מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

Master of Science

By

Yael Garten

Computational extraction and visualization of regulatory signals in the yeast transcriptional network from genome-wide data

Research conducted under the supervision of

Dr. Yitzhak Pilpel

February, 2005

Submitted to the Scientific Council of the

Weizmann Institute of Science

Rehovot, Israel

מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

חבור לשם קבלת התואר

מוסמך מדעים

מאת

יעל גרטן

נושא

בהדרכת:

ד"ר יצחק פלפל

פברואר, 2005

מוגש למועצה המדעית של

מכון ויצמן למדע

רחובות, ישראל

Abstract 1

1 INTRODUCTION 3

1.1 TRANSCRIPTION REGULATION 3
1.2 GENE EXPRESSION DATA 4
1.2.1 DNA Microarrays 4
1.2.2 Various experimental conditions 8
1.2.3 Noise in expression data 9
1.2.4 Role of gene expression data in unraveling regulation 9
1.3 GENOME-WIDE "LOCATION DATA" 10
1.3.1 Background 10
1.3.2 Description of the location analysis dataset 12
1.3.3 Noise in the location data 16
1.4 MOTIVATION FOR THIS WORK 17

2 RESULTS 19

2.1 EXPRESSION COHERENCE OF A GENE SET 19
2.2 DECOMPOSITION OF DATA VIA CLUSTERING 22
2.2.1 QT_clust clustering algorithm 24
2.2.2 QT_clust applications: expression- and sequence-based clustering 25
2.3 FILTERING EXPRESSION NOISE FROM THE LOCATION DATA 26
2.3.1 The p-value trade-off in the original location data 26
2.3.2 Method (i): Decomposition of gene expression profiles 27
2.3.3 Method (ii): Regulatory motifs analysis 35
2.3.4 Method (iii): Synergistic interactions between TFs 37
2.3.5 Method (iv): Co-localization of TFs in shared promoters 41
2.3.6 Relationship among the four methods of filtration 43
2.3.7 Interactive GUI on web server 47
2.3.8 New location analysis dataset by Harbison et al. 48
2.3.9 Comparison of our work to other studies 51
2.4 REGULATORY MOTIF DICTIONARIES PROJECT 55
2.4.1 Graphical User Interface for viewing dictionary data 56

3 METHODS 63

3.1 MRNA EXPRESSION DATA 63
3.2 LOCATION DATA 63
3.3 ALIGNACE, SCANACE, AND GROUP SPECIFICITY SCORE 64
3.3.1 AlignACE 64
3.3.2 MAP score 65
3.3.3 ScanACE 65
3.3.4 Group specificity 66
3.4 CALCULATING THE FALSE DISCOVERY RATE 66
3.5 STATISTICAL SIGNIFICANCE OF THE EC SCORE 72
3.6 STATISTICAL SIGNIFICANCE OF TF SYNERGIES 73
3.7 CLUSTERING PARAMETERS 73
3.8 COMPARISON OF QT_CLUST TO ADAP_CLUSTER 74
3.9 DICTIONARY-GENERATION PROCEDURE 77
3.9.1 Exhaustive genome scan 77

3.9.2 Score the k-mers and FDR *77*
3.9.3 Cluster the motifs *78*

4 DISCUSSION 79

4.1 FUTURE DIRECTIONS 82

5 REFERENCES 84

6 APPENDICES 88

6.1 APPENDIX A 88

Abstract

One of the most interesting and essential areas of biological research today is the study of transcriptional regulation of organisms. On a systems level, much remains to be understood about the network of regulatory functions that control cells' transcriptional responses to changes in their environment. What makes cells behave differently under different conditions? Who are the key regulators of the genome? Do they work alone, or in combinations to achieve higher-order coordination and fine-tuned regulation?

In recent years, novel genome-wide technologies have allowed high-throughput measurement of gene expression and of specific binding of proteins to DNA. Additionally, a variety of computational algorithms have been devised to mine genomic sequence for conserved regulatory motifs in co-regulated genes. These advances allow us now to ask such questions, utilizing the large datasets that have emerged from these studies. Great power lies in the combination of these datasets with one another, for the purpose of enhancing our understanding of transcriptional regulation.

However, because these new technologies are often quite noisy computational methodologies must be developed to extract signal amidst the noise. Additionally, useful, intuitive, interactive visualization tools must be developed in order to allow biologists and computer scientists alike to analyze the huge datasets and become familiarized with the challenges and biological insights lying within them.

The present study achieves exactly that: we have developed a computational methodology which combines genome-wide gene expression data, sequence data, and a comprehensive set of *in*-vivo transcription factor-DNA binding information in *S. cerevisiae*. This combination allowed us to provide a well-supported set of connections between regulators and regulatory genes, which we hope will become widely used by the community. We have also developed novel techniques that allow us to predict synergies between pairs of transcription factors which regulate common sets of genes. With this information, we were able to construct a network of regulatory connections within the yeast, and extract interesting biological results. We also developed visualization tools which allow one to view the datasets mentioned above in an

intuitive user-friendly manner, either to study specific genes of interest, or to browse the data in its entirety to grasp a more global picture of the regulatory controls functioning in yeast. These tools have been made publicly available on the world-wide web.

Most results described here were recently published in *Nucleic Acids Research* in an article titled "Extraction of transcription regulatory signals from genome-wide DNA-protein interaction data" by Garten, Kaplan and Pilpel, January 2005.

In recent years great advances have been made in our understanding of the transcriptional regulatory networks that control gene expression *S. cerevisiae*. The development and use of a number of genomic tools, including expression analysis and genome-wide location analysis, has stimulated this progress. Additionally, a variety of computational algorithms have been devised to mine genomic sequence for conserved regulatory motifs in co-regulated genes. The “location data" in yeast is a comprehensive resource that provides transcription factor-DNA interaction information *in-vivo*. Here we provide two contributions: firstly, we developed means to assess the extent of noise in the location data, and consequently for extracting signals from it. Secondly, we couple signal extraction with better characterization of the genetic network architecture. We apply two methods for detection of combinatorial associations between transcription factors, the integration of which provides a global map of combinatorial regulatory interactions. We discover the capacity of regulatory motifs and transcription factor partnerships to dictate fine-tuned expression patterns of subsets of genes, which are clearly distinct from those displayed by most genes assigned to the same transcription factor. Our findings provide carefully prioritized, high-quality assignments between regulators and regulated genes and as such should prove useful for experimental and computational biologists alike. We have also made available an interactive graphical user interface which allows the visual integration of sequence, expression, and binding data in an intuitive, interactive manner.

Most results described here were recently published in *Nucleic Acids Research* in an article titled "Extraction of transcription regulatory signals from genome-wide DNA-protein interaction data" by Garten, Kaplan and Pilpel, January 2005.

1 Introduction

1.1 Transcription Regulation

The genomes of most organisms contain thousands of genes, each of which has its own specific program of control. The complexity is immense: a large variety of organisms exist, with protein-coding and non-coding genes, which undergo transcriptional, translational, and post-translational control, as well as other types of control. We will focus here on transcriptional control. Genome sequences play a key role in specifying the gene expression programs that produce and maintain living cells, but the way in which cells control global expression programs is far from understood. The specificity of the programs controlling gene expression is essential to the proper functioning of the organism. These transcriptional control programs are modified as cells respond to various changes in the external environment, as they progress through the cell cycle, and during organismal development (DeRisi et al. 1997; Cho et al. 1998; Spellman et al. 1998; Gasch et al. 2000; Causton et al. 2001). Much of the specificity of these programs is affected by sequence-specific DNA binding proteins that bind to the proximal promoter and distal transcriptional regulatory regions. These sequence-specific DNA-binding transcription factors (TFs) interpret and transmit the information encoded in the DNA sequence to the various factors and cofactors that mediate RNA transcript synthesis from the DNA template. In this way, TFs function as the key interface connecting the vast array of genetic regulatory information encoded in the genome and the transcription system.

Transcription is an extremely complex process. It relies on the cooperative action of many components: the TFs with their binding sites along the DNA, the TFs with the RNA polymerase II transcriptional machinery, many coregulators that associate the DNA binding factors with the transcriptional machinery,

chromatin-remodeling factors that mobilize the nucleosomes, and many enzymes that catalyze the covalent modification (such as methylation, phosphorylation, acetylation) of histones and other proteins (Kadonaga 2004). Each of these components exert control and constraints on the system, and only in understanding each individually, and then as a network of interactions between regulation levels, will we truly grasp the underlying control mechanisms. We must eventually understand how all these factors work in concert to potentiate the transcriptional signals that emanate from the TFs.

An important question that arises is how transcription factors actually work. Current evidence indicates that the TFs function mainly by recruiting transcriptional coactivators and corepressors to the DNA template via protein-protein interactions (Ptashne and Gann 1997). These cofactors then act directly and indirectly, in order to regulate the activity of the RNA polymerase II transcriptional machinery at the core promoter. It appears that TFs also recruit chromatin-remodeling factors and histone-modifying enzymes, which in turn function to rearrange chromatin structure as well as to modify histones in a specific fashion that promotes the desired gene activation or repression.

The binding of a TF to the regulatory region of a gene may cause either activation or repression of the gene. In addition, more complex behaviors may be dictated by combinatorial interaction of sets of two or more transcription factors binding to gene promoters. Since individual TFs may bind generally to DNA, precise control of complex gene transcription programs is achieved by the binding of combinations of TFs to DNA. In this way, even though a single recognition site may be common in the genome, composite binding sites will be rarer. The current work begins with the study of the effects on gene expression of binding of single TFs, and then continues to investigate TF combinatorial interactions of pairs of TFs.

1.2 Gene Expression Data

1.2.1 DNA Microarrays

DNA microarrays are powerful tools to study gene expression. The primary use of DNA array technologies is gene expression monitoring. Arrays of nucleic

acids have been used for many years, but only in the last few years has it become possible to miniaturize nucleic acid arrays, and monitor the abundance of tens of thousands of mRNA molecules simultaneously (Lockhart et al. 1996). The ability to look at an enormous number of genes in parallel gives a broad viewpoint.

A DNA microarray (also known as a gene expression array) is a wafer similar to a computer chip, on which a densely packed array of thousands of defined DNA sequences is printed. RNA extract from a cell is converted to cDNA and hybridized to the microarray, in order to measure the expression level of the corresponding mRNAs. The microarrays often used in genome-wide expression experiments contain probes complementary to nearly all possible mRNA molecules of the organism of interest (alternatively spliced variants are not always known and thus not always probed specifically). Thus, a snapshot of the entire cell transcriptome is obtained at a particular time point, in a particular experimental or natural condition.

Once the cDNA is hybridized to the chip, in order to derive biologically meaningful results from the hybridization intensities measured, the intensity values corresponding to each transcript is summarized into one number, representing the amount of bound mRNA transcript that was measured in the experiment.

Although many different microarray systems have been developed by academic groups and commercial suppliers, the most commonly used systems today can be divided into two groups, according to the array material: complementary (cDNA) and oligonucleotide microarrays (see Figure 1). In the present study, we used yeast expression data gathered by both cDNA microarrays and oligonucleotide microarrays.

1.2.1.1 cDNA microarrays

Array preparation: Probes for cDNA arrays (double strand DNA at average size of 1000 mer) are usually products of a polymerase chain reaction (PCR). Each probe represents a gene, and can be generated from the gene's cDNA clone, and amplified by PCR. A micro-sample of each cDNA is deposited and

bonded on a glass surface, with each gene occupying a unique location on the microarray chip (see Figure 1). Spotted cDNA arrays allow a greater degree of flexibility in the choice of arrayed elements, particularly for the preparation of smaller, customized arrays for specific investigations. As a result, cDNA arrays have so far been the technique most frequently used in academic labs.

Target preparation: mRNA molecules are extracted from the control sample and are reverse transcribed to generate cDNA probes; fluorescent-labeled nucleotides are incorporated in the cDNA during synthesis, thus labeling them. Different mRNAs are extracted from the experimental sample (e.g. cells at different time points, cells exposed to a drug or toxic substance). The fluorescent labeling step is repeated to generate a second cDNA probe using a fluorescent molecule of different color. Generally, green label is used for the control and red for the experiment.

Hybridization: In the hybridization step the two fluorescent target samples are applied simultaneously to a single microarray, where they react competitively with the arrayed cDNA molecules. Next, each element of the chip is scanned for both fluorescent colors, and the signal intensity at each position gives a measure of the number of bound molecules of each type, and hence the gene expression level of the gene. The ratio between the two intensities measured (red and green) provides a quantitative measurement of the relative gene expression level in the two cell samples.

1.2.1.2 Oligonucleotide microarrays

Array preparation: the array is made by synthesis *in situ*, of short oligonucleotide (single strand, generally 20-25mer), deposited either by photolithography onto silicon wafers (high-density-oligonucleotide array from Affymetrix, www.affymetrix.com) or by ink-jet technology (developed by Rosetta Inpharmatics, www.rii.com). No time-consuming handling of cDNA resources is required; sequence information alone is sufficient to generate the DNA to be arrayed. Also, probes can be designed to represent the most unique and specific parts of a given transcript, making the simultaneous detection of closely related genes or splice variants possible.

The oligonucleotide (oligo) array contains collections of approximately 20 pairs of probes for each of the RNAs being monitored. Each probe pair consists of two patches. One contains copies of a selected oligonucleotide (usually 20 to 25 nucleotides in length) that is perfectly complementary (referred to as a Perfect Match, PM) to a subsequence of a particular RNA. The second, companion patch contains identical oligonucleotides, except for a single base difference in a central position (referred to as a Mismatch, MM). The MM probe of each pair serves as an internal control for hybridization specificity. The analysis of PM/MM pairs allows low-intensity hybridization patterns from rare RNAs to be sensitively and accurately recognized in the presence of cross-hybridization signals. Hence, each gene is represented by usually 20 pairs of PM and MM of specific oligos, as opposed to the cDNA array, in which each gene is represented by copies of a single cDNA, deposited in one spot.

Target preparation: Total mRNAs are extracted from different tissues or cell populations; the cytoplasmic transcripts that have an adenine chain (polyA) undergo a hybridization reaction with dT primers (oligos of thimidines). After the primers hybridize, reverse transcriptase leads the synthesis of the cDNA strand from the mRNA template. In the final step, a transcription reaction is carried out by an RNA polymerase enzyme, while biotin-labeled nucleotides are incorporated into the synthesized cRNA molecules.

The cRNA molecules, which contain biotin-labeled nucleotides, are hybridized to the array. A scanning microscope performs fluorescence imaging of the arrays. Since every gene is represented by 20 PM-MM pairs of specific oligos, the average intensity and background must be calculated, all 20 pairs

are taken into consideration, and the resulting number used is the absolute value of intensity for the particular RNA transcript.

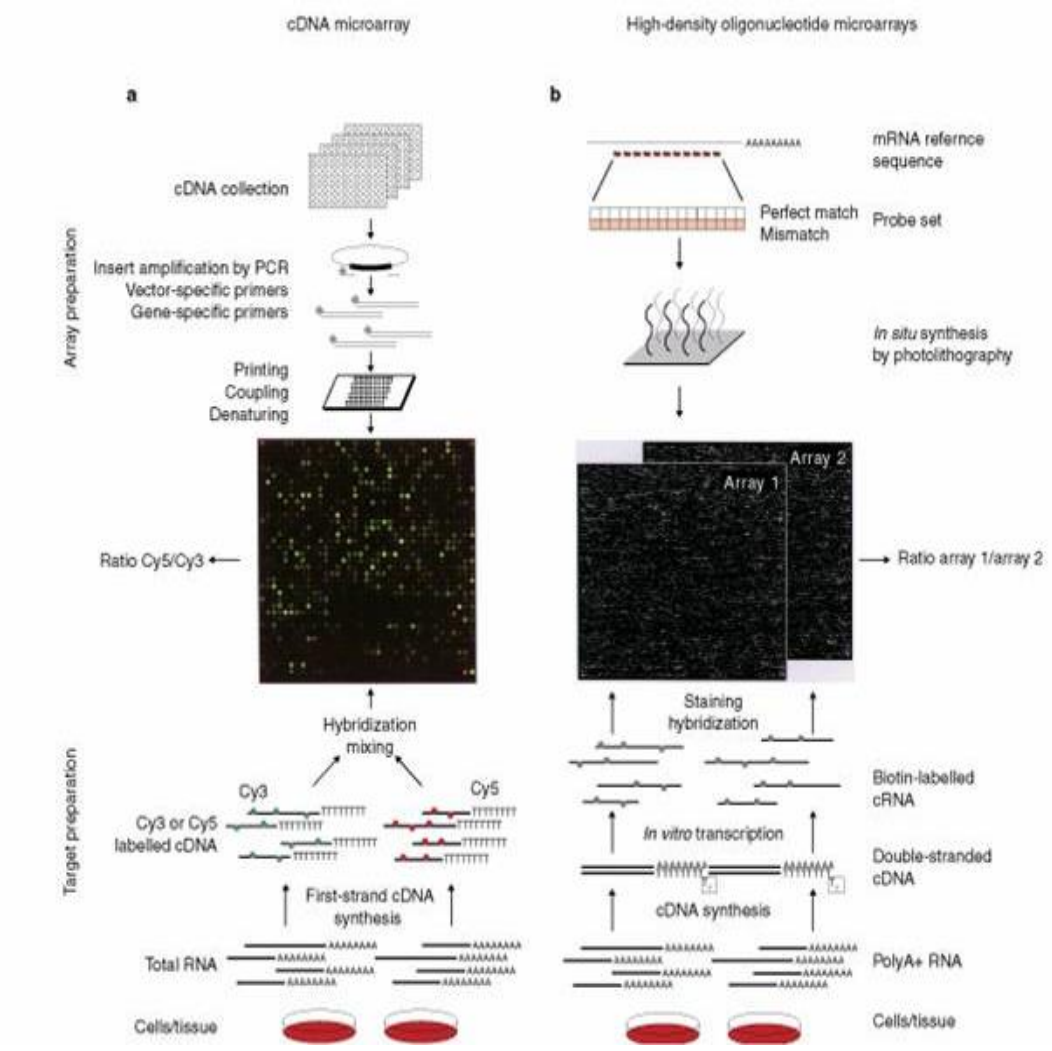


Figure 1: Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays.

1.2.2 Various experimental conditions

Genome-wide expression experiments are widely performed today using DNA microarrays. Many laboratories have examined the expression levels of transcripts in the cell by performing such experiments in a specific experimental or natural condition of interest, such as heat shock, nutrient starvation, exposure to reducing elements, cell cycle. These experiments have been carried out in a variety of organisms such as yeast, worm, human, etc. (DeRisi

et al. 1997; Cho et al. 1998; Spellman et al. 1998; Iyer et al. 1999; Gasch et al. 2000; Causton et al. 2001; Cho et al. 2001; Kim et al. 2001) While being subjected to the condition of interest, cells are extracted at various time points, their cDNA hybridized to microarrays, and thus mRNA transcription levels are quantified. These experiments are known as time-series experiments. By varying the time point under which a sample is taken, multiple arrays can be used to construct a vector of expression levels (the expression profile) for each gene.

1.2.3 Noise in expression data

The DNA microarray technology is a noisy and complex one; there are many technical challenges at various stages of the experiment. Quality assurance is performed to control for problematic spots printed on the chip and differences in amounts of wet substrate placed on the chips, also known as spot effects, such as differences in the concentration and amount of cDNA immobilized from one array to the next. Cross-hybridization and non-specific binding are biological problems that must be dealt with. In addition, the software which analyzes the read-outs of the chip results applies normalization procedures to the data, and different normalization procedures may result in different signals in the data. Some examples of normalization procedures are subtraction per measurement of mean expression level per gene or per chip, division by standard deviation across all time points, etc.

One must keep in mind, when using expression data to understand the regulatory control of genes, that not all levels of regulation are seen when analyzing the levels of RNA transcripts of genes. For example, post-translational modifications play a major role in regulation, the effects of which cannot be seen when examining RNA levels.

1.2.4 Role of gene expression data in unraveling regulation

Once expression levels have been determined by experimental means, it is important to find genes with similar expression profiles or patterns (co-expressed genes). There are two main reasons for the interest in co-expressed genes. Firstly, there is evidence showing that functionally related genes are co-

expressed (Eisen et al. 1998; Spellman et al. 1998; Tavazoie et al. 1999). For example, if several proteins are necessary to create a complex, it is logical that the genes coding for these proteins will be co-expressed, i.e. will have similar expression profiles. Consequently, grouping together genes with similar expression profiles may allow characterization of the function of previously uncharacterized genes (Eisen et al. 1998; Tamayo et al. 1999; Tavazoie et al. 1999). Secondly, we may be able to reveal the regulatory systems by clustering co-expressed genes. If a group of genes is controlled by a certain regulatory program, we may expect them to be co-expressed, especially when probed at the relevant condition. Thus, by locating co-expressed genes, we may be able to infer co-regulation, and thus understand the regulatory control of these genes. By clustering genes according to expression profiles, we hope to group together genes whose *cis*-regulatory elements are bound by the same proteins *in vivo*.

1.3 Genome-wide "Location Data"

1.3.1 Background

The **ch**romatin **i**mmuno-**p**recipitation (ChIP) procedure was developed in the late 1980s and used to study protein-DNA interactions at a small number of specific DNA sites (Solomon et al. 1988; Orlando and Paro 1993). Briefly, a DNA-binding protein of interest is allowed to bind to its *in vivo* DNA targets, and subsequently formaldehyde is added to the cells, causing fixation. Antibodies directed against the protein of interest allow immunoselection of all genomic binding sites. Cross-linking is fully reversed, and the immuno-precipitated DNA targets are amplified by PCR, and then sequenced. The combination of formaldehyde fixation and ChIP offered the ability to detect any protein at its *in vivo* binding site directly.

The genome-wide location analysis method developed by Ren et al. (Ren et al. 2000) allows protein-DNA interactions to be monitored across the entire yeast genome by combining ChIP with DNA microarray chip analysis, making use of microarrays imprinted with intergenic genomic regions. This combination

is also known as the ChIP-chip methodology (described in detail in Section 1.3.2).

