Communication between Layers in Biological Transcriptional Networks

by

Alex Tsankov

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

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2005

Author .....
Department of Electrical Engineering and Computer Science
May 19, 2005

Certified by .....
Moe Win
Associate Professor, Massachusetts Institute of Technology
Thesis Supervisor

Certified by .....
Pamela Silver
Professor, Harvard Medical School
Thesis Supervisor

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

Communication between Layers in Biological Transcriptional Networks

by

Alex Tsankov

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Chromatin-immunoprecipitation experiments in combination with microarrays (known as ChIP-chip) have recently allowed biologists to map where proteins bind in the yeast genome. The combinatorial binding of different proteins at or near a gene controls the transcription (copying) of a gene and the production of the functional RNA or protein that the gene encodes. Therefore, ChIP-chip data provides powerful insight on how genes and gene products (i.e., proteins, RNA) interact and regulate one another in the underlying network of the cell. Much of the current work in modeling yeast transcriptional networks focuses on the regulatory effect of a class of proteins known as transcription factors (TF). However, other sets of factors also influence transcription, including histone modifications and states (HS), histone modifiers (HM) and remodelers, nuclear processing (NP), and nuclear transport (NT) proteins. In order to gain a holistic understanding of the non-linear process of transcription, our work examines the communication between all five forementioned classes (or layers) of regulators. We use vastly available rich-media ChIP-chip data for various proteins within the five classes to model a multi-layered transcriptional network of the yeast species *Saccharomyces cerevisiae*. Following the introduction in Chapter 1, Chapter 2 describes the non-trivial process of incorporating the different sources of data into a coherent set and normalizing the heterogeneous data to improve biological accuracy. Using the normalized data, Chapter 3 finds biologically meaningful pairwise statistics between proteins, including filtered correlation coefficient, and mutual information p-values. It then combines the p-values of the two complementary approaches in order to increase the reliability of our predictions. Chapter 4 uncovers group-wise relationships between proteins using a novel semi-supervised clustering algorithm that preserves information about elements of a cluster in order to better capture group-wise dependencies. Throughout the theoretical analysis, we confirm various known biological processes and uncover several novel hypotheses. Based on the developed methodology, Chapter 5 builds a multi-layered transcriptional network and quantifies the communication between levels in biological transcriptional networks.

Thesis Supervisor: Moe Win
Title: Associate Professor, Massachusetts Institute of Technology

Thesis Supervisor: Pamela Silver
Title: Professor, Harvard Medical School

Acknowledgments

The fruition of this project would not have been possible without the cumulative contributions of many people. I first want to express my gratitude for my advisor, Moe Win. You are the most thoughtful professor I have ever me, and a boss I can truly call a friend. Thank you for your trust and loyal support during both happy times and tough moments at MIT. Thank you for being open-minded about my research direction and for your encouragements as I first started learning biology. I also want to acknowledge Ae Suwansantisuk and Hyundong Shin for their valuable contributions in proofreading the final document. In addition, I want to thank Professor Jaakola, Sourav Dey, and Obrad Scepanovic for their insights during Machine Learning class.

This project required the bridging of two disciplines and would not have even taken place without all the help form Pam Silver's lab at Harvard Medical School (HMS). Thank you Pam for bringing me into the family, granting me computing resources, and for taking a chance on someone who knew nearly nothing about biology nine months ago. Mike, thank you for your patience and experimental validations, without which the project would be incomplete. And finally, I want to especially thank Jason Casolari, who's creative insight inspired this undertaking. Thank you for always being honest with me, even during times when everything was foreign to me. Thank you for being my mentor in this new field, for explaining what YPD meant every Sunday afternooon, and for guiding me with your ingenious insight and biological intuition–I will never forget it.

In the end, I want to express my eternal gratitude towards my family. Thank you Dona, for reaching out to me as your younger brother and for understanding me and accepting me with all my flaws. Fetzi, thank you for becoming my best friend at a time when I had no other friends. And Sparti, thank you for your unconditional love and care from the day I was born. You are all a large part of who I am.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Intricate communication between genes regulate a complicated network within and between millions of cells that make our body function. The recent emergence of novel, data-rich experimental techniques that can simultaneously monitor the behavior of all genes within a cell has enabled computational biologists to study gene networks quantitatively. Understanding the process of transcription, or how genes are turned "on" and copied into more functional forms, is central to modeling gene regulatory networks.

1.1 Regulation of Transcription

Ordered sequences of deoxyribonucleic acid (DNA) base pairs, protected within the nucleus of each cell, encode the unique hereditary information of each individual. Genes are segments of DNA located on different chromosomes (large bundles of continuous DNA) that encode a unique product with a specific cellular function. During transcription, a gene is copied (or expressed) to a more mobile strand of information, known as RNA. For most genes, their RNA products are mere messengers (or mRNA) of the DNA code. The mRNA transports the DNA's information outside of the nucleus where it is translated into proteins, another even more functional string composed of amino acid molecules. The final product of some genes, however, is the RNA itself. Such strands execute specific tasks within the body in the

same manner as proteins. The repertoire of RNAs and proteins expressed from "on" genes in each cell allows for the myriad of functions necessary for cell survival and, on a larger scale, for the many different cell types found in complex organisms.

The transcription of genes is regulated by various RNAs and proteins, or products of other genes. Hence, genes regulate each other's activity through their respective products. Several classes of proteins or factors, which we refer to as layers or levels, collaborate to regulate the process of transcription. A set of proteins, known as transcription factors (TF), can enhance or suppress how actively a gene g works to produce its corresponding protein by binding to the DNA preceding g, or the promoter of g. Nuclear transport factors (NTs) represent another layer of proteins that affects transcription by controling the flux of molecules (such as TFs, mRNAs, etc) going in and out of the nucleus. Nuclear processing factors (NPs) also control transcription by executing functions within the nucleus, such as the actual copying of DNA to RNA or post-transcriptional processing of mRNAs. Chromatin, a macromolecular complex consisting of DNA and proteins that is found in eukaryotes (nucleus-containing organisms), also has profound affects on transcription. Chromatin is organized into packed, inaccessible and unpacked, accessible regions of DNA by structural proteins such as histones; therefore, histones have the potential to alter the expression state of genes. A fourth class of proteins that regulate transcription consists of histone modifiers (HMs), which change the accessibility of a gene's DNA by adding or removing acetyl, methyl, or other molecular groups on specific histone amino acids sites. We also include nucleosome remodelers in this category, or protein complexes that displace and reposition macromolecules of eight histones called nucleosomes. And finally, the actual histone states (HSs) in terms of the acetylation and methylation levels on specific histone amino acids, form yet another layer of control in transcription. For example, some proteins bind to specific patterns of histone acetylation levels in order to regulate the nearby DNA.

The protein classes described above each contribute to the expression of a gene in unique and situation-specific manners. Indeed, large research fields are devoted to the study of each of the described gene/protein classes and how particular stimuli are able to alter the

transcriptional output of a gene. In order to better understand the non-linear process of transcription, it is necessary to examine the genome-wide interplay of all of these transcriptional regulators both within particular fields, such as how modification of a histone at one amino acid affects modification at other amino acids, and between fields of study, such as the relationship between histone modification and recruitment of transcription factors to particular genes. The budding yeast *Saccharomyces cerevisiae* offers a unique opportunity to achieve this essential next step in understanding transcriptional control as it has a fully sequenced and annotated genome as well as an extraordinary body of literature regarding functions for many of these genes and their resulting RNAs or proteins. Figure 1 summarizes the different layers of organization known to regulate the transcriptional network in *Saccharomyces cerevisiae* which we will model. In order to bettter understand the non-linear process of transcription, we need information about where different classes of proteins bind to in the genome.

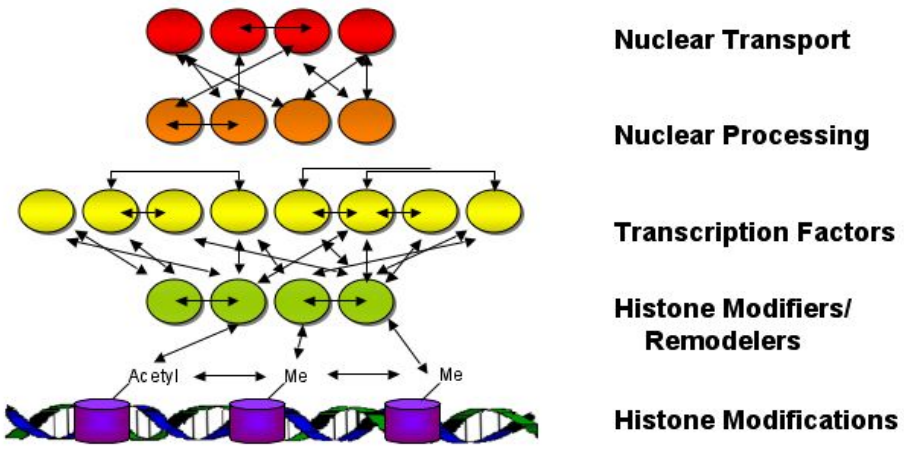


Figure 1-1: The five layers in the yeast transcriptional network that we aim to model. Two-sided arrows represent undirected functional connections between components within a level as well as interplay of components between layers of organization.

1.2 ChIP-chip

A biochemical technique known as chromatin immunoprecipitation (ChIP) allows for the chemical tethering of particular proteins to chromatin. The recent advent of ChIP in combination with microarrays (known as ChIP-chip or genomic location analysis) allows biologists to adapt this technique to a genome-wide scale, identifying the DNA binding sites for particular proteins across the entire genome [10]. Each ChIP-chip experiment begins by harvesting a yeast strain in a carefully prepared solution simulating a specific environmental condition. Most ChIP-chip experiments use the standard YPD (Yeast extract Peptone Dextrose) solution containing the sugar dextrose, in order to simulate rich media conditions where the yeast population grows exponentially with time. Next, addition of formaldehyde crosslinks or "fixes" any proteins bound to DNA inside the normal growing cell, or *in vivo*. The inner contents of the cell are then extracted and the strands of DNA along with the bound proteins are sheared into fragments by sonication. Using an anti-body that binds specifically to the protein i of interest, one can isolate all DNA fragments bound to the designated protein i by precipitation (called immunoprecipitation). The DNA-(protein i) crosslink is then reversed in order to release the DNA and digest the protein. The free DNA, previously bound by protein i, is then amplified by Polymerase Chain Reaction (PCR) and labeled with a fluorescent dye (Cy5). In parallel, a sample of sheared DNA is taken as a control and is not enriched for binding with protein i using immunoprecipitation. This sample is also amplified using PCR and labeled using a different fluorescent dye (Cy3). Ultimately, the ratio of Cy5 and Cy3 fluorescence intensities at each gene will allow one to infer protein i's predilection for binding to g.

Microarray technology when coupled with ChIP enables biologists to simultaneously measure the binding preference of protein i at all genes in the genome relative to other genes in just one experiment. One first needs to prepare a microarray matrix of wells, where each well g contains a probe specific to gene g (for example the complementary DNA of gene g) . Next, each microarray well is filled with a sample from both the pool of sheared DNA enriched (Cy5) and unenriched (Cy3) for protein i. Inside each well g, the differently dyed fragments

originally sheared from gene g will competitively bind to the complementary DNA specific to g during the process of hybridization. The array of wells can then be scanned using a laser that detects the intensities of the fluorescent red (Cy5) and green (Cy3) dye [39]. The ratio of intensities (Cy5/Cy3), also known as a binding ratio, is proportional to the ratio of the number of gene g fragments bound by protein i to the number of g fragments randomly resulting from the same process without immunoprecipitation. Each ChIP-chip experiment is often repeated three or more times in order to reduce the inherent experimental noise and the resulting binding ratios from each experiment are usually combined using weighted averaging.

ChIP-chip provides powerful, *in vivo* measurements of how genes and gene products (i.e. proteins, RNA) interact and regulate one another in the complex underlying network of a cell. Much of the current work in modeling yeast networks focuses on the regulatory effect of TFs [2]. However, recent publications [4,6,7,13–20] show that other sets of proteins (including HMs, HSs, NPs, and NTs) also control gene expression. Hence, transcription has several levels of organization. The next four chapters use ChIP-chip data to infer the underlying communication between different layers in the transcriptional network of the yeast species *Saccharomyces cerevisiae*.

Following the introduction in Chapter 1, four chapters build up the theory behind our transcriptional network model. Chapter 2 describes the preliminary data preparation on which we will base all future analysis. It first describes the integration of the different ChIP-chip data sets into one coherent set and then explores the crucial issue of binding data normalization and representation. Using a normalized data set, Chapter 3 finds biologically meaningful pairwise statistics between binding profiles of proteins, including filtered correlation coefficient, mutual information, and combined p-values. The next chapter uncovers group-wise binding relationships between proteins using Principal Component Analysis (PCA) and clustering. Based on the measures developed in Chapter 3, we introduce a novel semi-supervised clustering algorithm that preserves information about elements of a cluster in order to better capture group-wise dependencies between proteins. Throughout the

theoretical analysis, we confirm various known biological processes and predict several novel hypotheses. And finally, Chapter 5 combines the methodology developed in the previous chapters in order to build a multi-layered transcriptional network of the nucleus. To the best of our knowledge, our finalized network is the first attempt in the literature to quantify the communication between layers in biological transcriptional networks.

Chapter 2

Data Preparation

This chapter explores the extremely important issue of data preparation. Our data comes from several different sources and in various forms. The first section explains the nontrivial procedure for integrating all the different sources of data into one coherent set. The second part of the chapter attempts to reconcile the discrepancies between ChIP-chip experiments done in different labs by normalizing the data.

2.1 Data Integration

Integrating the publicly available data sets proved to be a tedious but non-trivial part of the project. We downloaded published genome-wide binding data for several sets of proteins that may be involved in regulation of genes, including TFs [1, 17, 21], NTs [13, 14], NPs [8, 14, 16, 23, 24, 28], HMs [4, 6, 7, 15, 20, 21, 25], and HSs [17, 18, 26, 27]. All of the ChIP-chip experiments were performed on yeast strains grown in rich media conditions, where the sugar dextrose and other nutrients are readily present so that the yeast can quickly multiply. Since some genes only become active during specific environmental conditions, we would ideally want ChIP-chip binding data for yeast grown in various media. However, from our experience and as evidenced in [7], binding relationships in biological processes that are not turned "on" during rich media conditions can often still be detected using our ChIP-chip

data. For example, our analysis in Section 4.2.4 captures the collaborative effect of GAL3 and GAL80, two transcription factors that regulate genes involved in breakdown of the sugar galactose, using ChIP-chip data for yeast grown in dextrose-rich, YPD medium. In addition, the authors in [7] found that the RSC nucleosome remodeling complex associates with many genes involved in nitrogen regulation and non-fermentative carbohydrate metabolism using YPD ChIP-chip data. In order to explain the relationship fully, more experiments under different conditions were necessary, but the initial prediction came from looking at rich media ChIP-chip data. Thus, we felt that usage of the vastly more complete YPD datasets would still allow for the formation of hypotheses reflecting other growth conditions.

The data sets further differed in the microarray probes used to detect binding relationships, where some experiments used probes that target the open reading frames (ORF), or regions of genes that code for RNA, while other experiments used probes for intergenic regions, or the promoter or control regions of genes. We used a gene-centric approach, where each relevant probe was assigned to its corresponding gene(s). For ORF arrays, we simply assigned the ChIP-chip information at each ORF to the corresponding gene.

For intergenic arrays, we assigned each DNA probe (or fragment) to the gene that it most likely regulates. Biologists commonly refer to the DNA region preceding the site where transcription starts as the upstream intergenic region of a gene, and the region following the site where transcription ends as the downstream intergenic region of a gene. In yeast, each intergenic region can control zero genes (e.g., telomeres at the end of chromosomes, intergenic regions at the downstream end of two genes), a single gene, or two genes (e.g., intergenic regions at the upstream end of two genes). Moreover, there may exist several probes for a single intergenic region. For example, small genes that encode tRNAs (a class of RNAs that have a role in protein production) are often contained within the intergenic region upstream of a longer gene; therefore, tRNA gene probes also measure the binding of factors that regulate the longer gene. To further complicate matters, for some genes it is still not known which intergenic region controls them. Hence, we needed to develop a many-to-many mapping from intergenic fragments to the genes they might control. The

mapping algorithm uses the union of intergenic probe-gene assignment pairs as defined by several authors [1, 6, 8, 9, 15, 19, 21, 23, 24, 27], where we included assignment pairs from at least one author from each different lab when it was available. Moreover, when two or more intergenic fragments mapped to the same gene, the probe that contains the most amount of information was chosen. Since ChIP-chip experiments contain more information at the tails of the binding distribution, we chose the most bound fragment for multiple probes that were consistently bound and the least bound fragment for multiple probes that were consistently not bound.

For each experiment, we obtained binding ratio (BR) data or the combination of BR and p-value (PV) data. We first mapped the binding ratios and p-values to all annotated genes for which data was available using the assignment algorithm described above. We then integrated the data into a single BR matrix and a single PV matrix, where rows represent factors, columns represent unique genes and entries represent the data. For example, the (i, g)th entry of the matrix PV, $PV_{i,g}$, represents the p-value of factor i binding to gene g. Missing data was annotated with NaNs.

As previously explained, a binding ratio for protein i and gene g represents a weighted average of ratios of the number of immunoprecipitated DNA fragments from g enriched with protein i to the number of control (unenriched) DNA fragments from g that occur at random. The p-value for the binding ratio of protein i at gene g measures the probability of erroneously deciding that protein i binds to g when the null hypothesis (i does not bind to g) is in fact true. Hence, small p-values correspond to large binding ratios. The error model used for calculating the p-values varies as chosen by the authors of each paper. Since we are integrating binding data from various papers that use different ChIP-chip experimental protocols and different error models, we needed to derive a normalized representation of the binding data in order to compare binding profiles across papers.

2.2 Data Normalization

To normalize the binding data, we first sorted the protein-gene binding interactions for each protein (i.e., row-wise in our matrices) from most bound to least bound, as defined by the authors of each paper. We then mapped the sorted binding data to uniformly spaced discrete points in the interval $[0, 1]$, where 1 and 0 correspond to the most and least bound genes for each protein, respectively. Hence we transformed the information in the BR and PV matrices to a percentile rank PR matrix of the same dimensions. For example, $PR_{i,g} = 0.95$ means that gene g is in the top 5% of genes most bound by protein i. Moreover, since each protein binds to a varying number of genes, each author usually defines a strength of binding threshold, above which all protein-gene interactions are classified as bound. We further normalized the PR matrix by subtracting the strength of binding threshold from each entry, making all interactions classified as bound positive and all interactions classified as unbound negative. We called this matrix the normalized N data matrix. By thresholding N at 0, we also easily derived a bound/unbound B data representation matrix, where all bound interactions were mapped to 1 and all unbound interactions to 0.

Each data representation has inherent advantages and disadvantages that may prove more biologically useful or less informative depending on the nature of the analysis. The bound/unbound B representation of the data seems most natural for representing the presence or absence of a biological interaction. However, ChIP-chip data contains significant experimental noise that varies from lab to lab and protein to protein; therefore, the classified bound/unbound data contains a large number of false positives and false negatives. The data set from [1] estimates the false positive rate to be around 4-6% and the false negative rate around 24%, for a binding threshold at p-value = 0.001. The percentile rank PR matrix removes discrepancies between the different averaging techniques used to derive binding ratios BR and the different error models used to calculate p-values PV. The PR matrix measures the pairwise binding strengths of gene-protein interactions consistently across data from different papers; however, it contains no information about which percentile determines significant binding for a given factor and the number of gene targets vary greatly for

different proteins. The normalized N matrix improves the PR matrix by subtracting the percentile threshold that each author used to classify interactions as bound and unbound. This achieves a soft, unclassified representation than the B matrix, where the more negative or positive an entry in the N matrix is, the more confident one can be that the gene-protein interaction is unbound or bound, respectively. While the PR and the N matrices measure binding strength relative to other interactions, they fail to capture the absolute strength of binding. To remedy this last shortcoming, we derive the SD matrix in the next section by finding the missing p-values in our data sets.

