Uncovering Transcriptional Regulators by Combined Relevance Analysis of Genome-Wide Transcription-Factor Binding data and Gene Expression

by

Reina Riemann

B.A., Mathematics, B.A., Biological Sciences, Columbia University
(1998)

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

S.M., Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2003

BARKER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
MAY 1 2 2003
LIBRARIES

©MASSACHUSETTS INSTITUTE OF TECHNOLOGY, MMIII.
All rights reserved.

Author ....
Department of Electrical Engineering and Computer Science
January, 2003

Certified by...................
David K. Gifford
Professor
Thesis Supervisor

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

Uncovering Transcriptional Regulators by Combined Relevance Analysis of Genome-Wide Transcription-Factor Binding data and Gene Expression

by

Reina Riemann

Submitted to the Department of Electrical Engineering and Computer Science
on January, 2003, in partial fulfillment of the
requirements for the degrees of
S.M., Electrical Engineering and Computer Science

Abstract

We present a method for integrating data from genome-wide location of DNA-binding proteins and gene expression to uncover relevant transcriptional regulators without any previous biological knowledge. We show that this method correctly identifies known regulators for the biological processes under study in three expression datasets. While other computational methods have been proposed to suggest transcriptional regulators, most have not integrated genome-wide binding data, which describes physical interactions between transcription factors and the regions upstream from a gene. Our method suggests likely regulators of the biological process under study in an expression dataset and is not limited to coexpressed regulators. The method comprises three steps: (i) assign a relevance score to the binding information of a transcription factor with respect to the correlation of all genes under study in the expression dataset; (ii) calculate the significance of the score; (iii) analyze the over-representation of the number of such significant interactions for a given transcription factor when compared to randomized datasets. This tool allows researchers to uncover those regulators whose genome-wide binding events are predictive of an expression dataset. It suggests the biological processes in which the regulators are active. The synergistic approach of this work represents a valuable tool for researchers in functional genomics.

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

I would like to thank David K. Gifford and Tommi Jaakkola for advice, suggestions and support for this work. I would also like to thank Claire Monteleoni and John Barnett for initial brainstorming and development of the idea; Ziv Bar-Joseph, for continuing discussions and comments; Duncan Odon, for guiding me through the biological datasets and for collaborating on the biological interpretation of the results. Finally, I would like to specially thank Stephen G. Johnson for his help and insightful advice at the different stages of this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We present computational methods for integrating different types of genome-wide data for the discovery of the transcriptional regulators of a biological process under study. This introduction gives the necessary background in order to understand this work in its biological context. In following sections, we present the method, compare it with other related work in computational biology and present the results obtained from applying it to genomic datasets.

1.1 Background

1.1.1 Brief biological background

We first present the minimal biological background needed to understand this work.

Cells are comprised of a variety of molecules whose coordinate behavior form a complex system that is able to adapt to the environment and reproduce. Some of the most important molecules in a cell are DNA *(deoxyribonucleic acid)*, RNA (*ribonucleic acid*) and proteins. The central dogma of biology posits that DNA, considered a stable information storage polymer, is transcribed into RNA, an intermediate short-lived polymer, that may then be translated into proteins. Proteins are the ubiquitous polymers in the cell, involved in most chemical reactions. They can serve as biochemical catalysts, intra- and extra-cellular signalling molecules or as structural elements among other functions.

The flow of genetic information within a cell is a complex multilevel process. DNA is coiled together with other proteins into what is called the genome of an organism. The short DNA segments that are transcribed into messenger RNA are called genes, although the formal definition of a gene is still subject to some controversy. Not all DNA is comprised of coding regions. The so-called intergenic regions may or may not have a known function. It is currently believed that short DNA regions close to genes are involved in recruiting DNA-binding proteins, known as transcription factors, involved in the transcriptional regulation of genes. There are also other known levels of biological regulation. The RNA polymers are subject to degradation, and a single polymer can be translated multiple times. Proteins are additionally subject to post-translational modifications. The understanding of gene regulation is foundational for most biological work.

1.1.2 Genomics

The recent sequencing of the genome of several organisms is revolutionizing biology. These efforts have ushered in the "post-genomic" era, in which we have the possibility to understand organisms from a new perspective. Genomics has fostered several computational subdisciplines, traditionally computational sequence analysis and structural genomics, which have developed different tools and models for the analysis of genomic sequences. A variety of resulting algorithmic proposals address the computational recognition of genes (open reading frames) and the prediction of secondary and tertiary structures of proteins, among other biological problems. Nevertheless, the tools developed for biological sequence analysis are limited by the information held in the sequences, lacking insight into the dynamic nature of the cell and without the perspective offered by other increasingly available sources of biological information. The need for understanding gene regulation and the biological context in which genes are active has spawned the emergence of another genomic field, functional genomics. Functional genomics makes use of data from emerging high-throughput technologies for gathering data at a genome-wide scale. In this work, we combine such datasets in order to extract the transcriptional regulators of various biological processes without applying any prior knowledge of their relevance.