The most recent advancement in the employment of the ChIP-chip methodology has been the use of tiling microarrays following the immunoprecipitation step, when hybridizing the DNA targets of binding proteins (reviewed in Johnson et al. 2005). Tiling microarrays are imprinted with all genomic regions, assaying regular intervals throughout the genome, thus covering not only promoter regions, but also introns, 3' UTR regions of genes, and intergenic regions. These arrays are unbiased to the positions of known and predicted genes, and address the possibility of transcription factor binding sites in locations other than upstream regions. This issue is particularly relevant in mammalian cells because the mRNA and protein coding sequences represent a small percentage of the total genome, and because transcriptional regulatory proteins can function at long and variable distances from transcriptional initiation sites (Cawley et al. 2004).

Alternative methods have measured protein-DNA binding interactions *in vitro* rather than *in vivo*. For example, Bulyk et al. quantified such interactions by allowing a DNA-binding protein to bind directly to a microarray imprinted with double-stranded DNA (Bulyk et al. 1999; Bulyk et al. 2001; Mukherjee et al. 2004). Another method for for determining the DNA-binding specificity of proteins *in vitro* is via DIP-chip (DNA immunoprecipitation with microarray detection) (Liu et al. 2005). In DIP-chip, protein-DNA complexes are isolated from an in vitro mixture of purified protein and naked genomic DNA. Whole-genome DNA microarrays are used to identify the protein-bound DNA fragments, and the sequence of the identified fragments is used to derive binding-site descriptions. Yet another *in* vitro method which can be used for similar purposes is commonly known as "in vitro selection" or "SELEX" (**s**ystematic **e**volution of **l**igands by **ex**ponential enrichment), which was developed in the laboratory of L.W. Szostak in 1990 (Famulok and Szostak 1992). Using the SELEX technique, large random pools of nucleic acids can be screened for a particular functionality, such as the binding to a particular protein. Functional molecules are selected from the mainly non-functional pool of DNA by column chromatography or other selection techniques.

1.3.2 Description of the location analysis dataset

Genome-wide location analysis, which is also known as genome-wide binding analysis, was developed and first introduced by Ren et al (Ren et al. 2000). It produces a dataset widely referred to as "location data" (Ren et al. 2000; Iyer et al. 2001; Lieb et al. 2001; Simon et al. 2001; Lee et al. 2002; Harbison et al. 2004). The method is a microarray method that reveals the genome-wide location of DNA-bound proteins. Ren et al. used this method to monitor binding of several gene-specific transcription activators in all intergenic regions of yeast. Lee et al. later used the method to identify genomic binding sites for many transcriptional regulators in living *S. cerevisiae* yeast cells under a single growth condition.

The method combines a modified chromatin immunoprecipitation (ChIP) procedure, which had previously been used to study DNA-protein interactions at a small number of specific DNA sites (Orlando 2000), with DNA microarray analysis.

The process of generating genome-wide location analysis data is as follows (Figure 2 shows a schematic depiction of the generation of such data):

Cells are fixed with formaldehyde, such that cross-links are formed between DNA and any proteins bound to it.

The cells are harvested, and chromatin is disrupted and sheared by sonication, fragmenting the DNA into short segments (length distribution of usually between 200-600 bp).

The DNA fragmenrts which are cross-linked to the protein of interest are enriched by immunoprecipitation (IP) with a specific antibody which recognizes an epitope tag of the TF protein.

The cross-links are reversed, and the enriched DNA is amplified and labeled with a fluorescent dye (Cy5; red), by use of ligation-mediated polymerase chain reaction (LM-PCR).

In addition, a sample of DNA that was not enriched by immunoprecipitation is subjected to LM-PCR in the presence of a different fluorophore (Cy3; green).

Both IP-enriched and –unenriched pools of labeled DNA are hybridized to a single DNA microarray. (In the studies of Ren et al. and Lee et al. described here, a chip containing all yeast intergenic sequences was used.)

The ratio of immunoprecipitated to control (unenriched) DNA is determined for each array spot. A confidence value (p-value) for binding for each spot is calculated from each array by using an error model. The ratio of fluorescence intensity obtained from three independent experiments is used with a weighted average analysis method to calculate the relative binding of the protein of interest to each sequence represented on the array. The result is one final p-value per sequence.

Note that the DNA array used in this protocol is not identical to those expression arrays described in section 1.2.1. In this protocol, the sequences printed on the array correspond to all intergenic regions of the yeast genome, as we are interested in knowing to which regulatory (promoter) genomic regions the TF of interest binds *in vivo*. An additional difference is that it is amplified genomic DNA that will bind to the array, and not RNA or cDNA.

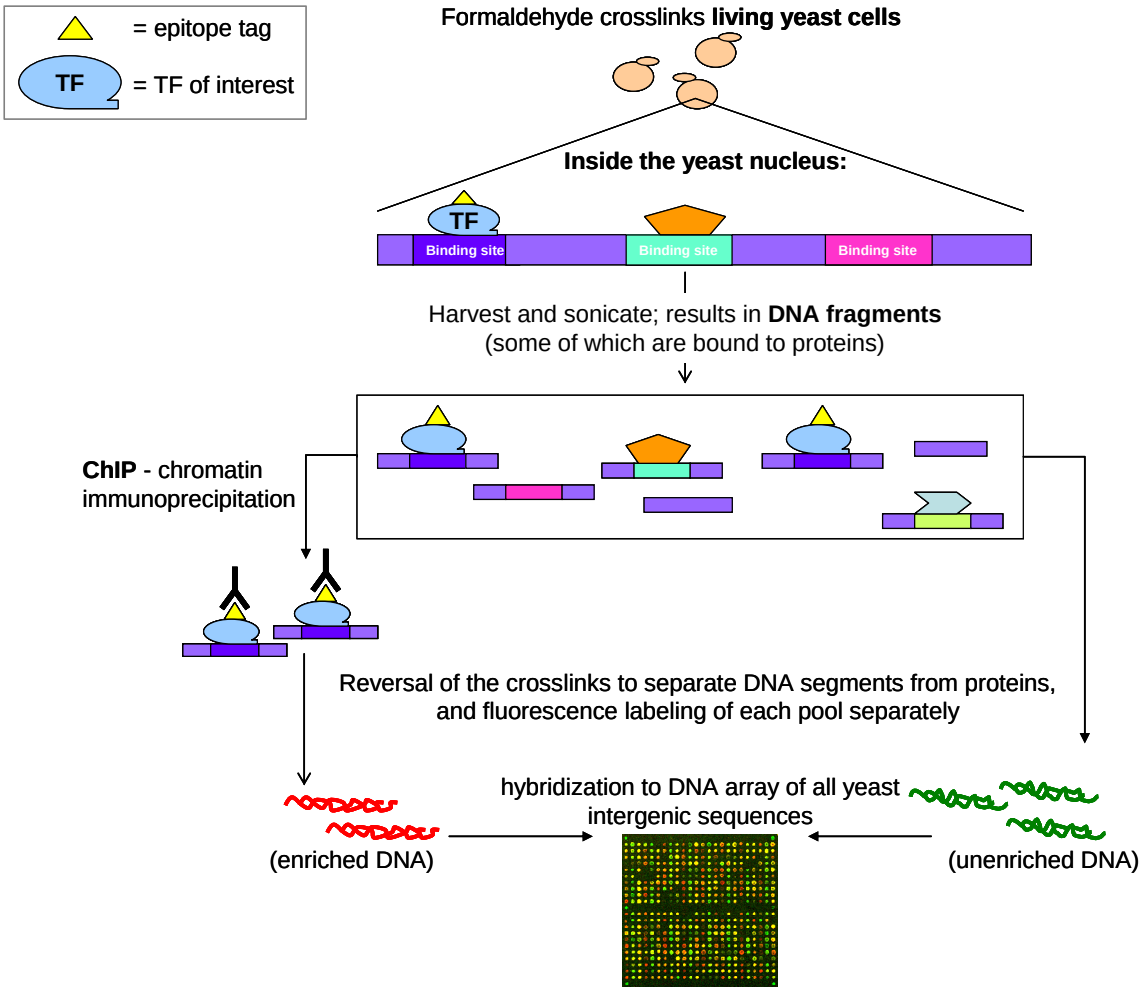


Figure 2: Schematic depiction of generation of genome-wide location analysis data. Detailed account appears in the text. Red spots on the DNA array represent IP-enriched DNA, green spots represent unenriched DNA, and yellow spots are merged spots.

Lee et al. generated genome-wide location data for 113 transcription factors *in vivo* in yeast cells. In order to obtain such a high-throughput dataset, they used a method which allowed them to employ one single antibody which could immunoprecipitate any TF of interest, thus overriding the necessity of having one specific antibody for each of the 113 TFs. To this end, a yeast strain was constructed for each TF such that the transcription factor protein, when expressed and translated, would have a myc epitope tag. An epitope tag coding sequence was introduced into the genomic sequence encoding the carboxyl-terminus of the TF. Appropriate insertion of the tag, and expression of the tagged protein, was then confirmed. Subsequently, the genome-wide location analysis was carried out on these constructed cells, and the

immunoprecipitation step was done with antibodies against the myc epitope tag. It is important to note that the attachment of an epitope tag could result in the loss or reduction of DNA binding interactions in the analyzed TFs. The authors made the assumption that the lack of serious growth defects and the presence of positive binding results together indicate that the epitope modification does not alter binding activity substantially. This assumption is based on experimental validations performed by the authors: essentially identical results were obtained for specific regulators when immunoprecipitation was performed with epitope-tagged regulators or when it was performed with polyclonal antibodies against those regulators (Ren et al. 2000). Thus, the epitope tag does not appear to modify binding interactions in these specific cases. Additional experimental confirmation of data included conventional, independent chromatin immunoprecipitation experiments conducted at a gene-specific level, which confirmed 93 of 99 binding interactions involving 29 different regulators, which were identified by location analysis data at a threshold p-value of 0.001.

The 113 TFs analyzed in this study were selected as follows: All 141 TFs that were listed in the Yeast Proteome Database (Costanzo et al. 2000) at the time, and were reported to have DNA binding and transcriptional activity, were selected for the study at the onset. For 17 of the 141 TFs, viable tagged cells were not obtained. Of the remaining 124 TFs, 106 TFs were expressed at levels that could be detected by immunoblot analysis. An additional seven TFs were later analyzed, thus the data was obtained for 113 TFs. For this location dataset, each tagged strain was grown in three independent rich medium cultures (yeast extract, peptone, and dextrose). Results of the three independent experiments were combined in a weighted average analysis method to calculate the relative binding of the protein of interest to each sequence represented on the array.

A p-value threshold was chosen in order to facilitate discussion of a subset of the data at a high confidence level. Of course, this thresholding artificially imposes a binary "bound or not bound" decision for each DNA-protein interaction. The stringent p-value threshold chosen was 0.001, which maximizes inclusion of true regulator-DNA interactions, while minimizing false positives. Following the authors of the original paper (Lee et al. 2002), when we discuss in

this study the location dataset, we refer to those binding events reported which have a p-value <= 0.001.

It is important to note that one of the major advantages of this method is that it is an *in-vivo* assay of the state of the yeast nucleus under normal conditions. For example, in these conditions, the chromatin, which is the natural state of the DNA template *in-vivo*, is unaffected.

1.3.3 Noise in the location data

The location dataset is a noisy one (Lee et al. 2002). It is difficult to estimate the number of false positives and false negatives in the dataset, as there are various types of noise. Several examples of noise are as follows:

experimental binding noise: binding events that were stated to take place (i.e. given a significant p-value) but do not result from actual *in-vivo* binding of the TF to the intergenic region (the authors estimate based on low-throughput validation tests that at the 0.001 p-value threshold, 6-10% of the predicted interactions are false binding reports)

regulatory noise: binding events that did in fact take place, but did not cause regulatory effects such as transcriptional activation of a gene.

indirect regulatory noise: binding event was reported to occur between a protein X and DNA sequence Y which actually occurred by a secondary protein (co-factor) Z binding to Y, and X binding to Z. Because of the physical proximity and due to the experimental protocol, cross-linking occurred and predicted X to bind Y when actually Z binds to Y.

Quantifying the true number of false negative reports is of course impossible, for to establish this, we would need perfect knowledge of the interactions that actually occur. One method of estimation is by performing a literature confirmation of the data. The authors did such a confirmation, and found that the location data generally agrees with the published literature in 41 of 50 cases. Since 9 out of 50 interactions reported in the literature were not detected by the location data, this suggests that the dataset has an 18% false negative rate. It is clear that loosening the p-value threshold of 0.001 to a less

stringent value would allow false negatives (i.e. true interactions) to enter the dataset. This of course would be at the expense of gaining false positives.

The false positives were also estimated by the authors. As stated above, ChIP experiments conducted at a gene-specific level confirmed 93 of 99 binding interactions involving 29 different regulators, which were identified by location analysis data at a threshold p-value of 0.001. This hints at a false positive rate of only 6%. Note, however, that only 99 interactions were tested in this validation, and these interactions only include a subset of the regulators which were studied. We contend that the actual false positive rate may in fact be much higher (see Section 2.3.1).

Several works have addressed the problem of noise in the location data (Banerjee and Zhang 2003; Bar-Joseph et al. 2003; Gao et al. 2004) and have attempted to extract regulatory signal by combining other data sources. Some address the issue of false positives (Gao et al. 2004); others attempt to recover false negatives (Bar-Joseph et al. 2003). We compare our method with that of others in section 2.3.9.

1.4 Motivation for this work

Many studies have focused on investigating the transcriptional regulatory network of various organisms (Tavazoie et al. 1999; Kim et al. 2001; Ihmels et al. 2002; Shen-Orr et al. 2002; Segal et al. 2003; Beer and Tavazoie 2004). The building blocks of such genetic networks are the transcriptional regulatory proteins and their connections to the regulated genes. The mass of our understanding of gene regulation and global expression patterns on a systems level will arise from analyzing genome-wide data such as the location data, which gives us the spectrum of *in vivo* binding of a large portion of the transcriptional regulators in the cell.

Several biocomputational works have utilized data sources such as gene expression data (Bar-Joseph et al. 2003; Gao et al. 2004) or sequence data (Segal et al. 2002) in conjunction with the genome-wide location analysis data. However, none have combined both these sources, as well as additional statistical analyses of TF combinatorics, as we have done in this study to address the problem of noise in the location data and to extract regulatory

network structural features out of it. The integration and intersection of these different data types can extract important biological signals from the noisy independent data sets. Interestingly, such additional sources may themselves be noisy and yet serve the purpose of filtering, provided that the noise in the different methods is not trivially correlated. We hypothesized that the analysis of the following may serve for filtering noise in the location data: (i) the coherent expression of genes regulated by the TFs (ii) the identification of regulatory motifs among the genes assigned to each TF (in similarity to other recent publications (Segal et al. 2003; Gao et al. 2004)), and (iii) combinatorial partnerships between sets of TFs. For each of the filtration methods we present: (a) the rational underlying the method, along with the relevant algorithm or computation, (b) exemplifying figures, (c) a “birds-eye view” of the results of application of the method to the entire dataset, and (d) the subset of the TF-gene assignments reported by the location data, that is supported by this method (on the supplementary website). In the datasets we provide in our website, genes are assigned to TFs only if expression, sequence, and/or combinatorial TF interaction support such assignments (http://longitude.weizmann.ac.il/TFLocation/TFLocation.html).

We further appreciate that appropriate utilization of genome-wide data requires visualization tools to analyze it. We have thus aimed at developing a graphical user interface (GUI) tool that would be accessible via the world-wide web, which would allow users to view the gene expression data of genes bound by various transcription factors. We required that users will be able to analyze the sequence motifs found in the promoters of the genes, and examine which genes contain these motifs and how coherent their expression profiles are, as well as analyze combinations of TFs, and the role of pairs of TFs which co-bind in exerting regulatory effects on the genes.

As part of the ongoing effort in the lab to create a comprehensive collection of regulatory motifs, the present work has contributed by providing a visualization tool and clustering algorithm implementation that has aided in fine-tuning our definition of a motif, and aided in our understanding of how expression and sequence should be taken into consideration when defining significant regulatory motifs.

2 Results

2.1 Expression Coherence of a Gene set

We set out to utilize genome-wide mRNA expression data to aid in deciphering key elements of the transcriptional regulatory network in yeast. To this end, we used publicly available data attained by DNA microarray experiments. The experimental data span a diverse set of 40 conditions, both natural (Cho et al. 1998; Spellman et al. 1998; Roberts et al. 2000) and perturbed (Chu et al. 1998; Eisen et al. 1998; Gasch et al. 2000; Jelinsky et al. 2000; Causton et al. 2001). Here and in all subsequent analyses we refer to a whole time series (such as exposure to heat shock, or progression through the cell cycle) as a "condition", which is composed of 3-28 time points (each time point corresponds to one microarray).

As a case study, we inspected the mRNA expression profiles of genes associated with a TF by the location data (for description of the "location data", see section 1.3.2). It may be expected that a set of genes, whose promoters are bound by a specific regulatory TF, be coherent throughout a time series experiment. An example of such coherent behavior is shown in Figure 3.

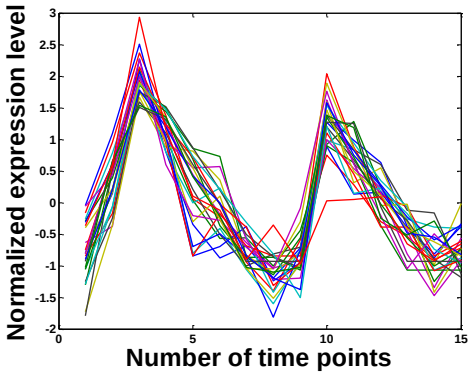


Figure 3: Normalized expression profiles of 23 genes; each line represents the normalized mRNA level of one gene across each of the different time points. The set of genes is clearly coherent across all of the 15 time points of the experiment.

One measure of coherence of a gene set is the expression coherence (EC) score developed by Pilpel *et al* (Pilpel et al. 2001). The EC score is a measure of how clustered a set of genes is in expression space. This score may be defined for any gene set for which expression profiles are available. If we have

an expression profile consisting of N time points for each of M genes, then each gene can be thought of as a point in an N-dimensional space, where the i-th dimension is the expression level of the gene at the i-th time point. In order to obtain the level of clustering of the M genes in this space, one might calculate the 'center of mass' of the cloud of genes, and then sum over distances of each gene from it. An alternative may be to sum over squares of such distances, or take the standard deviation around the mean, etc. Yet, these measures have one clear shortcoming: in a case where the set of genes is split, for example, into two tight clusters, that are remote from one another, any score based on deviation from the mean will be very low. However, the above-mentioned EC score developed by Pilpel *et al.* will give high scores in such cases.

The EC score is defined as the fraction of pairs in the gene set, whose Euclidean distance between expression profiles is under a certain threshold. Given a set of M genes, there are P=M*(M-1)/2 gene pairs. We calculate the Euclidean distances between the normalized expression profiles of each of the P pairs of genes. The EC score equals p/P, where p is the number of gene pairs whose distance is smaller than a threshold distance D. See section 3.5 for details on determination of D, and on calculation of a p-value assessing the significance of the EC score of a gene set.

When examining sets of genes with high EC scores, it is impossible to know whether the set is comprised of one or several tight clusters, based on the EC score and its p-value alone. Figure 4 shows two toy examples of gene sets which are very different from one another in structure: one set contains a large cluster, and several genes which are all dissimilar from one another, while the second set contains several small clusters. The two examples of equal set size, fundamentally different from one another, receive the same EC score, and thus the same p-value.

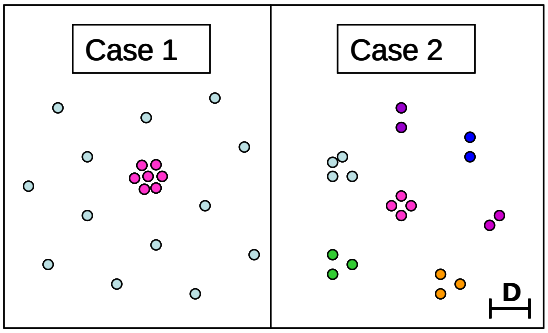


Figure 4: Different clustering scenarios may have identical EC scores. Each circle portrays the expression profile of one gene in expression space (shown here in only two dimensions). Both cases shown display the clustering of 20 genes. Case 1 shows one large cluster of 7 genes, and 13 genes which are all dissimilar from one another. Case 2 shows 7 small clusters. In each of the cases, 21 of the 190 possible pairs of genes have a Euclidean distance smaller than a certain threshold D. Thus, the EC score is 21/190 = 0.11.

Two sets of genes which receive identical EC scores may have very different biology underlying their expression coherence. For example, case 1 of Figure 4 may represent 20 genes regulated by a transcription factor, seven of which follow a tight, strict expression profile dictated by the TF (for example, upon high-affinity binding of the TF to the promoters of the genes). The remaining 13 genes may have promoters with low-affinity binding sites for the TF, for example sequence motifs which differ from the consensus binding site by several nucleotides. Case 2 may represent various distinct expression profiles dictated by the TF, dependent on other factors such as additional TFs which each co-regulate small subsets of the genes, thus bringing about different expression profiles. These are very simplistic examples of possible explanations for the differences between expression profiles of such gene sets. One can think of many more explanations. In fact, case 1 may also be explained as a case which easily allows us to distinguish between true and false positive assignments of genes to a regulatory protein. For example, it may be that only the 7 highly clustered genes are actually regulated by the TF, whereby the remaining 13 genes were falsely assigned to this TF, or are perhaps unbound and thus unregulated under the experimental condition studied. The EC score does not allow us to distinguish between cases such as the ones portrayed in Figure 4. We thus went on to examine ways which allow us to differentiate between such cases.