2.2.1 Finding Missing P-values

We consider the PV matrix as the most reliable source of information about making decisions on the presence or absence of a binding interaction. Further, nearly all of data sets with p-values base their calculation on adaptations of the single array error model [11]. Specifically, the data sets from [13, 14, 16] use a two-sided model, which can easily be converted to the equivalent right-sided single array model by the following conversion:

$$BR_{i,g} > 1 : PV_{i,g|1-\text{sided}} = \frac{1}{2} PV_{i,g|2-\text{sided}} \tag{2.1}$$

$$BR_{i,g} \leq 1 : PV_{i,g|1-\text{sided}} = 1 - \frac{1}{2} PV_{i,g|2-\text{sided}} \, . \tag{2.2}$$

Reconciling these discrepancies provides reliable, consistent p-values for two thirds of the proteins in the data. We need to find the missing p-values for the other third based on the available BR information in order to obtain a complete PV matrix. Figure 2-1 shows the binding disribution of protein Polymerase III (POL3) across the natural logarithm of its binding ratios. The figure reveals that the distribution of binding ratios consists of two heterogeneous classes: i) binding ratios measured at genes that are not bound by POL3 and ii) binding ratios at genes that are bound by POL3, congregated at the right tail of the distribution. We need to model the distribution of the unbound genes in order to find

appropriate p-values.

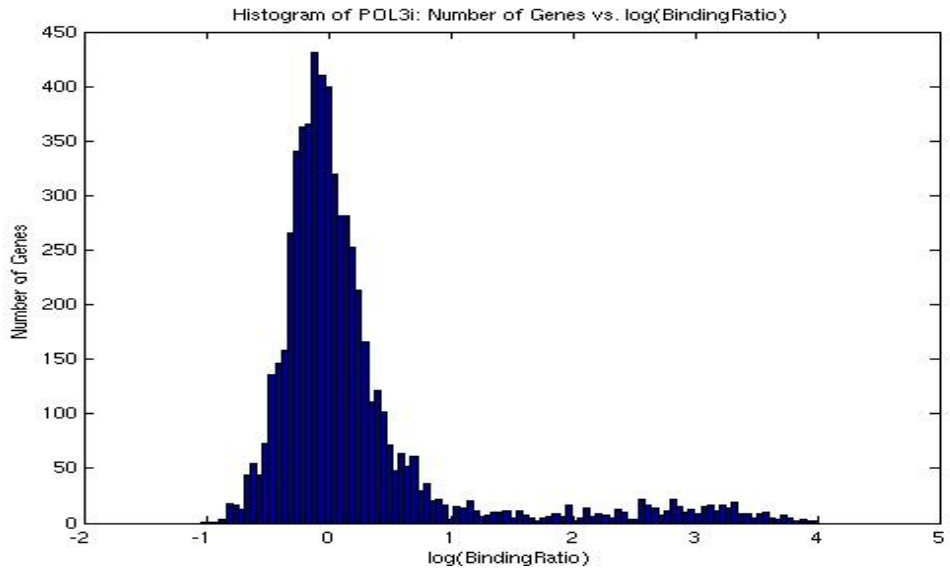


Figure 2-1: A histogram of the binding profile of protein Polymerase III: the number of genes bound by Pol III versus the natural logarithm of the binding ratios.

Mathematically, let $X_{i,g}$ denote random variables describing the natural logarithm of the binding ratios of protein i (POL3 in our example) at genes g and lets assume that the binding of i at each gene g, or $X_{i,g}$, is independent and identically distributed (i.i.d.). Since there is inherent noise in ChIP-chip experiments, it makes sense to use random variables to model our data. Following the convention of using capital letters for random variables and lowercase letters for their realizations, we use $x_{i,g}$ to denote the measured outcomes of random variable $X_{i,g}$ at each gene g. Since random variables $X_{i,g}$ are identically distributed at all genes g, let X_i denote the underlying binding tendency of protein i that we want to describe. Therefore, Figure 2-1 shows a representative, empirical example of the estimated distribution of X_i, or $\hat{P}_{X_i}(x_i)$. In order to extract p-values, we want to estimate the distribution of the natural logarithm of binding ratios at all genes g under the null hypothesis H_0 that protein i does not bind to g, or the noise distribution $\hat{P}_{X_i|H_0}(x_i|H_0)$. ChIP-chip experiments introduce various, independent sources of multiplicative noise, which corresponds to several independent sources of additive noise in the logarithm domain. Therefore, the Central Limit Theorem makes it

reasonable to assume that the distribution of the logarithm of binding ratios for the class of unbound genes is Gaussian [39]. Since logarithm of binding ratios in our data result from weighted averages of logarithm of binding ratios from single independent ChIP-chip experiments and since linear combinations of independent Gaussian random variables is also Gaussian, we can reasonably model $P_{X_i|H_0}(x_i|H_0)$ using a Gaussian distribution.

The Gaussian noise distribution $P_{X_i|H_0}(x_i|H_0)$ is uniquely defined by its mean and variance. We use the peak or mode of the overall distribution $\hat{P}_{X_i}(x_i)$ (e.g., the peak at $\log(BR) = 0$ in Figure 2-1) to estimate the mean of the noise distribution, $\hat{\mu}_{X_i|H_0}$. This follows from the fact that the alternative hypothesis that protein i binds to gene g, H_1, is is significantly less likely than the null hypothesis (on average, about 4% of gene targets are classified as bound, or $\Pr(H_1) \approx 0.04$ and $\Pr(H_0) \approx 0.96$). Since the overall distribution decomposes as follows,

$$\begin{aligned} P_{X_i}(x_i) &= \Pr(H_0)P_{X_i|H_0}(x_i|H_0) + \Pr(H_1)P_{X_i|H_1}(x_i|H_1) \\ &\approx 0.96P_{X_i|H_0}(x_i|H_0) + 0.04P_{X_i|H_1}(x_i|H_1)\,, \end{aligned} \tag{2.3}$$

the peak of the unbound distribution should roughly correspond to the observed mode of the overall distribution $\hat{P}_{X_i}(x_i)$. Moreover, since the noise distribution is Gaussian, the mostly unaffected peak of $P_{X_i|H_0}(x_i|H_0)$ corresponds to the mean $\hat{\mu}_{X_i|H_0}$. Hence, we can write

$$\hat{\mu}_{X_i|H_0} = \text{mode}_{g\in G_i}(x_{i,g})\,, \tag{2.4}$$

where G_i is the set of all genes with measured binding information for protein i. In order to estimate the variance of the noise distribution, $\hat{\sigma^2}_{X_i|H_0}$, we consider genes with a binding ratio smaller than $\hat{\mu}_{X_i|H_0}$ to be extremely unlikely targets of protein i. Hence, we use U_i to denote the set of genes to the left of the mode of $\hat{P}_{X_i}(x_i)$, or the genes the make up the left side of the unbound distribution (the distribution to the left of the peak at $\log(BR) = 0$ in Figure 2-1). Due to the symmetry of Gaussian distributions, we estimate the variance

of $P_{X_i|H_0}(x_i|H_0)$ only using the left side of the noise distribution. The maximum likelihood estimator for variance of Gaussian random variables states that

$$\hat{\sigma^2}_{X_i|H_0} = \frac{1}{|U_i|} \sum_{g \in U_i} (x_{i,g} - \hat{\mu}_{X_i|H_0})^2 \,, \tag{2.5}$$

where $|U_i|$ denotes the number of elements in set U_i, or the cardinality of U_i. The p-value $PV_{i,g}$ corresponds to the probability that just as extreme or more extreme of an observation could occur if we assume the null hypothesis H_0 that the factor i does not bind to gene g. To calculate the missing p-value for observation $x_{i,g}$, or $PV_{i,g}$, we integrate our estimated noise distribution over the interval $[x_{i,g}, \infty)$:

$$PV_{i,g} = \int_{x_{i,g}}^{\infty} \hat{P}_{X_i|H_0}(x_i|H_0) dx_i = \int_{x_{i,g}}^{\infty} \frac{1}{\sqrt{2\pi\hat{\sigma^2}_{X_i}}} \exp[\frac{-(x_i - \hat{\mu}_{X_i})^2}{2\hat{\sigma^2}_{X_i}}] dx_i \,. \tag{2.6}$$

After obtaining a complete PV matrix, we wanted to normalize the PV matrix so that we can assume a nearly Gaussian binding distribution for entries across rows for in the following analysis. Using the inverse of the Gaussian Q-function, we obtained the normalized SD matrix, where each entry $SD_{i,g}$ represents the number of standard deviations of confidence in rejecting the null hypothesis that protein i binds to gene g. Mathematically,

$$Q(z) = \int_{z}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[\frac{-w^2}{2}] dw \tag{2.7}$$

$$SD_{i,g} = Q^{-1}(PV_{i,g}) \,. \tag{2.8}$$

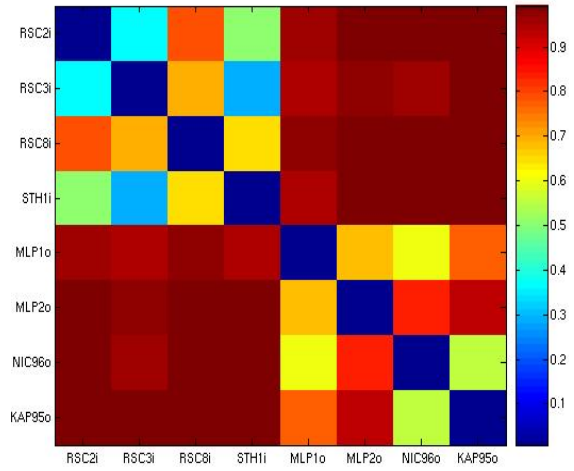
Since each row in the SD matrix approximates a Gaussian distribution with zero mean and unit variance, the SD matrix, just like the N matrix, fixes the problem in normalizing the data set so that binding levels can be compared across different experiments. Moreover, since the entries in the SD matrix represent standard deviation of confidence in rejecting the null hypothesis, the SD matrix preserves the absolute strength of binding of gene-protein interactions in a continuous manner, which the N matrix fails to capture. In the following

section we explore the advantages of the SD data matrix in relation to correlation analysis.

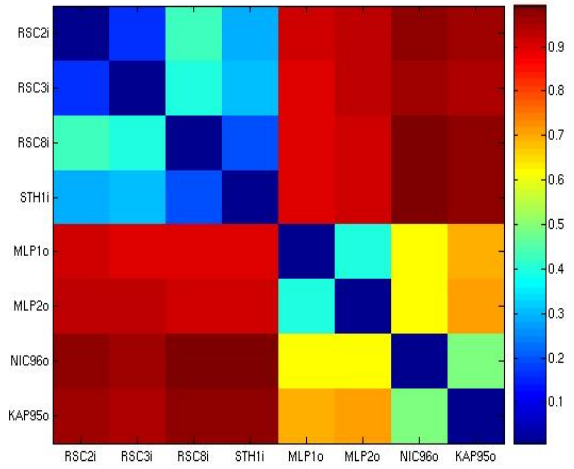
2.2.2 Evaluating Data Normalization: Correlation Analysis

We use the measure of correlation coefficient distance to gauge the performance of the various data representations in relation to correlation analysis. We define correlation coefficient distance as one minus the correlation coefficient; therefore, a distance of 0 represents full positive linear dependence and a distance of 1 denotes no linear relationship [**?**]. Figure 2-2 illustrates the concepts from the previous paragraph in Section 2.2.1. It shows four correlation coefficient distance matrices using the BR, $N\&PR$, PV, and SD data representations of a test set of proteins. Note that since the N and PR rows are simply shifted versions of one another and since correlation analysis is invariant to scales and shifts, the two matrices produce an identical correlation coefficient distance matrix, which we denote as resulting from $N\&PR$. Two known biological processes are shown: RSC2, RSC3, RSC8, and STH1 are all components of the RSC nucleosome remodeling complex [7] and MLP1, MLP2, NIC96, and KAP95 all play role in nuclear transport [13]. We would expect the members of each biological processes to share similar binding profiles. This is shown by small correlation distance values in the two squares along the diagonal. The BR matrix clearly performs worse than the PV matrix in confirming the known relationships, corroborating the fact that enrichment ratios contain less reliable information than p-values. The N and PR matrices are ranked version of the PV matrix, so the three, naturally, share very similar results. Since correlation coefficient analysis assumes Gaussian binding profiles, the SD matrix improves on the PV matrix even though (2.8) shows that the two are mere transformations of each other. Moreover, the SD matrix outperforms the N matrix not only because its row vector entries approximate a normal distribution, but also because it preserves information about the absolute strength of binding.

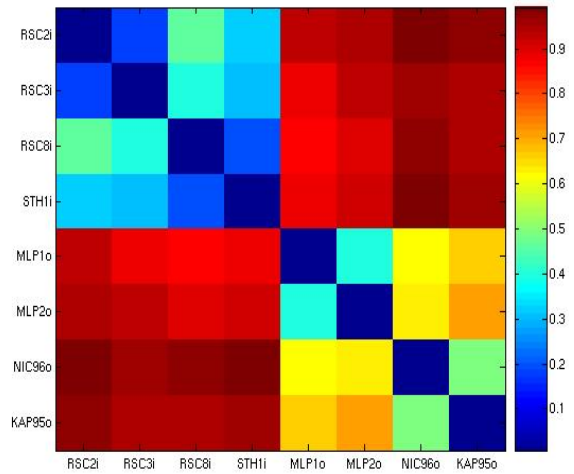
The performance of the different data representation in correlation analysis also extends to the whole data. We generated a test set of 1447 probable interactions, most of which were confirmed or suggested by published literature. Therefore, the known relationships discussed



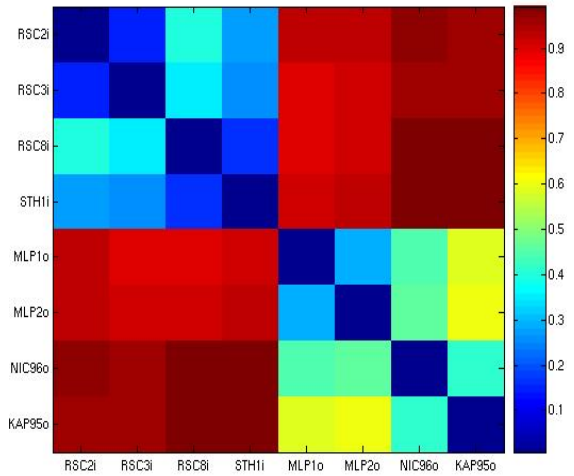
(a) BR correlation coefficient distance.



(b) $N\&PR$ correlation coefficient distance.



(c) PV correlation coefficient distance.



(d) SD correlation coefficient distance.

Figure 2-2: Correlation coefficient distance matrix using the (a) BR, (b) $N\&PR$, (c) PV, and (d) SD data representations of a test set of proteins. We define correlation coefficient distance as one minus the correlation coefficient; therefore, a distance of 0 represents full positive linear dependence and a distance of 1 denotes no linear relationship [?]. Protein names are listed on the x and y axes with an "o" or i at the end of the names signifying that the data came from an ORF or intergenic array, respectively. Two known biological processes are shown, namely RSC2-RSC3-RSC8-STH1 [7] and MLP1-MLP2-NIC96-KAP95 [13]. The SD matrix performs the best in confirming the known relationships followed closely by the PV and $N\&PR$ matrices, while the BR matrix clearly performs the worst.

Matrix	Total Distance at Probable Interactions	Total Distance at Unlikely Interactions
BR	978.5	5496.9
$\log(BR)$	974.3	5415.7
$N\&PR$	945.8	5341.4
PV	966.3	5763.9
SD	879.7	5374.5

Table 2.1: Total correlation coefficient distance for a test set of likely and unlikely interactions. A smaller total distance at known interactions should correspond to a smaller false negative rate for a given data representation, while a large total distance at unlikely interactions should lead to a smaller false positive rate. The SD matrix performs best in normalizing the data since it has the lowest total distance at known interactions and a comparable total distance to all the other matrices at unlikely interactions. The PV, N, and PR matrices do slightly worse but still better than both the $\log(BR)$, and BR data representations.

in Figure 2-2 represent a subset of this test set. We also created a test set of 5926 extremely unlikely interactions. Table 2.1 shows the computed total correlation coefficient distance for a test set of likely and unlikely interactions. A smaller total distance at known interactions should correspond to a smaller false negative rate for a given data representation, while a large total distance at unlikely interactions should lead to a smaller false positive rate. The SD matrix again performs best in normalizing the data since it has the lowest total distance at known interactions and a comparable total distance to all the other matrices at unlikely interactions. The PV, N, and PR matrices do slightly worse but still better than both the $\log(BR)$, and BR data representations. Note that the $\log(BR)$ data representation has a near Gaussian distribution for entries across rows. The fact that it slightly outperforms the BR matrix shows that the Gaussian data distribution accounts for some but not all of the improvement in normalizing the data. These results substantiate our decision to use the SD matrix for the correlation analysis that follows. The following chapter utilizes the different data representations in building a statistical method for detecting pairwise relationships between proteins.

Chapter 3

Pairwise Statistics

With an integrated and normalized data set, this chapter tries to find pairwise binding relationships between two proteins. We introduce two biologically meaningful pairwise measures of binding dependencies: filtered correlation coefficient and mutual information. We also find consistent methods for evaluating the p-values of the two analyses. In the end, we combine the p-values from the two complementary pairwise approaches in order to reduce the number of false positive and false negative biological predictions and increase the reliability of our overall analysis.

3.1 Filtered Correlation Coefficient

This section introduces the pairwise measure of filtered correlation coefficient, an extension of the standard correlation analysis commonly encountered in biology. This technique extracts the binding relationships between two factors with great accuracy, by isolating the analysis on the pertinent dimensions in ChIP-chip data. To estimate the filtered correlation coefficient between two proteins i and j, we use the SD matrix (as justified in Section 2.2.2). As in Section 2.2.1, we consider binding of factors i and j at genes g, denoted as $X_{i,g}$ and $X_{j,g}$, as i.i.d. random variables with measured outcomes $x_{i,g}$ and $x_{j,g}$, respectively. Moreover, we again denote the underlying binding tendency of the two factors i and j as random variables

X_i and X_j. Since the rows of the SD data representation closely resemble samples from a Gaussian distribution, we assume that X_i and X_j are jointly Gaussian. For factors i and j, let G_i and G_j represent the set of all genes for which we have binding information, respectively. Moreover, we use F_i and F_j to denote the filtered sets of genes bound by proteins i and j, respectively, and $F_{i,j} = F_i \cup F_j$ to represent the overall filtered set of genes, or the set of genes classified as bound by at least one of the two factors. Using the SD data matrix, the following equations find the means, filtered variances and filtered covariance of the binding tendencies of two proteins using maximum-likelihood (ML) estimators for jointly Gaussian random variables as shown below:

$$\hat{\mu}_{X_i} = \frac{1}{|G_i|} \sum_{g \in G_i} x_{i,g} \tag{3.1}$$

$$\hat{\mu}_{X_j} = \frac{1}{|G_j|} \sum_{g \in G_j} x_{j,g} \tag{3.2}$$

$$\hat{\sigma^2}_{X_i} = \frac{1}{|F_{i,j}|} \sum_{g \in F_{i,j}} (x_{i,g} - \hat{\mu}_{X_i})^2 \tag{3.3}$$

$$\hat{\sigma^2}_{X_j} = \frac{1}{|F_{i,j}|} \sum_{g \in F_{i,j}} (x_{j,g} - \hat{\mu}_{X_j})^2 \tag{3.4}$$

$$\hat{\sigma}_{X_i,X_j} = \frac{1}{|F_{i,j}|} \sum_{g \in F_{i,j}} (x_{i,g} - \hat{\mu}_{X_i})(x_{j,g} - \hat{\mu}_{X_j}) \,. \tag{3.5}$$

Note that the estimates of the means, $\hat{\mu}_{X_i}$ and $\hat{\mu}_{X_j}$, consider all genes for which we have binding information, while the estimates of the variances, $\hat{\sigma^2}_{X_i}$ and $\hat{\sigma^2}_{X_j}$, and covariance, $\hat{\sigma}_{X_i,X_j}$, consider only the genes classified as bound by at least one of the two factors. We call these estimates the filtered variances and filtered covariance of binding profiles X_i and X_j. We again use the ML estimator for jointly Gaussian random variables to estimate the filtered correlation coefficient between binding profiles X_i and X_j, or $\hat{\rho}_{X_i,X_j}$, as follows:

$$\hat{\rho}_{X_i,X_j} = \frac{\hat{\sigma}_{X_i,X_j}}{\sqrt{\hat{\sigma^2}_{X_i}\hat{\sigma^2}_{X_j}}}\,. \tag{3.6}$$

The difference between the filtered correlation coefficient and the standard correlation coefficient is that the estimates of the filtered variances and covariance of binding profiles X_i and X_j consider just a filtered subset of genes. As we would expect, the filtered correlation coefficient still retains some of the important properties of the correlation coefficient. For example, $0 \leq |\hat{\rho}_{X_i,X_j}| \leq 1$, with $\hat{\rho}_{X_i,X_j} = 0$ if and only if two data vectors have no linear dependence (uncorrelated) and $\hat{\rho}_{X_i,X_j} = 1$ if and only if one data vector is a shifted and scaled version of the other. This directly follows from Schwartz's inequality, which states that

$$|\frac{1}{|F_{i,j}|}\sum_{g\in F_{i,j}}(x_{i,g}-\hat{\mu}_{X_i})(x_{j,g}-\hat{\mu}_{X_j})| \leq \sqrt{(\frac{1}{|F_{i,j}|}\sum_{g\in F_{i,j}}(x_{i,g}-\hat{\mu}_{X_i})^2)(\frac{1}{|F_{i,j}|}\sum_{g\in F_{i,j}}(x_{j,g}-\hat{\mu}_{X_j})^2)}\,, \tag{3.7}$$

or equivalently that $|\hat{\sigma}_{X_i,X_j}| \leq \sqrt{\hat{\sigma^2}_{X_i}\hat{\sigma^2}_{X_j}}$, where the equality holds true if and only if at every g, $x_{i,g} = \alpha x_{j,g} + \beta$ for some choice of constants α and β.