DNA-microarrays have become a particularly widespread high-throughput tool. They provide genome-wide insight into how genes are expressed by giving the simultaneous measurement of the mRNA levels of the transcribed genes under different experimental conditions. Two microarray technologies have emerged: oligonucleutide and cDNA arrays. Nevertheless, these technologies depend on similar principles, relying on the base pairing recognition of complementary nucleutide pairs. Unfortunately, error and noise appear at different levels of the fabrication process, arising from the sequences used as probes on the chip, from experimental variation introduced by the bench scientists, and from the image scanning software, among others.

Microarray expression analysis has already uncovered known and candidate interactions in biological pathways [8]. In addition, these technologies show promise in the diagnosis and treatment of diseases, as well as the understanding of the basic biology of the organism under study [8]. Recently, new research has analyzed the expression levels of genes from different cancer tissue samples, uncovering candidate sets of genes that can serve for cancer diagnosis [11]. Additionally, microarray expression analysis promises to help in identifying drug targets and may eventually revolutionalize pharmacology, forming a new subdiscipline: "pharmacogenomics" [15].

Recently, Ren *et al.* developed a chromatin-immunoprecipitation microarray assay for the genome-wide detection of DNA-binding proteins [16]. This new source of high-throughput data gives additional insight into biological processes by providing knowledge about potential transcriptional regulators of a gene, that is, proteins that bind to its upstream intergenic region. These DNA-binding proteins are known as transcription factors. Transcription factors mediate biological processes in that they can be modified in response to cellular activity, and they influence the expression of target genes. The understanding of gene regulation is fundamental for most biological work.

Ren's assay has been successsfully used to investigate the genome-wide binding of transcriptional factors known to be involved in specific biological processes such as the cell cycle, signaling, stress, and metabolism [18, 21]. Additionally, Young *et al.* have applied Ren's assay in order to characterize the comprehensive binding of the set of all known transcriptional factors in *Saccharomyces Cerevisiae* (brewing yeast) [14]. They collected this genome-wide binding ("location") data from the wild-type strain under normal growth conditions. It is believed that some of these transcription factors are bound to DNA, becoming active under certain biological processes, enviromental conditions, or stresses. Our work analyzes the Young binding dataset combined with various expression datasets in order to uncover transcription factors involved in a given biological process.

1.2 Overview

The present work combines data from two high-throughput technologies DNA microarray expression data and the location of genome-wide DNA-binding proteins to identify potential transcriptional regulators to the biological process studied in the expression dataset. In the following sections of this work we present our method. We contrast this approach with other candidate approaches and present a literature review of previous approaches for combining microarray expression data with genome-wide location of DNA-binding proteins. We then apply the methods to real biological datasets and analyze the biological significance of the results. We conclude that this approach is able to uncover well-known regulators, and represents a valuable tool for functional genomics.

Chapter 2

Method

This chapter presents the motivation of the method and comparison with previous approaches.

2.1 Motivation

Our method combines genome-wide binding and expression datasets in order to uncover transcriptional regulators for which the binding information of a given factor is predictive of the correlation for an over-represented number of genes. Uncovering the regulators of a biological process is a major step in understanding the underlying mechanism: a candidate list of such regulators can guide biologists in the design of experiments in order to confirm their relevance. Biologically speaking, relevance implies that the regulators coordinate the behavior of a number of genes involved in the process. Our identification of an over-represented number of significant interactions for a given transcription factor suggests that it is potentially involved in the biological process under study. This approach is based on three steps:

- assigning a relevance score to the binding information of a transcription factor with respect to the coexpression of genes in the expression dataset.
- calculating the statistical significance of the relevance scores.

- estimating the significance of the number of relevant interactions found for a given transcription factor.

In other words, the method can be understood as a dual-filtering process. The first filter is the significance of the relevance score of a given gene with respect to the binding information of a transcription factor. The second is the significance of the number of such relevant interactions per transcription factor.

In section 2.2, we present formal mathematical definitions for our information-theoretic measure of relevance as well as for our procedure to calculate its significance. Similar ideas have been successfully applied to the discovery of cancer subclasses and in cancer classification by Friedman and Ben-Dor *et al.* [3, 2, 4].

Transcriptional regulators of a biological process may affect only a subset of all the genes (one does not necessarily expect the expression levels of all genes of an organism to be affected by a given biological process) and some of the subsets affected by different transcription factors might overlap. Our method uncovers the transcription factors for which there is a higher number of significant interactions than that obtained by randomization, suggesting that these transcription factors play a role in the regulation of the biological process under study in the expression dataset.