2.2 Decomposition of data via clustering

Decomposition of gene sets, via clustering based on expression profiles, allows us to gain a deeper understanding of transcriptional regulation when analyzing DNA microarray data. When researching transcriptional regulation, we are often interested in analyzing a set of genes which take part in some common pathway or response. The expression profiles of these genes may be quite different from one another, albeit an often highly significant EC score of the set, hinting at different response mechanisms at the transcriptional level. This decomposition allows us to distinguish between different subsets inside the larger set, and enables us to analyze related genes, rather than the group as a

whole, searching for the underlying biological mechanism differentiating the various subsets. Clustering the expression profiles of the gene set allows us to differentiate between cases where a gene set contains one large, tight cluster from those cases in which a gene set contains several smaller, tight clusters. In addition, it allows us to visually view the data as coherent subsets, and recognize a strong signal which appeared subdued amongst the entire group of genes.

Figure 5 shows the results of decomposing three different gene sets into clusters, and demonstrates the various trends that may exist in such data: (1) a set of genes which appears non-coherent and is in fact non-coherent according to the EC score significance test, (2) a set of genes which appears non-coherent but is coherent, and (3) a set of genes which appear coherent and is in fact coherent. When a set of genes appears incoherent, decomposition into clusters allows us to understand whether this gene set falls into category (1) or (2).

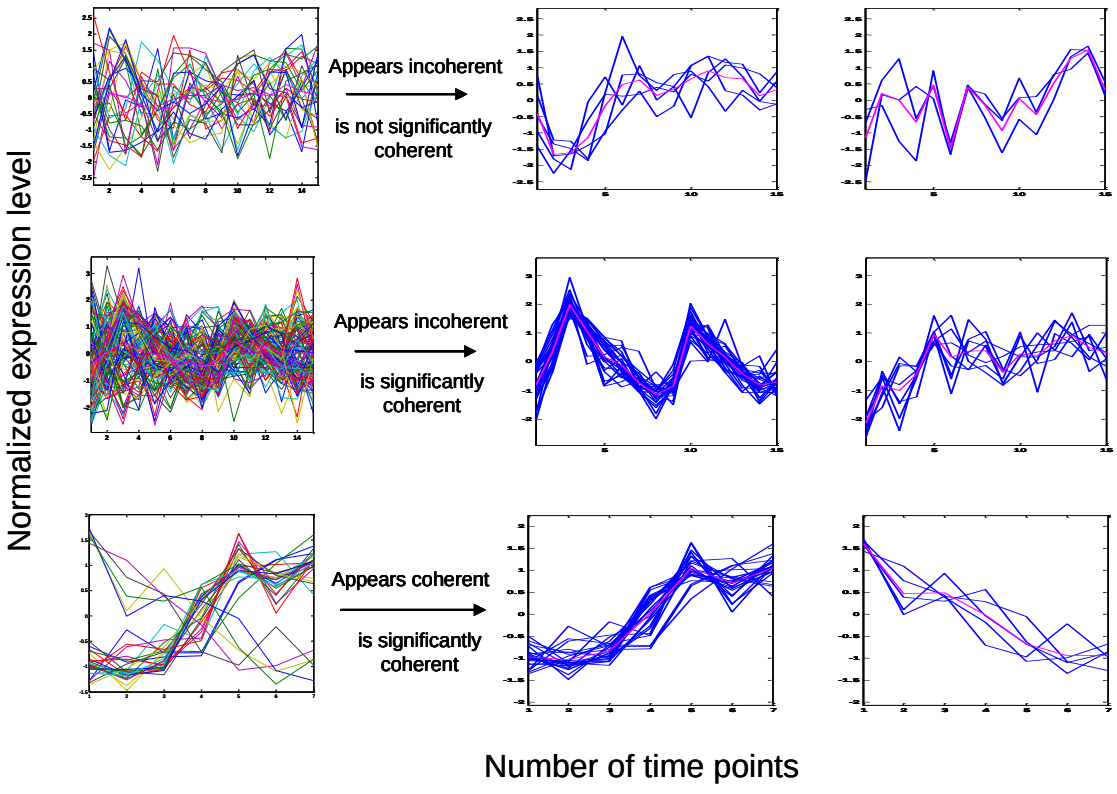


Figure 5: Three examples showing expression profiles of gene sets before and after clustering. A clustering algorithm clustered the genes based on the Euclidean distance between expression profiles. The expression profiles of the genes in the picture on the left were clustered, and the two largest clusters formed are shown on the right.

2.2.1 QT_clust clustering algorithm

After realizing the potential of such expression profile clustering, in order to extract regulatory signal from large datasets, we implemented the QT_clust clustering algorithm developed by Heyer *et al* (Heyer et al. 1999). Unlike many clustering algorithms such as k-means, that require in advance the determination of number of clusters, and that give rise to clusters of various extents of tightness (Tavazoie et al. 1999), in this algorithm the only inputs are: (i) 'maximal cluster diameter', or the maximal distance tolerated between any two entities in a cluster, and (ii) the distance matrix of all entities to be clustered. The output is the number of clusters, along with the cluster assignments of all entities.

Especially when dealing with gene expression data, quite often we do not know *a priori* the number of clusters existing in the data. On the other hand, we may often have an idea about the tolerated distance among co-regulated genes. The important contribution of the QT_clust algorithm, which makes it particularly appropriate in clustering expression data, is that the number of clusters is not predefined by the user; the user defines only the maximal cluster diameter. The pseudo code of the algorithm is shown in Figure 6. In our implementation, we have modified the algorithm such that it does not require two entities to reside in one cluster if they are not close enough to one another – an entity can populate a cluster alone, if it differs from all other entities. Additionally, in our use of QT_clust, we do not use an arbitrary distance as the diameter input parameter, but rather infer which distance should be used based on the data, as described later in the text. These modifications make QT_clust well-suited to deal with expression data.

```
QT_clust (pd, d)
while still have entities to cluster in G
{
    for each of the N entities (in G) left to cluster
    {
        //make candidate cluster around it that doesn't pass threshold diameter:
        • add entity which when added, gives lowest cluster diameter
        • continue to add entities to this cluster while diameter <= d threshold
    }
    //results in N candidate clusters
    pick C, the largest cluster made, out of the N clusters
    remove the entities of C from G; C is your newest cluster
}
```

Figure 6: Pseudo code of *QT_Clust* algorithm, which takes two inputs: (i) pd; the distance matrix between all pairs of entities in G and (ii) *d*; a diameter threshold, and returns a set of clusters.

2.2.2 QT_clust applications: expression- and sequence-based clustering

We have extensively applied the QT_clust algorithm to cluster various sets of genes, clustering them based on their expression profiles, to aide in elucidation of the signal often hidden within genome-wide expression data. Section 2.3.2 describes some of this work in great detail. Suffice it to say here that we subjected the mRNA expression profiles of genes assigned by the location data to each of 113 TFs to decomposition by the QT_clust clustering algorithm (Heyer et al. 1999) and extensively researched the results, uncovering promising biological results. Following the original publication of the location data, we use the notation of a gene being “assigned to a TF” if the p-value on the interaction between the TF and the gene’s promoter is below a threshold.

Another use of the QT_clust algorithm on expression data is described in Section 2.4 (section titled "Regulatory Motif Dictionaries project"). A major ongoing project in our lab, headed by doctorate student Michal Lapidot, has been to create 'regulatory motif dictionaries' for various organisms. Within the scope of this project, the QT_clust clustering algorithm was very useful. We developed a Graphical User Interface (GUI) which allowed easy viewing,

handling, and analysis of the huge amounts of biological data created by this effort. In this project, QT_clust was used not only to cluster genes based on similarity of their expression profiles, but also to cluster together sequence motifs based on sequence similarity. Since the input of the algorithm is a distance matrix between all pairs of entities to be clustered, the distance can be Euclidean distance between gene expression profiles when genes are to be clustered, or a measure of sequence distance, when sequences are to be clustered.

2.3 Filtering expression noise from the location data

With the experience we gained in utilizing QT_clust as a useful algorithm to extract signals from genome-wide expression data, we went on to analyze such data as crossed with the location data, which gives us experimental predictions of gene sets regulated by TFs. We analyzed all 113 TFs of the location data, throughout 40 conditions. By the integration of a number of filtration methods, we were able to produce a cleaner version of the location data.

2.3.1 The p-value trade-off in the original location data

The location data assigns a p-value on the hypothesis that transcription factor X binds to promoter Y and thus contains p-values for a set of multiple hypotheses. In order to determine which hypothesis is true, a p-value threshold is selected and only those hypotheses that pass this threshold are assumed to be correct. This thresholding, while probably capturing true assignments of TFs to promoters, results in a yet-to-be-determined amount of false assignments. We started by statistical assessment of the false discovery rate in the location data with the strictest p-value used by its authors (p-value = 0.001). Our calculations suggest that using this threshold the expected amount of false assignments of intergenic regions to TFs is 763 of the 4177 assignments (i.e., the false discovery rate, known as q-value, equals 0.18), hence there are 3414 expected true positives (4177-763=3414). See 3.4 for details on calculations. Without additional sources of information it is thus impossible to establish which of the added hypotheses are likely to be true TF-promoter assignments.

With the goal of a cleaner version of the location data in mind, which will allow better deciphering of the genetic regulatory map, we applied a number of methods which each produces a matrix, identical in size to that of the location data, of regulatory connections between the regulators in the cells and their regulated genes. These matrices were then used in combination to produce the noise-filtered version of the location data. In each matrix, a gene i was marked as regulated by TF j if and only if it was (a) assigned to TF j in the original location data, and (b) it also had evidence strengthening this assignment as found by one of the following methods of detection: (i) clustering of gene expression profiles, (ii) regulatory motif detection, (iii) synergy interactions, and (iv) co-localization of TFs.

2.3.2 Method (i): Decomposition of gene expression profiles

We began by inspection of the mRNA expression profiles of genes associated by the location dataset to each of the 113 TFs in a diverse set of 40 conditions. Here and in all subsequent analyses we refer to a whole time series (such as exposure to heat shock, or progression through the cell cycle) as a "condition", which is composed of 3-28 time points. An intuitive expectation from a set of genes that are indeed regulated by a shared TF is that they display similar expression profiles at least in the conditions in which the TF exerts a significant regulatory effect. Yet we need not necessarily anticipate one coherent cluster, an alternative may be that some TFs will give rise to several distinct expression patterns. The expression coherence (EC) score is thus a suitable measure of the extent to which a set of genes is clustered into one or more clusters in expression space (see section 3.5 for definition of EC score). We explored various thresholds that correspond to different extents of expression similarities that may be dictated by various regulators.

We have examined the expression profiles of the genes assigned to each TF in the location data in 40 time-series experiments that span a broad range of natural (Cho et al. 1998; Spellman et al. 1998; Roberts et al. 2000) and perturbed (Chu et al. 1998; Eisen et al. 1998; Gasch et al. 2000; Jelinsky et al. 2000; Causton et al. 2001) conditions. We performed expression coherence analyses (Pilpel et al. 2001; Lapidot and Pilpel 2003) on each gene set in each

condition and evaluated their statistical significance using a formalism recently proposed by our group (Lapidot and Pilpel 2003). We used the FDR theorem (Benjamini and Hochberg 1995) to account for the multiplicity of hypotheses tested and determined a p-value threshold that guaranteed a desired false discovery rate.

Figure 7 is a matrix depicting significant expression coherence of particular TFs in particular conditions (see section 3.5 for details on statistical significance of EC score). We assume that a transcription factor regulates the gene set assigned to it in the location data in a particular condition if these genes are significantly coherent in that condition. Figure 8and Figure 9 depict distributions of the number of TFs regulating each condition and number of conditions regulated by each TF respectively. The conditions that are controlled by the largest number of TFs are the Cho cell cycle experiment (Cho et al. 1998; Spellman et al. 1998), the MAPK signaling experiment (Roberts et al. 2000), and the nitrogen depletion experiment (Gasch et al. 2000). These conditions are subject to the regulation of 30-34 TFs. The two ribosomal protein regulators, Rap1 and Fhl1, show regulation in many of the conditions. Conversely, 16 out of the 113 TFs in the dataset which had three or more genes assigned to them, had no condition in which the genes assigned to them show significant coherence. Some of these TFs may be involved in AND-gated combinatorial regulation, and only when inspecting them along with their partners may coherence emerge. Alternatively, it may be that such TFs represent multiple cases of false TF-promoter assignments. It is also possible that some of the low-scoring TFs are in fact regulating conditions not examined here.

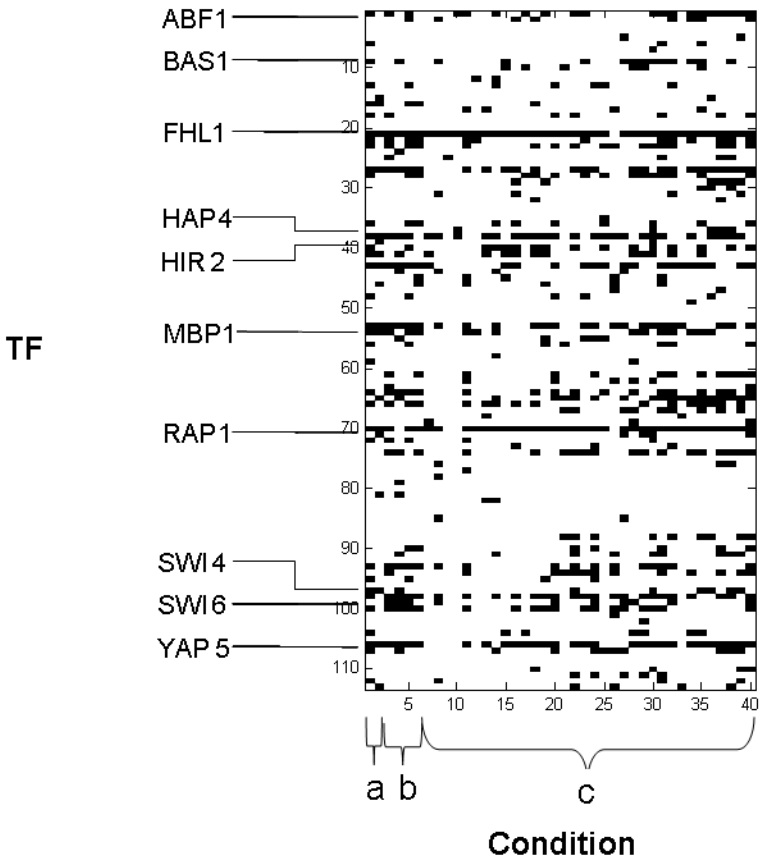


Figure 7: A matrix depicting expression coherence of each TF in each condition. An ij-th entry in the matrix is colored black if the i-th TF was significantly coherent in the j-th condition, and white otherwise. Conditions marked as 'a' are Cho's cell cycle (Cho et al. 1998) and Chu's sporulation (Chu et al. 1998), in 'b' are Spellman's 4 cell cycle conditions (Spellman et al. 1998), and in 'c' are predominantly stress responses (Eisen et al. 1998; Gasch et al. 2000; Jelinsky et al. 2000; Causton et al. 2001). Selected TFs are designated by their names; all TF and condition names are available in Appendix A.

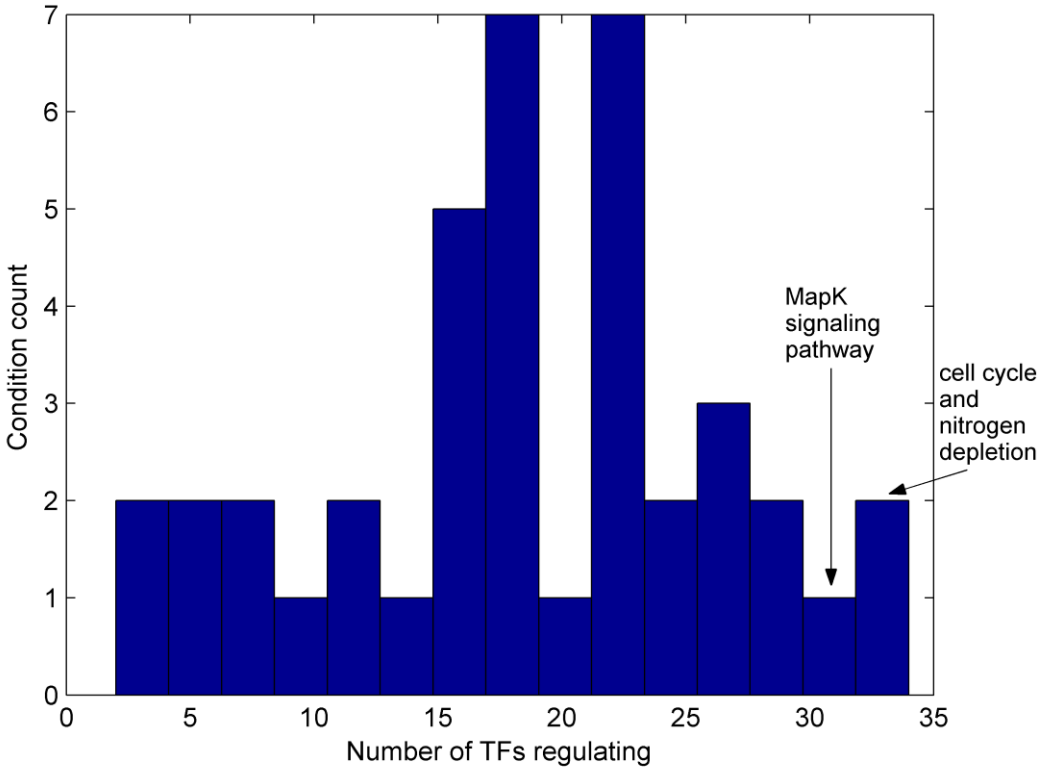


Figure 8: A histogram with the number of TFs regulating each condition.

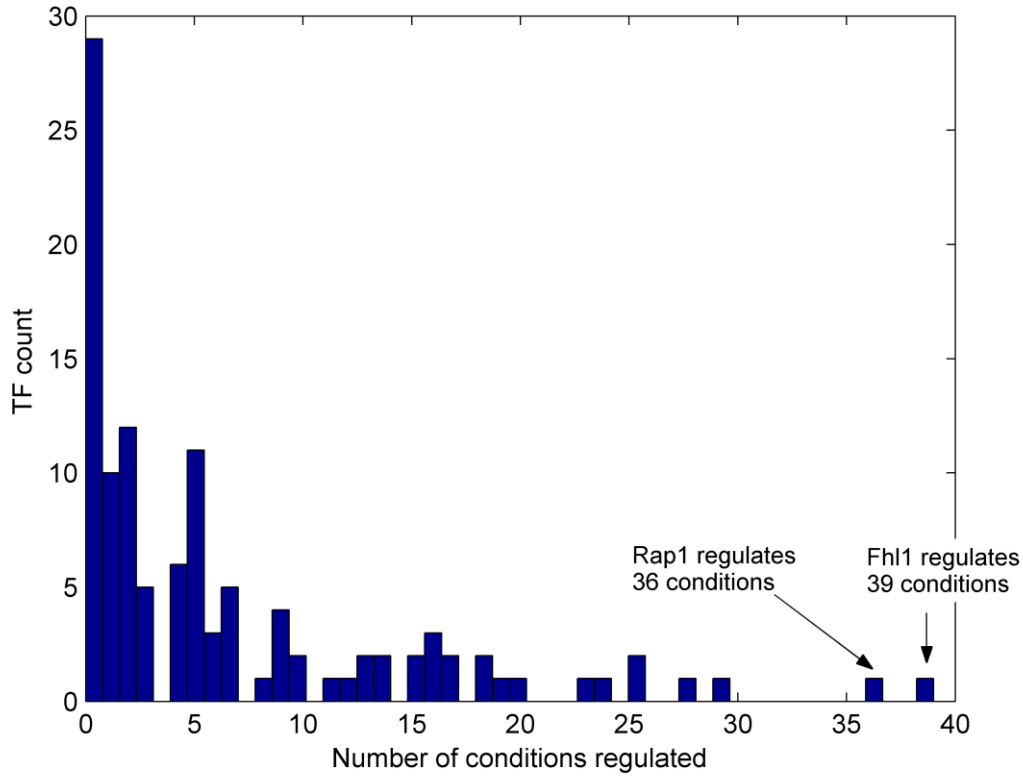


Figure 9: A histogram depicting the number of conditions regulated by each TF.

The EC score was deliberately designed such that TFs that predominantly give rise to one or a few tight clusters of genes (when clustered by expression profiles) can score highly, while a significant amount of genes with no clear cluster-assignment may be tolerated. In order to detect such behaviors we subjected the mRNA expression profiles of genes assigned to each of the TFs to decomposition by the QT_clust clustering algorithm (Heyer et al. 1999). Unlike many clustering algorithms such as k-means, that require in advance the determination of number of clusters and that give rise to clusters of various extents of tightness (Tavazoie et al. 1999), in this algorithm the only input is the minimal cluster tightness, and the output is the number of clusters along with the gene-cluster assignments. In all present analyses we used a relative, rather than absolute, measure of cluster tightness. The distance between each two genes in a cluster was required to be lower than a distance D, such that the probability of two random genes from that experiment to be at distance D or lower was p. For each TF we experimented with a range of values of p, from

0.05 to 0.5. Figure 10A shows the result of running QT_clust over the set of genes assigned to Abf1 using Chu's sporulation expression data (Chu et al. 1998). This is a clear example of a TF whose associated genes display various different expression patterns (colors of expression profiles are only relevant later in the text). Our analyses show such situations where the genes regulated by a TF may be decomposed into several distinct expression profiles, even in conditions in which the genes assigned to the TF are significantly coherent. For example, in only 288 of the 738 cases in which a TF scored highly at a condition, the most populated cluster is at least three times larger than the second largest cluster; the rest of the TFs represent cases in which the genes assigned to the TF give rise to several sizeable well-separated clusters.

