SFL1 is another gene which has not yet been associated with the cell cycle, but it has a strong positive association with the EM2 cluster and a negative association with EM3. It has been identified as a transcriptional repressor that acts through interactions with the Srb/mediator proteins to inhibit transcription [Song and Carlson, 1998] with involvement in pseudohyphal differentiation through interaction with protein kinase A [Pan and Heitman, 2002]. SFL1 repression might be important in regulating some G1 genes.

The associations highlighted by analysis of the weights matrix of the ANNs are highly overlapping with the cell cycle modules identified by different techniques of Gifford, Young and colleagues [Lee et al., 2002, Bar-Joseph et al., 2003, Harbison et al., 2004]. Using an algorithm referred to as GRAM (Genetic Regulator Modules). It works by using RNA co-expression to add support to regulatory interactions as measured by ChIP/chip measurements that would have otherwise been slightly below statistically significance [Lee et al., 2002, Bar-Joseph et al., 2003, Harbison et al., 2004]. As they focused on identifying more gen-

eral transcriptional modules our results are not directly comparable. However, we identified and properly associated the all core cell cycle regulators that they identified, with the exceptions of Swi4 and Skn7.

4.1.5 Prediction Accuracy

The average-of-best ANN was able to achieve an in-sample prediction accuracy of 86% (Figure 4.2). We also tested the out-of-sample accuracy, or the ability of our training paradigm to generalize to another set of independently collected binding measurements. When we trained an average-of-bests network using only binding measurements from the 111 regulators available in both the Harbison et al. (2004) study and the independent Lee et al., 2002 study, the proper expression class for 56% of the genes was correctly predicted (Figure 4.12). This is still a highly significant 17 standard deviations from the average linear assignment score (0.27 ± 0.017) of a random partitioning of the genes where class sizes are determined by drawing from a multinomial distribution based on the EM MoDG cluster sizes.

The prediction accuracy of any specific ANN was much lower (Figure 4.3). About one half of the genes were correctly classified in a small percentage of the trained ANNs, and around one quarter were never classified correctly. Likewise, roughly one third of the genes were always classified correctly and many more were classified correctly in the majority of ANNs trained. These observations are interesting as they may suggest that these "unpredictable" genes whose expression class could not be predicted from their regulator binding data may be subject to more complicated regulation involving other mechanisms than cis-regulation, such as chromatin remodeling or post-transcriptional regulation. The EM2 cluster has the highest proportion of highly predictable genes (4.4). This might be explained by the fact that the expression signature of EM2 is more well defined than the other clusters. This observation parallels those of chapters 2 and 3 where we also found the expression signature of EM2 the most distinct and robust to additional noise. Although we see a bias in predictability dependent on expression class we see no observable bias for predictability based on simple statistics such as absolute expression level (Figure 4.3c)

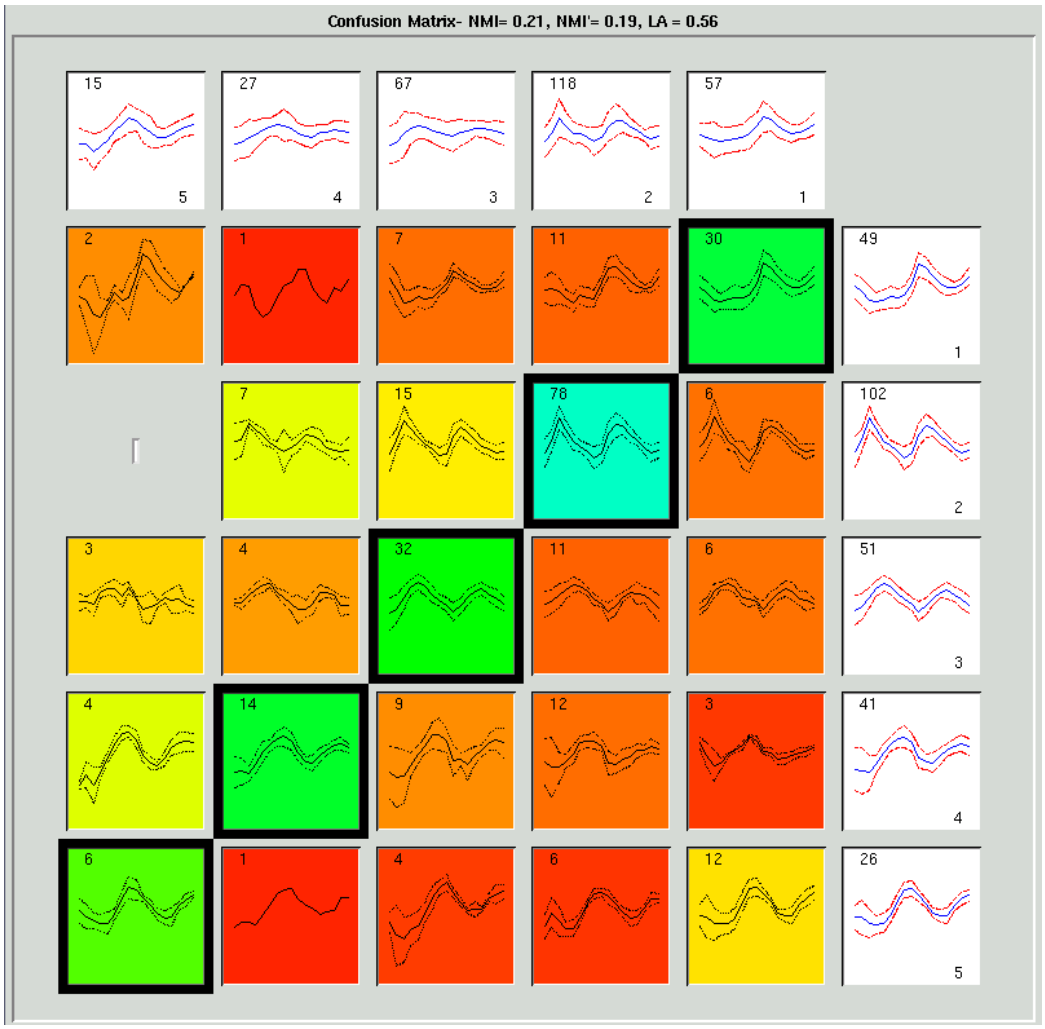


Figure 4.12: **ANN prediction accuracy using a validation dataset.** As in figure 4.2 a confusion array is shown between prediction classes (columns) and the EM MoDG expression classes (rows) used in training. In this case, the average-of-bests network was trained using only binding measurements from the 111 regulators available in both the Harbison et al., 2004 study and the independent Lee et al., 2002 study. After training the average-of-bests network using data from the Harbison et al., 2004 study, predictions were calculated based on the independent measurements from the Lee et al., 2002 study. Using the Lee et al, 2002 data as a validation dataset our training paradigm achieved a linear assignment score of .56 when compared to the EM MoDG classification.

or binding level (Figure 4.3d). Although not evident from the reported data, the high predictability of EM2 genes, and the poor predictability of EM5 genes could also be a result of better, or worse, chIP/chip measurements. Since the duration of each phase of the cell cycle is not equal, the longer phases of the cell cycle, such as G1, will be represented by more cells in freely cycling cultures. This is likely to result in more reliable chIP/chip measurements, especially for transient interactions.