Our technique is tolerant of noise because it identifies transcription factors that simply have more than the expected number of interactions when compared with randomization. Moreover, it does not depend, as other methods do [20], on scoring only single gene-to-gene interactions, an approach that may not be reliable given the noise levels, the limited number of data points in a time series dataset, and the few replicate experiments. Of course, if a factor regulates only a small number of genes or if the genes are not coexpressed, its relevance will be difficult to distinguish from chance. Overall, our method offers a list of candidate regulators with an over-represented number of significant interactions with respect to the expression of genes studied during a given biological process. It is a robust and valuable tool for the discovery of transcriptional regulators and the elucidation of regulatory circuits.

Clustering approaches, which have been successful in the analysis of expression datasets [9], do not offer an appropriate solution for combining different data sources in the discovery of relevant transcriptional regulators. It is important to note that the genome-wide binding data of DNA-binding proteins was collected from the wild-type strain of *Cerevisiae* under normal growth conditions,[1] not necessarily under the conditions in the expression dataset. Moreover, binding *per se* does not imply regulation. Thus, if one clusters only the binding data, the clusters do not offer context-specific regulation information. On the other hand, if one merges the clusters from the binding data together with the clusters of an expression dataset and if the distances between two given elements in the two datasets disagree, there does not seem to be a principled way of resolving the conflict. Furthermore, since the number of transcription factors expected to regulate a given biological process is small, most of the binding clusters (when using all of the binding data from all transcription factors) offer irrelevant information. Even if these obstacles were surmounted, clustering approaches would produce clusters in which the expression and binding data are both highly similar, and would not uncover more subtle relationship in which the expression and binding data are not highly similar. It is necessary to look for other approaches, such as the ones presented in this work, in order to filter out the relevant transcriptional regulators from the set of all candidate transcription factors.

2.1.1 Previous approaches for combining genome-wide binding and expression data

Previously, biologists have performed a limited number of experiments on the genome-wide binding of a small subset of transcription factors already known to regulate specific biological processes [16, 18]. Other research has been stimulated by this available binding data, merging it with additional data sources in order to build regulatory network models. Nevertheless, because of the limited amount of binding data, they could only use transcription factors already known to regulate a given biological process, in contrast with the approach taken in this work that aims at uncovering these regulators themselves.

---

[1] There are plans for experiments in additional conditions, but this data is not yet available.

Hartemink *et al.* build genetic regulatory networks in the *Cerevisiae* pheromone response pathway by using binding data for Ste12 and expression data. They use heuristic search methods combined with Bayesian Networks [13]. Segal *et al.* infer a predictive model of the cell cycle based on probabilistic relational models, merging promoter sequence analysis, transcription factor binding data for the set of all known cell-cycle regulators, and Spellman's cell-cycle gene expression data [17].

The recent availability of the genome-wide binding data for several known transcription factors in *S. Cerevisiae* stimulates the study of computational approaches for uncovering transcription factors that are related to a given biological process. Lee *et al.* [14] use this data in order to build regulatory abstractions that they call "multi-input motifs refined for common expression (MIM-CE)." MIM-CEs are comprised of genes that share both a strong correlation in the expression datasets and a subpattern of bound transcription factors in the location dataset within some very stringent thresholds. They apply their algorithm to the available genome-wide location data and to expression data from over 500 experiments, selecting those MIM-CEs that have a significant number of cycling genes from Spellman's cell-cycle expression dataset[19]. They discover known cell-cycle regulators from these enriched MIM-CEs.

In contrast, our approach to uncover regulators does not rely on the use of a large number of expression datasets. We are able to recover several known transcriptional regulators by combining only a given time-series dataset with the comprehensive genome-wide transcription-factor location dataset. This is a valuable tool for researchers in functional genomics. In addition, one could envision that applying our methods to uncover relevant regulators can serve as a preliminary step for automatically building regulatory networks with models such as Bayesian networks and probabilistic relational models. Nevertheless, those tools by themselves would not be able to uncover the regulators due to the combinatorial explosion of the search space.

2.1.2 Previous relevance and overabundance analysis of high-throughput microarray data

Researchers have previously successfully applied relevance scores in order to uncover relevant biological information from high-throughput microarray data. Butte *et al.* use pair-wise mutual information to analyze microarray datasets and threshold the results to build relevance networks which they find to be biologically meaningful [6]. Xing *et al.* apply a mixture of filter and wrapper methods using information-gain ranking and Markov-blanket filtering in order to select genes that are informative in cancer classification [22]. They further extend their method combining information-theoretic feature selection with spectral clustering in an algorithm, CLIFF [23]. Particularly influential to this work has been a series of papers by Nir Friedman and Ben-Dor *et al.*, in which they develop scores for ranking gene relevance using mutual information and threshold number of misclassification. They evaluate the significance of their scores by calculating p-values, and apply their method in several papers in tissue and cancer classification [3, 2, 4, 5].