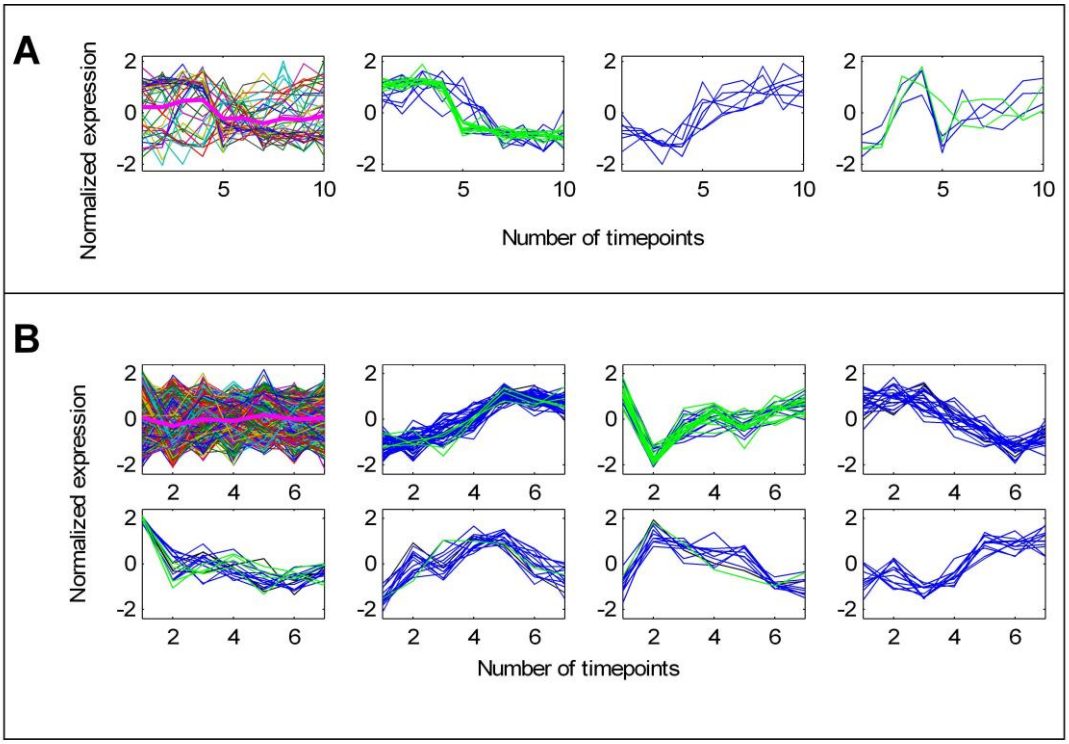


Figure 10: Expression profiles of genes regulated by Abf1 during sporulation (A) and Bas1 during nitrogen depletion (B). The first box on the left in each panel represents the expression profiles of all the genes assigned by the location data to the respective TFs. The rest of the boxes represent the results of decomposition of these genes into the most populated clusters generated by QT_clust. In A genes containing NCGTNNNNARTGAT and CGATGAGMTK are colored green, genes with only the first motif are colored blue, genes with only the second motif are colored red, and genes with none of the motifs are black. In B genes containing the RNMRGAGTCA motif in their promoter are colored green, the rest are blue.

Clustering of the expression profiles often yields several major clusters which are highly populated and have a clear, distinct expression pattern, and many more clusters which are lowly populated, often containing only one or few genes whose expression profile was dissimilar from that of all other genes. Additionally, it is possible and even likely that the lowly populated clusters consist of genes that were mis-assigned to the TF, since they have a profile so different than that of the genes which appear to be tightly regulated by the TF. Such a decomposition of the expression signal allows us to view the genes assigned to each TF and distinguish the signal from noise in the data. For example, refer again to Figure 10A. It seems that this TF gives rise to several different temporal patterns. In addition, 8 out of 28 of the clusters not shown contain only one or two genes. Thus expression-based data cannot support the proposed assignments of these genes to the corresponding TFs. On the other hand, the genes in the substantially populated clusters are the most likely true assignments, and recalculating the EC score of these genes alone may show that the TFs do in fact give rise to significantly coherent expression profiles, which were undetectable amidst the noise.

Hence, *filtration method (i)* yields a matrix in which gene i is assigned to TF j only if it was assigned in the original location data, and also if it belongs to a cluster of at least 3 genes, in at least one of the conditions in which the set of genes assigned to TF j is significantly coherent. This choice of minimal cluster size reflects a balance between the desire to include as many assignments as possible and the tendency to remove those that seem to be outliers (see our supplementary website).

Following this filtration, the EC score of each TF's genes was recalculated. Of the TF-condition pairs that did not pass the EC significance test on the original data, 96 TF-condition pairs were significantly coherent after this filtration.

We examined the relationship between the results of clustering the genes assigned to a TF, in multiple experimental conditions. For each TF, we calculated the number of common genes in the largest cluster, in all pairs of conditions in which the EC score of that TF was significantly coherent. Figure

11 shows a plot of the distribution of these pairwise overlaps, which are significantly larger than that expected at random. Thus the sets of coherent genes of the same TF in different conditions significantly overlap.

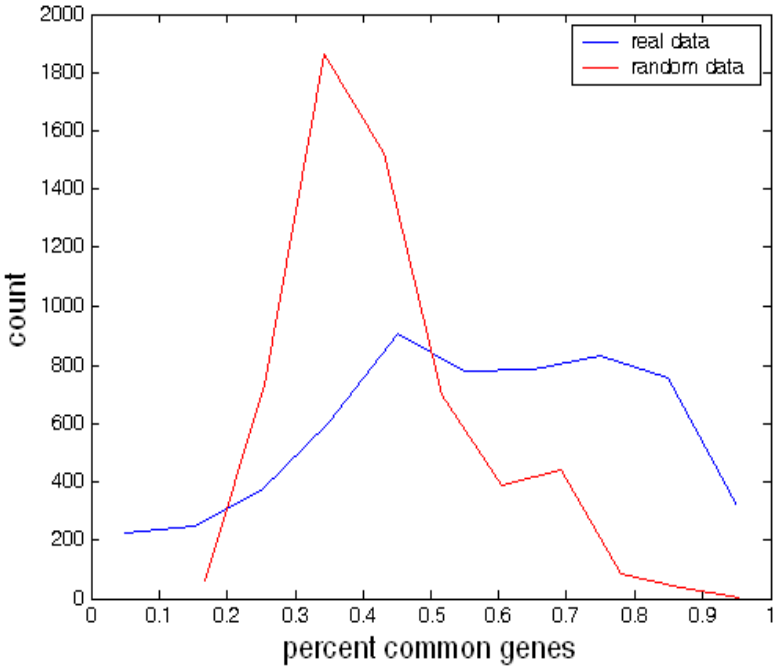


Figure 11: The relationship between the clustering of multiple conditions that correspond to the same TF, compared to that expected from random data. For each TF, we calculated the number of common genes in the largest cluster, in all pairs of conditions in which the EC score of that TF was significant. Plot shows the distribution of these pairwise overlaps (blue), compared to that expected by random (red). The random data was formed as follows: for each pair of conditions, the percent of expected common genes was obtained by randomly choosing two sets of genes (the sizes of the major cluster of each of the two conditions), out of the genes assigned to the TF, and calculating the percent of their overlap.

It is noteworthy that the location data was generated from yeast grown in rich medium, a growth condition quite different from many of the conditions for which we have expression data. Yet our analysis shows that often genes associated with many of the TFs display good coherence in multiple, diverse conditions. This may be taken to indicate that the TF is localized in the vicinity of its binding site, perhaps somewhat statically, and some additional modifications may render it active in the appropriate condition.

We next turned to investigate reasons underlying the existence of one or several large tightly controlled clusters of genes for each TF. In order to detect such behaviors, we inspected the results of the clustering of the mRNA expression profiles of genes assigned to each of the TFs. In order to

understand what may be responsible for a unique expression pattern of a subset of the genes assigned to the TF, we turned first to analyze regulatory sequence motifs.

2.3.3 Method (ii): Regulatory motifs analysis

It is generally thought that TFs bind short sequence elements, between six to 20 nucleotides long, found in the promoters of genes regulated by the TFs. We used AlignACE (Hughes et al. 2000), a Gibbs-sampler that searches for over-represented motifs in a set of DNA sequences, to derive regulatory motifs from promoters of sets of genes assigned to each TF in the location data. We identified a total of 567 significant motifs for 61 of the TFs. (A significant motif is one that had a MAP score > 10, as proposed in the original AlignACE paper (Hughes et al. 2000), and a group specificity score < 10^{-6}. In addition, we required that the ratio between the number of consecutive gaps and the nucleotides in the consensus sequence be <= 0.4, a threshold that reconciles removal of false motifs with maximization of the number of TFs for which motifs are derived.) See Sections 3.3.2 and 3.3.4 for definition of MAP and group specificity scores. We then turned to recalculate the EC score of genes assigned to each of these TFs, this time considering only a subset of these genes, namely the ones that contain the significant motif in their promoter. For each TF for which a motif was found we compared each such EC score to a distribution of 10,000 EC scores of random samples of genes assigned to the TF but that do not necessarily contain the motif. The sample size of each such random gene set was the number of genes assigned to the TF that also contained the motif in their promoter. We say that the motif improves the EC score of the TF in a given condition if its EC score is at the top 5% of the random scores distribution for that condition. Of the 738 TF-condition pairs shown as significant in Figure 7 across all 113 TFs, 641 pairs represent 61 TFs for which we found significant motifs. For 421 out of the 641 TF-condition pairs, we obtained a motif that significantly improves the EC score (data not shown). In these cases we hypothesize that the genes that contain a motif and belong to the cluster it dictates are the more likely targets of the TF.

An example of such behavior is the histidine and adenine biosynthesis regulator Bas1 that gives rise to incoherent expression profiles in the nitrogen

depletion condition (Figure 10B) (Gasch et al. 2000). Yet a motif we discovered by AlignACE, using the promoters assigned to this TF, whose consensus is RNMRGAGTCA (MAP score 24, group specificity score $3.8*10^{-10}$), is most highly over-represented in only one of the two major clusters of this TF. This motif is highly similar to a motif experimentally shown to be bound by Bas1 (Springer et al. 1996). While it is still possible that some of the genes in the other clusters are also targets of Bas1, by reassigning to this regulator only the genes that contain the motif found, we may have filtered a significant amount of false assignments.

Another interesting behavior is displayed by the genes assigned to the chromatin remodeling factor, Abf1 (see Figure 10A). AlignACE run on the promoters of 282 genes assigned to this TF resulted in two regulatory motifs: NCGTNNNNARTGAT (MAP score 390, group specificity score $1.6*10^{-98}$) that occurs in 262 of the TF targets and CGATGAGNTK (MAP score 26, group specificity score $9.9*10^{-6}$) that occurs in 37 of the targets. The latter motif is also known as the PAC motif, whose binding TF remains elusive (Dequard-Chablat et al. 1991). All of the 37 genes that contain the second motif in their promoters contain the first as well, and a possible interpretation is that these genes are under the regulation of at least two TFs. Interestingly, while a significant portion of the genes that have both motifs (green in Figure 10A) co-cluster in the sporulation condition shown here, across many conditions they display more complex behavior (not shown) that probably reflects condition-dependent dominance of either of the motifs.

When examining the significant motifs found, it is important to bear in mind that not all of the genes assigned a TF in the location data contain the significant motif found for that TF. The average ratio between the number of genes containing the motif and the number of genes assigned to the TF is about 39% (see Figure 12). This may indicate the level of noise in the data, although alternative motif-searching algorithms may change the exact picture.

The significant motifs discovered gave rise to a matrix in which gene i is assigned to TF j only if it was assigned in the original location data, and also the promoter of gene i contained a significant motif which was found for TF j. This matrix portrays filtration method (ii).

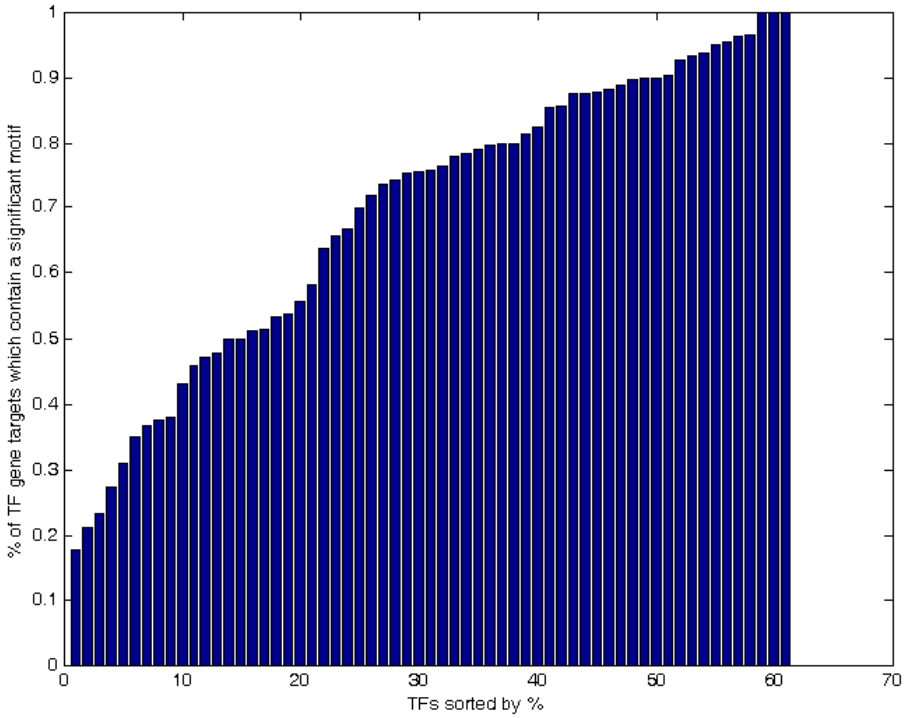


Figure 12: The figure shows the percent of genes assigned to a TF whose promoters contain at least one significant motif found for that TF, for the 61 TFs for which significant motifs were found.

2.3.4 Method (iii): Synergistic interactions between TFs

Figure 13 shows the expression profiles of the genes assigned to the regulator Ndd1 in the Carbon-1 medium in the environmental stress experiment (Gasch et al. 2000). Here again the expression of these genes is not coherent, yet the clustering shows that the gene expression profiles may be decomposed predominantly into two coherent clusters. Interestingly, we have identified two TFs, Swi5 and Mcm1, such that half of the genes bound by both Ndd1 and Swi5 fall in the largest cluster, and over a quarter of the genes bound by both Ndd1 and Mcm1 fall in the second largest cluster (see Figure 13). It thus appears that with alternative partners Ndd1 may participate in regulation of completely different responses. Interestingly, all three regulators in this set are known as cell cycle regulators, yet we provide here an indication that they are involved in the regulation of the response to nitrogen depletion, a process that evokes meiosis in yeast. This is another demonstration (Bussemaker et al. 2001; Pilpel

et al. 2001) of the extensive regulatory cross-talk between the meiotic and mitotic cell-division processes.

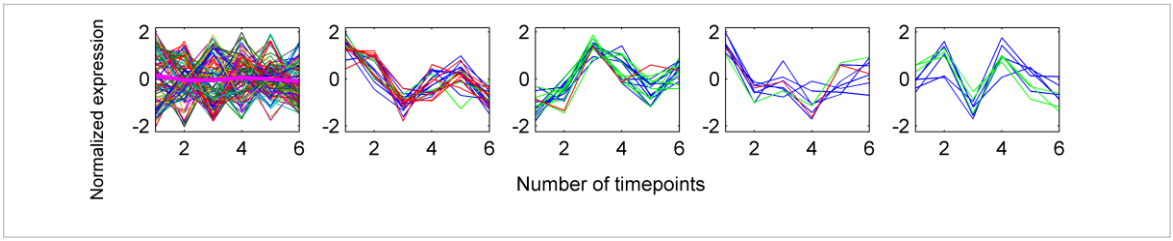


Figure 13: Genes assigned to Ndd1 in the Carbon-1 medium in the environmental stress experiment (Gasch et al. 2000), with the same QT_clust-based clustering as in Figure 10. Genes that are assigned to Ndd1 and Swi5 are colored red, while genes that are assigned to Ndd1 and Mcm1 are colored green. Genes assigned to Ndd1, but not to Swi5 and not to Mcm1 are colored blue.

In the detection of such combinatorial interactions between regulatory motifs Pilpel et al. previously defined motif synergy (Pilpel et al. 2001; Sudarsanam et al. 2002). A pair of regulatory motifs is considered synergistic if the EC score of the genes containing the two motifs together was significantly higher than that of the genes that contain either of the motifs alone. Zhang and co-workers have recently adopted this definition and explored synergistic interaction in the location data during cell cycle (Banerjee and Zhang 2003). We report here the detection of synergistic interactions among all pairs of TFs in the location data, in each of the above 40 conditions. A pair of TFs is considered synergistic if the EC score of the genes assigned to both TFs was significantly higher than that of the genes assigned to either of the TFs alone. We used the previous statistical formalism for calculating a p-value on the hypothesis that two TFs are synergistic (Pilpel et al. 2001; Sudarsanam et al. 2002). See section 3.6 for details on calculation of significance of synergy. For each condition, we derived a list of all pairs of TFs which are synergistic in that condition. This resulted in a total of 279 unique significant synergistic interactions across all 40 conditions.

An example of two synergistic TFs is given in Figure 14, which shows the expression profiles of the genes assigned to the regulator Yap5 during exposure to the reducing agent dtt (Gasch et al. 2000). The genes that are assigned also to Fhl1 are colored red, and appear predominantly in the largest cluster. Thus it appears that when the promoter of a gene is bound by both

Yap5 and Fhl1, the expression profile of this gene is likely to be very specific, and distinct from the expression profile of genes bound by Yap5 alone.

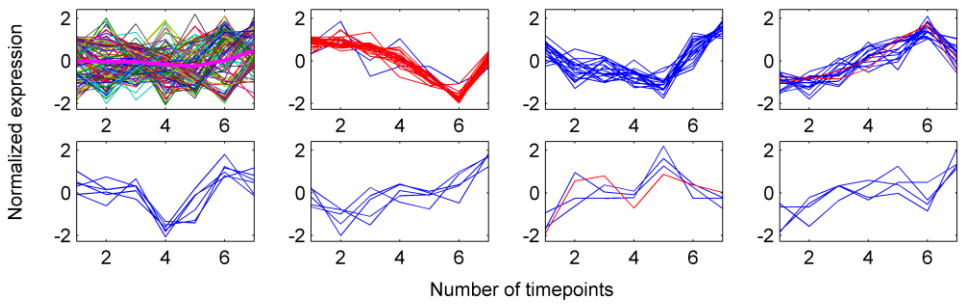


Figure 14: Genes assigned to Yap5 during exposure to the reducing agent dtt, with the same QT_clust-based clustering as in Figure 10 and Figure 13. Genes that are assigned also to Fhl1 are colored red, while genes only assigned to Yap5 are blue.

Figure 15 is a graph depicting all significant synergistic interactions of one of the 40 conditions, namely exposure to the dtt reducing agent (Gasch et al. 2000), in which synergism between the two TFs described in Figure 14 is highlighted. Similar maps in additional conditions, in addition to a combined map of all conditions, appear on the website.

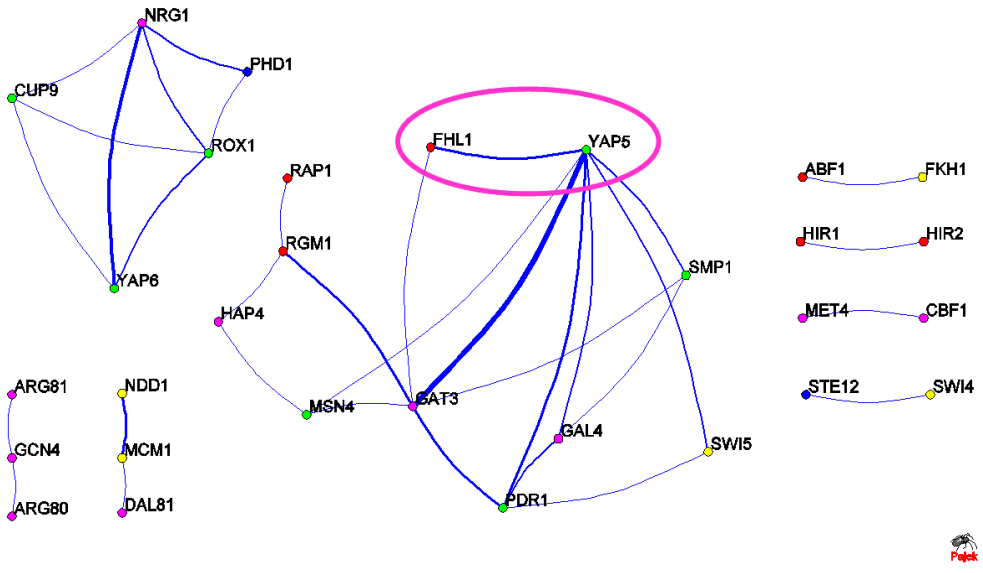


Figure 15: Graph depicts TF synergy during exposure to the reducing agent dtt. The nodes in the map represent TFs, an edge between two nodes represents significant synergy between the two corresponding TFs. Two nodes that are analyzed in detail in Figure 14, that correspond to Yap5 and Fhl1, are highlighted. Width of lines connecting two TFs reflects the number of genes assigned to both TFs; size of node reflects the number of genes assigned to the TF. Graph rendering was performed with Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm).

Synergistic interactions provide us with strengthened evidence of a true regulatory interaction. Thus this data of synergistic interactions produces the matrix of filtration method (iii). In this matrix, gene i is assigned to TF j only if it was assigned in the original location data, and also was assigned to another TF that shows synergy with TF j.

2.3.5 Method (iv): Co-localization of TFs in shared promoters

Another means to detect interactions between regulatory proteins, that does not involve expression data, is to detect their degree of co-localization in shared promoters. Two TFs are said to co-localize if they are shown in the location data to bind to the same promoter. Significant co-localization describes cases in which the number of promoters assigned to the two TFs is significantly large given the number of promoters assigned to each TF alone. The basic premise here is that if two or more TFs co-localize in a significantly high number of gene promoters, the genes in which the TFs co-localize are more likely to be true targets of the respective TFs compared to genes that are associated with each TF alone. We note, however, that high rate of co-localization of two TFs does not necessarily imply temporal co-localization, namely it may be that the two TFs are bound to the promoter in different conditions, perhaps even in a mutually exclusive manner.