4.1.6 Site Enrichment

The structure of the ANN also revealed principles of regulation that are conserved through several sequenced species of budding yeast. We find a conserved correlative enrichment for genes containing the respective binding sites for many of the strong positive and negative associations revealed in weight matrix of the ANN (Figure 4.6 and 4.7). For instance the weight matrix, as expected from previously known interactions, uncovers a very strong positive association with Swi6 and Mbp1 with the EM2 cluster (the G1 cluster). These two regulators form a heterodimeric complex which binds to the MCB binding site. We calculated how many genes from S. cerevisiae contain at least one site within the 1KB directly upstream from the coding domain within each of the EM clusters and compared it to how many we would expect by chance. We find that there is a conserved enrichment in the G1 cluster (EM2) across all species we compared. This suggests there is an extraordinary positive selective pressure for MBF sites in a specific subgroup of G1 genes.

To probe this hypothesis further we examined the fission yeast *S. pombe*, which is much more distantly related to *S. cerevisiae* than to any of the 7 sequenced budding yeast species. It is separated from *S. cerevisiae* by $\sim$400 mya. Starting from completely independent expression data and an independent clustering of this data we once again observe a significant enrichment of G1 genes that contain MCB sites (Figure 4.10). These results suggest, in the case of MCB, there are strong selective pressures acting to conserve the regulatory connection between MCB site presence and G1 RNA expression. Further, in plants and animals the MBF homologue E2F also shares a very similar binding site and is associated with G1 gene expression [Hateboer et al., 1998]. Although this level of conservation is probably

rare among binding sites, further experiments could ask how much functional complementation can be observed when either the DNA binding domain of MBF, or the whole complex from yeast is replaced with the corresponding domains or proteins from E2F.

Chapter 5

Conclusions and Directions

Most of the ideas regarding gene networks are based on studies of specific transcription factor interactions with one or a few "model" genes. A different approach regarding how regulatory connections both function and are maintained through evolution can be gained from the expansion of both the number of fully sequenced genomes and the availability of high throughput genome-scale functional assays. Network inferences based on these comprehensive, or nearly comprehensive, datasets provide the possibility to more globally map network connections in a direct and complete manner. However, these datasets introduce uncertainties and bioinformatic challenges associated with data quality and significance. Based on the work presented in the preceding chapters we have constructed a new map of transcriptional connections that are involved in the yeast cell cycle. This chapter discusses a few of the questions raised by the analysis described in the preceding chapters.

5.1 How Are Regulators Regulated?

Transcriptional regulation plays a significant role in the yeast cell cycle [Breeden, 2003]. In the previous chapters we developed computational methods to identify candidate transcriptional regulators from genome scale data. In figure 5.1 we show both the RNA expression pattern and the *in vivo* upstream binding activity for the summed set of the newly identified and previously known regulators of the cell cycle. We find the expression patterns for roughly half of these regulators of the cell cycle to have both lower amplitude dynamics and they do not fall as strictly into the identified expression classes as do many of their

target genes. Furthermore, the regulator binding occupancy for these genes do not show an obvious correlation with the bound regulator's RNA expression pattern. This suggests that the activities of these regulators are controlled through mechanisms that are not directly linked to their RNA expression.

As expected several of the regulators that function in the M/G1 transition show binding of Mcm1, Fkh1, Fkh2, and Ndd1, all of which have been shown to function together [Breeden, 2003]. The figure illustrates that although they are all known to function together at the protein level during one part of the cell cycle, their RNA expression is not coordinated. In this case, Mcm1 is constitutively expressed and, by extrapolating from detailed chromatin immunoprecipitation experiments of three target genes of MCM1 (CLN3, SWI4, and CDC6), it has been suggested that Mcm1 is bound to ECB sites throughout the cell cycle[Mai et al., 2002]. Its activity is modulated through its interactions with other regulators. Fkh1 and Fkh2 can bind to DNA in a phase independent manner, but recruitment of Ndd1 only occurs around M-phase [Koranda et al., 2000]. Interestingly, the DNA binding activity of Fkh1 and Fkh2 are thought to be interchangeable, and the functional differences have been ascribed to structural differences between the two proteins [Hollenhorst et al., 2000]. We find that Ace2, Yhp1, Fkh1 and Fkh2 are strong binders upstream of Swi5. Yet only Fkh2 binds upstream of Fkh1, possibly preventing a feedback loop that would have otherwise existed. Within these core cell cycle regulators we find Swi4, Sut1, and Tec1 are the only regulators that form direct feedback loops. Overall there are only 15 regulators out of the 206 regulators surveyed that show binding activity to their own upstream regulatory sequence at a $p \leq 0.001$.

In this core network, as expected from the known biology, Swi5 and Ace2 become active in late M-phase/early G1 through the regulation of Mcm1, Ndd1 and Fkh2. Swi5 then binds upstream of Ash1 and Tec1. Mcm1 also binds to Swi4 which binds upstream to Tec1, Yox1, Yhp1, Ndd1 and to itself. Yox1 and Yhp1 likely get expressed and begin to repress Mcm1 dependent M-phase genes. These observations are quite similar to our expectations that have been based on a survey of the current literature. Although, the set of regulators shown is likely an incomplete set of yeast cell cycle transcriptional regulators, it is still surprising that neither Mbp1 or Stb1 are bound significantly by any of these core

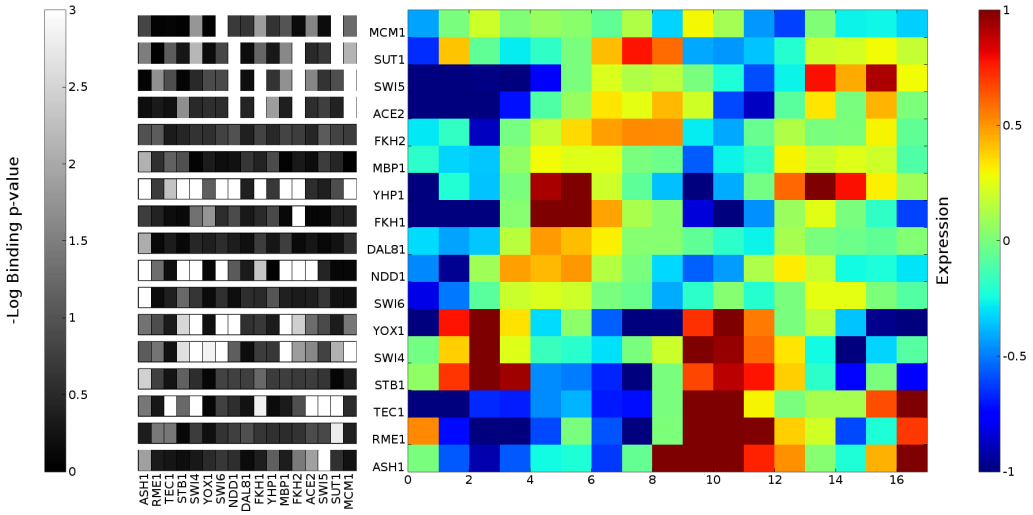


Figure 5.1: RNA Expression and Regulation of Cell Cycle Transcriptional Regulators. Shown on the right are the RNA expression trajectories for several previously known and hypothetical cell cycle transcriptional regulators. Genes are arranged along the rows and time is along the columns. On the left for each gene is a vector of binding values for each regulator of the cell cycle regulators. The gray scale is proportional to the $-log$ p-value of binding for each measurement as reported by Harbison et al., 2004. The order along both the rows and columns is fixed and based on the order of RNA expression through the cell cycle. Eleven of the 17 regulators were identified by both our ANN analysis and previously in the literature. These regulators are Mcm1, Swi5, Ace2, Fkh2, Mbp1, Fkh1, Ndd1, Swi6, Yox1, Swi4, and Stb1. Yhp1, Rme1 and Ah1 were identified only in the literature, and were not found by the ANNs. Tec1, Dal81 and Sut1 were suggested by our analysis to be cell cycle regulators, but have not, as of now, been recognized as such.