In addition, the filtered correlation coefficient is also shift and scale invariant. This is a particularly useful property for our analysis, since it normalizes for the fact that some binding profiles may vary more than others. To prove that this property still holds, let $X_i' = aX_i + b$ and $X_j' = cX_j + d$ represent linear transformations of random variables X_i and X_j with transformed observations $x_{i,g}' = ax_{i,g} + b$ and $x_{j,g}' = cx_{j,g} + d$, respectively. The estimates for the means, filtered variances, and filtered covariance change as follows:

$$\hat{\mu}_{X'_i} = \frac{1}{|G'_i|}\sum_{g\in G'_i} x'_{i,g} = \frac{1}{|G_i|}\sum_{g\in G_i}(ax_{i,g}+b) = \frac{a}{|G_i|}\sum_{g\in G_i}(x_{i,g}) + b = a\hat{\mu}_{X_i} + b \tag{3.8}$$

$$\hat{\mu}_{X'_j} = \frac{1}{|G'_j|}\sum_{g\in G'_j} x'_{j,g} = \frac{1}{|G_j|}\sum_{g\in G_j}(cx_{j,g}+d) = \frac{c}{|G_j|}\sum_{g\in G_j}(x_{j,g}) + d = c\hat{\mu}_{X_j} + d \tag{3.9}$$

$$\hat{\sigma^2}_{X'_i} = \frac{1}{|F'_{i,j}|}\sum_{g\in F'_{i,j}} (x'_{i,g} - \hat{\mu}_{X'_i})^2 = \frac{1}{|F_{i,j}|}\sum_{g\in F_{i,j}} (ax_{i,g}+b-a\hat{\mu}_{X_i}-b)^2 = a^2\hat{\sigma^2}_{X_i} \tag{3.10}$$

$$\hat{\sigma^2}_{X'_j} = \frac{1}{|F'_{i,j}|}\sum_{g\in F'_{i,j}} (x'_{j,g} - \hat{\mu}_{X'_j})^2 = \frac{1}{|F_{i,j}|}\sum_{g\in F_{i,j}} (cx_{j,g}+d-c\hat{\mu}_{X_j}-d)^2 = c^2\hat{\sigma^2}_{X_j} \tag{3.11}$$

$$\hat{\sigma}_{X'_i,X'_j} = \frac{1}{|F'_{i,j}|}\sum_{g\in F'_{i,j}} (x'_{i,g} - \hat{\mu}_{X'_i})(x'_{j,g} - \hat{\mu}_{X'_j}) = ac\hat{\sigma}_{X_i,X_j}\,. \tag{3.12}$$

Substituting the new estimates for filtered variances and covariance into 3.6 we see that the filtered correlation coefficient of the scaled and shifted versions of the data is the same as that of the original binding profiles X_i and X_j:

$$\hat{\rho}_{X'_i,X'_j} = \frac{\hat{\sigma}_{X'_i,X'_j}}{\sqrt{\hat{\sigma^2}_{X'_i}\hat{\sigma^2}_{X'_j}}} = \frac{ac\hat{\sigma}_{X_i,X_j}}{\sqrt{a^2\hat{\sigma^2}_{X_i}c^2\hat{\sigma^2}_{X_j}}} = \hat{\rho}_{X_i,X_j}\,. \tag{3.13}$$

Filtered correlation coefficient is a simple but powerful measure of binding relationships between two factors. It isolates the analysis on the pertinent dimensions (or genes) of the ChIP-chip data and predicts biological interaction between factors with great accuracy, as shown in the following sections and chapters. Moreover, it can compare vastly different data sets in a consistent manner. We explore the issue of significance in filtered correlation analysis next.

3.1.1 Filtered Correlation Coefficient P-values

Ultimately, we want to use our filtered correlation coefficient to make decisions on whether two factors are linearly related. Hence, under our null hypothesis H_0 the binding profiles of two factors are not linearly related (i.e., $\rho_{X_i,X_j} = 0$) and under our alternative hypothesis

H_1 they are linearly related (i.e., $\rho_{X_i,X_j} \neq 0$). For sample size n and filtered correlation coefficient $\hat{\rho}_{X_i,X_j}$, we want to evaluate the probability that we reject the null hypothesis when it is actually true, or the p-value. To evaluate the significance of our filtered correlation coefficient, we use the test statistic

$$T = \frac{\hat{\rho}_{X_i,X_j}\sqrt{n-2}}{1-\hat{\rho}^2_{X_i,X_j}}\,. \tag{3.14}$$

Assuming that binding profiles X_i and X_j are jointly Gaussian, [35] shows that T is a Student-T random variable of $n-2$ degrees of freedom. Further, T results from a generalized likelihood ratio test and hence defines the optimal decision boundary [35]. Let $t_{\hat{\rho},n-2}$ denote one positive outcome of the random variable T for a given $\hat{\rho}_{X_i,X_j} = \hat{\rho}$ and $n-2$. Also, let $\mathbf{x}_i = [x_{i,g_1} \ldots x_{i,g_{|G|}}]$ and $\mathbf{x}_j = [x_{j,g_1} \ldots x_{j,g_{|G|}}]$ represent the row vectors of binding data for proteins i and j across all genes g in the set G, following the convention of using boldface to represent vectors. Since, $t_{\hat{\rho},n-2}$ only depends on $n-2$ and the estimated value $\hat{\rho}$ found using $\mathbf{x}_i$ and $\mathbf{x}_j$, $t_{\hat{\rho},n-2}$ is completely determined by $\mathbf{x}_i$ and $\mathbf{x}_j$. We represent the filtered correlation coefficient p-value for a given $t_{\hat{\rho},n-2}$, or for a given $\mathbf{x}_i$ and $\mathbf{x}_j$, as $pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j)$. To find this quantity, we need to find the probability that test statistic t can have a value more extreme than $t_{\hat{\rho},n-2}$. This corresponds to integrating the distribution of T, $P_T(t)$, over the two disjoint intervals $[-\infty, -t_{\hat{\rho},n-2}] \cup [t_{\hat{\rho},n-2}, \infty]$ where T exceeds the outcome $t_{\hat{\rho},n-2}$. Due to the symmetry of the distribution of Student-T random variable T, we can write

$$pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j) = Pr(T \geq t_{\hat{\rho},n-2}) = 2\int_{t_{\hat{\rho},n-2}}^{\infty} P_T(t)\,dt\,. \tag{3.15}$$

Note that we implicitly assume that both highly positive and negative values of $\hat{\rho}$ are significant here. If we wanted to simply consider positive/negative correlation coefficients as significant, we would need to remove the factor of two and only integrate over the interval corresponding to the right/left tail of T's distribution.

The filtered correlation coefficient measures how consistently two binding profiles fluctuate above and below their respective means. It represents a normalized measure of the linear

relationship between two data vectors. However, two random entities can have no linear relationship (i.e., $\rho_{X_i,X_j} = 0$) but still dependent on each other in a non-linear fashion. We explore a more general notion of probabilistic dependence between random variables in the next section.

3.2 Mutual Information

Mutual information is a general measure of the probabilistic dependence between two random variables. As in the filtered correlation analysis, we again consider binding profiles X_i and X_j of proteins i and j as random variables, but now their outcomes only take on N_{X_i} and N_{X_j} discrete values. The entropy of binding profile X_i measures the amount of uncertainty in predicting the observations of X_i and forms the basis for calculating the mutual information. Given a probability mass distribution of binding profile X_i, $P_{X_i}(X_i = x_{i,r})$ for $r = 1, \ldots, N_{X_i}$, the entropy of X is define as

$$\begin{aligned} H(X_i) &= -\sum_{r=1}^{N_{X_i}} P_{X_i}(X_i = x_{i,r}) \log P_{X_i}(X_i = x_{i,r}) \\ &= -\sum_{r=1}^{N_{X_i}} \sum_{s=1}^{N_{X_j}} P_{X_i,X_j}(X_i = x_{i,r}, X_j = x_{j,s}) \log P_{X_i}(X_i = x_{i,r}) \,. \end{aligned} \tag{3.16}$$

Suppose now that we observe the binding profile X_j, which might be related to X_i. The conditional entropy of X_i given X_j measures the amount of uncertainty that X_i contains with prior knowledge of X_j and can be calculated as follows:

$$
\begin{aligned}
H(X_i|X_j) &= -\sum_{s=1}^{N_{X_j}} P_{X_j}(X_j = x_{j,s}) H(X_i|X_j = x_{j,s}) \\
&= -\sum_{s=1}^{N_{X_j}} P_{X_j}(X_j = x_{j,s}) \sum_{r=1}^{N_{X_i}} P_{X_i|X_j}(X_i = x_{i,r}|X_j = x_{j,s}) \log P_{X_i|X_j}(X_i = x_{i,r}|X_j = x_{j,s}) \\
&= -\sum_{r=1}^{N_{X_i}} \sum_{s=1}^{N_{X_j}} P_{X_i,X_j}(X_i = x_{i,r}, X_j = x_{j,s}) \log P_{X_i|X_j}(X_i = x_{i,r}|X_j = x_{j,s}) \,. \qquad (3.17)
\end{aligned}
$$

The amount that the uncertainty in X_i decreases with the observation of X_j; hence, the entropy reduction $H(X_i) - H(X_i|X_j)$ corresponds to the amount of information that X_j contains about X_i. Combining (3.16) and (3.17), the mutual information between random binding profiles X_i and X_j takes the form:

$$
\begin{aligned}
I(X_i; X_j) &= H(X_i) - H(X_i|X_j) = H(X_j) - H(X_j|X_i) \\
&= \sum_{r=1}^{N_{X_i}} \sum_{s=1}^{N_{X_j}} P_{X_i,X_j}(X_i = x_{i,r}, X_j = x_{j,s}) (\log P_{X_i|X_j}(X_i = x_{i,r}|X_j = x_{j,s}) - \log P_{X_i}(X_i = x_{i,r})) \\
&= \sum_{r=1}^{N_{X_i}} \sum_{s=1}^{N_{X_j}} P_{X_i,X_j}(X_i = x_{i,r}, X_j = x_{j,s}) \log \frac{P_{X_i,X_j}(X_i = x_{i,r}, X_j = x_{j,s})}{P_{X_i}(X_i = x_{i,r}) P_{X_j}(X_j = x_{j,s})} \,. \qquad (3.18)
\end{aligned}
$$

Note that the second equality in the first line shows that mutual information is symmetric, or $I(X_i; X_j) = I(X_j; X_i)$. This can easily be verified by observing that trading the places of all the X_i and X_j in (3.18) results in no change. Another property of the mutual information is that it is non-negative, or $I(X_i; X_j) \geq 0$.

In order to estimate the mutual information between the binding profiles X_i and X_j of proteins i and j, we need to estimate the marginal and joint probability mass functions of X_i and X_j based on the data. Using the PV and PR data representations (justification discussed in Section 3.2.2), we classify each data entry as a 1 or a 0, signifying the presence or absence of an interaction, respectively. Hence, we model our binding profiles as Bernoulli

random variables with i.i.d. random samples $X_{i,g}$ and $X_{j,g}$ and outcomes $x_{i,g}$ and $x_{j,g}$, where $x_{i,g}, x_{j,g} \in \{0,1\}$ at all genes g. Let $G_{i,j}$ denote the set of all genes with binding observations for both proteins i and j, and let $G_{i,j}$ have w gene members. We can use maximum likelihood estimators for the parameters of Bernoulli random variables to estimate the marginal and joint mass distributions using

$$\hat{P}(X_i = 1) = \frac{1}{w} \sum_{g \in \{g : x_{i,g} = 1\}} 1 = \frac{v}{w} \tag{3.19}$$

$$\hat{P}(X_i = 0) = 1 - \hat{P}(X_i = 1) = \frac{w - v}{w} \tag{3.20}$$

$$\hat{P}(X_j = 1) = \frac{1}{w} \sum_{g \in \{g : x_{j,g} = 1\}} 1 = \frac{u}{w} \tag{3.21}$$

$$\hat{P}(X_j = 0) = 1 - \hat{P}(X_j = 1) = \frac{w - u}{w} \tag{3.22}$$

$$\hat{P}(X_i = 1, X_j = 1) = \frac{1}{w} \sum_{g \in \{g : x_{i,g} = 1, x_{j,g} = 1\}} 1 = \frac{h}{w} \tag{3.23}$$

$$\hat{P}(X_i = 1, X_j = 0) = \frac{1}{w} \sum_{g \in \{g : x_{i,g} = 1, x_{j,g} = 0\}} 1 = \frac{v - h}{w} \tag{3.24}$$

$$\hat{P}(X_i = 0, X_j = 1) = \frac{1}{w} \sum_{g \in \{g : x_{i,g} = 0, x_{j,g} = 1\}} 1 = \frac{u - h}{w} \tag{3.25}$$

$$\hat{P}(X_i = 0, X_j = 0) = \frac{1}{w} \sum_{g \in \{g : x_{i,g} = 0, x_{j,g} = 0\}} 1 = \frac{w - v - u + h}{w}\,, \tag{3.26}$$

where h denotes the number of genes bound by both proteins, v the number of genes bound by protein with binding profile X_i and u the number of genes bound by protein with profile X_j. Setting $N_{X_i} = N_{X_j} = 2$ and substituting our estimated distributions in (3.18) gives us the estimated mutual information between binding profiles X_i and X_j, $\hat{I}(X_i; X_j)$.

Mutual information provides a more general framework for measuring the dependence between two random variables. Moreover, the following sections and chapters demonstrate that it is a very natural and biologically meaningful measure of protein dependence in ChIP-chip data. We ultimately want to use our estimate of $I(X_i; X_j)$ in order to make decisions

on whether two proteins with binding profiles X_i and X_j participate in the same biological process. In the next section, we introduce a method for finding p-values for the mutual information analysis.

3.2.1 Mutual Information P-values

P-values allow us to make unbiased decisions about biological dependence based on mutual information. In this scenario, under the null hypothesis H_0 the two factors have no binding dependence (i.e., $I(X_i; X_j) = 0$). The p-value measures the probability that an estimated mutual information of a given significance or greater can occur at random. Note that our estimate of the mutual information in (3.19)-(3.26) depends only on four parameters, namely h, u, v, and w. Hence, each $\hat{I}(X_i; X_j)$ maps to the Venn diagram shown in Figure 3-1, where X_i is the binding profile of the TATA-Box Protein (TBP) and X_j is the binding profile of POL3.

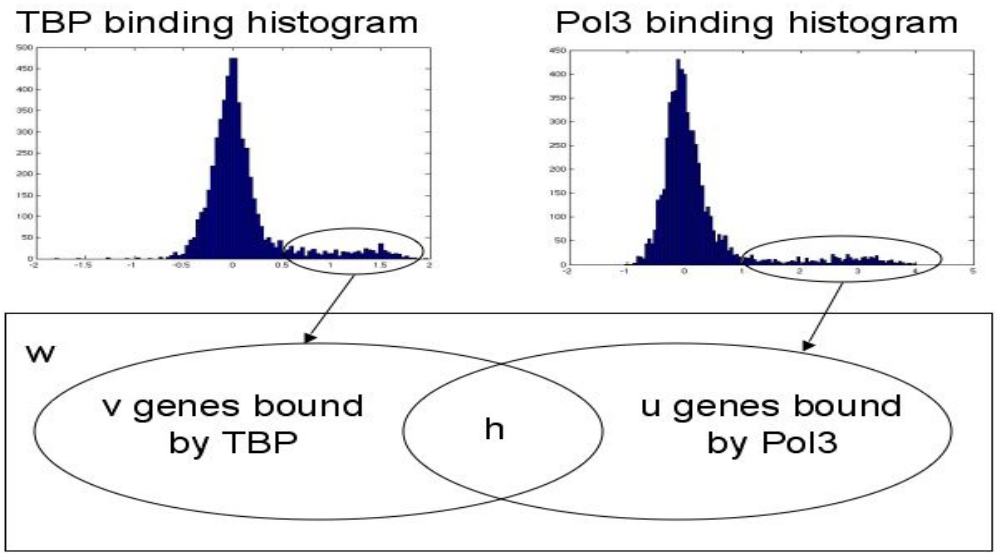


Figure 3-1: Venn diagram for the subsets of genes bound by proteins POL3 and TBP: w denotes the number of genes with observed binding information about both POL3 and TBP, v is the number of genes bound by TBP, u is the number of genes bound by POL3, and h is the number of genes bound by both. The classification of bound was chosen based on estimated p-values (see Section 2.2.1) below the threshold 0.001.

Given a superset of w genes and two subsets of u and v genes, the probability of the two subsets having an overlap of h elements at random has a hypergeometric distribution $P_{H|U,V,W}(h|u,v,w)$. Hence, we can calculate the probability of estimating a mutual information $\hat{I}(X_i;X_j) = \hat{i}_{h,u,v,w}$ at random using

$$\Pr(\hat{I}(X_i;X_j) = \hat{i}_{h,u,v,w}|H_0) = P_{H|U,V,W}(h|u,v,w) = \frac{\binom{w-v}{u-h}\binom{v}{h}}{\binom{w}{u}}\,. \tag{3.27}$$

The denominator in (3.27) represents the number of ways to choose $u-h$ non-overlaping elements from the complement of the set containing v objects, times the number of ways to choose the h overlapping elements from the set with v objects. The product hence represents the number of ways to choose a subset of u elements from a superset of w objects, such that exactly h of them overlap with a pre-designated subset of v elements. The numerator computes all the possible ways of choosing a subset of u elements from a set of size w, normalizing (3.27) in order to obtain a probability.