Analogous to a previous motif co-occurrence calculation by Sudarsanam et al. (Sudarsanam et al. 2002), we consider two TFs, TF1 and TF2, as potentially functionally interacting if the number of promoters in which they co-localize is significantly high considering the number of promoters assigned to each of them individually. To test the null-hypothesis that the observed or higher rate of co-localization of two TFs could be obtained by chance given the above priors, we use the cumulative hyper-geometric probability distribution.

$$P(X \geq \mathrm{tf12}) = \sum_{i=\mathrm{tf12}}^{\min(\mathrm{tf1},\mathrm{tf2})} \frac{\binom{\mathrm{tf1}}{i}\binom{\mathrm{g}-\mathrm{tf1}}{\mathrm{tf2}-i}}{\binom{\mathrm{g}}{\mathrm{tf2}}}$$

where g is the number of promoters in the genome, tf1 and tf2 is the number of promoters assigned to TF1 and 2 respectively, and tf12 is the number of promoters assigned to both TF1 and TF2.

We have generated a graph of all pair-wise interactions in the location dataset (see Figure 16). While co-occurrence analysis of regulatory motifs was introduced before (Sudarsanam et al. 2002), we now provide an analysis at the level of the TFs themselves and show that most such interactions occur within one highly connected graph. The nodes of the graph correspond to TFs, and edges connect between pairs of TFs if the p-value on the hypothesis that they significantly co-localize falls below a determined threshold. For clarity of the graph, and due to the high number of significant co-localizations, we set a p-value threshold of 10-10. The graph displays several interesting properties. Coloring the graph according to the biological function ascribed to each TF, we discover clustering of TFs according to their annotated function. (For details on derivation of biological functions, see legend of Figure 16). In particular, we discern a highly connected cluster of cell-cycle regulatory TFs (see cluster I in Figure 16). This observation is similar to the one Pilpel et al. initially made with cell-cycle regulatory motifs yet with a completely different criterion for regulatory interactions (Pilpel et al. 2001). This is another clear indication that the cell-cycle is one of the most tightly controlled processes in yeast, and that an intricate network of regulators is at work in its regulation. The map shows two other clusters that are also rather homogenous in terms of the functions of the TFs they contain. This clustering by function suggests, as in other biological networks (Schwikowski et al. 2000), that a “guilt-by-association” approach may be used for annotating the regulatory role of poorly-characterized TFs by ascribing them the role of their annotated partners, if they occur in such “functionally coherent” clusters.

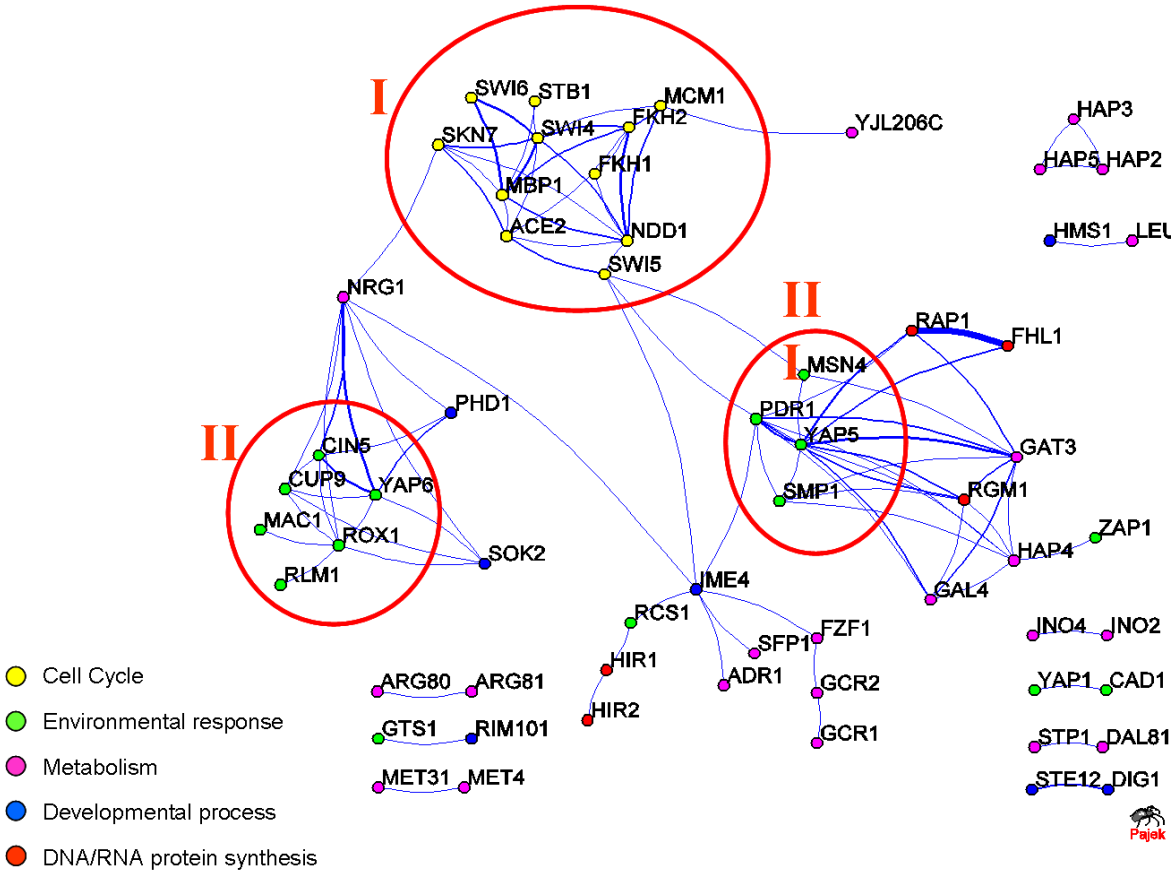


Figure 16: Graph depicts significant co-localization of TFs in common promoters. The nodes in the map represent TFs, an edge between two nodes represents significant (p-value < 10^{-10}) co-localization between the two corresponding TFs. Graph rendering was performed with Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.htm). The three main clusters of co-localized TFs are circled red. Nodes are colored according to the regulatory function of the TFs. Such functions were annotated in (Lee et al. 2002) according to the biological function of genes assigned to the TF. Width of lines connecting two TFs reflects the number of genes assigned to both TFs; size of node reflects the number of genes assigned to the TF.

Significant co-localization interactions provide us with strengthened evidence of a true regulatory interaction, and thus this data produces the matrix of filtration method (iv). In this matrix, gene i is assigned to TF j only if it was assigned in the original location data, and also was assigned to another TF which co-localizes significantly with TF j.

2.3.6 Relationship among the four methods of filtration

The four filtration methods discussed here each served to produce a higher quality matrix of TF-gene interactions. The numbers of interactions predicted by the single methods are 4044, 2795, 2313, and 2418 for clustering of gene

expression profiles followed by filtration of lowly populated clusters, motif detection, synergy, and co-localization analysis, respectively. Altogether, 1487 interactions were predicted by all four filtration methods presented in this study. Figure 17 shows the number of TF-gene assignments supported by each unique combination of methods.

# assignments		coherence	motifs	synergy	colocalized
AND[a]	OR[b]				
648	648	■			
49	49		■		
3	3			■	
63	63				■
1009	1706	■	■		
23	674	■		■	
106	817	■			■
0	52		■	■	
8	120		■		■
67	133			■	■
124	1856	■	■	■	
78	1961	■	■		■
569	1479	■		■	■
40	230		■	■	■
1487	4274	■	■	■	■

[a] Relationship between shaded methods is that of 'AND'. Number reports the interactions which are supported by all marked methods (but not by unmarked methods).

[b] Relationship between shaded methods is that of 'OR'. Number reports the interactions which are supported by any of the marked methods (but not by unmarked methods).

Figure 17: Table displaying number of TF-gene assignments supported by all possible combinations of methods. For each combination marked by gray boxes, the number of assignments supported by this unique combination of methods is reported. The first column reports the number of assignments supported by all of the methods marked in each row (an 'AND' relationship between the methods), while the second column reports assignments supported by any of the methods (an 'OR' relationship). For example, the fifth row reports that there are 1009 interactions supported by both the coherence and the motif method (and not supported by the other methods), and 1706 interactions supported by either the coherence or the motif method (and not supported by the other methods

Figure 18 shows an analysis of three of these methods: motifs, synergy, and co-localization. Each of the three methods utilizes a different type of data source – sequence analysis, expression data, or statistical analysis of common gene sets. The Venn diagram portrays the relationship between the cases of TF-gene interactions, and the methods which predict each interaction. For instance, there is a significant overlap between those interactions predicted by

the synergy method, with the interactions predicted by the co-localization method. In addition, there is a significantly large number of TF-gene interactions which were predicted by all three methods. A total of 3626 unique TF-gene interactions were predicted by the three methods.

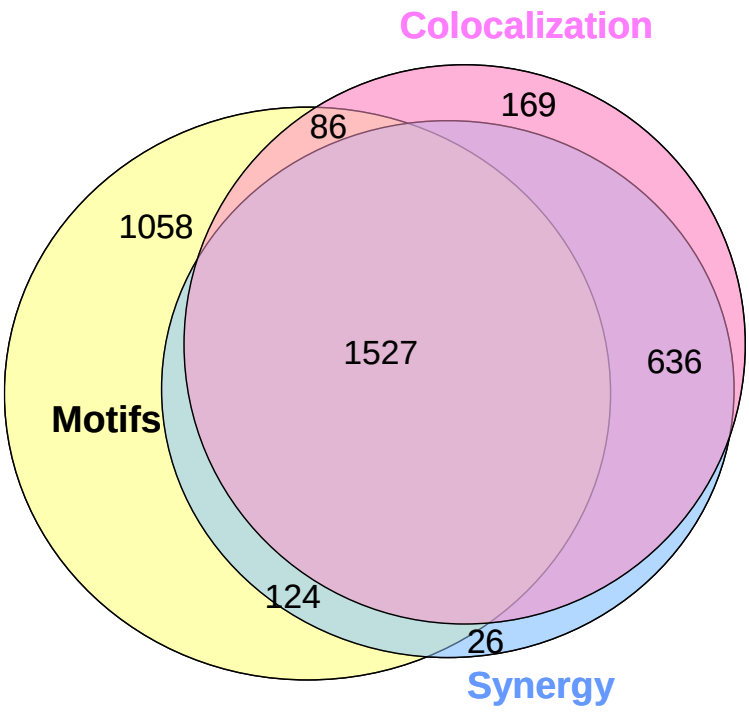


Figure 18: Venn diagram depicting the relationships among the TF-gene interaction predictions of three methods of filtration: motif detection, synergy, and co-localization. A total of 3626 unique interactions were predicted by at least one of the three methods, and 1527 interactions were predicted by all three methods.

Figure 19 and Figure 20 show in each of the four filtration methods, and in their union and intersection, per TF, the percent and absolute numbers of gene assignments not supported by the method, relative to the total number of genes assigned to the TF in the location data.

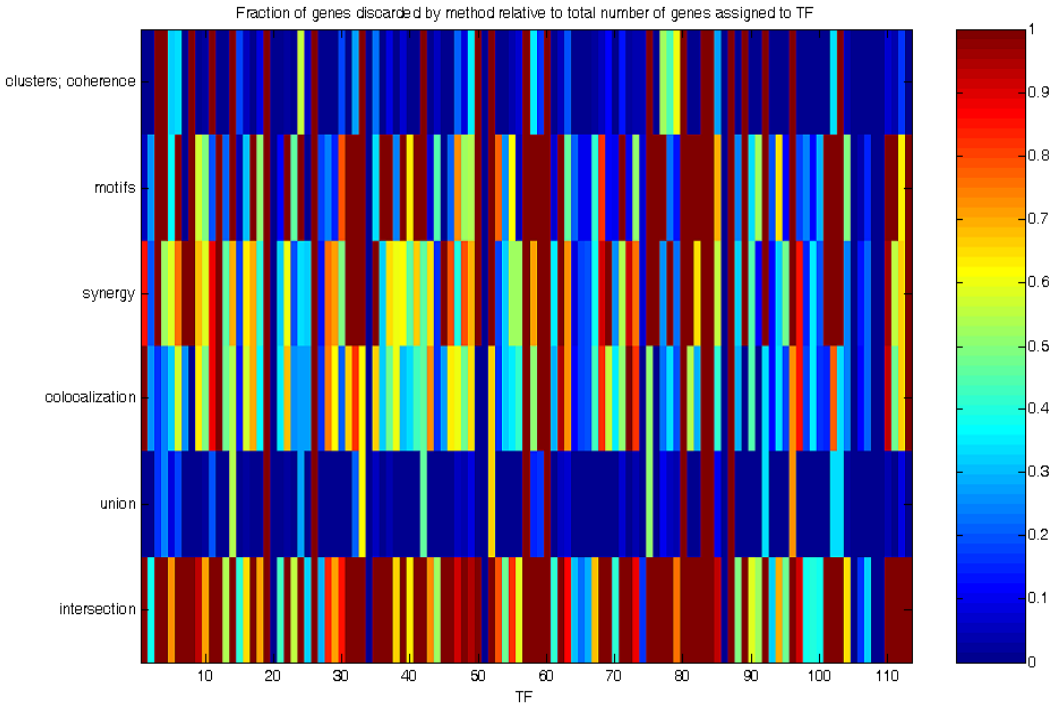


Figure 19: The image shows in each of the 4 filtration methods, and in their union and intersection, per TF, the fraction of genes discarded by the filtration, relative to the total number of genes assigned to the TF in the location data.

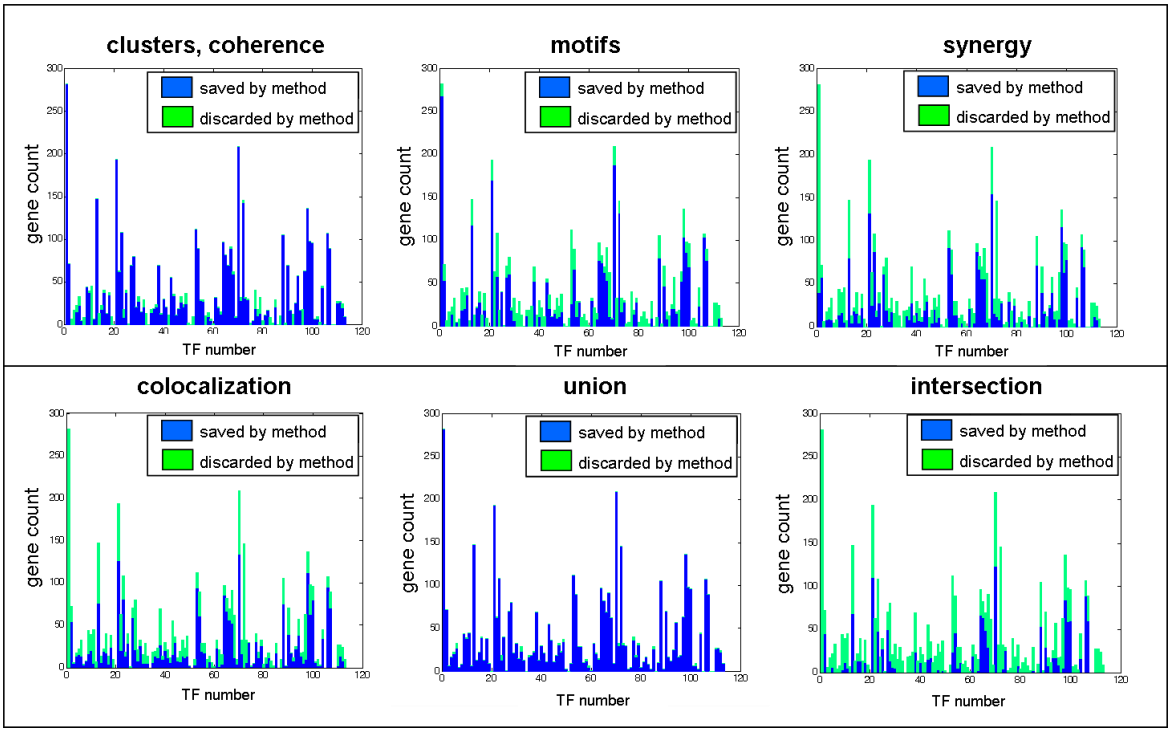


Figure 20: The bar diagrams below show in each of the 4 filtration methods, and in their union and intersection, per TF, the number of genes discarded by the filtration, and number of genes saved; for each TF the sum of these two numbers is the total number of genes assigned to the TF in the location data.

Finally we consider the matrices resulting from each filtration method as part of a more global prioritization scheme. On one extreme, the 1487 predictions supported by all four methods represent the highest-quality set of interactions. Nevertheless this set has the lowest coverage. The union of all four methods lies on the other end of the scale, and predicts 4274 interactions (all but 159 of the original TF-gene assignments in the location data). Between these two extremes are TF-gene assignments that are supported by various subsets of the filters. We have implemented a relatively simple prioritization scheme, offered on the supplemental website that ranks assignments based on the number of filters supporting them. In the future more sophisticated means will be offered that prioritize predictions according to the confidence of filter-specific scores supporting each assignment and partial dependencies between the different filters.

2.3.7 Interactive GUI on web server

A supplementary website for this work, which includes an interactive GUI, is available at http://longitude.weizmann.ac.il/TFLocation/TFLocation.html.

Included in the website is a Matlab GUI that allows exploration of the expression profiles of TFs in multiple conditions, detection of combinatorial interactions among them, and effect of regulatory motif on their coherence patterns. In addition, the website provides our noise-filtered version of the location database, and various interactive means for user-defined filtering strategies. Figure 21 shows a snapshot of the GUI.

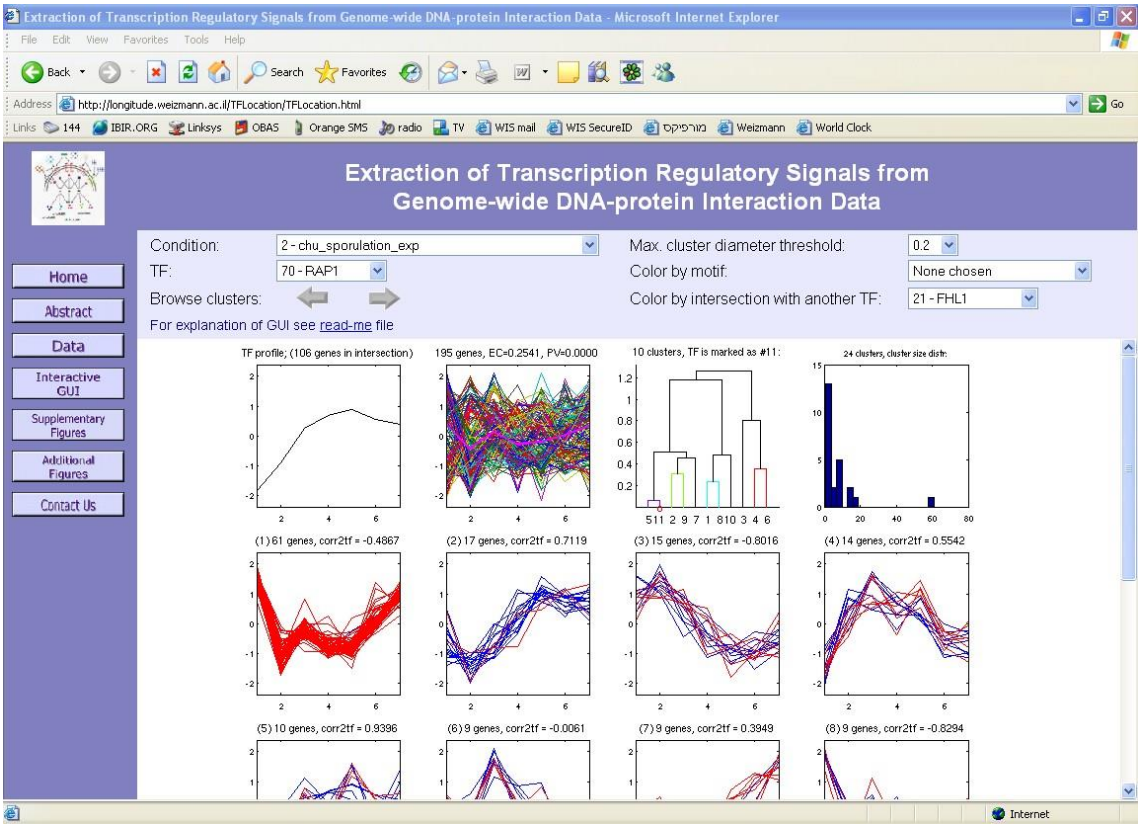


Figure 21: Snapshot of the GUI, available on the world-wide web. The current query shows the expression profiles of genes bound by Rap1, during the Chu sporulation experiment (Chu et al. 1998). Profiles of genes that are bound both by Rap1 and by Fhl1 appear in red, genes bound by Rap1 (but not by Fhl1) appear in blue.

2.3.8 New location analysis dataset by Harbison et al.

In September 2004, Harbison et al. published a new location analysis dataset which determined the genomic occupancy of 203 TFs (Harbison et al. 2004). This study was completed in Richard Young's laboratory at the Whitehead Institute, the same group which previously published the Lee et al. location data. As the infrastructure lay ready for analysis of any location

analysis dataset, we were excited to analyze and clean this new dataset using our noise filtration platform.

A bird's eye view of the data can be obtained by observing Figure 22, which shows the 203 TFs across all 40 conditions for which we have expression data. The figure shows which TF-condition pairs show significantly coherent expression profiles among the group of genes assigned to the TF. Compare this to Figure 7, which shows the same information for the 113 TFs studied in the Lee et al. study.