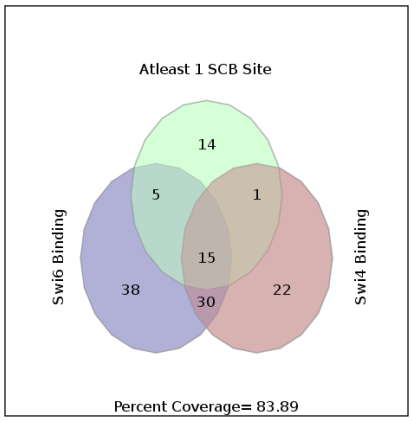
regulators. It is especially odd given they play a central role in regulating many G1 genes as discussed in chapter 4. Searching outside of these core regulators reveals that Pho2 is the regulator that binds most strongly (p=0.001) to Mbp1. Gcn4 and Rap1 are the regulators that bind most strongly to Stb1. This suggests that unlike most of the other regulators in this core network, the transcriptional regulation of both Mbp1 and Stb1 is primary regulating through connections that are indirectly linked to the core cell cycle network that we have defined.

5.2 Why are G1 genes co-expressed?

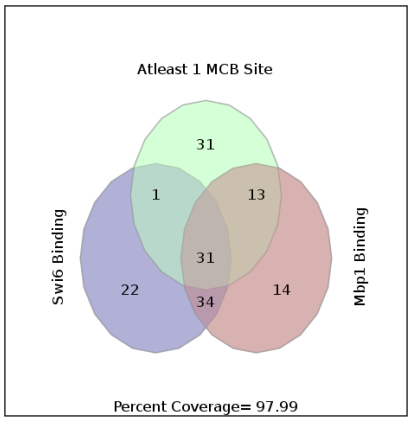
The most simple explanation for two genes being co-expressed is that they are also functionally co-regulated. For genes expressed coordinately during the late G1 phase of the cell cycle, detailed prior molecular and genetic studies have shown that two primary regulatory complexes have evolved; MBF (MCB binding factor) and SBF (SCB binding factor) [Andrews and Herskowitz, 1989, Koch et al., 1993]. Using artificial neural networks in chapter 4 and using simpler statistics in chapter 2 we find that both site presence and binding of these regulators are highly correlated with late G1 gene expression. Yet, in agreement with observations by Brown and colleagues [Iyer et al., 2001] we find only 60% of the genes in the G1 cluster (EM2) are bound by Swi6 ($p \leq .05$), the common component of both MBF and SBF (figure 5.2). In addition, Horak et al. (2002) showed that 40% of genes that are actively bound by either MBF or SBF only bind one of the two factors, providing further evidence of the distinct roles for each binding complex [Horak et al., 2002]. Although neither MBP1 or SWI4 deletions are lethal by themselves, double deletions are [Koch et al., 1993]. It is still unclear exactly what the distinctions are between the roles of MBF and SBF, and how mutant strains retain viability. Experiments directly surveying deletion strains for expression and binding patterns may help further our understanding.

As shown in figure 5.2, 95% of genes in EM2 are either bound by ($p \leq .05$), or contain the binding site for SBF or MBF within 1 KB upstream in *Saccharomyces cerevisiae*. Of the genes that contain at least a single instance of either MCB or SCB, 60% are also bound ($p \leq .05$) by at least one component of MBF or SBF (i.e. Swi4, Swi6, or Mbp1). This leaves 40% of genes with late G1 expression kinetics that contain at least one identified MCB or SCB binding site, but no significant observed binding to either MBF or SBF at a $p \leq .05$. These numbers may represent poor descriptions of the binding sites, limitations of the binding data, or the possibility that other transcriptional and/or post-transcriptional mechanisms are operating on these genes. Yet, even considering the potential shortcomings of the data, there is a large degree of agreement with expectations.

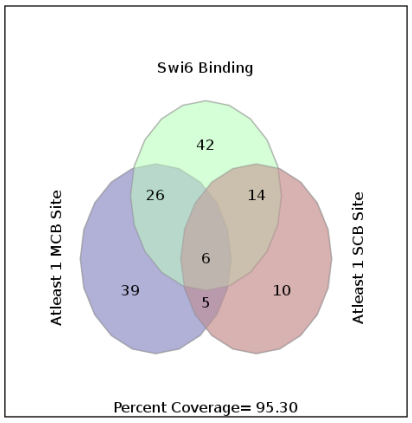
Given that we find a dramatic enrichment of MCB sites in the EM2 G1 group, relative to other cell cycle clusters and to all other genes in the genome, (figures 4.7 and 4.8) it is



(a) SBF with SCB



(b) MBF with MCB

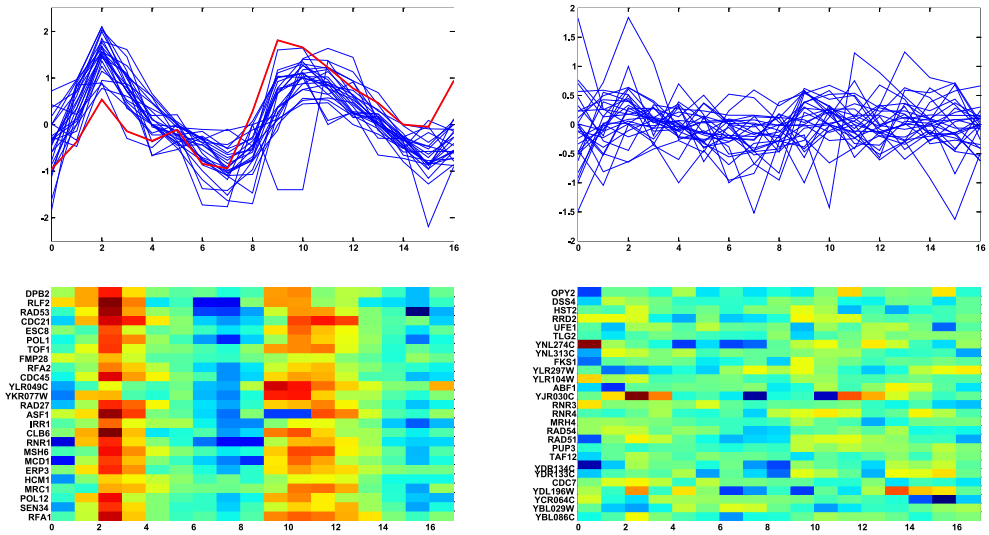


(c) Swi6 and SCB or MCB

Figure 5.2: **Observed Binding Site and in vivo binding Overlaps for 149 late G1 genes**. a and b) We compare the number of genes that have either individual or combined *in vivo* binding of SBF (Swi4 and Swi6) or MBF (Mbp1 and Swi6) with the presence of at least one copy of the consensus binding site within 1 kb upstream. c) Compares the number of genes that show binding of either MBF or SBF (by way of binding of Swi6 the common subunit of the two heterodimeric complexes) and presence of either MCB or SCB sites. In each case the percent coverage represents the total number of genes in EM2 that are represented in the union of all three sets.

possible that even a single site has biological significance. This is particularly interesting because one report showed that a single MCB site placed upstream of a gene failed to drive a reporter construct. However, two sites were able to drive the same reporter construct very well [Lowndes et al., 1991]. On the other hand, mutational analysis of a prominent MBF target, DNA polymerase I, showed that a single MBF site is necessary and sufficient for its cyclic expression [Gordon and Campbell, 1991]. The observation that two or more Mbp1 sites adjacent to a gene is a powerful predictor of strong G1 specific expression was supported by interrogating our cell cycle network. We find 24 out of the 25 cycling genes that have two or more MCB sites are in EM2. The one remaining gene that is not found in EM2 is found in EM1 and has an expression pattern that could be argued to be G1-like (figure 5.3a). In contrast, most of the genes that were not classified as being in the cycling set, but contain two or more MCB sites within 1KB upstream lack G1 expression (figure 5.3b). However, upon closer inspection, nearly all genes that have two MCB sites upstream, and exhibit no appreciable G1 RNA expression pattern, share intergenic space with a G1 gene. About 50% have obvious high amplitude G1 patterns, and the other 50% show much diminished, but still observable G1 patterns (figure 5.4). Interestingly, given that the canonical consensus is palindromic, this analysis shows there is a remarkable specificity for MCB sites to influence the transcription of only one of the two genes that share upstream regulatory sequence. This is even true when the two genes are very close to each other.