2.2 Information-theoretic relevance score and significance calculations

This section describes the measure of relevance used in this work. We use an information-theoretic score, $Info$ (information gain) for short. The following sections present our notation, the mathematical formulation of the score and the procedure to calculate its statistical significance as well as the significance of the number of such relevant interactions.

Let B be a binary matrix with elements representing significant binding events of transcription factors to genes. Specifically, $B_{ij} = 1$ whenever transcription factor i binds to gene j at the cut-off p-value of 0.001, the threshold that the experimentalists found to be significant [14]. We denote as B_i the vector of significant binding events between transcription factor i and all the genes in the dataset. Let PC be another binary matrix with elements representing significant pairwise correlation between the expression vectors of the genes. The significance of the correlation between two expression vectors was calculated by randomizing the time points. Specifically $PC_{ij} = 1$ whenever gene i is significantly correlated with gene j at the cut-off p-value of 0.001. We denote as PC_j the vector of significant correlations between gene j and all the genes in the dataset.

2.2.1 Transcription-factor information-theoretic relevance

The information gain between two random variables X and Y is $H(X) - H(X|Y)$, where $H(X) = E[-logP(X)]$ is the entropy of X and $H(X|Y) = E[-logP(X|Y)]$ is the conditional entropy of X given Y. Let $Info_{B,PC}(i,j)$ be a real-valued matrix whose elements are the information gain between B_i and PC_j. Informally, $Info_{B,PC}(i,j)$ is indicative of how predictive the binding data of transcription factor i is, in an information-theoretic sense, of the correlation of gene j with all other genes.

Formally,

$$Info_{B,PC}(i,j) = H(PC_j) - H(PC_j|B_i)$$

We calculate $Info_{B,PC}(i,j)$ by using the empirical distribution induced by B_i and PC_j, thus defining two jointly distributed random variables in the following way. Recall that B_i is the vector comprised of the significant binding events $B_{ig} \forall g$, and that PC_j is the vector comprised of the significant correlations $PC_{jg} \forall g$. For $b_1 \in \{0,1\}$ and $b_2 \in \{0,1\}$, let $N_{i,j}(b_1,b_2)$ be the number of genes g for which $B_{ig} = b_1$ and $PC_{jg} = b_2$. We then define the empirical distribution $P_{B_i,PC_j}(b_1,b_2) = \frac{N_{i,j}(b_1,b_2)}{G}$, where G is the number of genes in the dataset.

For example, if $B_i = \langle 1,0,1,0,1,1 \rangle$ and $PC_j = \langle 0,1,0,1,0,1 \rangle$ for some transcription factor i and gene j, then this implies that $G = 6$ and that $N_{i,j}(1,1) = 1$, $N_{i,j}(1,0) = 3$, $N_{i,j}(0,0) = 0$, $N_{i,j}(0,1) = 2$. Therefore, $P_{B_i,PC_j}(1,1) = \frac{1}{6}$, $P_{B_i,PC_j}(1,0) = \frac{3}{6}$, $P_{B_i,PC_j}(0,0) = 0$ and $P_{B_i,PC_j}(0,1) = \frac{2}{6}$. This allows us to calculate $H(PC_j) = 1$ and $H(PC_j|B_i) = 0.54$. From these values we calculate $Info_{B,PC}(i,j) =$

0.46.

2.2.2 Statistical significance of the *Info* score

We assign a statistical significance, or "p-value," to the $Info_{B,PC}(i,j)$ score by estimating the probability of obtaining the given value of the score at random. In order to calculate a p-value, we randomly permute the vector of pair-wise correlation coefficients for gene j, PC_j. Let us denote the vector drawn from the distribution of random permutations of PC_j as RPC_j. Then, we calculate the number of times that the score $Info_{B,RPC}(i,j)$ is greater than or equal to the original score between B_i and PC_j.

Formally:

$$p(Info_{B,PC}(i,j)) = \\ Pr(Info_{B,RPC}(i,j) \geq Info_{B,PC}(i,j))$$

Informally, the lower $p(Info_{B,PC}(i,j))$, the less likely it is to get $Info_{B,PC}(i,j)$ by chance, which suggests that interaction between transcription factor i and gene j is significant.