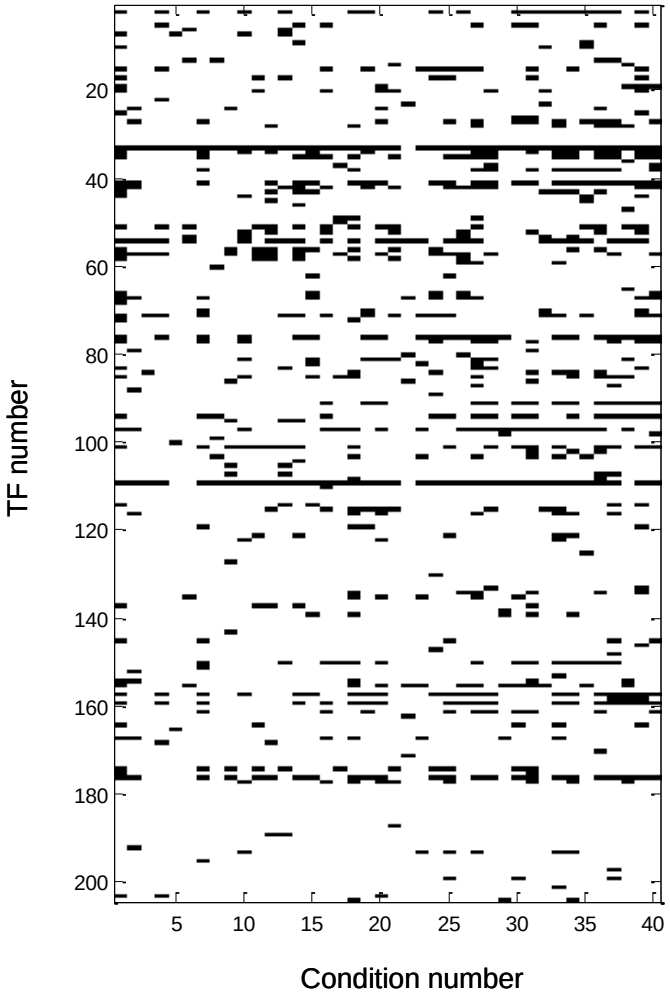


Figure 22: A matrix depicting expression coherence of each TF in each condition. An ij-th entry in the matrix is colored black if the i-th TF was significantly coherent in the j-th condition, and white otherwise. Data shown is the data from the Harbison et al. study (Harbison et al. 2004); all TF and condition names are available in Appendix A.

In the new study, 203 TFs were studied, of which there were 73 TFs whose set of assigned genes were not significantly coherent in any of the 40

conditions, and hence 131 TFs with at least one coherent condition. Altogether, 832 TF-condition pairs were significantly coherent. In the previous study by Lee et al., 29 of the 113 TFs had no condition for which they were coherent, and there were 738 TF-condition pairs which were significantly coherent. It is apparent from the surprisingly small rise in coherent TF-condition pairs, that the TFs included in the new study (that were not previously studied) do not have gene sets that are coherently expressed in the set of conditions we examine. In addition, it is interesting to note that the general trend of the number of conditions regulated per TF, and number of TFs regulating per condition, have remained the same, as visible by comparison of Figure 23 and Figure 24 with Figure 8and Figure 9, respectively.

Future research will include a complete analysis of the new dataset with the tools which were built during the present study. It will be interesting to see what level of regulatory noise exists in the data of the additional TFs that were included in the Harbison et al. dataset.

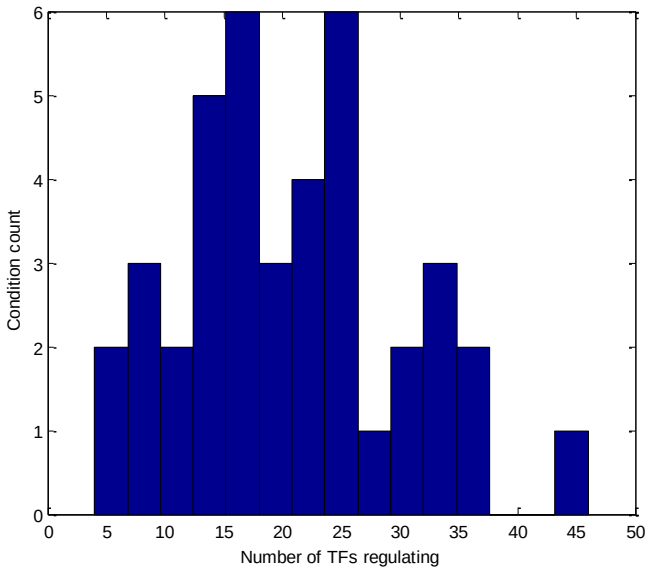


Figure 23: A histogram with the number of TFs regulating each condition (Harbison et al. data).

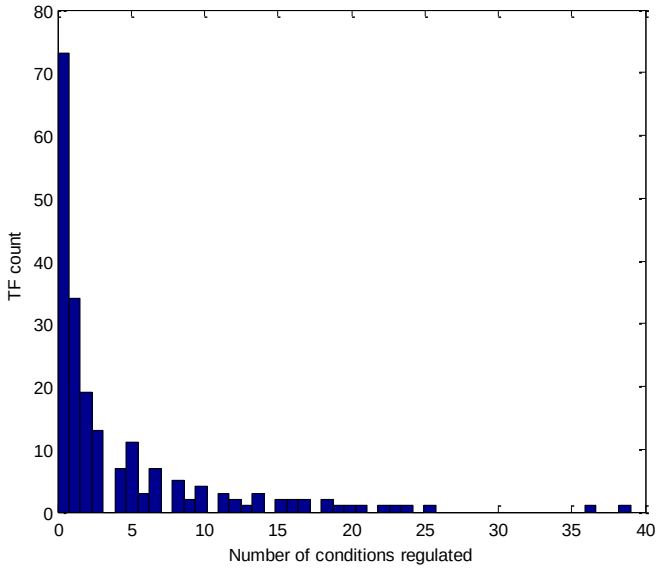


Figure 24: A histogram depicting the number of conditions regulated by each TF (Harbison et al. data).

2.3.9 Comparison of our work to other studies

In the work described here, we have significantly improved the accuracy of the DNA-protein location data and, through this process, have gained new insights on gene network design principles. An approach developed recently by Bar-Joseph et al. (Bar-Joseph et al. 2003) is most useful for adopting a more permissive p-value threshold on TF-gene assignments in the location data in order to reduce false negatives, when TF combinatorics and expression data support it.

Our methods, on the other hand, are mainly aimed at removing false positives. In that respect the two approaches are complementary to each other. Another method that prioritized TF-promoter interactions based on the location and expression data was that of Gao et al.(Gao et al. 2004). We have thus performed comparative analysis that gauged the extent of overlap between TF-gene assignments supported by the three studies, using in our study the intersection of assignments derived from all four filters (see Figure 25 and Figure 26). We found that the three studies produce significantly overlapping sets of assignments, yet each study identifies unique assignments that the other studies do not provide support for. Among the possible reasons for lack of

congruence are Bar-Joseph's algorithm's sensitivity towards assignments with higher than 0.001 p-value, our explicit reliance on TF synergies, co-localization and sequence motifs, and Gao's emphasis on contribution of the TF to the expression fold change of regulated genes at individual time points (as opposed to effect across an entire time series).

We calculated the Meet/Min and Jaccard coefficients (Goldberg and Roth 2003), two measures of overlap between sets, between the lists of genes assigned to each TF by Gao et al. (Gao et al. 2004), Bar-Joseph et al. (Bar-Joseph et al. 2003), and by the intersection of our four methods (intersection matrix). These coefficients are respectively defined as the size of the intersection of two sets divided by the size of the smaller of the two sets, and the size of the intersection of the two sets divided by the size of their union.

We have only analyzed TFs for which there exists at least one gene assignment by all three works (Gao, Bar-Joseph, and our own). The figures color-code the Meet/Min and Jaccard coefficients between each pair of studies.

	Average Meet/Min	Average Jaccard
Gao vs Bar-Joseph	0.6181	0. 2795
Gao vs Ours	0.7082	0. 3848
Bar-Joseph vs Ours	0.4575	0. 2106

It is clear from the figure that the work of Gao and ourselves are highly congruent and the average Meet/Min coefficient across 28 TFs is 0.7082. On the other hand, the Bar-Joseph assignments show somewhat lower congruence with Gao's study and an even lower similarity with the present work.

Note that the Meet/Min coefficient minimizes the differences between the sets of genes assigned to a TF by each pair of studies, which stem directly from the addition by Bar-Joseph et al. of gene targets assigned p-values greater than 0.001 in the location data.

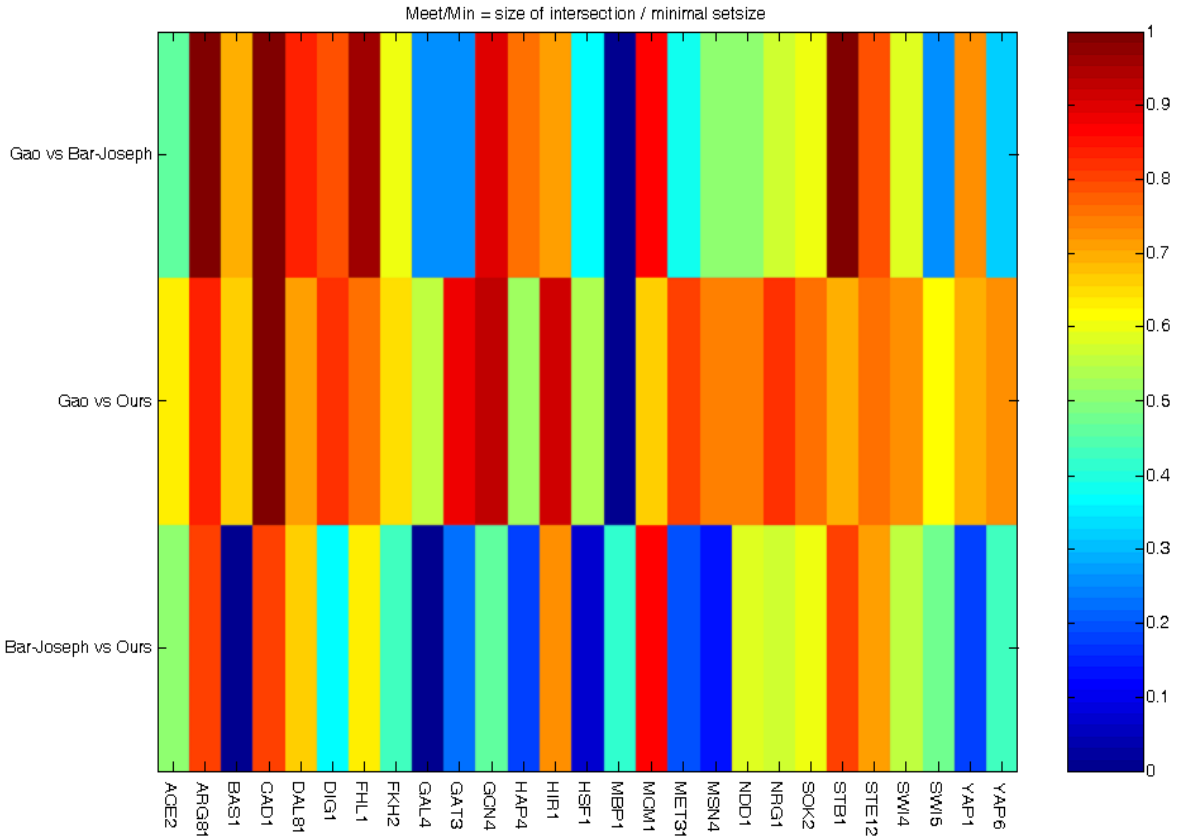


Figure 25: Meet/Min matrix: each row of the matrix colorcodes the Meet/Min coefficient between the specified pair of studies: Gao and Bar-Joseph, Gao and ours, and Bar-Joseph and ours for rows 1-3 respectively. Each column represents a different TF, as specified in the x-axis.

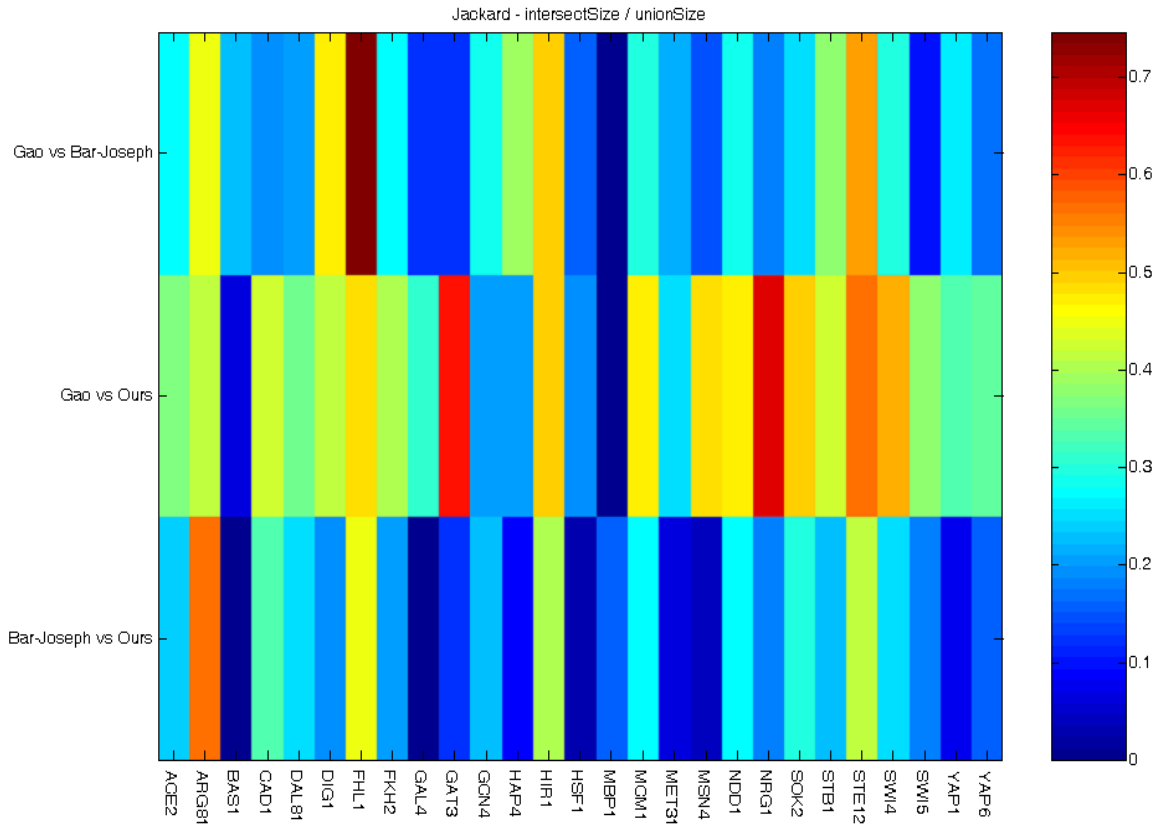


Figure 26: Jackard index matrix: each row of the matrix colorcodes the Jackard coefficient between the specified pair of studies: Gao and Bar-Joseph, Gao and ours, and Bar-Joseph and ours for rows 1-3 respectively. Each column represents a different TF, as specified in the x-axis.

2.4 Regulatory Motif Dictionaries project

As briefly described in section 2.2.2, QT_clust was used to cluster gene groups based on expression profiles and also based on sequence in the context of the 'regulatory motif dictionaries' project. This project is a novel effort to create a comprehensive catalog of regulatory motifs which control the transcriptional program of the cells of an organism. The project aims at cataloguing all those motifs which exert regulatory control over genes. Regulatory control in this case is measured by expression coherence of the set of genes which contain the motif (presumably a binding site for TFs) in their promoter.

The dictionary generation flow is as follows: in the first stage of creating the dictionary, we exhaustively scan genomic sequences of gene promoters for all possible k-mers (with k ranging from 7 to 11). For each k-mer, we obtain a list of all genes which contain the k-mer in their promoter regions. In the second stage, we calculate the EC score of the set of genes associated to each k-mer. The EC score is calculated in each of 40 different conditions, each of which corresponds to a time-series experiment. In the third step, we identify which specific k-mers are significant by performing a multiple hypothesis control, applying the false discover rate (FDR) method to the results (Benjamini and Hochberg 1995). In this context, a motif is considered significant if the genes in whose regulatory region it appears, display statistically significantly coherent mRNA transcript expression. The fourth stage is a clustering stage, where k-mers with sequence similarity are combined, thus constructing expression specific sequence matrices (ESSMs). These matrices describe both candidate motifs and the effect on expression of substitutions from the motif. By combining sequence and expression data, we are able to distinguish true, strong regulatory motifs, and to assign condition dependence to the discovered motifs.

The result of the third stage of this flow is a list of candidate motifs. When run on expression data of yeast in cell cycle (Cho et al. 1998), the result was 1102 significant motifs. Each of the 1102 motifs found in this condition appears

in the promoters of between 5 and 47 coherently expressed genes. In order to complete the fourth stage we must cluster the motifs into 'sequence-clusters'. To this end, we used a simplistic sequence distance measure which aligns, without gaps, two sequences such that the fraction of nucleotides that do not match perfectly is minimal. When calculating the fraction, the length of the shorter of the two sequences is used as the denominator, since this is the maximal number of nucleotides which may be perfectly matched. This measure also allows offsetting the sequences relative to one another, and allows alignment of one sequence with the reverse complement of the other if it acquires a better score. Thus, if dealing with 10-mers, a score of 0.3 means that 3 of the 10 nucleotides are not perfectly aligned. Using this sequence distance measure, we built a distance matrix between all 1102 significant motifs, and then submitted this matrix to the QT_clust algorithm. The 1102 significant motifs clustered into 301 sequence-clusters.

2.4.1 Graphical User Interface for viewing dictionary data

The decomposition performed by QT_clust allows us to analyze subsets of data with similar properties. However, it is essential to view the data in order to further analyze the massive amounts of data. When clustering together motifs based on sequence, we would like to verify that we do not combine motifs that seem to exert very contrasting transcription expression profiles. Thus, we built a graphical tool which was very helpful in analysis of the results of generation of the dictionaries. See Figure 27 for a snapshot of the GUI. This tool was built with the intention of enabling the extraction of signal of many sorts, from a variety of data types.

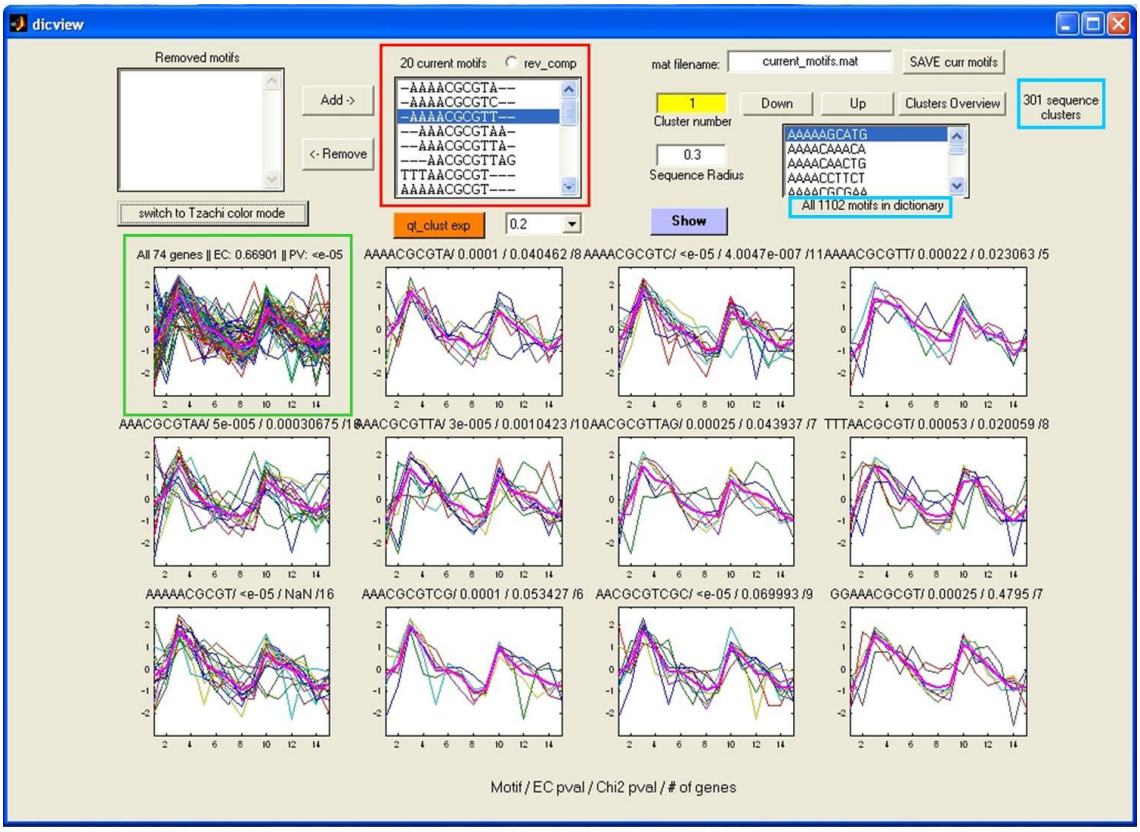


Figure 27: Snapshot of GUI used to view regulatory motif dictionaries.

As seen in the blue boxes of Figure 27, this dictionary contains 1102 motifs, which cluster into 301 sequence-clusters. The sequence-cluster diameter in this example was set as 0.3, meaning that any pair of sequences in one cluster may have a maximum of 3 of the 10 nucleotides mismatched. The snapshot displays sequence-cluster number 1 of the dictionary (the current cluster number is displayed in the yellow box). In the red box, the title states that there are 20 motifs which populate cluster number 1. The aligned motifs themselves can be seen in the red box. The panel in the green box shows all 74 genes which contain any one of the 20 motifs in their promoters. It is evident that this sequence-cluster, comprised of 20 motifs appearing in 74 genes' promoters, is also very tightly clustered in expression space. Therefore, in this example it is quite clear that it is biologically meaningful to cluster together these 20 motifs and build an ESSM of them; the presence of a motif matching this ESSM in the promoter of a gene in the genome confers a coherent expression pattern.