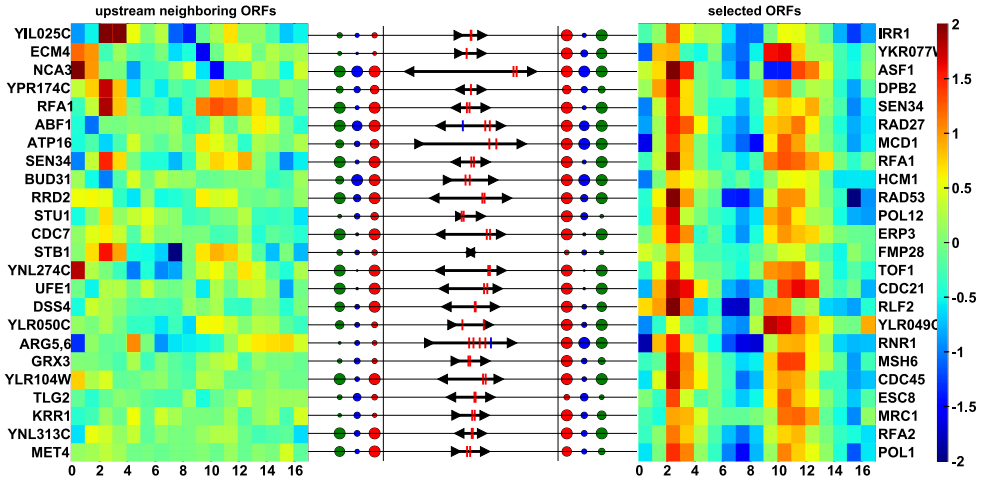
Does a single hexameric MCB site have functional consequences? Of the cycling genes that have exactly one MCB site, 52 are in EM2. Of the remaining, 14 are in EM1, 9 are in EM3, 2 are in EM4, but no genes in EM5 are found to have a single MCB site. Thus a primary conclusion is that there is strong evidence that a single MCB site can be functionally important. This is particularly striking in the EM2 G1 group of genes, where the enrichment of MCB containing genes is highly significant. In addition many of the MCB site occurrences in the EM2 genes are also evolutionarily conserved (figures 4.7 and 4.8). This does not, however, prove that the MCB site is acting alone, and future studies of how MBF possibly used in combination with the other cell cycle regulators will address that issue. The complete absence of MCB sites from EM5, which precedes G1 kinetically, raises the question of whether they are devoid of these sites, because expression of many



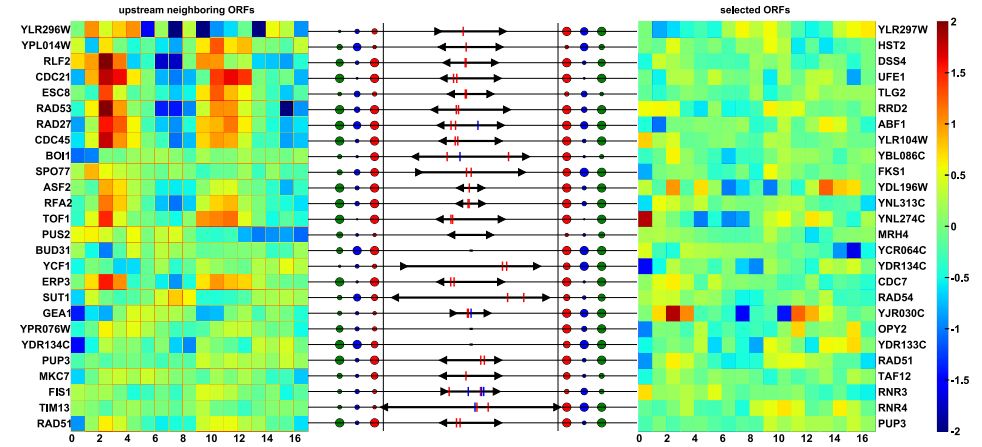
(a) Cycling with $\geq$ 2 MCB Sites

(b) Non-cycling with $\geq$ 2 MCB Sites

Figure 5.3: Gene Expression Trajectories for genes with 2 MCB Sites. As in previous figures trajectory summaries are shown with time along the x-axis and expression level on the y-axis. a) Shown are the 25 trajectories for all genes in the cycling set as defined by Cho et al., 1998. Highlighted in red is the one gene that was not classified by EM into the late G1 group (EM2). b) Trajectory summaries for the 27 genes in the yeast genome that are not in the cycling set, but do contain two or more MCB sites. In each case, the bottom panel is a heatmap representation of the same genes.



(a) Cycling with $\geq$ 2 MCB Sites



(b) Non-cycling with $\geq$ 2 MCB Sites

Figure 5.4: **MCB sites have directional specificity** In Panel (a) the selected ORFs are "cycling genes" that contain two MCB sites with in the upstream 1KB. In panel (b) the selected ORFs are all genes in the genome that have 2 MCB sites upstream but are not in the "cycling genes". In each case, these gene sets are the same as in figure 5.3. In both panel a and b the heat map displays on the right and left are drawn as in figure 5.3. The right panel shows the expression profiles for the selected ORFs. The left panel shows the expression profile for the ORF directly upstream of each ORF on the right panel. The center diagram represents a scaled version of shared intergenic sequence. The scale is set such that the two vertical lines span 1 KB. Red ticks are MCB sites, blue ticks are SCB sites, and the arrows at the end of each line indicate the direction of transcription. The circles adjacent to each line indicate observed *in vivo* binding activity for either Mbp1 (red), Swi4 (blue), or Swi6 (green) where the area is proportional to the measured -log(p-value) with a maximal size set at $p \leq 0.05$. As shown above, nearly all genes with two MCB sites that do not exhibit an G1 RNA expression pattern share intergenic space with a gene that does.

of these during G1 would be deleterious. It would be particularly interesting if insertion of MCB sites within M-phase genes had a more dramatic effect on cell cycle progression than a insertion elsewhere in the genome.

5.3 Putting Networks Together

Progression through cellular states is often manifested through processes that modulate regulatory connections. For example, Swi5 enters the nucleus after it becomes dephosphorylated by cdc14 in response to the degradation of Clb5 during the transition between M and G1. Thereby, Swi5 binds to target genes creating regulatory connections between it and its targets. These newly formed connections, in part, pushes the cell from M-phase into G1. In this example, Swi5 recruits the chromatin modifying complexes Swi/Snf and SAGA [Neely et al., 1999]. This is thought to facilitate interaction with other sequence specific factors such as Swi4. A specific gene where this has been shown is HO [Cosma et al., 1999, Cosma et al., 2001].

Each phase of the cell cycle could be considered a unique state, discernible by a distinct gene expression signature [Cho et al., 1998, Spellman et al., 1998]. If you consider the regulatory network of the cell to be dynamic, where connections are created and destroyed as they become active or inactive, then there would also be a unique transcriptional regulatory network underlying each of the the cell cycle phases. Understanding the relationships between kinetically adjacent states and the changes in regulatory connections between them should lead to testable hypotheses regarding mechanisms that might underlie the transitions. For instance, in chapter 4 we identified Sut1 to be strongly associated with EM1 (early G1). Figure 5.1 shows that Sut1 is expressed during M-phase and many of the cell cycle regulators bind upstream to it. Also, Sut1 binds to many of the early G1/Late M-phase regulators. As such, Sut1 is likely to play a role in the regulation of G1 genes.