The mapping from mutual information estimate $\hat{I}(X_i;X_j) = \hat{i}_{h,u,v,w}$ to hypergeometric probabilities is not one to one. For example, the parameters ($h=1$, $u=2$, $v=2$, $w=20$) and ($h=300$, $u=600$, $v=600$, $w=6000$) have the same $\hat{i}_{h,u,v,w}$ but hypergeometric probabilities of 0.1895 and 2.3028×10^{-165}, respectively. This exaggerated example shows that the mutual information estimate does not take into account the sample size w in judging the significance of the dependence between two random variables, unlike the corresponding hypergeometric probability $P_{H|U,V,W}(h|u,v,w)$. Since a larger value of w should increase the confidence in our estimate $\hat{I}(X_i;X_j)$, random variable H when conditioned on $U=u$, $V=v$, and $W=w$ is a more appropriate test statistic for evaluating mutual information significance. As before, outcomes h, u, v, and w are completely determined by discrete vectors of binding data $\mathbf{x}_i$ and $\mathbf{x}_j$ for proteins i and j. To find the p-value, or the probability of randomly estimating a mutual information of a positive binding relationship equally or more significant than $\hat{i}_{h,u,v,w}$, we sum over the right tail of the hyper-geometric test statistic $H|U=u,V=v,W=w$ for the outcome h as follows:

$$pv_{MI}(\mathbf{x}_i, \mathbf{x}_j) = \Pr(H \geq h | u, v, w) = \sum_{r=h}^{\min(u,v)} \frac{\binom{w-v}{u-r}\binom{v}{r}}{\binom{w}{u}} . \tag{3.28}$$

As before, we express our mutual information p-value $pv_{MI}(\mathbf{x}_i, \mathbf{x}_j)$ as a function of the binding data $\mathbf{x}_i$ and $\mathbf{x}_j$. The hyper-geometric p-values above only evaluate the significance of synergistic binding between two factors, while mutual information can capture both positive and negative binding relationships. To evaluate the p-value of a negative binding relationship, or the probability of having an overlap of h or smaller at random, we would need to sum from 0 to h in (3.28). However, due to the high sensitivity at $h = 0$ and due to the large amount of false positives and false negatives in our binding data, this evaluation proves unreliable. Opposing binding relationships can evince interesting biological phenomena, as well, but it becomes clear later why such relationships complicate future analyses and their biological interpretation. Hence, we avoid using mutual information in finding negative binding relationships and will exclusively use the filtered correlation coefficients for that purpose. Now that the we have developed a framework for evaluating the significance of our mutual information analysis, we can substantiate our decision for using the PV and PR data representations in classifying the data and finding $\hat{I}(X_i; X_j)$.

3.2.2 Evaluating Data Classification: Mutual Information Analysis

This section gauges how accurately the different data representations classify their binding information into 0s and 1s, in order to estimate the mutual information and find p-values as in Sections 3.2 and 3.2.1. We generated a test set of 1447 probable protein-protein binding dependencies, most of which were confirmed or suggested by published literature. We also created a test set of 20707 extremely unlikely relationships. In order to compare the different matrices in an unbiased manner, we designated reasonable thresholds for classifying the data sets into a similar number of bound (1s) and unbound (0s) gene-protein interactions. Table 3.1 lists the data representations, their corresponding threshold test and the number of

Data Matrix	Classification Test for Binding	Total Entries Classified Bound	Hit Rate (captured /probable links)	Total Noise at Unlikely Links
BR	$BR \geq 1.9$	81450	738/1447	5283.9
PR	$PR \geq .965$	77377	1402/1447	3784.5
N	$N \geq -.002$	79639	1409/1447	1460.5
PV	$PV \leq .015$	80948	1445/1447	1021.1
$N + PR$	$N \geq -.002$ or $PR \geq .99$	80215	1414/1447	1394.7
$PV + PR$	$PV \leq .015$ or $PR \geq .99$	81461	1446/1447	1021.0

Table 3.1: Hit rate (capture/probable links) at probable interactions and total noise at unlikely interactions based on mutual information estimates using different data representations (see text).

entries classified as bound (all around 80,000) in the first three columns. The classified data for each data representation was then used to find mutual information p-values, pv_{MI}, as in Section 3.2.1. The next column of Table 3.1 enumerates the hit rate, or the number of relationships captured at a significance level of 10^{-10} divided by the total number of probable links in our test set. The last column lists the total noise at unlikely links, computed by accumulating the total significance at interactions in second test set. The significance of an interaction between proteins i and j was calculated by replacing pv_C with pv_{MI} in (4.4). A higher hit rate should correspond to fewer false negatives for a given data representation, while less total noise at unlikely relationships should lead to fewer false positives.

The PV (equivalent to thresholding on SD) matrix performs best in individually classifying the data, since it has the highest hit rate at known interactions and the lowest total noise at unlikely interactions. The N and PR matrices perform slighly worse in confirming probable relationships. However, the N and PR matrices introduce 43% and 271% more noise than the PV matrix, respectively, which will undoubtedly lead to more false positive claims. Since some proteins have very few gene targets at the chosen thresholds, the classified, binary data would provide little information about their behavior. In order to use mutual information to compare these proteins with the rest, the last two rows in Table 3.1 classify the data using a combination of the top two data representations (PV and N) and the PR data representation, guaranteeing that at least the top 1% of most immunoprecipitated gene probes is considered bound for each factor. The number of entries classified

as bound increase negligibly, since most publications already considered the strongest 1% of gene-protein interactions for each protein as bound. Creating a buffer of bound targets slightly improves the hit rate of both the individual N and PV matrix at no expense of added noise. These results substantiate our decision to use the combination of the PV and PR matrices for classifying the data in order to estimate the mutual information p-values. Now that we have a method for evaluating the significance of both our filtered correlation coefficient and mutual information analyses, we can combine the evidence from the two approaches in order to improve the reliability of our biological predictions.

3.3 Combining P-values

The p-value calculations for both the filtered correlation coefficient and mutual information estimation test similar null hypotheses. In Section 3.1.1, we test the null hypothesis that two binding profiles X_i and X_j are not linearly dependent ($\rho_{x,y} = 0$) while in the last section we considered the null hypothesis that the two profiles are not probabilistically dependent ($I(X_i; X_j) = 0$). Ultimately, we aim to find out whether two proteins belong in the same biological process. Hence we want to combine the p-values from the two analyses in order to incorporate both sources of evidence in testing the overall null hypothesis that two proteins with binding profiles X_i and X_j do not share a biological function.

We combine p-values using Fisher's method. It assumes that the p-values result from independent studies that use continuous random variables as test statistics to challenge similar null hypotheses. Since we use the same data source to estimate the filtered correlation coefficient and mutual information, our studies are not completely independent. Moreover, the hyper-geometric test statistic is not continuous. Despite these approximations, the authors in [37] and Section 3.3.1 demonstrate how combining p-values using Fisher's method improves biological prediction.

In order to derive Fisher's method for combining p-values for synergistic binding relationships, we only consider positive correlations as significant for the following. Hence, in this section we calculate our p-values for the filtered correlation coefficients using a one-sided test, or

using (3.15) without the factor of 2. Let $p_1 = \Pr(T \geq t_{\hat{\rho},n-2})$ and $p_2 = \Pr(H \geq h|u,v,w)$ represent two p-value observations that result from test statistics T and $H|U = u, V = v, W = w$ and observations $(\hat{\rho}, n-2)$ and h, respectively. Using the assumption of independent studies, the joint probability of observing events $T \geq t_{\hat{\rho},n-2}$ and $H \geq h|u,v,w$ under the null hypothesis equals

$$\Pr(T \geq t_{\hat{\rho},n-2}, H \geq h|u,v,w) = \Pr(T \geq t_{\hat{\rho},n-2}) \Pr(H \geq h|u,v,w) = p_1 p_2 \,. \tag{3.29}$$

From the equation above, it seems natural to consider the observation of p-value pairs $(p_1 = 0.01, p_2 = 0.01)$ and $(p_1 = 0.1, p_2 = 0.001)$ with equivalent products $p_1 p_2$ as evenly significant sources of evidence for rejecting the null hypothesis H_0. Intuitively, this refers to having two equally strong sources of evidence for dependence between two proteins or having a slightly stronger and a slightly weaker source of evidence. Hence, we want to make decisions about the strength of our combined evidence using the test statistic $K = P_1 P_2$, where P_1 and P_2 are the p-values resulting from the filtered correlation coefficient and mutual information estimations, respectively. Since P_1 and P_2 depend on assumed independent random variables T and H, they are themselves independent random variables. The distributions of p-values P_1 and P_2 resulting from a continuous test statistics is exactly uniform on the interval $[0,1]$ [37]. Although, our test statistic for mutual information is discrete, assuming a continuous uniform distribution for P_2 is just an approximation that only slightly affects the accuracy of our evaluation [37].

To find the overall p-value for a given p-value product $p_1 p_2$, we need to evaluate the probability of randomly arriving at a p-value product more significant than the observation $k_2 = p_1 p_2$. This corresponds to the region in probability space where $p_1 p_2 \leq k_2$. Hence, we need to find the volume of the region $p_1 p_2 \leq k_2$ over the joint distribution of P_1 and P_2. Since P_1 and P_2 were shown to be independent and uniformly distributed random variables, their joint distribution $P_{P_1,P_2}(p_1, p_2)$ is a 2-dimensional unit cube defined on $\{(p_1, p_2) : p_1 \in [0,1], p_2 \in [0,1]\}$. The combined p-value from the two analyses for observed p-value product

k_2, $pv_C(k_2)$, follows from the integration below:

$$\begin{aligned} pv_C(k_2) &= \iint_{p_1 p_2 \le k_2,\ 0 \le p_1, p_2 \le 1} P_{P_1,P_2}(p_1,p_2)dp_1 dp_2 \\ &= \int_0^{k_2} 1 dp_1 + \int_{k_2}^1 \frac{k_2}{p_1} dp_1 = p_1 \mid_0^{k_2} + k_2 \ln p_1 \mid_{k_2}^1 = k_2 - k_2 \ln k_2 \,. \end{aligned} \tag{3.30}$$

Since $p_1 = pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j)$ and $p_2 = pv_{MI}(\mathbf{x}_i, \mathbf{x}_j)$ for two proteins i and j with binding data vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the combined p-value for the binding relationship between proteins i and j only depends on $\mathbf{x}_i$ and $\mathbf{x}_j$ and can be equivalently denoted as $pv_C(\mathbf{x}_i, \mathbf{x}_j)$. The equation in (3.30) only depends on the number of dimensions, 2, and the product of our observed p-values, k_2. The extension for combining p-values from n independent experiments, also depends only on n and the observation $k_n = p_1 p_2 \cdots p_n$. Now, the integration takes place over the n dimensional region $p_1 p_2 \cdots p_n \le k_n$ of the joint distribution $P_{P_1,\ldots,P_n}(p_1,\ldots,p_n)$. Again from independence, the joint distribution is an n-dimensional unit hypercube defined on $\{(p_1,\ldots,p_n) : p_1 \in [0,1], \ldots, p_n \in [0,1]\}$ and the combined p-value is

$$pv_C(k_n) = k_n \sum_{r=0}^{n-1} \frac{(-\ln k_n)^r}{r!} \,. \tag{3.31}$$

The following section mitigates the problem of assuming independent analyses and illustrates how combining p-values can improve biological prediction.

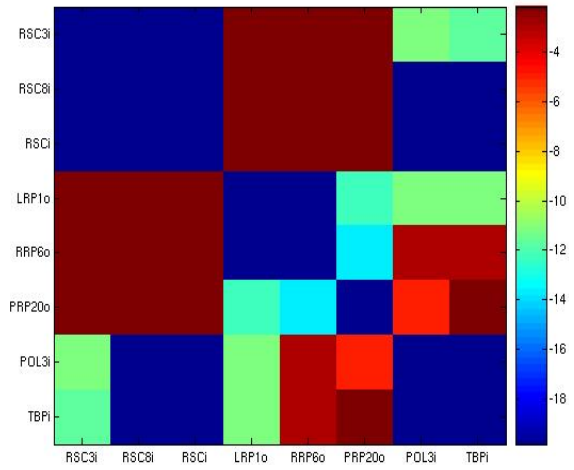
3.3.1 Biological Predictions using Combined P-values

To show the benefit of combining p-values we present an example of a biological prediction that is representative of the rest of the pairwise protein analysis. Figure 3-2 shows filtered correlation coefficient, mutual information, and combined p-values for a test set of proteins. The filtered correlation coefficient and mutual information p-values were created using one-sided models that only considered positive relationships as significant. The scale on the right is in units of $\log_{10}(pv)$ ranging from -2 (or p-value = .01) to -20 (or p-value = 10^{-20}). Protein

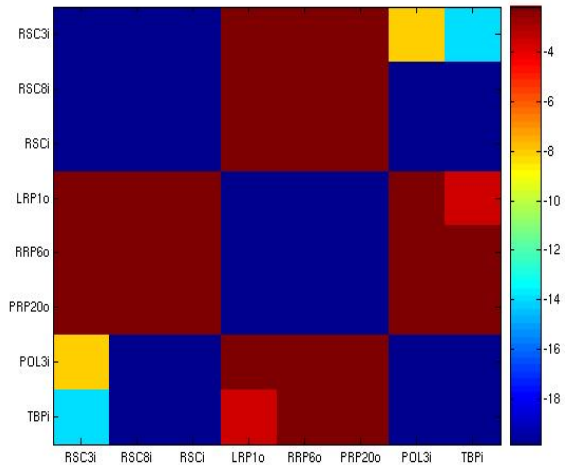
names are listed on the x and y axis with and "o" or "i" at the end of the names signifying that the data came from an ORF or intergenic array, respectively. Since combining p-values integrates evidence from the two complementary pairwise analyses, it should lead to fewer errors in deciding whether two factors have a significant binding dependence.

The first 3 factors, RSC3, RSC8, and RSC, are components of the RSC nucleosome remodeling complex. Since these proteins carry out tasks as part of the same complex, we would expect a high degree of similarity in their binding profiles, as shown in [7]. Indeed, each subfigure shows a p-value $\leq 10^{-20}$ for the interactions within the entire complex (dark blue box on the top left). For individual analyses, we consider a p-value of 10^{-10} as significant for a pairwise interaction. However, since we are combining two studies that are not fully independent, we consider a p-value of 10^{-20} as significant for the combined p-values. If our two analyses were completely dependent on one another, the p-values testing the same null hypothesis should be identical and a significance of 10^{-10} would translate to a significance of 10^{-20} when the p-values are combined. However, since our pairwise studies are not fully dependent on one another, a p-value threshold of 10^{-20} for combined p-values is more stringent than a p-value threshold of 10^{-10} for a single analysis.

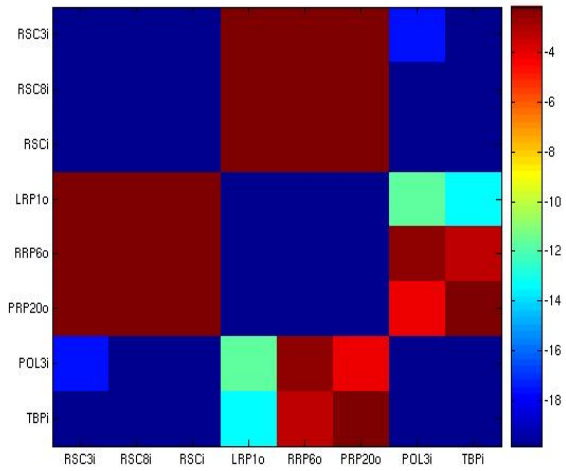
The next two proteins LRP1 and RRP6, are involved in mRNA degradation and surveillence, respectively. In [16], the authors show that LRP1 depends on RRP6 for recruitment to a large fraction of its gene targets. Therefore, we would expect similarity in their binding profiles and both analyses capture this interaction with a low p-value. Curiously, protein PRP20 also shows high similarity with the LRP1-RRP6 biological process. PRP20, also known as RanGEF in humans, binds preferentially to inactive genes [13]. This leads us to a novel biological hypothesis. It is known that the exosome (a complex of proteins which also contains LRP1 and RRP6) controls protein synthesis by degrading the mRNAs of genes that are improperly formed, processed, or transcribed at a given time. For example, we would not want the *GAL* proteins, involved in galactose breakdown, to be synthesized when the cell is grown in a dextrose-rich YPD medium. Moreover, [13] shows that PRP20 binds to genes that should be turned off during a given condition, such as GAL genes in YPD.



(a) Filtered correlation coefficient p-values for a test set of proteins.



(b) Mutual information p-values for test a set of proteins.



(c) Combined p-values for test a set of proteins.

Figure 3-2: Filtered correlation coefficient (a), mutual information (b), and combined (c) p-values for a test set of proteins. The filtered correlation coefficient and mutual information p-values were created using one-sided models that only considered positive relationships as significant. Protein names are listed on the x and y axis with an "o" or "i" at the end of the names signifying that the data came from an ORF or intergenic array, respectively. The scale on the right is in units of $\log_{10}(pv)$ ranging from -2 (or p-value = .01) to -20 (or p-value = 10^{-20}).

The fact that the exosome binds to a similar set of gene targets as PRP20 suggests that inactive genes are actually not completely turned off. It may be possible that the process of transcription is leaky, in which case inactive genes that should not be transcribed in a given condition may be erroneously transcribed. Indeed, there is a growing body of evidence in the transcription field to suggest that RNA polymerase is promiscuous in its induction of transcription at non-ORF sites within the genome, the functional consequences of which are still unclear. This suggests that the exosome may have a role in quality control at inactive genes, binding to them in order to readily degrade their unwanted mRNA.

The last 2 proteins, Polymerase III (POL3) and the TATA-Box Protein (TBP), also associate significantly for all three p-value calculations. In a recent paper [28], the authors show that the TBP preferentially associates with gene targets of POL3 in various environmental conditions. Although, the POL3 and TBP experiments were done in different laboratories, our normalized analyses confirm the results in [28].

After identifying connections within 3 biological processes, we consider the interplay between the three complexes. The authors in [7] show that the RSC nucleosome remodeling complex has a preference for POL3 gene targets. Hence, we should expect significant interaction in the top right corner of Figures 3-2(a), 3-2(b), and 3-2(c). The filtered correlation coefficient analysis finds all 6 interactions significant (p-value $\leq 10^{-10}$) but the mutual information and combined p-value calculation finds 5/6 relationships as significant (p-value $\leq 10^{-10}$ and p-value $\leq 10^{-20}$, respectively), making an error on RSC3-POL3. The filtered correlation coefficient analysis additionally predicts interactions LRP1-POL3 and LRP1-TBP but no relationships between its recruitment factor RRP6 and POL3 or TBP. Due to this discrepancy and due to the lack of literature supporting this interaction, we consider these two predictions as likely false positives. The mutual information and the combined p-value do not make this probable mistake. In summary, the filtered correlation makes 2 potentially false positive claims while the mutual information and combined p-values make one false negative error. Moreover, if we reduce the combined p-value threshold for significance to a less conservative and more accurate 10^{-18} (since our complementary analyses are not

completely dependent), we obtain 0 errors.

This example accurately represents the overall trend that correlation coefficient and mutual information tests will generate a higher number of false positives and false negatives, respectively. This is because mutual information makes a hard decision on classifying the data as 0s and 1s prior to the analysis and cannot capture relationships when a significant number of errors are made in the classification, resulting in false negatives. In contrast, the filtered correlations cost function considers a soft and continuous version of the data in terms of standard deviations of confidence and does capture these relationships. However, this also leads to considering unbound genes as partially bound and creates a higher number of false positives. Combining the two complementary analyses leads to fewer errors or to a more optimal operational point on an ROC curve of false negative rate versus false positive rate. We use this concept in building a more reliable network of our nucleus in Chapter 5. The next section attempts to reduce the false positive rate of predictions based on mutual information.

3.4 Minimized Mutual Information P-value

The previous section revealed that mutual information p-values are highly sensitive to the chosen cutoff for classifying the binding data into 1s and 0s. To alleviate this problem, this section introduces minimized mutual information p-values. This method allows for the binding cutoffs to take on several values from an allowable set and finds the cutoffs that minimize the mutual information p-value between two proteins. Section 3.2.2 shows that transforming data into 0s and 1s seems most appropriate when using the PV and PR matrices. Let $A = \{0.001, 0.005, 0.01, 0.02\}$ denote the allowable set of p-value cutoffs for thresholding the PV matrix. Moreover, let $c_i, c_j \in A$ denote the p-value cutoffs of proteins i and j for transforming continuous row vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ from the PV matrix to binary vectors, respectively. For each combination of cutoffs $c_i, c_j \in A$, lets also consider the strongest 1% of gene-protein interactions for proteins i and j as bound. Then the minimized mutual information p-value, pv_{MMI}, between binding vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ takes the form

$$pv_{MMI}(\mathbf{x}_i, \mathbf{x}_j) = \min_{c_i \in A, c_j \in A} pv_{MI}(\mathbf{x}_i, \mathbf{x}_j)\,, \tag{3.32}$$

where pv_{MI} is the mutual information p-value previously calculated in (3.28). The minimized mutual information p-value mitigates the sensitivity in the mutual information analysis due to hard classification of the data. However, since this analysis allows for the same protein to have different bound cutoffs, it is only suited for making decisions on pairwise interactions between two proteins and does not allow for unbiased group-wise comparisons. Hence, we will use the minimized mutual information p-values for making decisions about adding links in a network between two proteins (Chapter 5) but refrain from using it for inferring group-wise relationships between proteins (Chapter 4).