2.2.3 Statistical significance of the transcription factor over-representation

We define the number of significant interactions for B_i with genes in the dataset as the integer valued matrix $N_{B,PC}(i) = \sum_j [(Pr(Info_{B,RPC}(i,j) \geq Info_{B,PC}(i,j)) \leq 0.001]$. In order to assign a p-value to $N_{B,PC}(i)$, we calculate the number of times that the number of significant interactions for B_i was higher in the randomizations in the following way.

$$
\begin{aligned}
&p(N_{B,PC}(i)) = \\
&\quad Pr[\Sigma_j((Info_{B,RPC}(i,j) \geq Info_{B,PC}(i,j))) \\
&\quad \geq N_{B,PC}(i)]
\end{aligned}
$$

Informally, the lower the $p(N_{B,PC}(i))$, the less likely it is to get $N_{B,PC}(i)$ at random, which suggests that transcription factor i regulates a larger number of genes than expected by chance.

Chapter 3

Empirical Evaluation

The comprehensive genome-wide location data was collected in rich-media conditions. It is unclear how to extract the biological regulators of a given biological process without the use of additional datasources. We apply the information-theoretic analysis described in Section 2.2 in order to combine different expression datasets with the comprehensive genome-wide location of the DNA-binding protein dataset. We use three expression datasets, Spellman's cell-cycle expression dataset [19], Gasch's oxidative-stress-response dataset [10], and Causton's oxidative-stress-response dataset [7] in combination with Young's comprehensive DNA-binding of transcriptional regulators [14]. First, we calculate the $Info$ score of the transcription factors with respect to the correlation vector of each gene in the expression datasets. Then, we calculate the statistical significance of the $Info$ scores, as described in Section 2.2.2, obtaining the distributions shown in Figure 3-1. There, we find that there are few transcription factors with a large number of significant interactions.

The information-theoretic relevance analysis yields significant biological results by uncovering known regulators of the biological processes studied in the expression datasets. Listing the top transcription factors according to the number of significant interactions in Table 3.1, the following known cell-cycle regulators (**Swi6**, **Ndd1**, **Mbp1**, **Fkh2**, **Mcm1**, **Swi4** and **Swi5**) appear among the first few regulators. Seven out of the nine known cell-cycle regulators [18] are thereby uncovered without the use of any previous biological knowledge. Our method additionally uncovers

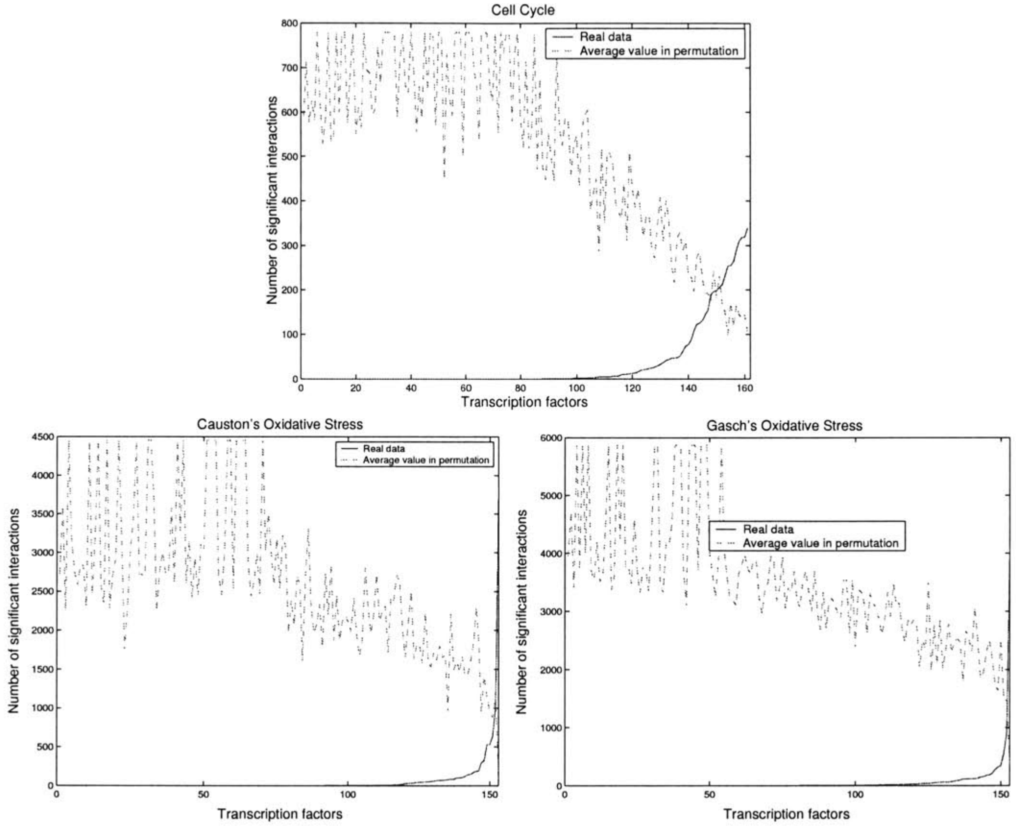

Figure 3-1: Number of significant interactions for each transcription factor for cell-cycle and hyperoxia expression datasets.The transcription factors were sorted by the number of relevant interactions. Note that only a few transcription factors are significantly predictive of the expression of a large number of genes in the cell cycle and oxidative-stress datasets.