The artificial neural networks (ANNs) constructed in chapter 4 are an example of a state-centric modeling approach. The weights in the network were inferred based on a classification of genes that focused on grouping genes into cell cycle phases. These groups, or clusters, represent the five most prominent expression patterns in the data. The synchro-

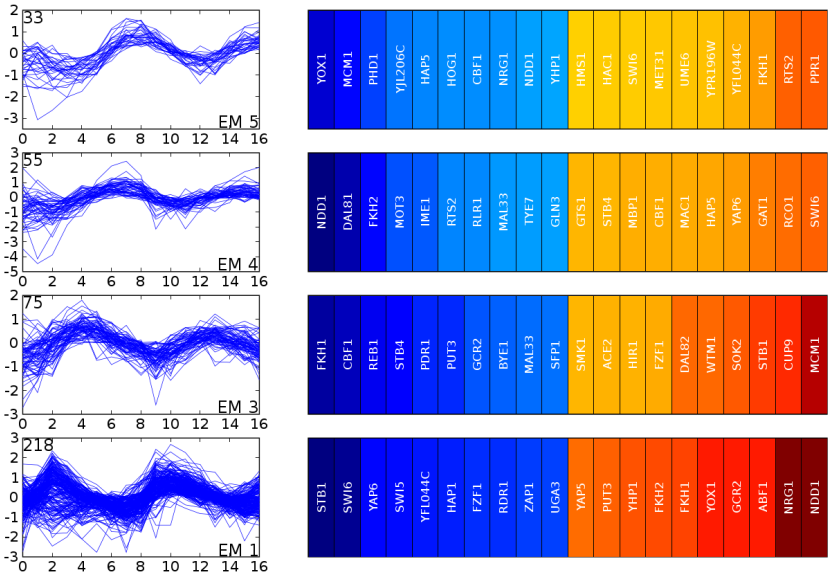


Figure 5.5: **ANN weights sorted on an expression class basis after merging EM1 and EM2**. Shown are the ANN weights from the average-of-best network as in figure 4.6. EM1 represents genes that only peak during the G1 phase of the second observed cell cycle, while EM2 represents genes that peak during both observed G1 phases. When merged, many of the known regulatory connections are changed. Notably Mbp1 and Ace2 are not recovered as important regulators of this merged G1 cluster

nization method of the Cho et al. (1998) microarray data was based on a temperature sensitive cdc28 mutant which arrests the cells at the edge of *Start*. As mentioned in chapter 2 we could have clustered the data by ignoring the effects of this synchronization method, but by using it we were able to resolve aspects of the regulatory differences between early G1 as found in EM1 and late G1 as found in EM2. Indeed, if we perform analogous neural network experiments but group all of EM1 genes and EM2 genes together many of the inferred regulatory connections change, for instance Mbp1 is no longer identified as a G1 regulator (Figure 5.5).

Although the regulators identified in the ANN included many of the known cell cycle regulators that are important in facilitating the transition between cell cycle phases, Swi4 was a notable exception. This is interesting in terms of regulatory networks because Swi4 is known to be important in regulating G1 expression. On closer inspection, the binding

data available for Swi4 shows it has diffuse binding in genes throughout EM1, EM2 and EM3. 47% of the genes that have significant binding ($p \leq .05$) to Swi4 are not in EM2. Thus the networks found little predictive power in using Swi4 in classifying genes into a single expression group. Likely this is an indication that Swi4's role during the cell cycle spans multiple phases. Figure 5.1 lends support to this as Swi4 binds to Yox1 and Yhp1 which likely aids in the shutting down of M-phase genes during early G1. Swi4 also binds to Ndd1 which according to our analysis in chapter 4 plays a central role in regulating early M-phase.

The conserved presence of transcription factor binding sites within groups of co-expressed genes reveal part of the regulatory network that has been maintained through evolution. As discussed in chapter 4 the MCB binding site is dramatically enriched in G1 genes (Figure 4.7). Further the enrichment patterns were conserved across most of the sequenced Saccharomyces and even persisted in S. pombe (Figure 4.10). SCB, the binding site for SBF, does not show nearly the same level of conserved enrichment. This parallels observations from the binding data as well (Figure 4.11), and furthers the argument that Swi4 functions less in the regulation of G1 specific genes, but instead functions throughout the G1 and S transition.

In part I of this thesis we focused on developing and testing methods and tools for analyzing large-scale gene expression data. This allowed us to evaluate the reproducibility of commonly used clustering algorithms in the presence of different types of noise. We also showed an example of manually mining expression data, protein:DNA binding data, and sequence data with the assistance of comparative clustering techniques to more finely define the G1 gene expression cluster. In part II, using the tools and techniques introduced in Part I, we constructed a new, more comprehensive, transcriptional connectivity map for the yeast cell cycle based primarily on genome-scale data. This map included nearly all previously known regulators of the cell cycle and also several new candidate regulators. These connections inferred using artificial neural networks are based on the idea that changes in cellular behavior are, in part, the result of connectivity changes in the underlying regulatory networks.

Appendix A

Average-of-Bests Regulators

Gene Descriptions for the top ten positively and negatively associated regulators for each cluster as determined by the ANN weights matrix figure 4.6. (Source http://www.yeastgenome.org)

Cluster	±	Regulator	Description
EM 1	+	SUT1	Involved in sterol uptake
	+	SWI5	transcriptional activator
	+	HAP5	Regulates respiratory functions; subunit of a heterotrimeric complex required for CCAAT binding
	+	ACE2	involved in transcriptional regulation of CUP1. enters nucleus only at the end of mitosis.
	+	RTS2	similar to mouse KIN7 protein
	+	DAL80	Negative regulator of multiple nitrogen catabolic genes
	+	TEC1	transcription factor of the TEA/ATTS DNA-binding domain family, regulator of Ty1 expression
	+	AZF1	probable transcription factor, suppressor of mutation in the nuclear gene for the core subunit of mitochondrial RNA polymerase
	+	YFL044C	None
	+	MOT3	DNA-binding protein implicated in heme-dependent repression, repression of a subset of hypoxic genes by Rox1p, repression of several DAN/TIR genes during aerobic growth, and regulation of membrane-related genes
	-	NDD1	Nuclear Division Defective 1
	-	YJL206C	None
	-	HAP2	Global regulator of respiratory genes
	-	STB4	binds Sin3p in two-hybrid assay
	-	GLN3	Responsible for nitrogen catabolite repression (NCR)-sensitive transcription. During nitrogen starvation, Gln3 is nuclear. Under excess nitrogen, Gln3 is cytoplasmic. Also regulates glutamine-repressible gene products.
	-	YAP3	bZIP protein; transcription factor
	-	WAR1	
	-	SFL1	Transcription factor with domains homologous to myc oncoprotein and yeast Hsf1p required for normal cell surface assembly and flocculence
	-	CAD1	Transcriptional activator involved in resistance to 1,10-phenanthroline; member of yeast Jun-family of transcription factors related to mammalian c-jun
	-	PHO2	Regulation of phosphate metabolism