Filtered correlation coefficient, mutual information, and combined p-values allow us to judge the strength of pairwise interactions between two proteins. The following chapter introduces clustering and PCA, which can evince group-wise relationships between factors that regulate transcription.

Chapter 4

Group-wise Relationships: PCA and Clustering

Biological processes depend on the collaborative interaction of several proteins. In the previous chapter we developed pair-wise statistics in order to describe the relationship between two proteins. This chapter introduces Principal Component Analysis (PCA) and clustering, two methods that can evince group-wise dependencies between proteins. The two techniques verify known biological mechanisms and uncover novel cellular processes.

4.1 Principal Components Analysis

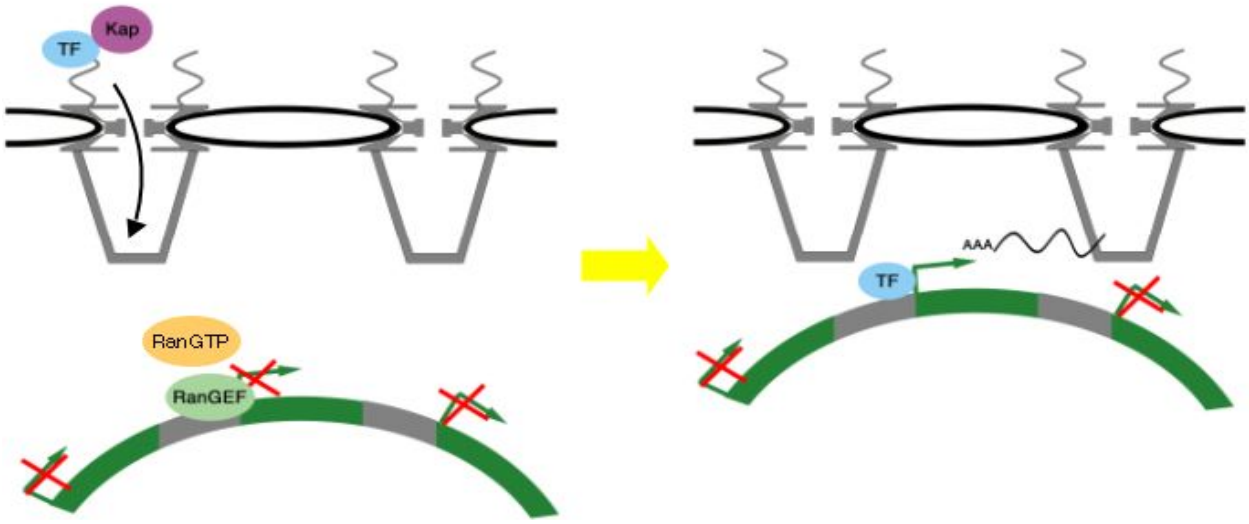
PCA is a common technique for data reduction and visualization. It determines the directions of the greatest variance within a data matrix by calculating the orthonormal eigenvectors associated with the largest eigenvalues of the scatter matrix of the data [36]. The principal components, usually referred to as scores, are row vectors of the same dimensionality as row vectors of the data matrix. The data can be modeled by taking linear combinations of the scores; the associated weights with each score are called loads. Since most of the variance within a data set can usually be captured with a small number of principal components, PCA exploits the common information within the data in order to reduce its dimensionality.

One can choose the number of principal components one wishes to calculate; more principal components can characterize the data better at the cost of having to keep track of more information. For visualization purposes, two or three principle components are often used. For a given choice of data representation, let $\mathbf{x}_i = [x_{i,g_1} \dots x_{i,g_{|G|}}]$ denote a row vector of binding data for protein i across all genes g in the set G. If performing PCA on the binding profiles of each protein i, we can write

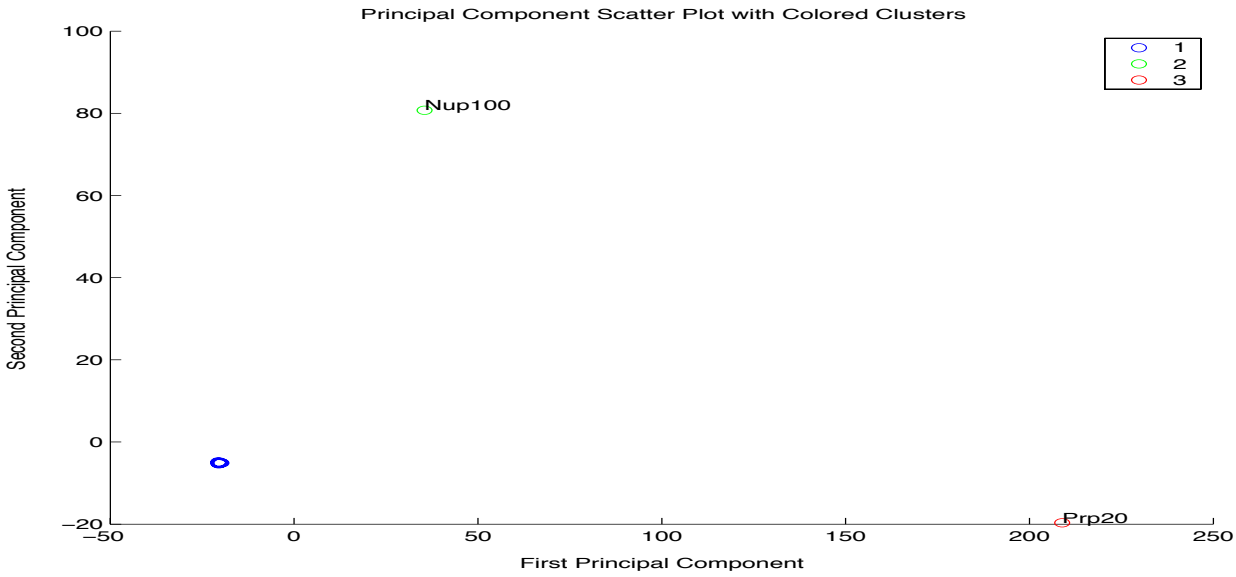
$$\mathbf{x}_i = \hat{\mu}_i \mathbf{1} + \sum_{n=1}^{N} w_n^i \mathbf{s}_n \,. \tag{4.1}$$

The PCA equation (4.1) uses N principal components, with row vector scores $\mathbf{s}_1, \dots, \mathbf{s}_N$ and scalar loads $\{w_1^i, \dots, w_N^i\}$ specific to each $\mathbf{x}_i$, where $\hat{\mu}$ is a scalar that is the mean across the elements of $\mathbf{x}_i$, and $\mathbf{1}$ is row a vector of all ones.

Figure 4-1 shows how PCA can provide useful biological insight. The large isolation of protein Prp20 and Nup100 from the rest of the nuclear pore related proteins in Figure 4-1(b) supports a model introduced in [13]. Casolari et al. shows that Prp20, known as RanGEF in humans, binds preferentially to transcriptionally inactive genes. RanGEF is also known to catalyze the unloading of transcription factors (TF) transported into the nucleus by nuclear import proteins, such as Kap. The left part of Figure 4-1(a) illustrates how RanGEF converts RanGDP to RanGTP which in turn unloads the TF transported into the nucleus by Kap. Hence, the authors suggest that Prp20 may contribute to transcriptional activation by facilitating the release of transcription factors at genes requiring fast activation. Accordingly, induction of said genes leads to experimentally confirmed loss of RanGEF association and a gain in binding at the nuclear pore on the periphery of the nucleus, as illustrated on the right part of Figure 4-1(a). Thus, Prp20 and the rest of the nuclear pore proteins should bind to very few genes in common at any one time as shown by the great separation between the two in Figure 4-1(b). Further, the results suggest that the other aberrant nuclear factor, Nup100, also has a role that is orthogonal to the workings of the majority of the nuclear pore proteins. The following section discusses another method for finding biological processes using ChIP-chip data.



(a) Casolari et al. Model for Nuclear Transport (picture from [13])



(b) Principal Component Analysis Validates the Model

Figure 4-1: The subfigures show (a) Casolari et al. model for nuclear transport and (b) its validation using Principal Component Analysis (PCA) [13]. Figure (b) plots the loads on the first two principal components. The colors denote clusters constructed using hierarchical clustering based on correlation coefficient distance, as defined in Section 2.2.2, and a cutoff for three clusters.

4.2 Clustering

Clustering of genomic data can evince group-wise relationships between proteins or genes. We first introduce K-means and hierarchical clustering and then adapt the algorithms in order to use the pair-wise distance metrics developed in the previous chapter. We then develop a novel semi-supervised clustering algorithm that preserves information about elements within each cluster in order to better capture group-wise dependencies between proteins.

4.2.1 K-means Clustering

K-means clustering is one of the most commonly encountered algorithms in biology, due to its ease of implementation. K-means clustering uses the expectation maximization (EM) algorithm to partition (cluster) N elements into K disjoint subsets S_k, $k = 1, \ldots, K$. Given an appropriate choice of data representation, let $\mathbf{x}_i$ denote a row vector of binding data for protein i, as in Section 4.1. In order to find groups of related proteins, we represent binding profiles as elements that we want to cluster. The K-means algorithm finds K disjoint subsets S_k, $k = 1, \ldots, K$ that minimize the overall cost function

$$J = \sum_{k=1}^{K} \sum_{i \in S_k} d^2(\mathbf{x}_i, \mathbf{c}_k)\,, \tag{4.2}$$

where $d(\mathbf{x}_i, \mathbf{c}_k)$ is the distance between binding profile $\mathbf{x}_i$ and centroid $\mathbf{c}_k$ of subset S_k. The algorithm usually uses Euclidean distance for d and defines $\mathbf{c}_k$ as the mean of all the elements in S_k:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{i \in S_k} \mathbf{x}_i\,. \tag{4.3}$$

The user initiates the clustering by designating the K underlying groups of objects. Then the K centroids are initialized, usually by assigning each element to one of the K clusters at random and using (4.3). The algorithm then iterates between (i) assigning all the elements to subset S_k with the "closest" (in terms of distance d) centroid $\mathbf{c}_k$ (E-step) and (ii) reestimating

the cluster centroids based on the new assignments using (4.3) (M-step). The procedure stops once no further change in assignments occurs.

We adapted the K-means algorithm in order to incorporate the pairwise statistical analysis from Chapter 3. The following formula converts the combined p-values from Section 3.3 to a non-negative distance, where a small p-value corresponds to a small distance between the binding profiles $\mathbf{x}_i$ and $\mathbf{x}_j$ of two proteins i and j:

$$d(\mathbf{x}_i, \mathbf{x}_j) = -\log_{10}(1 - pv_C(\mathbf{x}_i, \mathbf{x}_j)) \,. \tag{4.4}$$

K-means clustering (as any EM based algorithm) does not guarantee global minimization of the cost function in (4.2), and its performance heavily depends on the initialization of the K centroids. Due to this problem, the K-means algorithm often fails to identify known protein clusters (e.g., the RSC complex). Other EM based partitioning algorithms, such as fuzzy membership clustering, would also suffer in performance due to random initialization; therefore, the next section introduces a different family of algorithms based on hierarchical clustering.

4.2.2 Hierarchical Clustering

Hierarchical clustering has wide applicability in biology, as well. In addition, hierarchical partitioning does not require *a priori* knowledge of the number of clusters and can be easily visualized using a tree structure (called dendrogram), making it often preferable to K-means clustering. Hierarchical clustering is subdivided into agglomerative methods, which proceed by series of fusions of the N objects into groups, and divisive methods, which separate N elements successively into finer subsets.

We adapted the agglomerative hierarchical clustering algorithm in order to incorporate the pairwise statistical analysis from Chapter 3. As in the previous section, we use binding profiles as elements and define all the pairwise distances between elements using (4.4). At the start, the algorithm treats each element as a cluster, and proceeds for $N - 1$ iterations. At each iteration, the algorithm links the two most similar clusters (represented as linking

two child nodes to a parent node on a dendrogram) until all N elements are unified into one partition. Hierarchical clustering algorithms can define distance (or similarity) between clusters in various ways. The following equations describe several common distances for linking two clusters C_k and C_l:

$$\text{Nearest Neighbor}: \quad d(C_k, C_l) = \min_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j) \tag{4.5}$$

$$\text{Farthest Neighbor}: \quad d(C_k, C_l) = \max_{i \in C_k, j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j) \tag{4.6}$$

$$\text{Average}: \quad d(C_k, C_l) = \frac{1}{|C_k||C_l|} \sum_{i \in C_k} \sum_{j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)\,. \tag{4.7}$$

Hierarchical clustering based on combined p-values from Section 3.3 confirms known biological processes with great accuracy, especially when using the average distance for linking clusters. However, hierarchical clustering only considers pairwise relationships between binding profiles. Figure 4-2 uses an example to illustrate the potential pitfalls with clustering solely on pairwise distances. The Venn diagrams on the left and on the right describe two hypothetical relationships between three factors (proteins). Each circle or oval represents the subset of genes classified as bound by the six different factors. In the left scenario, each possible pair of factors (i.e., A↔B, A↔C, and B↔C) have a number of genes that they bind to in common. But when considering the group-wise relationship (i.e., A↔B↔C), we find no common intersection in bound genes between all three factors. In contrast, the scenario on the right shows that factors X, Y, and Z share a large common intersection and thus interact in a pairwise as well as group-wise manner. Since both sets of three factors have equivalent pairwise distances, hierarchical clustering cannot distinguish any difference between the two scenarios. However, the much stronger group-wise interaction of the scenario on the right makes it more likely for factors X-Y-Z to share a common biological process than for factors A-B-C. In the next section, we develop a novel semi-supervised clustering algorithm that preserves information about elements within each cluster in order to better capture group-wise dependencies between proteins.

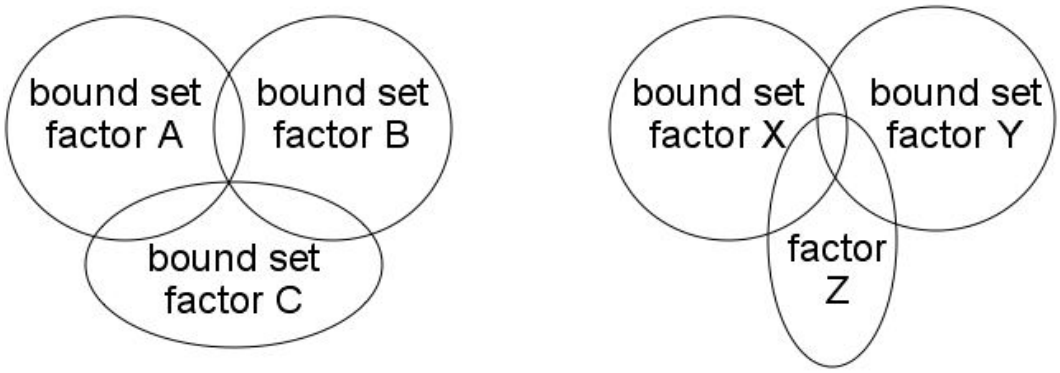


Figure 4-2: The Venn diagrams on the left and on the right describe two hypothetical relationships between three factors. Since both groups of three factors have equivalent pairwise overlaps in bound genes, hierarchical clustering cannot distinguish any difference between the two scenarios (see text). However, the much stronger group-wise interaction of the scenario on the right makes it more likely for factors X-Y-Z to share a common biological process than for factors A-B-C.

4.2.3 Semi-Supervised Clustering

Semi-supervised clustering derives its name from the fact that it retains information about elements within each cluster as it partitions the objects. In order to preserve information about the elements of cluster C_k the algorithm maintains two vectors, $\mathbf{f}_k$ and $\mathbf{x}_k$. Vector $\mathbf{f}_k = [f_{k,g_1} \dots f_{k,g_{|G|}}]$ records the fraction of elements that bind to each gene g in the set G. Vector $\mathbf{x}_k = [x_{k,g_1} \dots x_{k,g_{|G|}}]$ represents the averaged binding profile at gene $g \in G$ for all members within the partition, based on the SD data representation. When merging two clusters C_k and C_l into C_o, the resulting two vectors for cluster C_o are weighted combinations of the vectors for C_k and C_l:

$$\mathbf{x}_o = \frac{1}{|C_k| + |C_l|} \left(|C_k|\mathbf{x}_k + |C_l|\mathbf{x}_l\right) \tag{4.8}$$

$$\mathbf{f}_o = \frac{1}{|C_k| + |C_l|} \left(|C_k|\mathbf{f}_k + |C_l|\mathbf{f}_l\right) . \tag{4.9}$$

Note that merged vector $\mathbf{f}_o$ still maintains the fraction of genes bound by the elements of

the new cluster and $\mathbf{x}_o$ still represent the average binding profile of all the joined objects. To define similarity between clusters, we again use the distance based on combined p-values in (4.4). Although there are many choices for how to define similarity, we wanted to incorporate the robust measures we derived in the previous chapter. Appendix A develops the algorithm with a more theoretically intuitive but less biologically meaningful distance based on Kullback-Leibler(KL) divergence.

To find combined p-values for the relationship between two clusters, the algorithm evaluates the filtered correlation coefficient and mutual information but incorporates the group-wise dependence by modifying how to select the pertinent sets. For filtered correlation, let F_k and F_l represent the filtered subsets of genes for clusters C_k and C_l, respectively, and let $F_{k,l} = F_k \cup F_l$ again denote the filtered subset of genes over which we will estimate the variances and covariance of two partitions using (3.1) - (3.5). The filtered subsets F_k and F_l no longer consist of the set of genes bound by one factor but now represent the sets of genes bound by a fraction of objects within a cluster. Mathematically, letting f_{thresh} represent the fraction of proteins that need to bind to a gene in the filtered subsets, we define $F_k = \{g : f_{k,g} \geq f_{\text{thresh}}\}$ and $F_l = \{g : f_{l,g} \geq f_{\text{thresh}}\}$, respectively. Having defined $F_{k,l} = F_k \cup F_l$, the algorithm uses the averaged binding vectors $\mathbf{x}_k$ and $\mathbf{x}_l$ in (3.1) - (3.5) to compute the means, filtered variances, and filtered covariance of clusters C_k and C_l. Next, using (3.6), (3.14), and (3.15), the algorithm derives the filtered correlation coefficient p-value between the two partitions. For finding mutual information p-values, we can now assign $u = |F_k|$, $v = |F_l|$, $h = |F_k \cap F_l|$, and w equal to the number of genes with binding information for both clusters. Using these quantities, we can use (3.28) to find the mutual information p-value between two clusters. And finally, we again combine p-values using (3.30).