Spellman's Cell Cycle	
Ntf	tf
336	Gat3
324	**Swi6**
318	**Ndd1**
316	**Mbp1**
296	**Fkh2**
261	Yap5
251	Mac1
235	Cup9
213	Rgm1
208	Msn4
206	**Mcm1**
195	Pdr1
180	**Swi4**
161	Zap1
144	**Swi5**

Table 3.1: Top transcription factors for the cell cycle dataset, sorted according to their number of information-theoretic relevant interactions with respect to the respective expression dataset. Known regulators are shown in bold. These relationships were uncovered without the use of any previous knowledge of their biological relevance to the processes.

other transcription factors. For the genes that are in non-bold type, the Yeast Proteome Database (YPD) [1] suggests that Mac1 is involved in stress response, in copper binding, and in cell growth and maintenance; Yap5 and Msn4 are involved in stress response. One possible explanation for the significance of the non-bold type regulators is that given the experimental process to which the yeast cells were subjected, namely cell-cycle arrest, one might expect that stress-response mechanisms were activated. Gat3 has no specific biological function listed in YPD; Pdr1 is involved in drug susceptibility and resistance; Cup9 is involved in copper homeostasis and in the regulation of peptide import; Zap1 is involved with zinc homeostasis. This suggests that cell-cycle regulation might play a relevant role in homeostasis. Transcription factors whose functions are imperfectly understood become candidate genes for future experimentation to understand their potential involvement in the cell cycle.

We apply the same information-theoretic relevance analysis to two independent hyperoxia expression datasets [7, 10]. Listing the top transcription factors according

Causton's Hyperoxia		Gasch's Hyperoxia	
Ntf	*tf*	*Ntf*	*tf*
3872	Fhl1*	4092	Fhl1*
1007	**Yap1***	1393	**Cad1***
638	Swi4	912	**Yap1***
528	**Hsf1***	552	**Yap5**
526	Rgm1*	345	Swi5
496	**Cad1***	319	Bas1
323	**Cin5**	275	Ndd1
294	Abf1	212	Phd1
191	Leu3	183	Rgm1*
185	Stb1	179	Gcn4
174	**Skn7**	153	**Pdr1**
149	Gal4	148	**Mac1**
141	Ste12	148	**Hsf1***

Table 3.2: Top transcription factors from the location datasets, sorted according to their number of information-theoretic relevant interactions with respect to the respective expression dataset. Known regulators are shown in bold. Transcription factors marked with asterisks are present in the lists for the two oxidative-stress response expression datasets. These relationships were uncovered without the use of any previous knowledge of their biological relevance to the processes.

to the number of significant interactions in Table 3.2, the following known stress-response regulators (**Yap1***, **Yap5**, **Cad1***, **Skn7**, **Mac1**, **Hsf1***, **Pdr1**) [10, 7] appear among the regulators in the ranking, as can be seen in Table 3.2. **Yap1***, **Skn7** and **Cad1*** are among the few well known oxidative-stress-response regulators. The method additionally uncovers other transcription factors. Fhl1 is the top regulator in both data sets. Recently, it has been reported that Fhl1 is *the* central regulator responsible for ribosomal gene regulation [14]; inspection of both hyperoxia datasets used here shows that almost every ribosomal mRNA level is strongly repressed after exposure to hydrogen peroxide [10, 7]. Other regulators appear to be involved in numerous processes including for example nuceotide metabolism (Bas1), cell-cycle (Swi5, Ndd1), and filamentous growth (Phd1). The method also identifies numerous stress-response factors. This result suggests that, consistent with previous reports, stress-response mechanisms might share a similar pattern of expression regardless of the stimulus [10, 7]. After applying out method to the two oxidative-stress response