Figure 28 below shows the result of applying QT_clust to the expression profiles of all 74 genes. It is evident that at the cluster diameter parameter

chosen, there is in fact one large, extremely coherent cluster of genes, while the remaining outlying genes are quite dissimilar in expression profile. These genes are perhaps either loosely regulated by this motif, or perhaps the motif is not in fact bound by a regulatory protein in the condition under which this motif exerts active regulatory control.

The remaining 11 panels visible in Figure 27 above show information regarding 11 of the motifs in sequence-cluster number 1 of the dictionary. The motif described appears in the title above each panel, along with its EC score and the number of genes whose promoters contain it, and the panel shows the expression profiles of the genes which contained this specific motif in their promoter.

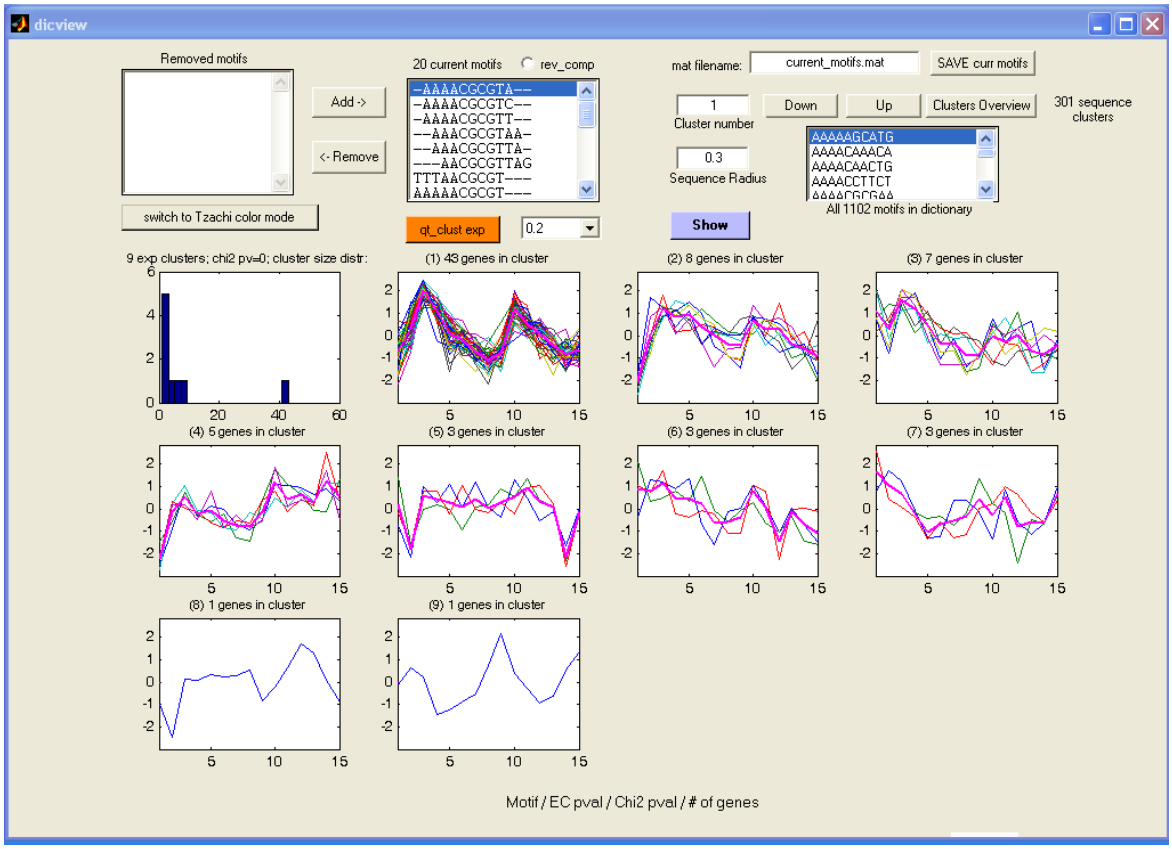


Figure 28: Snapshot of activating QT_clustering of expression profiles of genes whose promoters contain instances of the motifs shown in Figure 27. There is one major cluster of 43 genes, and the remaining genes are dissimilar in profile from this cluster.

Cluster number 1, described in detail above, is an example in which both sequence and expression profiles are highly similar among the various genes

related to the cluster, although they contain different motifs in their promoters. However, in principle two motifs can give similar expression profiles such as the ones in the example, but may also yield different ones. Using this GUI, one can ask questions such as: 'Are the genes associated with a particular motif activated or repressed?' or 'How many peaks are there during the cell cycle?' It is clear that the majority of the genes in cluster number 1 have periodic expression profiles; their expression level peaks twice during cell cycle, during the G1 and G2 phases.

In order to see what all sequence-clusters of the dictionary look like, one can use the GUI in order to see a "Clusters Overview", in which one can browse the various sequence-clusters, viewing per cluster, the expression profiles of all genes associated with the cluster (see Figure 29 below for a snapshot of this view).

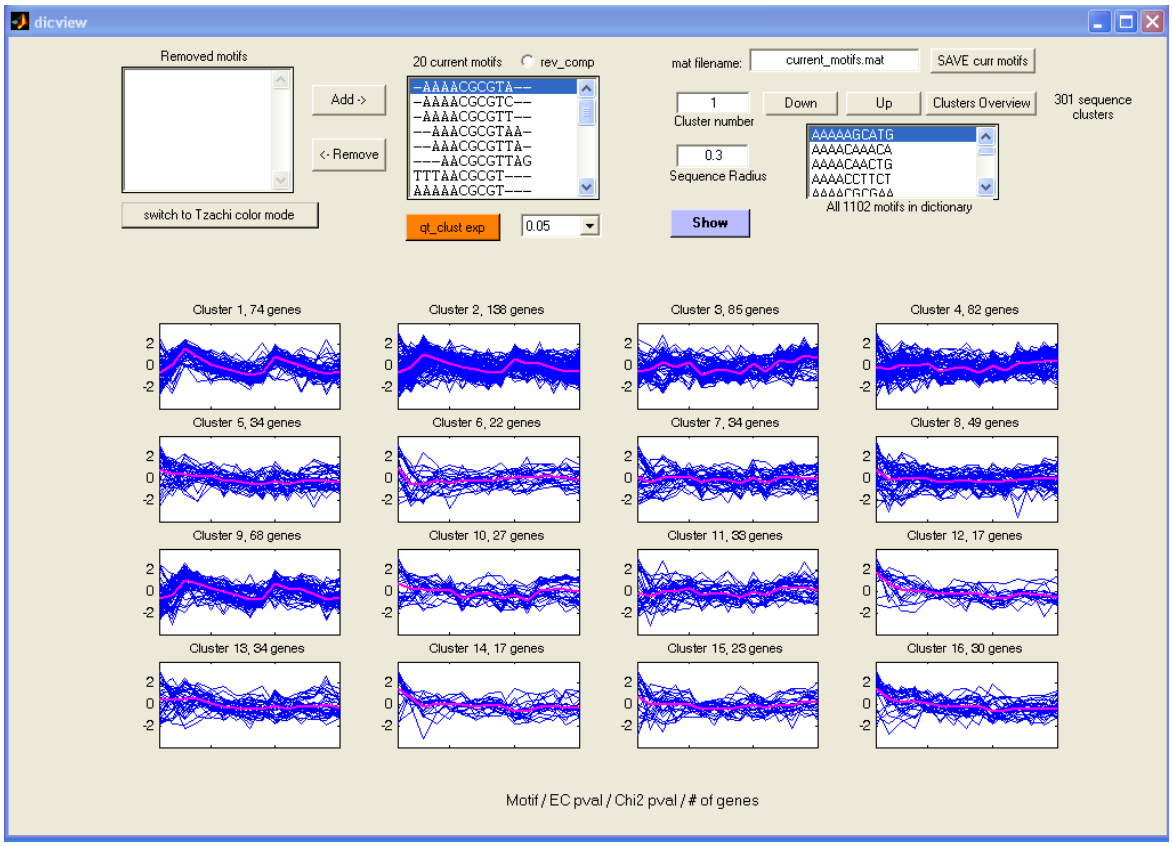


Figure 29: Overview of first 16 sequence-clusters in the dictionary. For each cluster, the expression profiles of all genes associated with the cluster is shown. One may browse through all 301 sequence clusters in this dictionary.

The strength of this tool is that in using it, one can easily view all sequence-clusters of a dictionary, and visually consider both the sequence similarity and expression similarity of the motifs composing each cluster. One can also update the clusters by, for example, removing certain motifs from the sequence-cluster, and then analyzing sequence similarity and even performing QT_clustering on the expression profiles, in order to see if the sequence-cluster is more coherent without the questionable motifs. In this way, one can refine the definition of a particular regulating ESSM. Figure 30 shows a snapshot of cluster number 65 of the dictionary. It is visible from the first panel that not all of the genes whose promoters contain motifs in this sequence-cluster have a coherent expression pattern. They seem to compose two or three main expression patterns. Using the GUI, we can remove the last two motifs from the cluster (the gene expression profiles of the genes are shown in the the last two panels). The result is shown in Figure 31. Now the first panel shows all genes associated with the 3 motifs remaining in the cluster, and they are significantly coherent

(EC score p-value improved from 0.00011 to $<10^{-5}$). The GUI allows one to cluster the expression profiles of a subset of motifs using the QT_clust algorithm, and to save the results for further investigation.

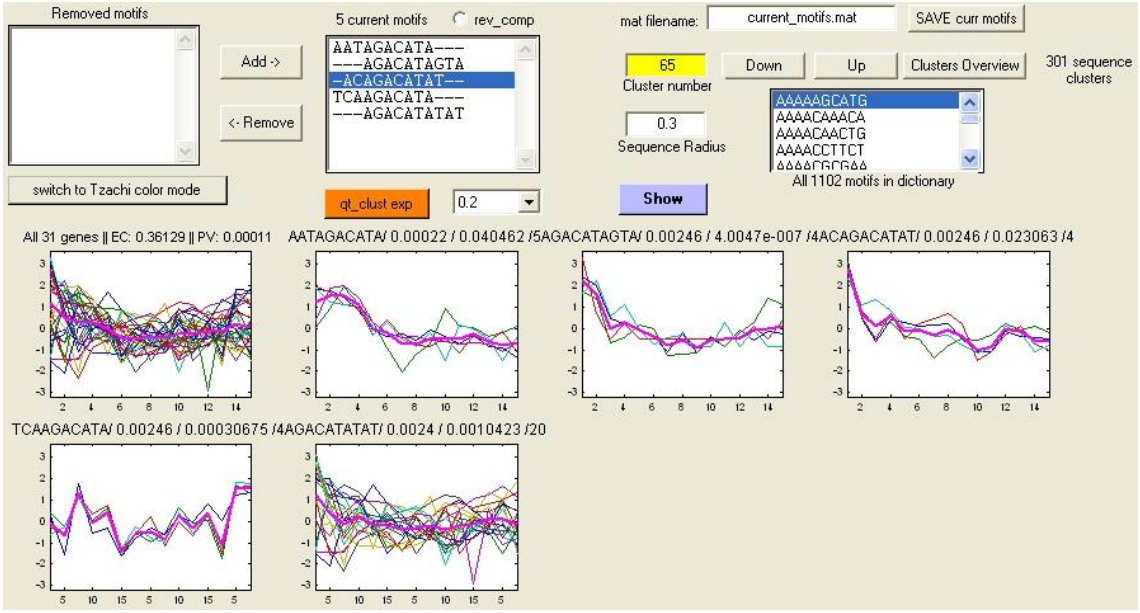


Figure 30: Sequence-luster number 65 of the dictionary contains 5 motifs. Expression profiles of all 31 genes containing any of the 5 motifs are shown in first panel of GUI, as well as in separate subsequent panels which show the genes per motif.

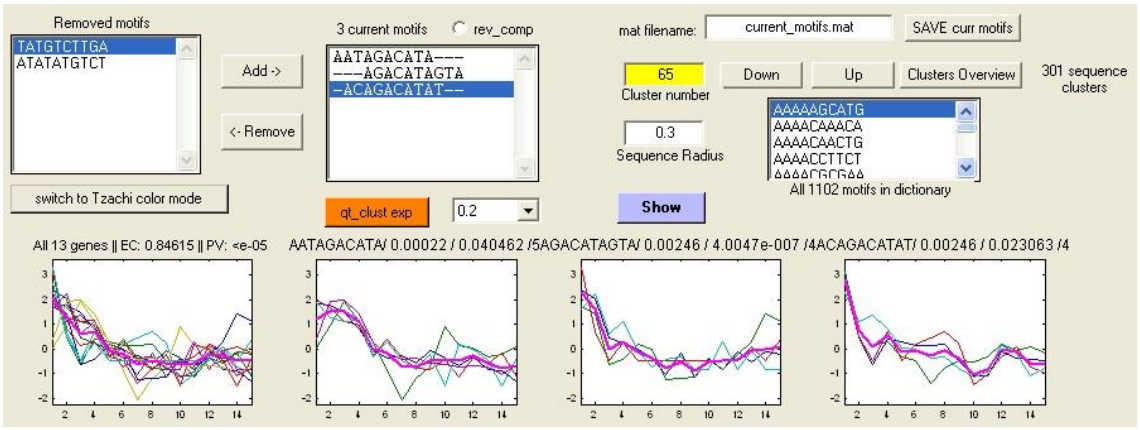


Figure 31: Sequence-luster number 65 of the dictionary now contains only 3 motifs (the last two were removed and appear in the top left box titled 'Removed motifs'). Expression profiles of genes containing any of the 3 motifs are shown in first panel of GUI, and are visibly more coherent than were the entire group of genes shown in Figure 30.

Often we are interested in a single motif, and would like to see the genes related to this motif, and perhaps ask which other motifs are highly related to it in sequence. One of the additional features that the GUI allows is to view the

data related to a single motif, as opposed to the whole cluster-oriented view previously described.

The decomposition performed by QT_clust combined with the strength of the visualization tool developed allows us to analyze subsets of data with similar properties. These subsets can be searched based on criteria (e.g. high correlation of a subset to the profile of the TF itself), and prioritized, or manually (visually) analyzed in order to extract significant biological insights and understanding.

3 Methods

3.1 mRNA expression data

Whole-genome mRNA expression data of 40 time series in yeast were obtained from ExpressDB (Aach et al. 2000). These time series represent a wide range of natural (e.g. cell cycle) (Cho et al. 1998; Spellman et al. 1998; Roberts et al. 2000) and perturbed (Chu et al. 1998; Eisen et al. 1998; Gasch et al. 2000; Jelinsky et al. 2000; Causton et al. 2001) conditions. The number of time points ranges from 3-28 in these experiments. Detailed description of all analyzed conditions appears in Appendix A. The data was normalized as follows: first, the intensity values were $\log_2$-transformed. Second, the mean of the transformed expression level of each gene was subtracted from all measurements of that gene, such that the mean of the gene expression level is zero. Finally, every centered measurement was divided by the standard deviation, such that its variance and standard deviation became unity.

3.2 Location Data

In-vivo TF location data provides a list, for each TF, of the promoters that are detected to be bound by it *in-vivo*. This is a result of an immuno-precipitation assay in which DNA-binding proteins are allowed to bind their target sites along the genome, followed by detection of the sites bound by each protein individually. We used the data produced by Lee et al. (Lee et al. 2002) obtained for yeast cells grown in rich medium. The TF-promoter assignments in that data are provided in the form of a p-value on the hypothesis that there exists an interaction between a TF and a promoter. In all analyses reported here we adopted the most restrictive p-value as suggested in the original publication (Lee et al. 2002), a p-value threshold of 0.001. For the purpose of our analysis, we only used intergenic region bindings that occurred upstream of an open reading frame.

The data of Lee et al. was downloaded in April 2003 from: http://staffa.wi.mit.edu/cgi-

bin/young_public/navframe.cgi?s=17&f=downloaddata. It provides data for 113 TFs, while the original Lee et al. paper included genome-wide location analysis experiments performed for only 106 yeast strains that expressed epitope-tagged regulators.

Since about a quarter of yeast genes are arranged in pairs transcribed from divergent promoters, the number of intergenic regions is considerably smaller than the number of ORFs. On the other hand, long intergenic regions were segmented in the location data chips. In total the number of printed probes was 6756 and the number of ORFs in this dataset is 6270.

Sequences of the intergenic regions printed on the chips were obtained from Richard Young's group at the Whitehead Institute (courtesy of Itamar Simon).

The data of Harbison et al. was downloaded in December 2004 from: http://jura.wi.mit.edu/young_public/regulatory_code/GWLD.html. It provides data for 203 TFs. This dataset includes the 113 TFs studied by Lee et al., however the experiment was repeated for these TFs and did not use the actually location binding data of Lee et al (Lee et al. 2002).

3.3 AlignACE, ScanACE, and group specificity score

3.3.1 AlignACE

AlignACE (**Align**s Nucleic **A**cid **C**onserved **E**lements) is a program which searches for sequence elements in a set of DNA sequences, using a Gibbs sampling strategy (Hughes et al. 2000). An iterative masking procedure is used to allow multiple distinct motifs to be found within a single data set. AlignACE defines a motif as the characteristic base-frequency patterns of the most information-rich columns of a set of aligned sites. The **m**aximal ***a***-***p****riori* (MAP) score is the criterion on which the final output motif is based (see below). A Linux version of AlignACE was obtained from Jason Hughes of George Church's Lab at Harvard Medical School (http://arep.med.harvard.edu), and AlignACE was run using default parameters.

For each TF examined in the location data, AlignACE received as input the FASTA-formatted sequence file of the promoter regions of the set of genes assigned to the TF (the intergenic region upstream of each gene).

3.3.2 MAP score

The MAP (maximum *a priori* log likelihood) score is used by AlignACE to judge different motifs sampled during the course of the algorithm. A crude approximation of the MAP score is given by the formula N*logR, where N is the number of aligned sites and R is the degree of over-representation of the motif in the input sequence. To summarize the general properties of this score, the following lead to higher MAP scores for otherwise similar motifs: greater numbers of aligned sites, less total input sequences, more tightly packed information-rich positions, more tightly conserved motifs, and enrichment of the motif with nucleotides that are less prevalent in the genome (the base frequencies in the genome are taken into consideration; 62% A+T in the case of *S. cerevisiae*).

3.3.3 ScanACE

ScanACE (**Scan**s for Nucleic **A**cid **C**onserved **E**lements) is a program which scans DNA sequences for elements which match a DNA motif found by AlignACE. It uses a weight matrix approach. The program finds the best matching sites for a motif in the target sequence. For consistency, it uses the same scoring mechanism that AlignACE uses in its sampling phase.

In our study, ScanACE was used to scan the FASTA file of all intergenic regions in the *S. cerevisiae* genome, and match it to the DNA motifs found by AlignACE.

ScanACE finds all sites scoring better than a cutoff based on the mean and standard deviation of the scores of the aligned sites of the motif. We used the default parameter; all sites scoring above the mean of the aligned sites that formed the motif itself were indexed. The positions of the sites are returned by ScanACE, and this information may then be used to generate the necessary data for calculating the group specificity score.

3.3.4 Group specificity

The group specificity score is a measure of how well (or how specifically) a given motif targets the genes whose upstream regions were used to find it. For each motif found by AlignACE, the ScanACE output is used to rank all intergenic region sequences according to the strength of the site best matching the scoring matrix in each intergenic sequence. The top 100 sequences in this list are compared to the sequences in the group used to find the motif. More than 100 intergenic sequences are included in the target list. Next, the probability that these sets would have the observed intersection or greater is calculated. This probability is what we refer to as the group specificity score. It is given by the formula:

$$S = \sum_{i=x}^{\min(s1,s2)} \frac{\binom{s1}{i}\binom{N-s1}{s2-i}}{\binom{N}{s2}}$$

where N is the total number of intergenic sequences, s1 and s2 are the numbers of intergenic sequences in the group used to find the motif and in the list of target sequences, respectively, and x is the number of intergenic sequences in the intersection of the two lists. This statistic quantifies the degree to which a motif is specific to the intergenic regions from which it was found.

3.4 Calculating the false discovery rate

In many instances of the present analyses we generate a multiplicity of hypotheses. We adjust p-value thresholds on the generated hypotheses by controlling the rate of false discovery as follows:

Let R denote the number of hypotheses rejected by a procedure.

Let V denote the number of true null hypotheses erroneously rejected (type I error).

Q = V/R when R>0 and 0 otherwise.

The false discovery rate (or q-value) is the expected proportion of false positives (type I error) among the rejected hypotheses. It is given by the following False Discovery Rate (FDR) theorem formula (Benjamini and Hochberg 1995): FDR = E(Q) .

In the context of the location data analysis, V is the number of expected false positive binding predictions. At a given p-value threshold p, V = p * Number of hypotheses. In the dataset produced by Lee et al:

'Number of hypotheses' = #TFs * #intergenic regions = 113 * 6756 = 763,428. Thus, V = 0.001 * 763,428 = 763 false positive binding predictions.

At the p-value threshold of 0.001, 4177 TF-intergenic region assignments were predicted by the location data. Of these 4177, 763 are expected to be false discoveries.

R is the number of predicted binding events at the current p-value threshold. In the location data the probability that R>0 is effectively equal to one at the p-value of 0.001. Thus, Q = V/R. Therefore, calculation of the FDR q-value for a p-value of 0.001 yields a q-value of 0.18, or 18% (q-value = 763 / 4177 = 0.18). The blue line in Figure 32 shows all hypotheses of the location data, plotted sorted by size. If a line is drawn from the origin, to 0.18 for the hypothesis with the largest p-value (the green line in the figure), it crosses the p-value curve at hypothesis number 4177. This means that for these 4177 hypotheses (those with p-value<=0.001), the expected false discovery rate is 18%. If we would like to allow less false positives, for example an expected FDR of 10%, then only the 2750 hypotheses with the lowest p-values should be accepted (marked by the point at which the blue line crosses the red line in the figure).