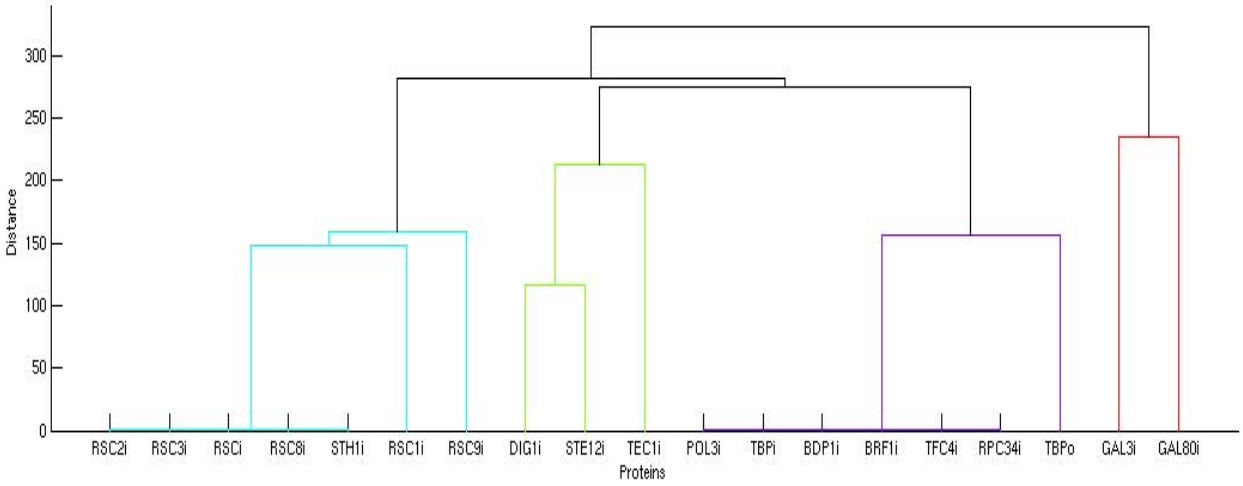
Similar to hierarchical clustering, the algorithm treats each element as a cluster at the start and proceeds for $N - 1$ iterations. At each iteration, the algorithm links the two most similar partitions, based on the combined p-value distance in (4.4), until all N elements are unified into one partition. However, the new method for finding the combined p-values

between clusters incorporates the group-wise dependence of elements, rectifying the potential problem with using hierarchical clustering. The next section validates the algorithm using several known biological processes.

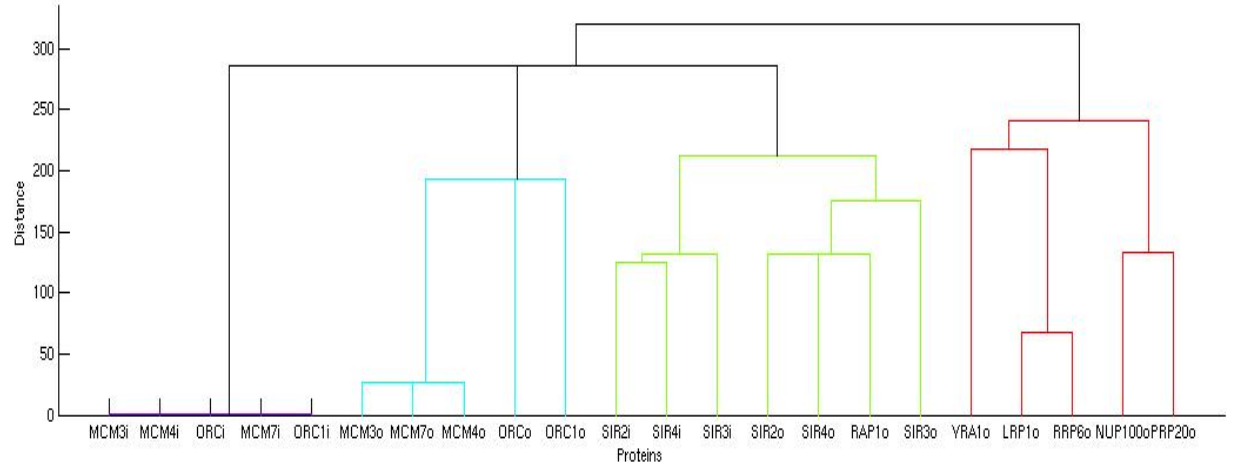
4.2.4 Biological Validation of Semi-Supervised Clustering

In order to evaluate the performance of the new algorithm, we clustered the SD data representation using a bound cutoff of $PV \leq .015 \vee PR \geq .99$, where $\vee$ denotes logical or, consistent with our results in Sections 2.2.2 and 3.2.2. Moreover, we set $f_{\text{thresh}} = 0.5$ in order to consider genes bound by half or more of the factors in a partition as representative of the biological process of the partition. We first clustered the whole data set and then extracted the produced clusters into separate plots to facilitate visualization. Figure 4-3 illustrates how the algorithm performs on a selected group of biological processes using tree structures or dendograms. The horizontal branches on the dendrogram represent the merging of two similar objects at an iteration in the algorithm, while the length of the vertical branches corresponds to the similarity distance (as in (4.4)) of the fused nodes. Unlike K-means, semi-supervised clustering does not decide the final number of clusters but allows the user to choose a significant threshold distance for a cluster. In Figure 4-3, we chose a distance of less than 303 as significant, which corresponds to a combined p-value of 10^{-20} in our implementation. For visualization purposes, however, we color coded clusters of distance ≤ 253 or of combined p-value $\leq 10^{-70}$.

Figure 4-3(a) shows the clustering of four known biological processes using different colors. It again illustrates the significant association of components of the RSC nucleosome remodeling complex (RSC, RSC1, RSC2, RSC3, RSC8, RSC9, and STH1) and of components of the Polymerase III transcriptional machinery (POL3, TBP, BRF, TRF4, and RPC34). It also demonstrates the significant clustering between transcription factors STE12, DIG1, and TEC1. In [5], the authors show that STE12 activates two different sets of genes under different conditions. In the presence of a nutrient limiting agent butanol, binding of STE12, DIG1, and TEC1 activates genes necessary for growth of filaments, or structures that expand



(a) Validation of the RSC, POL3, STE12, and GAL biological processes.



(b) Clustering of the MCM/ORC, SIR, and LRP1 biological processes.

Figure 4-3: Confirmation of known biological processes and new insight using semi-supervised clustering (see text). The two dendrograms above show a subset of the results from clustering the whole data set using the SD data representation and a bound cutoff of $PV \leq .015 \vee PR \geq .99$. We set $f_{\text{thresh}} = 0.5$ in order to consider genes bound by half or more of the factors in a partition as representative of the biological process of the partition. The horizontal branches signify the merging of two similar objects, while the length of the vertical branches corresponds to the similarity distance (as in (4.4)) of the fused nodes. For visualization purposes, different colors represent clusters of distance ≤ 253 or of combined p-value $\leq 10^{-70}$.

the size of the cell and facilitate storage of nutrients. During pheromone exposure, binding of STE12 and DIG1 (no TEC1) induces genes involved in mating. Although our data monitors the binding of the three factors under no treatment in YPD, the clustering still evinces the strong binding dependence between STE12 and DIG1, and their auxiliary relationship with TEC1. Moreover, Figure 4-3(a) also confirms the significant interdependence between the POL3 machinery and RSC complex (p-value $\leq 10^{-40}$). It further suggests a previously unknown relationship between the STE12 cluster and the POL3 and RSC biological processes. And finally, the right most partition shows significant similarity between transcription factors GAL3 and GAL80. As discussed in Section 2.1, the two transcription factors regulate the expression of genes that breakdown the sugar galactose. Similar to the STE12 analysis, using YPD data again uncovers the general relationship between the two galactose factors. However, we do not believe that the GAL transcription factors share significant functions with the previous three processes (p-value $\geq 10^{-5}$).

Figure 4-3(b) displays a dendogram of several other known and hypothesized biological mechanisms. The rightmost colored tree corroborates our hypothesis that LRP1, RRP6, and PRP20 may synergistically bind at inactive genes. As discussed in Sections 3.3.1 and 4.1, LRP1 and RRP6 may bind to inactive genes in order to readily degrade their unwanted mRNA, while PRP20 may bind at inactive genes to facilitate fast activation [13]. Moreover, the association LRP1 and RRP6 with mRNA export factor YRA1, supports the results in [16] that demonstrate coupling of mRNA production quality assurance to mRNA export. Further, the similar relationship between NUP100 and PRP20 suggests that NUP100 may also have a role that is orthogonal to the workings of the majority of the nuclear pore factors, as predicted in the PCA analysis. However, this cluster does not seem to share significant commonality with the rest of the colored trees on the figure.

The other three colored clusters in Figure 4-3(b) also provide validation of our analysis and new biological insight. During normal growing conditions, a yeast cell periodically cycles through several phases of growth. If a mother cell has sufficient nutrients in its environment it may decide to divide into two daughter cells, in a phase of the cell cycle called mitosis. To

make sure that the daughter cells inherit the genetic information of the mother cell, prior to mitosis and during the S phase of the cell cycle, the mother cell makes two identical copies of all of its genetic material in a process known as DNA replication. In [8], the authors show that the Origin Recognition Complex (ORC) and the MiniChromosome Maintenance (MCM) proteins bind together at origins of replication, or sites along the chromosomes where DNA replication is initiated. The two left most colored trees confirm the significant association of components of the ORC (ORC1, ORC1-6) and MCM (MCM3, MCM4, and MCM7) complex at both intergenic and open reading frame (ORF) regions of the genome, denoted by an "i" and "o" at the end of each protein name, respectively. Although DNA replication only takes place during the S phase of the cell cycle, and despite the fact that our unsynchronized ChIP-chip data samples cells during all phases of growth, the semi-supervised clustering still uncovers the relationship described in [8]. The next colored partition confirms the significant association between components of the SIR complex at both intergenic and ORF regions [21]. Moreover, the SIR complex associates significantly (p-value $\leq 10^{-35}$) with the MCM/ORC cluster, suggesting a mutual dependence between the two processes. The binding profile of SIR2o also has a close relationship with RAP1o, a TF involved in transcriptionally active processes, which we will explore in more detail in the next section.

4.2.5 Active Processes and SIR2

Transcriptionally active processes, or simply active processes in our context, refer to biological mechanisms that occur at highly expressed genes. For example, the model introduced in Section 4.1 [13] of how highly expressed genes bind to the nuclear pore describes an active process. To gauge how actively a gene is expressed in YPD, we downloaded data for the number of mRNAs produced per hour, or the transcriptional frequency of a gene, as measured in [12]. The results from [12] show that only about 21% of the genes produce more than 6 mRNAs/hr. Therefore, we define genes with transcriptional frequency in the top 20% as active. To determine whether a factor participates in active processes, we again calculate filtered correlation coefficient, mutual information, and combined p-values between

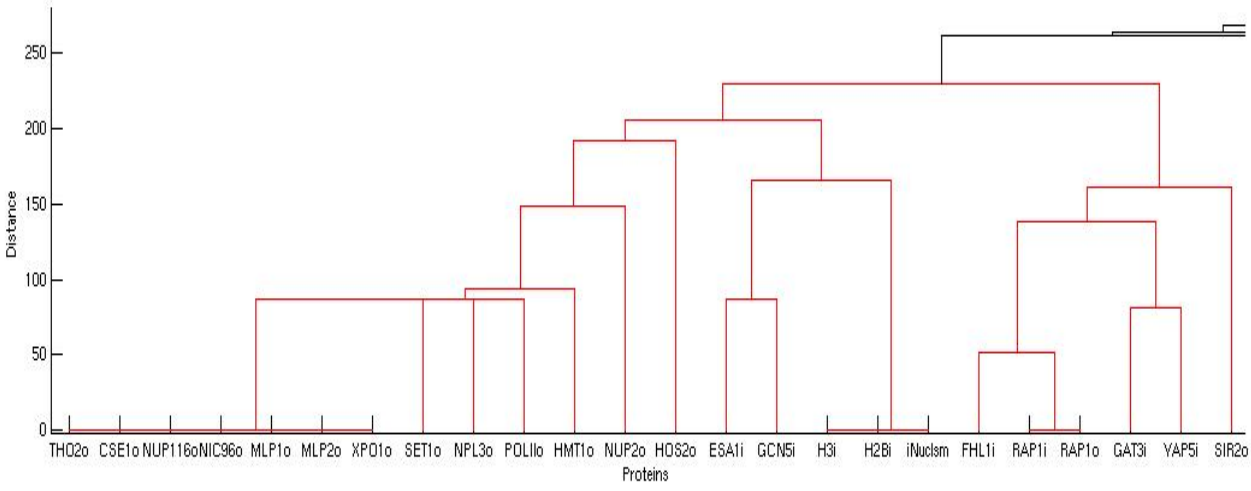
the bound set of a factor and the set of active genes. Mathematically, we let A represent the set of active genes and let F_i represent the set of genes classified as bound for a potential active factor i. Then we define $F_{i,j} = F_i \cup A$ in the filtered correlation coefficient equations (3.1) - (3.5), and use (3.6), (3.14), and (3.15) in succession to derive p-values. For finding mutual information p-values, we can now assign $u = |A|$, $v = |F_i|$, $h = |A \cap F_i|$, and w equal to the number of genes with observed information for both data sets. Using these quantities, we can use (3.28) to find the mutual information p-value. And finally, we again combine p-values using (3.30) and consider a factor as active if it has a combined p-value $\leq 10^{-20}$ or a mutual information p-value $\leq 10^{-10}$ (justified in Section 5.1).

In order to explore group-wise relationships between active factors, we categorized all proteins as active or non-active using the stringent criterion above. Figure 4-4 displays the results of clustering only the active factors using the SD data representation and a bound cutoff of $PV \leq .015 \vee PR \geq .99$ once again. Note that all the factors included have a **significant** similarity distance of ≤ 283, or cluster combined p-value $\leq 10^{-40}$. The results provide interesting biological validation and exciting new insight. Figure 4-4 includes all the nuclear proteins considered to have a preference for binding to active genes in [13] (CSE1, NUP116, NIC96, MLP1, MLP2, XPO1, NUP2, KAP95, and NUP60) and in [14] (THO2, NPL3, and HMT1), justifying the above criterion for finding active factors. Moreover, the tight clustering of the active nuclear pore factors once again validates the model in [13], which suggests that these factors share a biologically active mechanism. The large similarity between nuclear proteins THO2, NPL3, and HMT1 also seems to confirm the results in [14]. Yu et al. introduced a model showing that methyltransferase HMT1 plays a role in disassociating transcriptional elongation protein THO2 and mRNA export factor NPL3. Moreover, nonfunctional (mutated) HMT1 leads to failed export of mRNA produced at a selected active gene. Since all three factors bind preferentially to highly expressed genes, they seem to share an active biological process, whereby HMT1 dependent dissociation of THO2 and NPL3 may prove necessary for export of mRNAs produced by active genes. Further, their close relationship with the nuclear pore proteins implies that this mechanism might

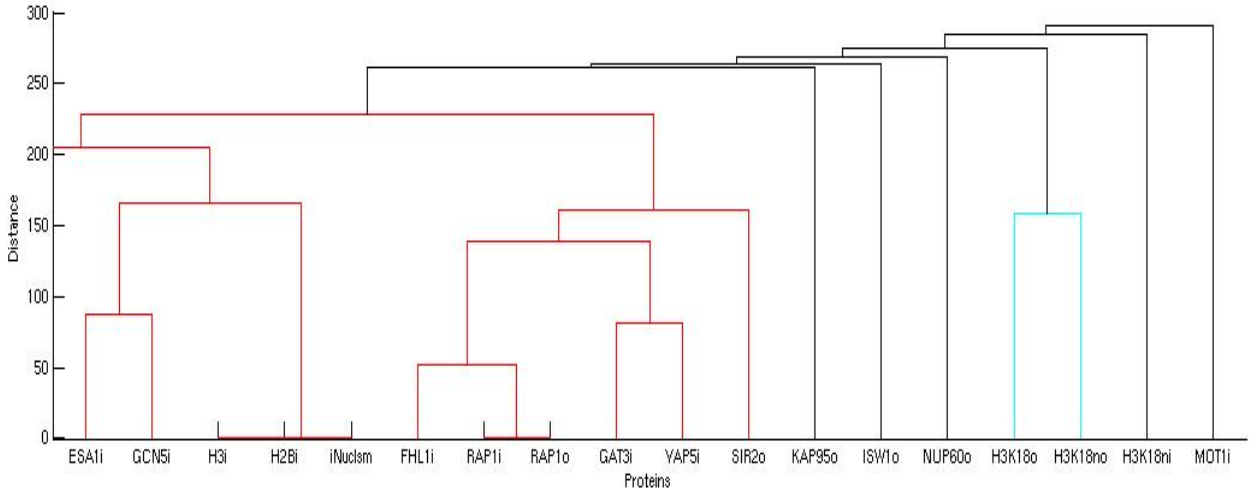
also occur at the periphery of the nucleus.

The red tree on the left part of Figure 4-4(b) illustrates another active biological process. Histone acetyltransferases (HATs) ESA1 and GCN5 have a predilection for binding to active genes [4], while data H3i, H2Bi, and iNuclsm measuring nucleosome depletion also overlaps significantly with promoters of active genes [27]. Moreover, transcription factors FHL1 and RAP1 regulate protein biosynthesis and other active processes [38]. The nearby congregation of RAP1, nucleosome depletion, and HATs ESA1 and GCN5 seems to support a conjecture mentioned in [27]. Berstein et al. showed that RAP1 is necessary but not sufficient for the mechanism that displaces nucleosomes at highly expressed genes. Hence, the authors conjectured that other factors, including ESA1 and GCN5, might be required along with RAP1 in order to reposition nucleosomes at induced genes. Moreover the large similarity between this process and the nuclear pore factors as seen in Figure 4-4(a) again implies that this active process takes place at the periphery of the nucleus.

Surprisingly, Figure 4-4(b) also uncovers a great similarity between the binding of SIR2 at ORF regions, denoted as SIR2o, and other active proteins. The SIR complex does not seem to bind to DNA directly but interacts with other factors, such as RAP1 or deacetylation at histone 4, in order to access the nearby genes. This explains the association of SIR2 with active factor RAP1 in Figure 4-3(b). However, SIR2 is a histone deacetylase that plays an important role in silencing transcription at telomeres (or ends of chromosomes) and mating-type gene loci HML and HMR [21]. Hence, the designation of SIR2o as an active factor and its significant binding similarity with all other active factors is unexpected. In Figure 4-4(a), SIR2o associates with several active transcription factors, including GAT3, FHL1, RAP1, and YAP5. It also has a significant binding similarity with semi-active (combined p-value $\leq 10^{-10}$ instead of $\leq 10^{-20}$) transcription factors PDR1, RGM1, and SMP1. These transcription factors (TFs) regulate various different biological pathways, including biosynthesis, environmental response, and metabolism. However, all the enumerated TFs bind preferentially to nucleosome depleted promoters in a functionally cooperative manner [27]. This may suggest a role for SIR2 in nucleosome displacement. Moreover, SIR2 deacetylates



(a) Active cluster, part I



(b) Active cluster, part II

Figure 4-4: Active biological processes and new insight using semi-supervised clustering. Figures 4-4(a) and 4-4(a) display the results of clustering all active factors (see text), using the SD data representation and a bound cutoff of $PV \leq .015 \vee PR \geq .99$ once again. We again set $f_{\text{thresh}} = 0.5$ in order to consider genes bound by half or more of the factors in a partition as representative of the biological process of the partition. All members of the cluster associate significantly (combined p-value $\leq 10^{-40}$), but for visualization purposes different colors represent clusters of distance ≤ 253 or of combined p-value $\leq 10^{-70}$.

lysine 16 on histone 4, which is generally associated with active processes. Finding the full extent of SIR2o's role in active processes, and why it only appears to take place at ORF and not intergenic regions may uncover exciting and new biological insight.

Figure 4-4 also includes several other factors with proposed roles in active processes. First, histone deacetylase HOS2 was shown to play a role in deacetylation at active genes in [20] and clusters tightly with the nuclear pore factors in Figure 4-4(a). Second, [6] shows that histone methyltransferase SET1, also in Figure 4-4(a), adds three methyl groups at lysine 4 of histone 3 during transcription (i.e., at active genes), which may serve as memory that a gene was recently turned "on". Third, Santos-Rosa et al. also shows an intimate connection between SET1 and nucleosome remodeling complex protein ISW1, shown in Figure 4-4(b) [25]. The authors show that ISW1 recruitment at active genes is dependent upon methylation by SET1. Further, the paper illustrates that proper transcription at selected genes depends upon the collaborative function of both ISW1 and SET1. And finally, [17] shows that high acetylation level at histone 3 lysine 18 at both intergenic and ORF regions correlates with gene activity. Figure 4-4(b) includes the three factors, designated as H3K18o, H3K18no, and H3K18ni, where "o", "no", and "ni" at the end refer to ORF, normalized ORF, and normalized intergenic regions [17]. Our clustering analysis captures the collective relationship between all the mentioned active factors for the first time in the literature. In the next chapter, we use the theory developed thus far to build a multi-layered transcriptional network of the nucleus.