Cluster	±	Regulator	Description
EM 2	+	SWI6	Involved in cell cycle dependent gene expression
	+	MBP1	transcription factor
	+	STB1	binds Sin3p in two-hybrid assay and is present in a large protein complex with Sin3p and Stb2p
	+	SFL1	Transcription factor with domains homologous to myc oncoprotein and yeast Hsf1p required for normal cell surface assembly and flocculence
	+	WTM1	WD repeat containing transcriptional modulator 1
	+	LEU3	Regulates genes involved in branched chain amino acid biosynthesis and in ammonia assimilation. Positively regulated by alpha-isopropylmalate, an intermediate in leucine biosynthesis.
	+	GAT1	activator of transcription of nitrogen-regulated genes; inactivated by increases in intracellular glutamate levels
	+	YPR196W	None
	+	HAP3	Regulates respiratory functions; encodes divergent overlapping transcripts
	+	NDT80	Meiosis-specific gene; mRNA is sporulation specific; required for exit from pachytene and for full meiotic recombination
	-	FKH2	Fork Head homolog two
	-	NRG1	involved in regulation of glucose repression
	-	PUT3	Positive regulator of PUT (proline utilization) genes
	-	USV1	None
	-	NDD1	Nuclear Division Defective 1
	-	YOX1	Homeodomain protein that binds leu-tRNA gene. acts as a repressor at early cell cycle boxes (ECBs) to restrict their activity to the M/G1 phase of the cell cycle.
	-	MIG3	
	-	UME6	Regulator of both repression and induction of early meiotic genes. Ume6p requires Ume4 for mitotic repression and interacts with and requires Ime1p and Rim11p for induction of meiosis-specific transcription
	-	SMP1	Second MEF2-like Protein 1 Transcription factor of the MADS (Mcm1p, Agamous, Deficiens, SRF) box family; closely related to RLM1
	-	ARO80	

Cluster	±	Regulator	Description
EM 3	+	FKH1	forkhead protein
	+	PUT3	Positive regulator of PUT (proline utilization) genes
	+	FKH2	Fork Head homolog two
	+	USV1	None
	+	ARR1	Similar to transcriptional regulatory elements YAP1 and cad1
	+	RLM1	serum response factor-like protein that may function downstream of MPK1 (SLT2) MAP-kinase pathway
	+	YKL222C	None
	+	WTM2	WD repeat containing transcriptional modulator 2
	+	BYE1	
	+	MAL33	Part of complex locus MAL3; nonfunctional in S288C, shows homology to both functional & nonfunctional MAL-activator proteins in other Sc strains & to other nonfunctional MAL-activator sequences from S288C (i.e. MAL33, YPR196W, & YFL052W)
	-	WTM1	WD repeat containing transcriptional modulator 1
	-	ACE2	involved in transcriptional regulation of CUP1. enters nucleus only at the end of mitosis.
	-	ARG81	Regulator of arginine-responsive genes with ARG80 and ARG82
	-	IFH1	Interacts with fork head protein. Protein controlling pre-rRNA processing machinery in conjunction with Fhl1p
	-	SMK1	SMK1 encodes a mitogen-activated protein kinase required for spore morphogenesis that is expressed as a middle sporulation-specific gene.
	-	RPI1	possesses a transcriptional activation domain and affects the mRNA levels of several cell wall metabolism genes.
	-	HAP4	Regulates respiratory functions; encodes divergent overlapping transcripts
	-	INO2	Transcription factor required for derepression of inositol-choline-regulated genes involved in phospholipid synthesis
	-	SFL1	Transcription factor with domains homologous to myc oncoprotein and yeast Hsf1p required for normal cell surface assembly and flocculence
	-	RAP1	DNA-binding protein involved in either activation or repression of transcription, depending on binding site context. Also binds telomere sequences and plays a role in telomeric position effect (silencing) and telomere structure.

Cluster	±	Regulator	Description
EM 4	+	NDD1	Nuclear Division Defective 1
	+	DAL81	Positive regulator of multiple nitrogen catabolic genes
	+	ACA1	contains an ATF/CREB-like bZIP domain; transcriptional activator
	+	PDC2	Regulates transcription of PDC1 and PDC5, which encode pyruvate decarboxylase
	+	FKH2	Fork Head homolog two
	+	IME4	IME4 appears to activate IME1 in response to cell-type and nutritional signals and thereby regulate meiosis
	+	MBF1	
	+	WAR1	
	+	INO4	Transcription factor required for derepression of inositol-choline-regulated genes involved in phospholipid synthesis
	+	UME6	Regulator of both repression and induction of early meiotic genes. Ume6p requires Ume4 for mitotic repression and interacts with and requires Ime1p and Rim11p for induction of meiosis-specific transcription
	-	SWI6	Involved in cell cycle dependent gene expression
	-	GAT1	activator of transcription of nitrogen-regulated genes; inactivated by increases in intracellular glutamate levels
	-	FAP7	
	-	MAC1	metal-binding transcriptional activator
	-	YAP6	bZIP protein
	-	HIR1	Involved in cell-cycle regulation of histone transcription
	-	HAP5	Regulates respiratory functions; subunit of a heterotrimeric complex required for CCAAT binding
	-	TEC1	transcription factor of the TEA/ATTS DNA-binding domain family, regulator of Ty1 expression
	-	SWI5	transcriptional activator
	-	MBP1	transcription factor

Cluster	±	Regulator	Description
EM 5	+	YOX1	Homeodomain protein that binds leu-tRNA gene. acts as a repressor at early cell cycle boxes (ECBs) to restrict their activity to the M/G1 phase of the cell cycle.
	+	MCM1	Involved in cell-type-specific transcription and pheromone response
	+	FAP7	
	+	CRZ1	calcineurin responsive zinc-finger
	+	NRG1	involved in regulation of glucose repression
	+	HAP5	Regulates respiratory functions; subunit of a heterotrimeric complex required for CCAAT binding
	+	PHO4	Transcription factor that activates expression of phosphate pathway
	+	YDR049W	None
	+	PHD1	protein similar to StuA of Aspergillus nidulans
	+	SPT23	Dosage dependent suppressor of Ty-induced promoter mutations. Homolog of Mga2. Spt23p and Mga2p differentially activate and regulate OLE1 transcription.
	-	HMS1	High-copy mep2 suppressor
	-	SWI6	Involved in cell cycle dependent gene expression
	-	HSF1	heat shock transcription factor
	-	LEU3	Regulates genes involved in branched chain amino acid biosynthesis and in ammonia assimilation. Positively regulated by alpha-isopropylmalate, an intermediate in leucine biosynthesis.
	-	STP2	Involved in pre-tRNA splicing and in uptake of branched-chain amino acids
	-	BAS1	Transcription factor regulating basal and induced activity of histidine and adenine biosynthesis genes
	-	MAL13	Part of complex locus MAL1; nonfunctional in S288C, shows homology to both functional & nonfunctional MAL-activator proteins in other Sc strains & to other nonfunctional MAL-activator sequences from S288C (i.e. MAL33, YPR196W, & YFL052W)
	-	HIR3	Involved in cell-cycle regulation of histone transcription
	-	UME6	Regulator of both repression and induction of early meiotic genes. Ume6p requires Ume4 for mitotic repression and interacts with and requires Ime1p and Rim11p for induction of meiosis-specific transcription
	-	PPR1	Positive regulator of URA1 and URA3

Table A.1:

Bibliography

[Amon et al., 1994] Amon, A., Irniger, S., and Nasmyth, K. (1994). Closing the cell cycle circle in yeast: G2 cyclin proteolysis initiated at mitosis persists until the activation of g1 cyclins in the next cycle. *Cell*, 77:1037–50.

[Andrews and Herskowitz, 1989] Andrews, B. and Herskowitz, I. (1989). The yeast swi4 protein contains a motif present in developmental regulators and is part of a complex involved in cell-cycle-dependent transcription. *Nature*, 342:830–3.

[Bar-Joseph et al., 2003] Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21:1337–42.

[Ben-Hur et al., 2002] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pac Symp Biocomput*, pages 6–17.