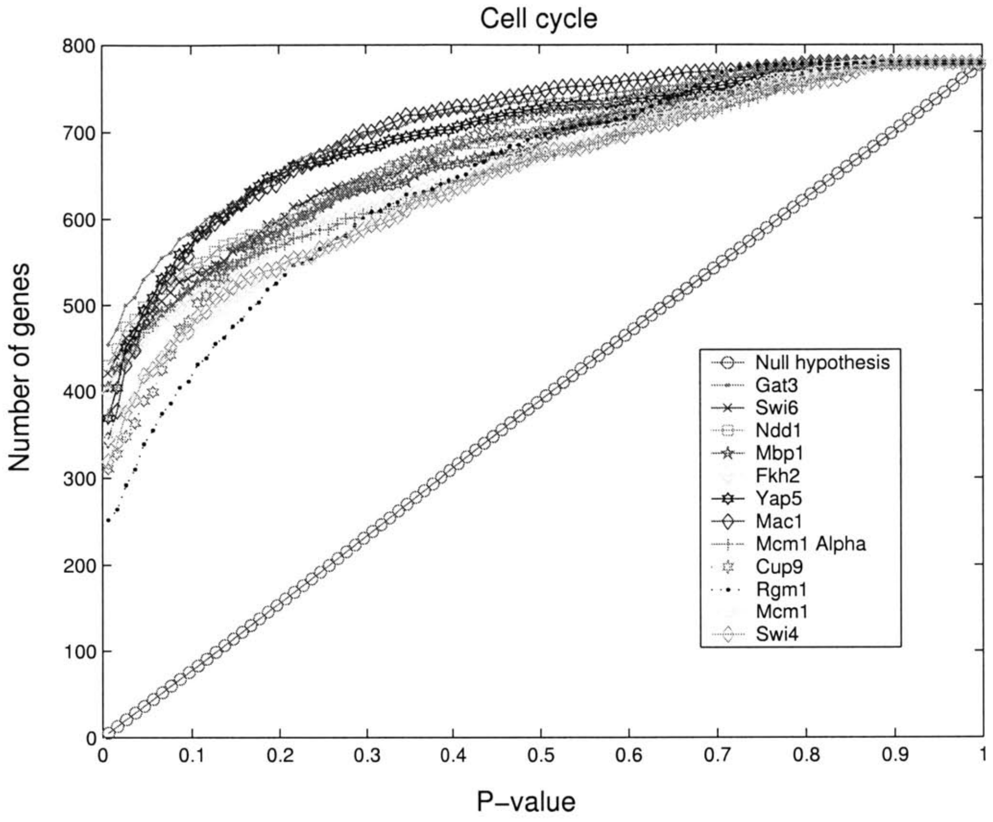

Figure 3-2: The number of genes influenced by a transcription factor at a given p-value is higher than that predicted by the null hypothesis (random labels). The figure shows plots from all the transcription factors that were found to be relevant.

expression datasets, we note that over a third of the top transcription factors are shared by both lists. These common transcription factors are marked by asterisks in Table 3.2. Their apparent relevance in two completely independent datasets is a strong argument for their involvement in oxidative-stress response.

Furthermore, we applied our over-representation analysis in order to obtain a final list of candidate regulators with a number of significant interactions larger than that obtained in the randomization. For the cell-cycle, all known transcription factors, except for Swi5, appear in the final list of regulators that can be seen in Table 3.3. We observe in Figure 3-2 that, for the transcription factors in the final candidate list for the cell cycle, there are a large number of relevant interactions with respect to the genes in the the cell-cycle expression dataset, further confirming their relevance.

Over-representation analysis of the hydrogen peroxide expression datasets reveals

$p(Ntf)$	Ntf	tf
.001	336	Gat3
.001	324	**Swi6**
.001	318	**Ndd1**
.001	316	**Mbp1**
.001	296	**Fkh2**
.001	261	Yap5
.001	251	Mac1
.001	235	Cup9
.001	213	Rgm1
.001	206	**Mcm1**
.001	180	**Swi4**

Table 3.3: This table shows the list of transcription factors that were found to have a significant number of information-theoretic relevant interactions with respect to the cell-cycle expression dataset. Known cell-cycle regulators are shown in bold.

one transcription factor, Fhl1 (see discussion above), which shows a highly significant number of relevant interactions when compared to those from the randomization. The remaining factors listed in Table 3.2 show a large number of significant interactions when compared to the other transcription factors in the dataset.

Chapter 4

Conclusions

This thesis presents methods for uncovering transcriptional regulators by combining the comprehensive genome-wide binding dataset of transcription factors with a given expression dataset. Our technique recovered known biological regulators for both the cell-cycle and oxidative-stress-response expression datasets. Such a tool has the ability to impact functional genomics by suggesting which transcription factors may be involved in regulating the biological process studied in an expression dataset.

It is remarkable that the cell-cycle analysis largely reiterated the transcription factors known to be involved in the cell cycle. Many of the remaining factors that we identified may be induced as a result of stress caused by the method used to obtain cell-cycle synchrony. Similarly, a comparable analysis on the hydrogen peroxide datasets revealed eight factors previously suggested to be involved in either oxidative response or a related stress response.