Chapter 5

Network of the Nucleus

This chapter builds on the methodology developed in the previous chapters in order to build a network of the nucleus. Our holistic approach is the first attempt in the literature to quantify the communication between layers in eukaryotic transcriptional networks. Analysis of the finalized network will hopefully unveil a general view of the non-linear process of transcription. The next section describes the model we used to determine the interplay between various proteins.

5.1 Network Model

In order to build a transcriptional network of the nucleus, we represent each factor as a node (or oval) and use the pairwise statistics developed in Chapter 3 to find significant binding relationships between factors, represented as edges (or links) between nodes. Using the convention of italicizing sets but not members of sets, we again define the following five categories of proteins: Transcription Factors (TF), Nuclear Transport proteins (NT), Nuclear Processing proteins (NP), Histone Modifiers and Nucleosome Remodelers (HM), and Histone State (HS) in terms of acetylation levels, methylation levels, and nucleosome distribution. We use an undirected network model, since our ChIP-chip data cannot evince directionality in binding dependencies, or whether factor i causes factor j to bind to gene g.

Therefore, an edge between two factors i and j contains zero or two arrows in our model, lacking information about directionality.

Our network model consists of two types of edges—a positive and a negative edge representing a significant synergistic and an opposing binding relationship, respectively. Since our mutual information p-values only evaluate the significance of positive binding relationships, decisions on extending negative edges will solely depend on correlation analysis. Hence, we use the two sided integration in (3.15) to evaluate the significance of finding both extremely positive or negative binding relationships. To avoid ambiguity when using correlation analysis, we set the p-value between proteins i and j to one if $\hat{\rho}_{X_i,X_j} < 0$ when considering positive edges and if $\hat{\rho}_{X_i,X_j} \geq 0$ when considering negative edges.

For simplicity, we first consider protein-protein interactions, excluding all factors within the HS level. Mathematically, let ν_i and ν_j denote the nodes (or vertices) for proteins i and j, respectively. Then $\epsilon^+_{i,j}$ and $\epsilon^-_{i,j}$ represent a positive and a negative binding relationship between i and j, respectively. Moreover, $\epsilon^+_{i,j}$ and $\epsilon^-_{i,j}$ can only equal to 1 or 0, corresponding to a presence or absence of a significant binding relationship between i and j. We add a positive edge ($\epsilon^+_{i,j} = 1$) between proteins i and j from all layers except the set HS if their respective binding profiles $\mathbf{x}_i$ and $\mathbf{x}_j$ have a combined p-value $pv_C \leq 10^{-20}$ as in (3.30) or a minimized mutual information p-value $pv_{MMI} \leq 10^{-10}$ as in (3.32):

$$i \notin HS, j \notin HS \ : \quad \epsilon^+_{i,j} = 1 \ \text{ if } \ (pv_{MMI}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-10}) \vee (pv_C(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-20}) \,. \quad (5.1)$$

The decision to use the or condition (denoted by the symbol $\vee$) above is substantiated by our observation in Section 3.3.1 that mutual information analysis leads to very few false positive claims at a significance level of 10^{-10}. In addition, the test incorporates the information from the filtered correlation analysis in the evaluation of the combined p-value. Since filtered correlation analysis leads to more false positives, it needs to combine its evidence with that of the mutual information analysis and meet a stricter significance threshold to create an edge.

Negative edges in the network prove much more complicated to assign and interpret. Section 3.2.1 discusses the inherent difficulty with using mutual information p-values for finding opposing binding relationships. Moreover, it is possible to consider the overlap between extremely bound and unbound sets of genes but the biological interpretation of significant overlap in that scenario is not clear. Hence, it seems reasonable to solely base our decision on assigning negative edges in the network using filtered correlation coefficient p-values. To assign a negative link ($\epsilon^-_{i,j} = 1$) between the binding profiles $\mathbf{x}_i$ and $\mathbf{x}_j$ of the proteins i and j not in the set HS, we use a filtered correlation coefficient p-value pv_{FCC} threshold of 10^{-10}:

$$i \notin HS, j \notin HS \ : \quad \epsilon^-_{i,j} = 1 \ \text{ if } \ (pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-10}) \, . \tag{5.2}$$

Data from the HS class of factors requires particular care. For proteins, it seems natural to classify binding data at a gene as 0s and 1s, representing an absence or presence of association with a gene's DNA. However, ChIP-chip data of histone acetylation levels, for example, can indicate gradations of acetylation (i.e., hypo, medium, hyperacetylation, etc.). Hence, it seems more natural to consider the continuum of data for members of the HS layer. Although mutual information analysis extends to continuous distributions, finding p-values for this scenario that are consistent with the previously developed pairwise statistics proves more difficult. Hence, we decided to use standard correlation coefficient p-values (using ChIP-chip data at all genes) to quantify the relationships between two factors within the HS layer. The p-values for standard correlation coefficient estimates, pv_{SCC}, can also be assigned using (3.6), (3.14), and (3.15) in succession. Mathematically, let i and j represent two factors from the HS layer with binding profiles $\mathbf{x}_i$ and $\mathbf{x}_j$. We extend positive and negative links in our network model if the p-value $pv_{SCC} \leq 10^{-20}$:

$$i \in HS, j \in HS : \quad \epsilon^+_{i,j} = 1 \ \text{ if } \ (pv_{SCC}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-20}) \wedge (\hat{\rho}_{X_i, X_j} \geq 0) \tag{5.3}$$

$$i \in HS, j \in HS : \quad \epsilon^-_{i,j} = 1 \ \text{ if } \ (pv_{SCC}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-20}) \wedge (\hat{\rho}_{X_i, X_j} < 0) \, . \tag{5.4}$$

In the above equations, $\wedge$ stands for a logical and, requiring that positive and negative interactions have a positive and negative estimate of the filtered correlation coefficient, respectively. Now let us consider relationships between one protein not in the set HS and one factor from the set HS. Ultimately, we want to make statements such as genes bound by transcription factor i have high acetylation. In order to use filtered correlation coefficient analysis in this context, we need to let the non-HS factor determine the pertinent set of dimensions or genes. Letting F_i denote the set of genes bound by the non-HS protein i, we define the filtered set $F_{i,j} = F_i$ in (3.1) - (3.5) and use (3.6), (3.14), and (3.15) in succession to derive p-values. Mutual information p-value analysis again proves difficult to extend in this context. For non-HS protein i and HS factor j, we assign positive and negative links in our network using the following thresholds on our filtered correlation coefficient p-value pv_{FCC}:

$$i \notin HS, j \in HS: \quad \epsilon^{+}_{i,j} = 1 \text{ if } (pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-4}) \wedge \hat{\rho}_{X_i,X_j} \geq 0 \tag{5.5}$$

$$i \notin HS, j \in HS: \quad \epsilon^{-}_{i,j} = 1 \text{ if } (pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j) \leq 10^{-4}) \wedge \hat{\rho}_{X_i,X_j} < 0\,. \tag{5.6}$$

Using a lower threshold for assigning links in the scenario above is necessary due to the lower dimensionality of the data. Since, our filtered set now only contains the set of genes defined as bound by a single factor, it often proves impossible to achieve such high significance thresholds as in the previous analyses. Moreover, we used an empirical, non-parametric method for calculating p-values for the above scenario in order to check if they matched our analytical p-value derivation in Section 3.1.1. The algorithm first finds the filtered correlation for the genes bound by the non-HS factor, $g \in F_i$, as described above. Then it picks 10000 random sets of genes of the same size, $|F_i|$, finds the filtered correlation coefficient for each random set and counts the n filtered correlation coefficients that exceed that of factor i. Then, the non-parametric p-value equals $\frac{n}{10000}$. Both approaches achieve very similar p-values, proving the reliability of our method for finding p-values introduced in Section 3.1.1. The next section considers how we can exploit clustering in order to uncover other significant

interactions omitted by our initial thresholds.

5.1.1 Assigning Links: Second Pass

The thresholding approach for finding significant binding relationships introduced in the previous section has an inherent sensitivity to the chosen cutoffs. To alleviate this problem and to minimize the chance of false negatives, this section describes a second pass in our algorithm for assigning links. As Section 4.2.4 illustrated, clustering ChIP-chip data at a stringent significance cutoff (combined p-value $\leq 10^{-70}$) evinces various known biological complexes. Moreover, we noticed in Section 3.3.1 that we can infer the binding relationship between POL3 and RSC3 based on the interactions between POL3 and other components of the RSC complex (i.e. RSC8, RSC). Generally, we can deduce that protein i's binding profile $\mathbf{x}_i$ has a significant relationship with protein j_1's binding profile $\mathbf{x}_{j_1}$ contained in cluster C_j, $\mathbf{x}_{j_1} \in C_j$, if the combined binding relationship between $\mathbf{x}_i$ and all the binding vector components of C_j is significant. Using (3.31), we can find the total p-value for the binding relationship between $\mathbf{x}_i$ and the cluster C_j, or $pv_T(\mathbf{x}_i, C_j)$, by combining the p-values for the individual binding dependencies between $\mathbf{x}_i$ and each component of C_j, where $C_j = \{\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_{|C_j|}}\}$:

$$pv_T(\mathbf{x}_i, C_j) = pv_C(pv(\mathbf{x}_i, \mathbf{x}_{j_1}) \cdots pv(\mathbf{x}_i, \mathbf{x}_{j_{|C_j|}})) = k_{|C_j|} \sum_{r=0}^{|C_j|-1} \frac{(-\ln k_{|C_j|})^r}{r!}\,. \tag{5.7}$$

Equation 5.7 only depends on the number of p-values combined, $|C_j|$, and the product of the p-values $k_{|C_j|} = \prod_{m=1}^{|C_j|} pv(\mathbf{x}_i, \mathbf{x}_{j_m})$, as in (3.31). However, (5.7) incorporates p-values based on binding relationships between independent data sets $\mathbf{x}_i, \mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_{|C_j|}}$. In this context, our use of (3.31) satisfies the independent studies assumption and we do not have to double the p-values as in Section 3.3.1. Moreover, (5.7) does not specify which type of p-values to combine. For a possible positive or negative link, our network algorithm incorprates the combined $pv_C(\mathbf{x}_i, \mathbf{x}_j)$ or filtered correlation coefficient $pv_{FCC}(\mathbf{x}_i, \mathbf{x}_j)$ p-values, respectively. For extending a potential positive or negative link, we consider a total p-value

$pv_T(\mathbf{x}_i, C_j) \leq 10^{-20}$ or $\leq 10^{-10}$ as significant, respectively. With the additional evidence that protein i's binding profile associates significantly with the biologically related components of C_j, the threshold for significant binding relationship between i and j_1 should decrease. Repeating this process for every possible combination of i and j_1, where $j_1 \in C_j$, the second pass of our network algorithm assigns positive and negative edges as follows:

$$\epsilon^{+}_{i,j_1} = 1 \; if \; [pv_T(\mathbf{x}_i, C_j) \leq 10^{-20}] \wedge [(pv_{MMI}(\mathbf{x}_i, \mathbf{x}_{j_1}) \leq 10^{-7.5}) \vee (pv_C(\mathbf{x}_i, \mathbf{x}_{j_1}) \leq 10^{-15})]$$
$$\epsilon^{-}_{i,j_1} = 1 \; if \; [pv_T(\mathbf{x}_i, C_j) \leq 10^{-10}] \wedge [(pv_{FCC}(\mathbf{x}_i, \mathbf{x}_{j_1}] \leq 10^{-7.5})] \,. \tag{5.8}$$

The first pass of the network algorithm found 4692 positive and 2629 negative binding relationships, while the second pass uncovered 806 and 574 more significant binding relationships, respectively. The larger number of positive links reflects the fact that significant positive binding relationships are easier to find using ChIP-chip data than significant negative binding relationships. Actual visualization of all the significant links proves unwieldy. Therefore, Figure 5-1 displays only highly significant positive binding relationships that have a minimized mutual information p-value $\leq 10^{-20}$ or combined p-value $\leq 10^{-40}$. Different colored ovals represent factors from the various levels and lines between ovals denote significant synergistic interactions between factors. Inspecting the graph, one can notice neighborhoods of nodes that share common biological processes. For example, the MCM-ORC cluster occupies the top left corner of Figure 5-1, with the SIR complex directly to its right. Continuing to the right from the SIR complex, one can observe a congregation of nuclear transport proteins, factors from the active cluster, the POL3 machinery, and the RSC complex at the middle right end of the figure. Having built a network of the nucleus, the next section verifies the ability of our network model to predict protein-protein interactions by comparing it with several previous protein-protein studies.

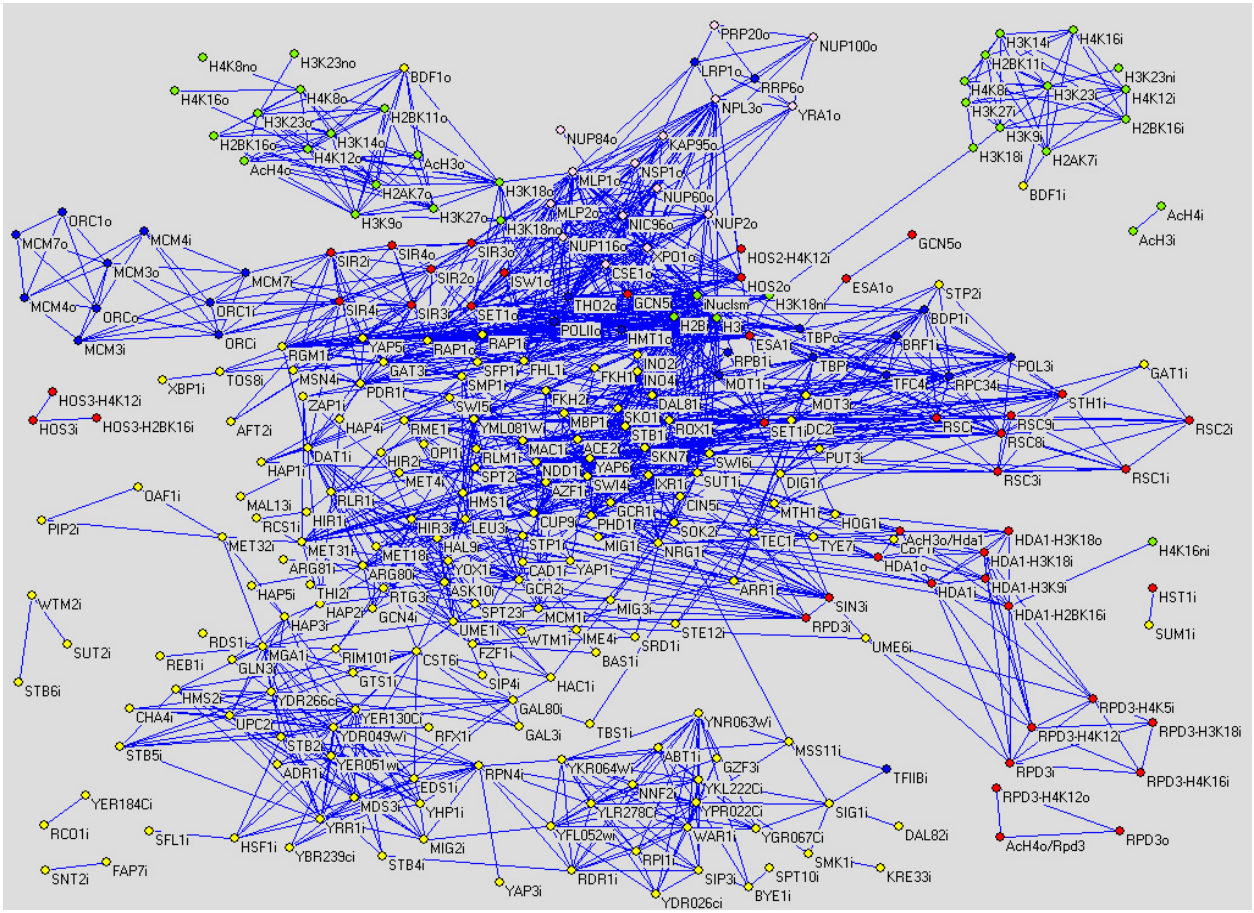


Figure 5-1: Network of highly significant positive binding relationships that have a minimized mutual information p-value $\leq 10^{-20}$ or combined p-value $\leq 10^{-40}$. Colored ovals represent factors from the various levels and lines between ovals denote significant synergistic interactions between factors. Different colors represent the five layers: yellow = TF, green = HS, red = HM, blue = NP, and pink = NT. Graph rendering was performed with Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm).

Data Set	Interactions	Overlap	P-Value
Yu et al.	395	134	7.42×10^{-46}
Gavin et al.	462	129	3.11×10^{-34}
Ho et al.	137	45	9.32×10^{-16}
Ito et al.	43	8	0.0280
Uetz et al.	17	4	0.0521

Table 5.1: Comparison between various protein-protein interaction data sets and our network. The first column lists the source of the data. The second column records the number of predictions from each data set between proteins considered in this work. The third column lists the overlap in predicted significant interactions between the five studies and our network. The last column finds p-values for the significance in the overlap using the hypergeometric test statistic discussed in Section 3.2.1 (see text).

5.1.2 Network Comparison

Numerous previous large scale and small scale methods have predicted protein-protein interactions with varied success. For comparison, we downloaded data from five protein-protein interaction studies [30–34]. Ito et al. and Uetz et al. used an experimental technique called Yeast Two-Hybrid (Y2H) to infer associations between proteins, while Gavin et al. and Ho et al. used a more accurate mass spectrometry method. Yu et al. incorporated the significant predictions from the above data sets along with interactions found using several small scale studies. Small scale studies generally have less noise in their predictions than the high-throughput methods in [31–34]; therefore, the Yu et al. data set is probably most reliable.

Table 5.1 demonstrates the results of the comparison. The first column lists the source of the data. The second column records the number of predictions from each data set between proteins considered in this work. The third column lists the overlap in predicted significant interactions between the five studies and our network. The last column finds p-values for the significance in the overlap using the hypergeometric test statistic discussed in Section 3.2.1. The p-values were calculated by setting u equal to entries in the second column, h to entries in the third column, $v = 3771$, or the number of significant positive protein-protein binding relationships in our network (i.e., excluding edges with factors in the HS layer), and $w = 43956$, or all possible protein-protein relationships that could be predicted.

The table above shows that the three most reliable data sets overlap significantly with the predictions made in our network. The Y2H data does not have enough predictions in common with our data set in order to accurately determine the significance in the overlap. Moreover, the Y2H technique is considered to have the most amount of noise. Generally, we do not expect a full overlap between the methods since they target different biological mechanisms. Our method finds biological complexes that associate at or near DNA but does not capture other cellular protein-protein interactions. For example, most of the above data sets show that protein Hrp1 and Nab2 form a complex but our analysis demonstrates that the two proteins have very different binding profiles. Moreover, our method can capture binding dependencies between proteins that consistently bind to the same gene targets and hence participate in similar biological processes, but that do not necessarily bind at the same time or associate with one another. And finally, it is generally agreed upon that the data sets above have a large amount of noise. In summary, this comparison validates the ability of our method to find protein-protein interactions predicted in previous studies. Having fully specified the network structure enables us to infer the underlying communication between layers in the eukaryotic transcriptional network in the next section.

5.2 Analysis of Network Topology

This section introduces several statistical methods for quantifying the interplay between the layers in our yeast transcriptional network. Table 5.2 shows the number of links within and between the five categories of factors (namely TF, HS, HM, NP, and NT) for the entire and the positive edge network. Each entry within the table represents the total number of links and the percentage of possible links realized (in parenthesis) between factors from the layers listed in the corresponding entry in the first row and column. For example, the first entry in Table 5.2(a) shows that there are 3659 links between pairs of transcription factors, which represents 17.2% of the total possible links that could be realized between all pairs of TFs. Percentage of links realized represents a normalized measure of the connectedness between two layers, which accounts for the number of factors in each layer. Given two layers

Layers	TF	HS	HM	NP	NT
TF	3659(17.2%)	1018(9.28%)	737(7.58%)	409(7.32%)	171(5.16%)
HS	1018(9.28%)	616(44.7%)	776(31.2%)	236(16.5%)	276(32.5%)
HM	737(7.58%)	776(31.2%)	198(18.3%)	208(16.4%)	100(13.3%)
NP	409(7.32%)	236(16.5%)	208(16.4%)	118(33.6%)	97(22.5%)
NT	171(5.16%)	276(32.5%)	100(13.3%)	97(22.5%)	95(79.2%)

(a) Total links and percentage of links captured for the entire network.