[Benos et al., 2002] Benos, P. V., Lapedes, A. S., and Stormo, G. D. (2002). Probabilistic code for dna recognition by proteins of the egr family. *J Mol Biol*, 323:701–27.

[Bishop, 1995] Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

[Bolouri and Davidson, 2002] Bolouri, H. and Davidson, E. H. (2002). Modeling dna sequence-based cis-regulatory gene networks. *Dev Biol*, 246(1):2–13.

[Breeden and Nasmyth, 1987] Breeden, L. and Nasmyth, K. (1987). Cell cycle control of the yeast ho gene: cis- and trans-acting regulators. *Cell*, 48:389–97.

[Breeden, 2000] Breeden, L. L. (2000). Cyclin transcription: Timing is everything. *Curr Biol*, 10:R586–8.

[Breeden, 2003] Breeden, L. L. (2003). Periodic transcription: A cycle within a cycle. *Curr Biol*, 13(1):R31–8.

[Cawley et al., 2004] Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116:499–509.

[Cho et al., 1998] Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2(1):65–73.

[Cliften et al., 2003] Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003). Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*.

[Cosma et al., 2001] Cosma, M. P., Panizza, S., and Nasmyth, K. (2001). Cdk1 triggers association of rna polymerase to cell cycle promoters only after recruitment of the mediator by sbf. *Mol Cell*, 7:1213–20.

[Cosma et al., 1999] Cosma, M. P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell*, 97:299–311.

[Costanzo et al., 2003] Costanzo, M., Schub, O., and Andrews, B. (2003). G1 transcription factors are differentially regulated in saccharomyces cerevisiae by the swi6-binding protein stb1. *Mol Cell Biol*, 23:5064–77.

[Csank et al., 2002] Csank, C., Costanzo, M. C., Hirschman, J., Hodges, P., Kranz, J. E., Mangan, M., O'Neill, K., Robertson, L. S., Skrzypek, M. S., Brooks, J., and Garrels, J. I. (2002). Three yeast proteome databases: Ypd, pombepd, and calpd (mycopathpd). *Methods Enzymol*, 350:347–73.

[Davidson, 2001] Davidson, E. H. (2001). *Genomic Regulatory Systems, Development and Evolution*. Academic Press.

[Davidson et al., 2003] Davidson, E. H., McClay, D. R., and Hood, L. (2003). Regulatory gene networks and the properties of the developmental process. *Proc Natl Acad Sci U S A*, 100:1475–80.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum lielihood from incomplete data via the em algorithm. *J. Royal Statistical Society, Series B*, 39:1–38.

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

[Dolinski et al., 2004] Dolinski, K., Balakrishnan, R., Christie, R., K., Costanzo, C., M., Dwight, S., S., Engel, R., S., Fisk, G., D., Hirschman, E., J., Hong, L., E., Nash, R., Oughtred, R., Theesfeld, L., C., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D., , Cherry, and M., J. (2004). Saccharomyces genome database. *http://www.yeastgenome.org/*.

[Doniger et al., 2003] Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome Biol*, 4:R7.

[Doolin et al., 2001] Doolin, M., Johnson, A., Johnston, L., and Butler, G. (2001). Overlapping and distinct roles of the duplicated yeast transcription factors ace2p and swi5p. *Mol Microbiol*, 40:422–32.

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8.

[Forbes, 1995] Forbes, A. (1995). Classification-algorithm evaluation - 5 performance-measures based on confusion matrices. *Journal Of Clinical Monitoring*, 11(3):189–206.

[Forsburg and Nurse, 1991] Forsburg, S. L. and Nurse, P. (1991). Cell cycle regulation in the yeasts saccharomyces cerevisiae and schizosaccharomyces pombe. *Annu Rev Cell Biol*, 7:227–56.

[Frenz et al., 2001] Frenz, L. M., Johnson, A. L., and Johnston, L. H. (2001). Rme1, which controls cln2 expression in saccharomyces cerevisiae, is a nuclear protein that is cell cycle regulated. *Mol Genet Genomics*, 266:374–84.

[Gabow, 1973] Gabow, H. (1973). PhD thesis, Stanford University Stanford, CA 94305.

[Galperin, 2004] Galperin, M. Y. (2004). The molecular biology database collection: 2004 update. *Nucleic Acids Res*, 32 Database issue:D3–22.

[Ghosh and Chinnaiyan, 2002] Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286.

[GOConsortium, 2001] GOConsortium, X. (2001). Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33.

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7.

[Gordon and Campbell, 1991] Gordon, C. B. and Campbell, J. L. (1991). A cell cycle-responsive transcriptional control element and a negative control element in the gene

encoding dna polymerase alpha in saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 88:6058–62.

[Green and Johnson, 2004] Green, S. R. and Johnson, A. D. (2004). Promoter-dependent roles for the srb10 cyclin-dependent kinase and the hda1 deacetylase in tup1-mediated repression in saccharomyces cerevisiae. *Mol Biol Cell*, 15:4191–202.

[Gusfield, 2002] Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*, 82(3):159–164.

[Harbison et al., 2004] Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104.

[Hateboer et al., 1998] Hateboer, G., Wobst, A., Petersen, B. O., Cam, L., Vigo, E., Sardet, C., and Helin, K. (1998). Cell cycle-regulated expression of mammalian cdc6 is dependent on e2f. *Mol Cell Biol*, 18:6679–97.

[Ho et al., 2002] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–3.

[Hollenhorst et al., 2000] Hollenhorst, P. C., Bose, M. E., Mielke, M. R., Muller, U., and Fox, C. A. (2000). Forkhead genes in transcriptional silencing, cell morphology and

the cell cycle. overlapping and distinct functions for fkh1 and fkh2 in saccharomyces cerevisiae. *Genetics*, 154:1533–48.

[Horak et al., 2002] Horak, C. E., Luscombe, N. M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M., and Snyder, M. (2002). Complex transcriptional circuitry at the g1/s transition in saccharomyces cerevisiae. *Genes Dev*, 16(23):3017–33.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(2-3):193–218.

[Hughes et al., 2000] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26.

[Ihmels et al., 2002] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–7.

[Iyer et al., 2001] Iyer, V., Horak, C., Scafe, C., Botstein, D., Snyder, M., and Brown, P. (2001). Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–8.

[Jones et al., 2004] Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B. B., Newport, G., Thorstenson, Y. R., Agabian, N., Magee, P. T., Davis, R. W., and Scherer, S. (2004). The diploid genome sequence of candida albicans. *Proc Natl Acad Sci U S A*, 101:7329–34.

[Kelley et al., 2004] Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B. R., and Ideker, T. (2004). Pathblast: a tool for alignment of protein interaction networks. *Nucleic Acids Res*, 32:W83–8.

[Kellis et al., 2003] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–54.

[Klug and Famulok, 1994] Klug, S. J. and Famulok, M. (1994). All you wanted to know about selex. *Mol Biol Rep*, 20:97–107.

[Knapp et al., 1996] Knapp, D., Bhoite, L., Stillman, D., and Nasmyth, K. (1996). The transcription factor swi5 regulates expression of the cyclin kinase inhibitor p40sic1. *Mol Cell Biol*, 16:5701–7.

[Koch et al., 1993] Koch, C., Moll, T., Neuberg, M., Ahorn, H., K, K. N., and a. s. m. y. t. h. K, N. (1993). A role for the transcription factors mbp1 and swi4 in progression from g1 to s phase. *Science*, 261:1551–7.

[Koranda et al., 2000] Koranda, M., Schleiffer, A., Endler, L., and Ammerer, G. (2000). Forkhead-like transcription factors recruit ndd1 to the chromatin of g2/m-specific promoters. *Nature*, 406:94–8.