Importantly, however, the candidates identified under oxidative-stress response do not make the same rigorous over-representation analysis as do the cell cycle regulators. One explanation of this may be that oxidative-stress regulators are not be fully active under rich media conditions. In contrast, the cell cycle regulators are known to be fully bound and functional under the rich-media conditions used to obtain the genome-wide binding data[18]. Much of the regulatory network responsible for stress-response is quiescent under nonstressful conditions, like rich media[12]; and many stress-related transcription factors may not be present in the location analysis datasets because of the mechanisms used to regulate their function. For example, Msn2 and Msn4 are known to be largely excluded from the nucleus under unstressed conditions[12]. Thus, it is remarkable that our analysis identified numerous known stress-response genes. In particular over a third of the top transcription factors are common for both datasets, which were collected completely independently. These common factors can serve to initiate further study.

We further predict that complete characterization of the genome-wide binding of these regulators under the relevant conditions, when combined with expression analysis using our method, will allow us to identify regulators in those conditions. Furthermore, one can envision that when similar datasets from humans or from organisms used to model certain diseases become available, this tool could have a major impact on human health by identifying the regulators that affect the conditions under study.

Some possible extensions of this work include the study of the combinatorial interactions of these transcriptional regulators; that is, when several transcription factors jointly coordinate the behavior of a group of genes in a given biological process. A possible approach could be to study the overlapping genes that have significant scores with respect to the candidate regulators. These candidate combinatorial interactions suggest protein complexes or mechanisms by which the regulators jointly affect the regulation of the biological process. Another possible extension of this work would be the derivation of a closed form for the p-value calculations of the information-theoretic score. This would avoid the need for repeated randomizations of the dataset in order to obtain the significance values of the scores. Finally, one could investigate alternative statistical measures of relevance, correlation and over-representation.

Bibliography

[1] http://www.incyte.com/proteome/YPD/.

[2] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. In *RECOMB*, pages 54–64, 2000.

[3] A. Ben-Dor, N. Friedman, and Z. Yakhini. Scoring genes for relevance. Technical report, Agilent Labs, 2000.

[4] A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data. In *RECOMB*, pages 31–38, 2001.

[5] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. NSampas, E. Dougherty, E. Wang, F. Marincola, Gooden C., J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–40, Aug 3 2000.

[6] A. Butte and I. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29, 2000.

[7] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young. Remodeling of yeast genome

expression in response to environmental changes. *Mol Biol Cell*, 12(2):323–37, February 2001.

[8] M. E. Chicurel and D. D. Dalma-Weiszhausz. Microarrays in pharmacogenomics - advances and future promise. *Pharmacogenomics*, 3(5):589–601, September 2002.

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

[10] A. P. Gasch and M. Werner-Washburne. The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics*, 2(4-5):181–92, September 2002.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 15 1999.

[12] W. Gorner, E. Durchschlag, M. T. Martinez-Pastor, F. Estruch, G. Ammerer, B. Hamilton, H. Ruis, and C. Schuller. Nuclear localization of the c2h2 zinc finger protein msn2p is regulated by stress and protein kinase a activity. *Genes Dev*, 12(4):586–97, Feb 15 1998.

[13] A. J. Hartemink, D. K. Gifford, and T. Jaakkola. Combining location and expression data for principled discovery of genetic regulatory network models. In *PSBC*, 2001.

[14] T.I. Lee, N.J. Rinaldi, F. Robert, D. T. Odom, , G. K. Gerber, Gifford D. K., and R. A. Young. A transcriptional regulatory network map for saccharomyces cerevisiae. *Science*, 2002. in press.

[15] M. Nees and C. D. Woodworth. Microarrays: spotlight on gene function and pharmacogenomics. *Curr Cancer Drug Targets*, 1(2):155–75, August 2001.

[16] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, Dec 22 2000.

[17] E. Segal, Y. Barash, I. Simon, N. Friedman, and D. Koller. Class discovery in gene expression data. In *RECOMB*, pages 31–38, 2001.

[18] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106(6):697–708, Sep 21 2001.

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

[20] E. P. van Someren, L. F. Wessels, E. Backer, and M. J. Reinders. Genetic network modeling. *Pharmacogenomics*, 3(4):507–25, July 2002.

[21] J. J. Wyrick, J. G. Aparicio, T. Chen, J. D. Barnett, E. G. Jennings, R. A. Young, S. P. Bell, and O. M. Aparicio. Genome-wide distribution of orc and mcm proteins in s. cerevisiae: high- resolution mapping of replication origins. *Science*, 294(5550):2357–60, Dec 14 2001.

[22] E. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. In *PSBC*, 2001.

[23] E. P. Xing and R. M. Karp. CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. In *ISMB*, 2001.