Layers	TF	HS	HM	NP	NT
TF	2099(9.84%)	602(5.49%)	519(5.33%)	328(5.87%)	108(3.26%)
HS	602(5.49%)	513(37.2%)	323(13%)	151(10.6%)	138(16.3%)
HM	519(5.33%)	323(13%)	161(14.9%)	178(14%)	85(11.3%)
NP	328(5.87%)	151(10.6%)	178(14%)	113(32.2%)	93(21.5%)
NT	108(3.26%)	138(16.3%)	85(11.3%)	93(21.5%)	87(72.5%)

(b) Total links and percentage of links captured for the positive edge network.

Table 5.2: Total links and percentage of links realized within and between the five layers of the (a) entire and the (b) positive edge transcriptional network. Each entry within each table represents the total number of links and the percentage of possible links realized (in parenthesis) between factors from the layers listed in the corresponding entry in the first row and column. For example, the first entry in Table (a) shows that there are 3659 links between pairs of transcription factors, which represents 17.2% of the total possible links that could be realized between all pairs of TFs.

with n and m factors, respectively, and k mutual interactions the inter-level percentage of links realized is simply $100\frac{k}{nm}\%$. Moreover, if a layer with n elements has l interactions within its layer, the intra-level percentage of links realized is $100\frac{2l}{n(n-1)}\%$.

Table 5.2(a) measures the level of connectedness or communication within and between layers. By examining the percentage of links realized, we see that most of the communication occurs within layers. The diagonal in Table 5.2(a) shows that the five layers have 17.2%, 44.7%, 18.3%, 33.6%, and 79.2% of intra-connectedness. Moreover, inspecting the entries in row 1 of Table 5.2(a) from left to right, we see that the percentages gradually decrease, with TFs being most highly associated with HSs, followed by HMs, NPs, and NTs. These results confirm the current thought in biology that close collaboration between TFs, HSs, and HMs

induces specific classes of genes while NPs and NTs carry out more general mechanisms related to transcription. However, there are still a number of interactions between the levels, illustrating that significant amount of communication between layers of the yeast transcriptional network.

Several other network statistics can also quantify the behavior of the different classes of proteins. For, example we can count the number of links stemming from each node, or the number of degrees, and find a distribution of degrees for nodes within a layer. Computing the average of that distribution measures the average level of connectivity for members from each layer. Table 5.3 shows the average number of degrees for proteins within the five layers in both the entire and the positive edge network. Members of the HS level seem most promiscuous in their association with other factors, but, overall, all layers have a similar number of average degrees.

Clustering coefficient is a common method for measuring the cliquishness of each node in the network, or how much the node's nearest neighbors (adjacent nodes) interact with one another. Finding the average clustering coefficient for all proteins within a particular class measures the tendency for cliquishness in each layer. Let node i of layer L have k_i nearest neighbors. Moreover, let its k_i adjacent nodes have c_i connections between them out of a possible $\frac{k_i(k_i-1)}{2}$ links. Then the clustering coefficient at node i, or the fraction of links realized between i's nearest neighbors becomes $\frac{2c_i}{k_i(k_i-1)}$. Finally, the average clustering coefficient for all factors i in layer L, CC_L, takes the form:

$$CC_L = \frac{1}{|L|} \sum_{i \in L} \frac{2c_i}{k_i(k_i - 1)} \tag{5.9}$$

Table 5.3 shows the average clustering coefficient for proteins within the five layers in both the entire and the positive edge network. The preference for cliquish interaction between nearest neighbors of nodes seems strongest amongst members of the NT class of proteins. Given that almost all nuclear transport factors considered seem to associate with members of the nuclear pore, this measure seems consistent with biological mechanism of the NT class.

In summary, the network topology statistics discussed in this section confirm that the

Layer Name	Avg. Number of Degrees	Avg. Clustering Coefficient
TF	47.2	0.458
HS	68.9	0.427
HM	47.3	0.495
NP	43.9	0.539
NT	52.1	0.631
All	50.5	0.473

(a) Network topology statistics for the entire network.

Layer Name	Avg. Number of Degrees	Avg. Clustering Coefficient
TF	27.8	0.526
HS	42.3	0.482
HM	30.4	0.539
NP	36.1	0.575
NT	37.4	0.676
All	31.4	0.532

(b) Network topology statistics for the positive edge network.

Table 5.3: Network topology statistics for (a) the entire and (b) the positive edge transcriptional network.

TF, HM, and HS levels seem most tightly coupled. However, over a thousand significant interactions also occur between factors outside of these subsets, demonstrating that eukaryotic transcription depends on the intricate interplay between all five layers presented.

Chapter 6

Conclusion

Genome-wide binding data from ChIP-chip experiments provides a wealth of information about protein-gene interactions and about the underlying workings of eukaryotic cells. In order to gain a holistic understanding of the non-linear process of transcription, our work examines the communication between various classes of regulators in the yeast specie *Saccharomyces cerevisiae*. We used ChIP-chip data to quantify the interplay between five categories of factors that affect transcription: histone states, histone modifiers and nucleosome remodelers, transcription factors, nuclear processing proteins, and nuclear transport factors.

Following the introduction in Chapter 1, Chapter 2 described the non-trivial process of incorporating the different sources of ChIP-chip data into a coherent set. Due to the heterogeneity of the obtained data, this chapter further discussed the need for data normalization and showed the benefits of a normalized data set. Chapter 3 used the processed data to find pairwise statistics that test binding dependencies between two proteins. Combining the complementary pairwise measures of filtered correlation coefficient and mutual information p-values reduced the false positive and false negative rates in our biological predictions and increased the reliability of our analysis. Next, Chapter 4 uncovered group-wise relationships between factors using Principal Component Analysis (PCA) and clustering. PCA allowed us to visualize subsets of the data in 2-D. Based on the developed pairwise measures, Chapter 4 also introduced a novel semi-supervised clustering algorithm that preserves information

about elements of a cluster in order to better capture group-wise dependencies between proteins. And finally, Chapter 5 combined the methodology developed in the previous chapters in order to build a multi-layered transcriptional network of the nucleus.

Throughout the theoretical analysis, we validated various known biological processes that occur in dextrose rich conditions. Moreover, our analysis and previous literature [7] showed that usage ChIP-chip experiments in rich media YPD conditions can still provide insight about biological processes in other growth environments. Further, the fact that our data does not synchronize cells at a particular phase of growth, does not preclude formulation of hypothesis about biological processes that occur only in a specific phase of the cell cycle, as shown with the MCM-ORC complex in Section 4.2.4. Our theoretical analysis also uncovered several novel biological hypothesis. In particular, Chapter 3 suggests that proteins LRP1 and RRP6 might participate in a mechanism that ensures quality control at falsely transcribed genes, by binding to inactive genes in order to readily degrade unwanted production of their mRNAs. Moreover, Chapter 4 hypothesized that SIR2, long thought to silence the transcription of DNA, might have a role in transcriptional activation at the ORF region of genes. And finally, Chapter 5 quantified the communication between layers in biological transcriptional networks for the first time in the literature. The network topology statistics discussed in Chapter 5 confirm that the TF, HM, and HS levels seem most tightly coupled. However, over a thousand significant interactions also occur between factors outside of these three categories, demonstrating that eukaryotic transcription depends on the intricate interplay between all five layers presented.

Appendix A

Semi-Supervised Clustering based on KL-divergence

Semi-supervised clustering is more theoretically intuitive when objects are represented using distibutions. We chose to define the probability distribution of object i as the normalized binding profile of binding ratios across all genes, or normalized rows in the BR matrix. Specifically, if $\mathbf{x}_i = [x_{i,g_1} \dots x_{i,g_{|G|}}]$ denotes a row vector from the BR matrix for protein i across all genes g in the set G, the normalized binding profile (or binding distribution) for protein i is

$$P(g|i) = \frac{x_{i,g}}{\sum_{g \in G} x_{i,g}} \,. \tag{A.1}$$

We implemented a semi-supervised clustering algorithm that can perform hierarchical partitioning of distributions as defined in (A.1). We treat the binding distribution $P(g|i)$ for each protein i as an object (or node) and merge objects into clusters (aggregations of nodes) based on a distance metric using the Kullback-Leibler divergence.

The KL-divergence represents how different two distributions are. For the normalized binding profiles, it is defined as

$$D(P(g|C_k)\|P(g|C_k \cup C_l)) = \sum_{g \in G} P(g|C_k) \log \frac{P(g|C_k)}{P(g|C_k \cup C_l)}\,, \tag{A.2}$$

where C_k and C_l represent two clusters and $C_k \cup C_l$ denotes the merged partition of C_k and C_l. In information theoretic terms, the KL-divergence represents the information lost by joining C_k and C_l into a combined distribution for $C_k \cup C_l$. Note that the KL-divergence is not a distance metric; it does not satisfy the triangle inequality nor is it commutative. For the semi-supervised clustering algorithm, we define the following distance measure based on the KL-divergence:

$$d(C_k, C_l) = |C_k| D(P(g|C_k)\|P(g|C_k \cup C_l)) + |C_l| D(P(g|C_l)\|P(g|C_k \cup C_l))\,. \tag{A.3}$$

The above equation (A.3) can intuitively be interpreted as the total information lost in combining the distributions of C_k and C_l into $P(g|C_k \cup C_l)$, where the weights $|C_k|$ and $|C_l|$ account for the number of nodes in each cluster. When a partition C_k contains just a single protein, $|C_k| = 1$. Note that this distance, which we refer to as the KL-distance, is commutative.

The hierarchical clustering proceeds by initially treating all objects as clusters and then successively merging partitions that are "closest", in terms of the KL distance in (A.3). When objects are merged into clusters, the new clusters preserve information about their component nodes in their distribution. The combined distribution is a weighted average of the constituent distributions:

$$P(g|C_k \cup C_l) = \frac{1}{|C_k| + |C_l|} \left(|C_k| P(g|C_k) + |C_l| P(g|C_l)\right)\,. \tag{A.4}$$

The performance of the clustering algorithm is illustrated in Figure A-1. Figure A-1 partitions the nuclear processing and nuclear transport factors and shows similar results as the PCA analysis in Section 4.1. The related NTs cluster together once again. In addition, HMT1 (a protein that adds methyl groups to other proteins), THO2 (a protein that

starts transcription), and YRA1 (a protein that exports mRNA from the nucleus) also cluster tightly with the NTs in [13], suggesting a common regulatory mechanism. Moreover, proteins RRP6, an mRNA surveillance factor, and mRNA export factor NPL3 cluster together, supporting the results in [16] that demonstrate coupling of mRNA quality assurance to mRNA export. The algorithm also captures the relationship between LRP1 and PRP20 discussed in section 3.3.1. However, when the algorithm is run on the entire data, it fails to find commonality between several known biological processes. In the end, the pairwise statistics discussed in Chapter 3 prove much more biologically meaningful than KL-divergence.

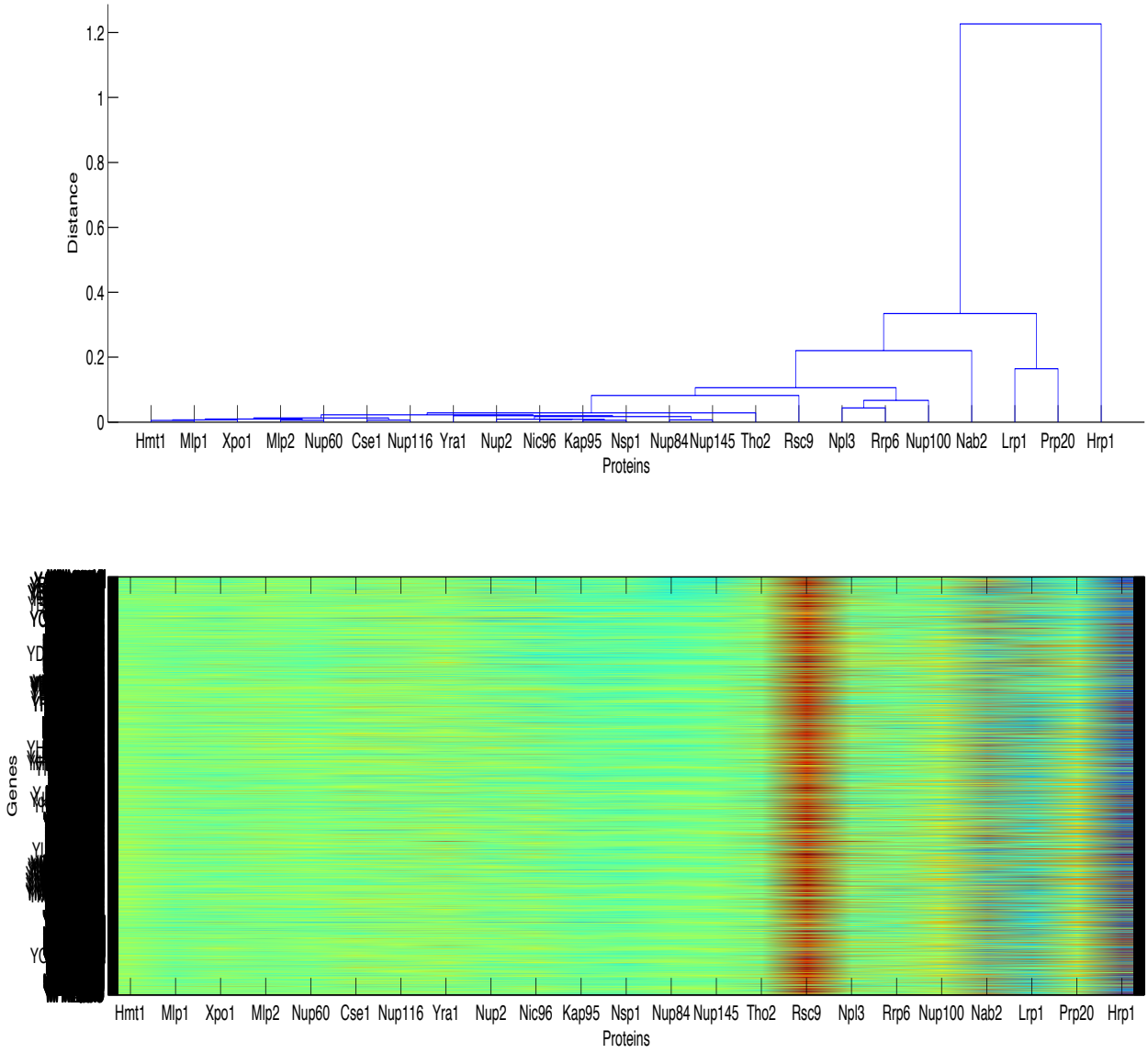


Figure A-1: Semi-supervised clustering of all nuclear transport and nuclear processing factors binding profiles [13–16] using the KL distance metric. The upper plot illustrates the dendrogram. The vertical length of each branch is proportional to the distances between clusters. The lower plot illustrates the vectors of binding ratios of each factor on a scale from 0 to 2. The binding vectors are arranged in the same order as their corresponding protein in the dendrogram.

Bibliography

[1] C. T. Harbinson et al. Transcrptional regulatory code of a eukaryotic genome. *Nature*, 431:99–103, September 2004.

[2] Z. Bar-Joseph et al. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342, November 2003.

[3] T. I. Lee et al. Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science*, 21(11):1337–1342, October 2002.

[4] F. Robert et al. Global Position and Recruitment of HATs and HDACs in the Yeast Genome. *Molecular Cell*, 2004.

[5] J. Zeitlinger et al. Program-Specific Distribution of a Transcription Factor Dependent on Partner Transcription Factor and MAPK Signaling. *Cell*, May 2, 2003.

[6] H. H. Ng et al. Targeted recruitment of Set1 Histone Methylase by Elongating Pol II Provides a Localized Mark and Memory of Recent Transcriptional Activity. *Molecular Cell*, March, 2003.

[7] H. H. Ng et al. Genome-wide location and regulated recruitement of the RSC nucleosome-remodeling complex. *Genes and Development*, 2002.

[8] J. J. Wyrick et al. Genome-Wide Distribution of ORC and MCM Proteins in *S. cerevisiae*: High-Resolution Mapping of Replication Origins. *Cell*, December 14, 2001.

[9] I. Simon et al. Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell*, September 21, 2001.

[10] B. Ren et al. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, December 22, 2000.

[11] C. J. Roberts et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 2000.

[12] F. C. Holstege et al. Dissecting the Regulatory Circuitry of a Eukaryotic Genome. *Cell*, November 25, 1998.

[13] J. M. Casolari et al. Genome-wide Localization of the Nuclear Transport Machinery Couples Transcriptional Status and Nuclear Organization. *Cell*, 117:427–439, May 2004.

[14] M. C. Yu et al. Arginine methyltransferase affects interactions and recruitment of mRNA processing and export factors. *Genes and Development*, 2004.

[15] M. Damelin et al. The Genome-wide Localization of Rsc9, a Component of the RSC Chromatin-Remodeling Complex, Changes in Response to Stress. *Molecular Cell*, 9:563–573, March 2002.

[16] H. Hieronymus et al. Genome-wide mRNA surveillance is coupled to mRNA transport. *Genes and Development*, 2004.

[17] S. K. Kurdistani et al. Mapping Global Histone Acetylation Patterns to Gene Expression. *Cell*, 117:721–733, June 11, 2004.

[18] D. Robyr et al. Microarray Deacetylation Maps Determine Genome-Wide Functions for Yeast Histone Deacetylases. *Cell*, 109:437–446, May 17, 2002.

[19] S. K. Kurdistani et al. Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nature Genetics*, June 24, 2002.

[20] A. Wang et al. Requirement of Hos2 Histone Deacetylase for Gene Activity in Yeast. *Science*, 2002.

[21] J. Lieb et al. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics*, August, 2001.

[22] P. L. Nagy et al. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *PNAS*, May 27, 2003.

[23] J. V. Geisberg et al. Cellular Stress Alters the Transcriptional Properties of Promoter-Bound Mot1-TBP Complexes. *Molecular Cell*, May 21, 2004.

[24] Z. Moqtaderi et al. Genome-Wide Occupancy Profile of the RNA Polymerase III Machinery in *Saccharomyces cerevisiae* Reveals Loci with Incomplete Transcription Complexes. *Molecular and Cellular Biology*, December, 2003.

[25] H. Santos-Rosa et al. Methylation of Histone H3 K4 Mediates Association of the Isw1p ATPase with Chromatin. *Molecular Cell*, November, 2003.

[26] B. E. Bernstein et al. Methylation of histone H3 Lys 4 in coding regions of active genes. *PNAS*, June 25, 2002.

[27] B. E. Bernstein et al. Global nucleosome occupancy in yeast. *Genome Biology*, August 20, 2004.

[28] J. Kim et al. Global Role of TATA Box-Binding Protein Recruitment to Promoters in Mediating Gene Expression Profiles. *Molecular and Cellular Biology*, September, 2004.

[29] N. M. Luscombe et al. Genomic analysis o regulatory network dynamics reveals large topological changes. *Letters to Nature*, September 16, 2004.

[30] H. Yu et al. Genomic analysis of essentiality within protein networks. *TRENDS in Genetics*, June, 2004.

[31] A. C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 2002.

[32] Y. Ho et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 2002.

[33] T. Ito et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *PNAS*, 2000.

[34] P. Uetz et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 2000.

[35] R. J. Larsen, M. L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, 2001.

[36] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[37] T. L. Bailey, M. Gribskov et al. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 1998.

[38] H. W. Mewes. *Munich Information Center for Protein Sequences: Comprehensive Yeast Genome Database*. http://mips.gsf.de/genre/proj/yeast/index.jsp, 2004.

[39] Steen Knudsen. *Guide to Analysis of Dna Microarray Data*. Wiley, 2004