[Kovacech et al., 1996] Kovacech, B., Nasmyth, K., and Schuster, T. (1996). Egt2 gene transcription is induced predominantly by swi5 in early g1. *Mol Cell Biol*, 16:3264–74.

[Lee et al., 2002] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

[Levine and Domany, 2001] Levine, E. and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593.

[Lichtenberg et al., 2004] Lichtenberg, U., Jensen, L. J., Fausboll, A., Jensen, T. S., Bork, P., and Brunak, S. (2004). Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics*.

[Lowndes et al., 1991] Lowndes, N. F., Johnson, A. L., and Johnston, L. H. (1991). Coordination of expression of dna synthesis genes in budding yeast by a cell-cycle regulated trans factor. *Nature*, 350:247–50.

[M. A. Beer, 2004] M. A. Beer, S. T. (2004). Predicting gene expression from sequence. *Cell*, 117:185–98.

[MacKay et al., 2001] MacKay, V. L., Mai, B., Waters, L., and Breeden, L. L. (2001). Early cell cycle box-mediated transcription of cln3 and swi4 contributes to the proper timing of the g(1)-to-s transition in budding yeast. *Mol Cell Biol*, 21:4140–8.

[Maclin et al., 1992] Maclin, R., Opitz, D., and Shavlik, J. W. (1992). University of wisconsin-madison backpropagation (uwbp).

[Mai et al., 2002] Mai, B., Miles, S., and Breeden, L. L. (2002). Characterization of the ecb binding complex responsible for the m/g(1)-specific transcription of cln3 and swi4. *Mol Cell Biol*, 22:430–41.

[Martone et al., 2003] Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M., Rinn, J. L., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S., and Snyder, M. (2003). Distribution of nf-kappab-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A*, 100:12247–52.

[Mendenhall and Hodge, 1998] Mendenhall, M. and Hodge, A. E. (1998). Regulation of cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, 62:1191–243.

[Miller et al., 2004] Miller, W., Makova, K. D., Nekrutenko, A., and Hardison, R. C. (2004). Comparative genomics. *Annu Rev Genomics Hum Genet*, 5:15–56.

[Mjolsness et al., 1991] Mjolsness, E., Sharp, D. H., and Reinitz, J. (1991). A connectionist model of development. *J Theor Biol*, 152(4):429–53.

[Nasmyth, 1985] Nasmyth, K. (1985). A repetitive dna sequence that confers cell-cycle start (cdc28)-dependent transcription of the ho gene in yeast. *Cell*, 42:225–35.

[Neely et al., 1999] Neely, K. E., Hassan, A. H., Wallberg, A. E., Steger, D. J., Cairns, B. R., Wright, A. P., and Workman, J. L. (1999). Activation domain-mediated targeting of the swi/snf complex to promoters stimulates transcription from nucleosome arrays. *Mol Cell*, 4:649–55.

[Novak et al., 2002] Novak, J. P., Sladek, R., and Hudson, T. J. (2002). Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, 79:104–13.

[Orlando, 2000] Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci*, 25:99–104.

[Pan and Heitman, 2002] Pan, X. and Heitman, J. (2002). Protein kinase a operates a molecular switch that governs yeast pseudohyphal differentiation. *Mol Cell Biol*, 22:3981–93.

[Pease et al., 1994] Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. P. (1994). Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proc Natl Acad Sci U S A*, 91:5022–6.

[Perou et al., 2000] Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn M, d. e. van, Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406:747–52.

[Peterson, 1954] Peterson, W. W. (1954). The theory of signal dectectability. *Ire Transactions on Information Theory*, (4):171–212.

[Pramila et al., 2002] Pramila, T., Miles, S., GuhaThakurta, D., Jemiolo, D., and Breeden, L. L. (2002). Conserved homeodomain proteins interact with mads box protein mcm1 to restrict ecb-dependent transcription to the m/g1 phase of the cell cycle. *Genes Dev*, 16:3034–45.

[Ptashne and Gann, 2002] Ptashne, M. and Gann, A. (2002). *Genes and Signals*. Cold Spring Harbor Laboratory Press.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for evaluation of clustering methods. *journal of the american statistical association*, 66(336):846–850.

[Regnacq et al., 2001] Regnacq, M., Alimardani, P., B, E. M., o. u. d. n. i. El, M., and Berges, T. (2001). Sut1p interaction with cyc8p(ssn6p) relieves hypoxic genes from cyc8p-tup1p repression in saccharomyces cerevisiae. *Mol Microbiol*, 40:1085–96.

[Ren et al., 2000] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of dna binding proteins. *Science*, 290:2306–9.

[Ross et al., 2000] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–35.

[S. Raychaudhuri, 2000] S. Raychaudhuri, J.M. Stuart, R. A. (2000). Principal components analysis to aummarize microarray experiments: Application to sporulationi time series. *Pacific Symposium of Biocomputing*.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray [see comments]. *Science*, 270(5235):467–70. Comment in: Science 1995 Oct 20;270(5235):368-9, 371.

[Schneider et al., 1998] Schneider, B. L., Patton, E. E., Lanker, S., Mendenhall, M. D., Wittenberg, C., Futcher, B., and Tyers, M. (1998). Yeast g1 cyclins are unstable in g1 phase. *Nature*, 395:86–9.

[Sherlock, 2000] Sherlock, G. (2000). Analysis of large-scale gene expression data. *Curr Opin Immunol*, 12(2):201–5.

[Song and Carlson, 1998] Song, W. and Carlson, M. (1998). Srb/mediator proteins interact functionally and physically with transcriptional repressor sfl1. *EMBO J*, 17:5757–65.

[Spellman et al., 1998] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97.

[Stuart et al., 2003] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–55.

[Swets, 1988] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–93.

[Tamayo et al., 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12.

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture [see comments]. *Nat Genet*, 22(3):281–5. Comment in: Nat Genet 1999 Jul;22(3):213-5.

[Tu et al., 2002] Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A*, 99:14031–6.

[Uetz et al., 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403:623–7.

[Visintin et al., 1998] Visintin, R., Craig, K., Hwang, E. S., Prinz, S., Tyers, M., and Amon, A. (1998). The phosphatase cdc14 triggers mitotic exit by reversal of cdk-dependent phosphorylation. *Mol Cell*, 2:709–18.

[Vohradsky, 2001] Vohradsky, J. (2001). Neural model of genetic network. *J Biol Chem*.

[Wang et al., 2002] Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., and Brown, P. O. (2002). Precision and functional specificity in mrna decay. *Proc Natl Acad Sci U S A*, 99(9):5860–5.

[Weaver et al., 1999] Weaver, D. C., Workman, C. T., and Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. *Pac Symp Biocomput*, pages 112–23.

[Wodicka et al., 1997] Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in saccharomyces cerevisiae. *Nat Biotechnol*, 15:1359–67.

[Wood et al., 2002] Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., Lelaure, V.,

Mottier, S., Galibert, F., Aves, S. J., Xiang, Z., Hunt, C., Moore, K., Hurst, S. M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V. A., Garzon, A., Thode, G., Daga, R. R., Cruzado, L., Jimenez, J., Sanchez, M., del Rey F, e. y. del, R., Benito, J., Dominguez, A., Revuelta, J. L., Moreno, S., Armstrong, J., Forsburg, S. L., Cerutti, L., Lowe, T., McCombie, W. R., Paulsen, I., Potashkin, J., Shpakovski, G. V., Ussery, D., Barrell, B. G., Nurse, P., and Cerrutti, L. (2002). The genome sequence of schizosaccharomyces pombe. *Nature*, 415:871–80.

[Yang et al., 2002] Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J., and Quackenbush, J. (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol*, 3:research0062.

[Zhu et al., 2000] Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N., and Futcher, B. (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406:90–4.