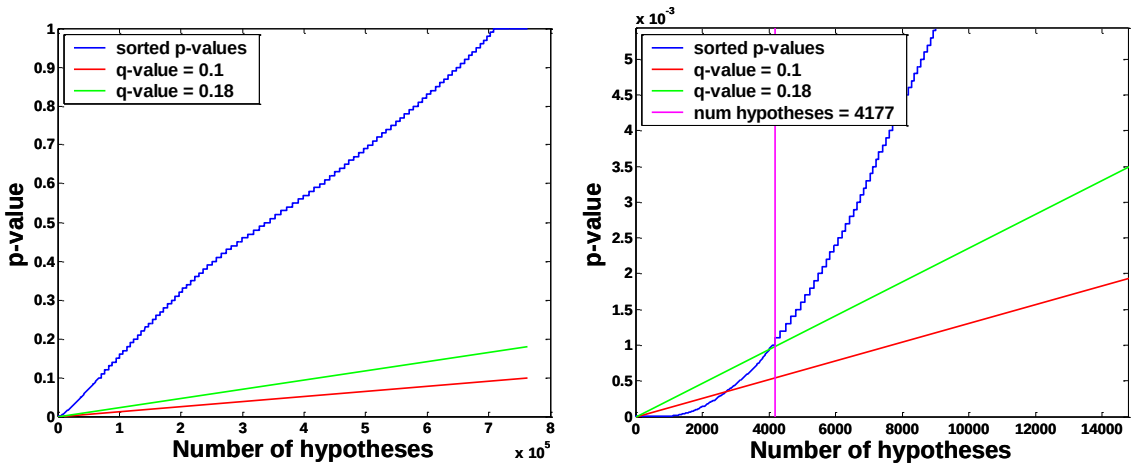


Figure 32: Plot demonstrating the use of FDR on the location data. Sorted p-values on hypotheses are plotted in blue (763,428 hypotheses), in red and green are lines corresponding to q-values 0.1 and 0.18 respectively. The point at which the lines intersect (p-values line and q-value line) is the actual number of accepted hypotheses.

Figure 33 shows an analysis of the effects of using various p-value thresholds to capture true assignments in this dataset. We examined the relationship between two conceivable methods of filtering the data: (1) using a p-value stricter than 0.001 on the binding predictions of the location data, and (2) using TF-gene assignments which have the support of at least three of the four filtration methods described in the text (Section 2.2). We show here that our methods save many assignments which have supporting evidence, which would be discarded by using a stricter p-value. The figure shows the ratio between the number of TF-gene assignments filtered out by both methods and those filtered out of the data by using a stricter p-value alone.

The histogram shows the following, as a function of changing the p-value threshold from 0.001 to the value shown on the x-axis:

In red, the number of specific assignments which are filtered out by both methods; using a p-value stricter than 0.001 on the location data predictions, and also by the requirement of support of at least three of the four filtration methods described.

In blue, the number of assignments lost when filtering only according to p-value.

The plot shows the ratio between these two (red/blue).

It is clear that at all of the p-values thresholds, there is a large number of hypotheses that we rediscover, that filtration using the p-values assigned in the location data would discard. The fraction of such hypotheses grows larger as the p-value selected is more stringent.

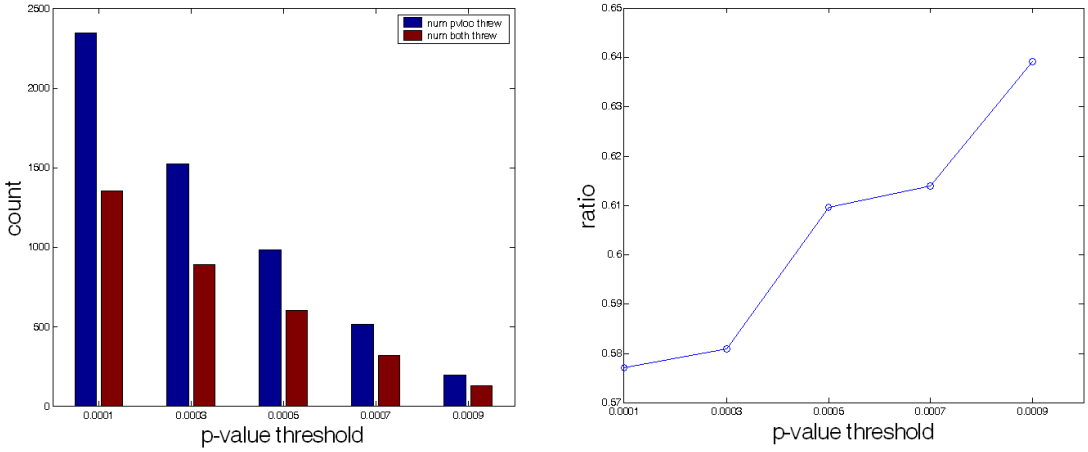


Figure 33: Relationship between two conceivable methods of filtering the data: (1) using a p-value stricter than 0.001 on the binding predictions of the location data, and (2) using TF-gene assignments which have the support of at least three of the four filtration methods described in the text (Section 2.2). Figure shows ratio between the number of TF-gene assignments filtered out by both methods and those filtered out of the data by using a stricter p-value alone.

Figure 34 contains an analysis of the trade-off between false discovery rate and the location data p-value threshold.

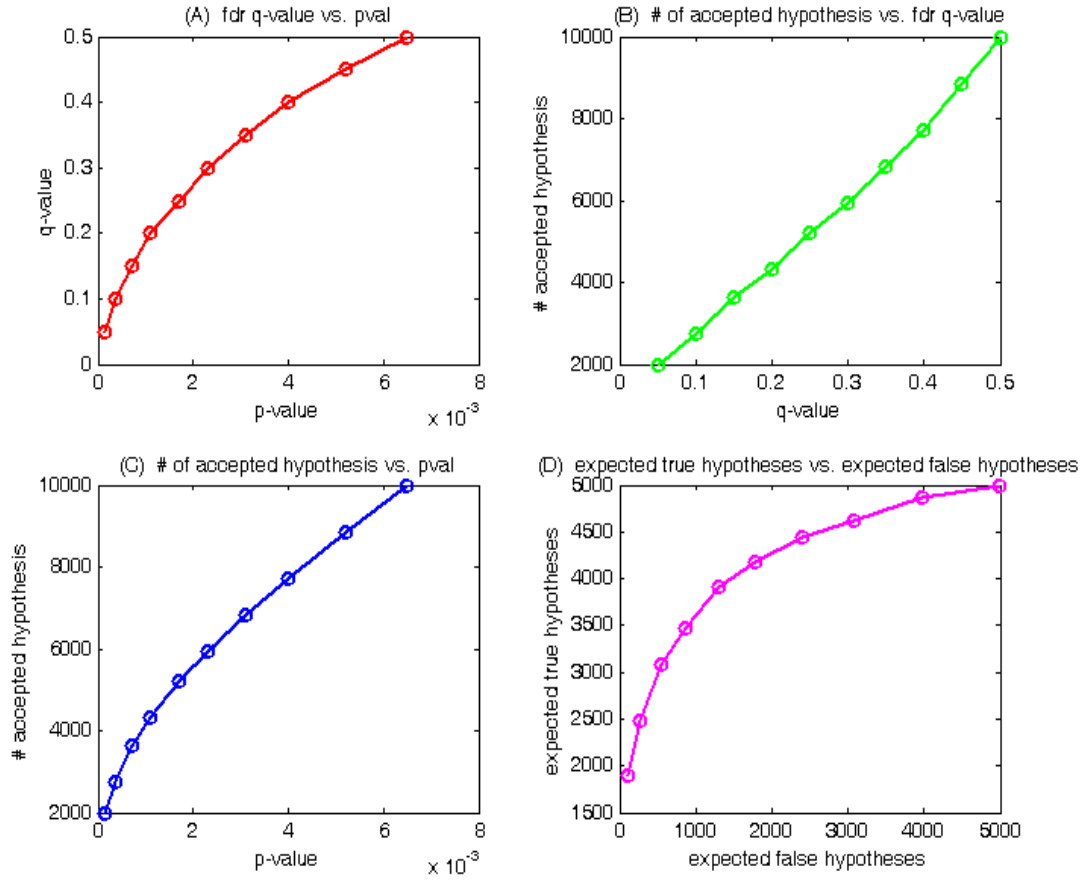


Figure 34: An FDR analysis of the original location data.

(A) shows relationship between the expected false discovery rate (q-value) and the p-value. It can be seen that at a p-value of 0.001, the q-value (FDR) is ~18%.

(B & C) show the relationship between the number of accepted hypotheses as a function of the q-value and p-value respectively. In order to achieve a false discovery rate of 10%, the number of hypotheses drops significantly to ~3400, significantly lower than the ~4200 hypotheses that were accepted at a p-value threshold of 0.001.

(D) shows the relationship between the number of expected true hypotheses and the number of expected false hypotheses at various p-value

thresholds. This graph displays the tradeoff which occurs at the different p-value thresholds. As the p-value threshold is raised (moving from left to right on the x-axis, and from lower to higher values on the y-axis), in the region where the slope of the graph is less than 1, more false hypotheses are added than true hypotheses. On the contrary, when the p-value is strict, in the region where the slope is greater than 1, mostly true hypotheses are gained by relaxing the p-value

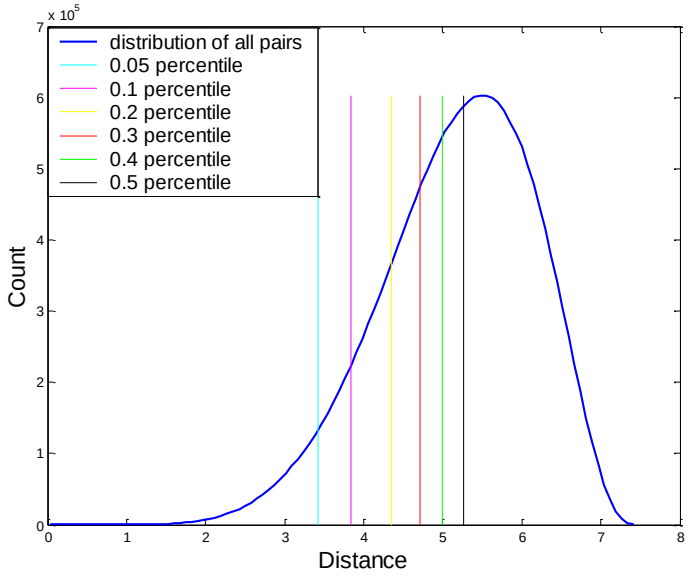
In all FDR analyses in this study we have fixed a permissive expected rate of false discovery of 10%.

3.5 Statistical significance of the EC score

The expression coherence (EC) score is a measure of the extent to which a set of genes is clustered into one or more clusters in expression space, and is equal to the fraction of gene pairs in the set whose normalized Euclidean distance (between expression profiles) falls below a threshold (Pilpel et al. 2001).

In this study, we explored various thresholds when using the EC score. The thresholds were chosen as follows: For each experimental condition, we calculated the pairwise distances between each pair of genes measured, yielding a distribution of pairwise distances. The six thresholds we explored were the distances associated with the top 5, 10, 20, 30, 40, and 50 percentile of this distance distribution. (See Figure 35 for a depiction of such thresholds.)

Figure 35: Thresholds used as in calculation of EC score. Blue line portrays distribution of all pairwise



distances between genes measured in one experimental condition; the Cho cell cycle in this example (Cho et al. 1998). Vertical lines correspond to the 5, 10, 20, 30, 40, and 50 percentile of the distance distribution (the height of these vertical lines is insignificant and was chosen for illustration purposes).

The significance of the EC score of a set of X genes is measured by randomly sampling 10^5 sets of X genes, and calculating the EC score for each sampled set. The fraction of sets which have an EC score greater than or equal to the score of the original set of genes is the approximated upper bound on the p-value of the score (Lapidot and Pilpel 2003).

3.6 Statistical significance of TF synergies

Significantly synergistic TF pairs were detected in a way similar to the original definitions (Pilpel et al. 2001; Sudarsanam et al. 2002). Let G1 be the set of genes assigned to TF1, and let G2 be the set of genes assigned to TF2. G12 is the set of genes in the intersection of G1 and G2, that is, the set of genes assigned to both TF1 and TF2. We define and calculate the "intersection set EC score" as the EC score of the N genes in G12. We then randomly sample 1000 sets of N genes from G1, and also 1000 sets of N genes from G2, and calculate their EC scores. A pair of TFs is synergistic if its intersection set EC score is at the top 5% of each of the two distributions of random EC scores. The use of a relatively permissive 5% threshold is justified for two reasons. First, each of the two sets of genes already has a relatively high EC score, since each set is bound by a regulating TF. Therefore a subset of these coherent sets, which is at their top fifth percentile in EC score, is even more significant. Second, since the score of the intersection set, in synergistic pairs, is at the top of *both* gene sets' distribution of scores, the effective significance of two events with p-value of 0.05 should typically be even better than 0.05.

3.7 Clustering parameters

The QT_clust algorithm receives as an input diameter a maximal cluster diameter, which is defined as the maximal distance between any two entities

within a cluster. When clustering the expression profiles of genes, we define the distance between two genes as the Euclidean distance between their expression profiles.

In clustering the expression profiles of the genes assigned to each TF using the QT_clust algorithm in each expression condition, in order to choose this diameter in a non-arbitrary fashion, we explored various thresholds of cluster diameters. These diameters were derived from the data itself, and were chosen as follows: For each condition, all pairwise distances between each pair of genes measured on the DNA chip were calculated, yielding a distribution of pairwise distances. The six diameter thresholds we explored were the distances associated with the top 5, 10, 20, 30, 40, and 50 percentile of this distance distribution. Per condition, these diameter thresholds are actually the same set of exploratory thresholds we used when calculating the EC score of a set of genes.

3.8 Comparison of QT_clust to Adap_Cluster

An adaptive clustering algorithm was recently developed by De Smet et al (De Smet et al. 2002) for purposes similar to those of the QT_clust algorithm. This algorithm, called Adap_Cluster, is a clustering method that starts from the principles described in the original QT_clust paper by Heyer et al (Heyer et al. 1999). Adap_Cluster has a faster running time than QT_clust, and aimed to improve upon QT_clust in its use of an input parameter which is both meaningful (e.g. using a probability instead of a cluster diameter) and also non-arbitrary, in that it is based on the data. The algorithm is a heuristic, two-step approach that defines the clusters sequentially (the number of clusters is not known in advance). In the first step, a cluster is located ("quality-based step"), and in the second step the quality of the cluster is derived ("adaptive approach"). As explained by DeSmet et al., the quality of a cluster is actually also called the radius of that cluster (De Smet et al. 2002). Adap_Cluster receives two input parameters from the user: a minimum number of genes per cluster and a significance level S, which symbolizes the probability of a point assigned to a cluster to actually belong to the cluster. The stated strength of

Adap_Cluster over QT_clust is that it uses this probability as the user input, instead of an arbitrary cluster diameter size used by QT_clust. In the Adap_Cluster algorithm, the radius parameter is calculated per cluster and not set to a fixed value.

We have thoroughly examined the use of Adap_Cluster in comparison to the QT_clust clustering algorithm. For the reasons detailed below we decided to use the QT_clust algorithm.

Firstly, we note that we have modified the QT_clust algorithm, and thus the major disadvantage of QT_clust does not pertain: the diameters we use are not arbitrarily chosen. Instead, the chosen diameters have a probabilistic meaning (see section 3.7). In addition, and most importantly, they are adaptively learned from specific datasets based on the distribution of all gene expression profiles they contain, i.e. in a way that appropriately accounts for factors such as the dimension and complexity of the data.

Secondly, we have overcome another disadvantage of classical clustering methods that Adap_Cluster tried to overcome; genes are allowed to fall into singleton clusters in our modified version of the QT_clust algorithm. One of the disadvantages of classical algorithms like k-means and Self-Organizing Maps is that genes are forced into clusters despite a low correlation with other cluster members. The modified algorithm we use does not do this, and allows genes to fall into singleton clusters if their distance from all other genes is larger than the learned diameter parameter.

We carried out a thorough comparison of the two algorithms. Figure 36 shows a specific transcription factor, Bas1, as an example. Similar results were obtained with other transcription factors. As shown in the figure, due to our modifications of the QT_clust method it outperforms the adaptive algorithm in that it generates a wider range of sizes of coherent clusters, each corresponding to a different choice of threshold. In contrast in many cases the adaptive clustering yields a constant cluster regardless of the 'significance level' chosen as input parameter. We are interested in having the ability to allow looser constraints and to consequently receive larger clusters, because the biological expression data at hand is noisy and sometimes produces clusters which do not include certain genes which are farther from the cluster center

than other genes, but nonetheless still belong to the cluster and maintain a highly similar expression profile.

The figure shows the most populated cluster obtained by clustering the genes assigned to the TF Bas1, when using the QT_clust and Adap_Cluster algorithms. The size of the most populated cluster obtained with Adap_Cluster is constant across the entire range [0-1] of the significance parameter S. In contrast, using QT_clust with a range of input diameter values yields a wider dynamic range of cluster sizes, corresponding to significantly coherent gene sets. We report here that regardless of the significance level the adaptive algorithm generates a largest cluster of constant size of 16 genes. On the other hand, QT_clust obtains this size (with same set of genes), yet in addition it also obtains other sizes that correspond to alternative thresholds. Thus, use of the QT_clust algorithm allows additional results to be obtained which cannot be obtained by the adaptive algorithm, namely a larger major cluster of coherent genes.

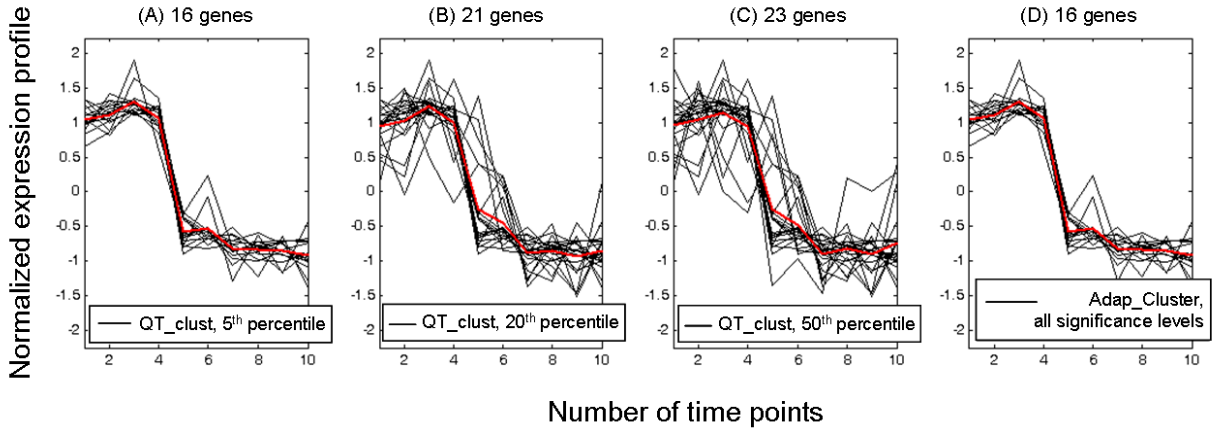


Figure 36: Figure shows the largest cluster obtained by clustering the genes assigned to the TF Bas1, when using the QT_clust and Adap_Cluster algorithms, with various values of input parameters as input to both algorithms. (A-C) show the result of the QT_clust algorithm, with diameters calculated directly from the expression data (obtained with 5th, 20th and 50th distance percentiles for A-C respectively, see section 3.7 for details on diameter calculations). (D) shows results obtained with Adap_Cluster that were obtained with any of 20 significance level values that evenly span the range [0-1]. Regardless of significance level, algorithm generates a largest cluster of constant size of 16 genes. QT_clust obtains variable sizes and can obtain a larger major cluster of coherent genes. Qualitatively similar results were obtained with other transcription factors (not shown).

3.9 Dictionary-generation procedure

The generation of the dictionary consists of four stages:

Exhaustive genome scan: For each fixed k-mer, find all genes which contain the exact k-mer in their regulatory regions.

Score the k-mers: assign an EC score and corresponding p-value to each k-mer, according to the genes which were associated with it in step 1.

FDR – Use the FDR method to select k-mers with a statistically significant p-value on the EC-score.

Cluster motifs – group together k-mers which have sequence similarity

3.9.1 Exhaustive genome scan

Genomic sequences were downloaded and Fasta files were generated for the promoter regions of *S. cerevisiae* genes. The files were downloaded from the *Saccharomyces* Genome Database (SGD) (ftp://genome-ftp.stanford.edu/pub/yeast/). Each promoter was between 250 and 1000bp long; taken upstream of the transcription start site (TSS).

The Fasta files were scanned for the occurrence of every possible k-mer (with k ranging from 7 to 11) using a perl program. The program created index files: for each k-mer, the index file contained the list of all genes this k-mer appears in. The 'k' usually ranged from 7 to 11.

3.9.2 Score the k-mers and FDR

After creation of the index files, a MATLAB program calculated the EC score and the corresponding p-value of the set of genes containing each k-mer (see Section 3.5 for details on p-value assessment). The results were then corrected for multiple hypotheses testing using an FDR test (Benjamini and Hochberg 1995) with q-value 0.1. Finally, the motifs that passed the p-value threshold determined by FDR constituted the dictionary's first draft, which was then clustered into groups of motifs.

3.9.3 Cluster the motifs

The final stage of dictionary creation is the clustering of motifs, based on their sequence similarity. The regulatory effects exerted by the motifs on genes whose promoters contain them, may also be used as a criterion in clustering. This stage is described in detail in the text in Section 2.4. Briefly, motifs are clustered according to a simplistic sequence distance measure which aligns, without gaps, two sequences such that the fraction of nucleotides that do not match perfectly is minimal. If this distance between all members of a group of motifs is under a certain threshold, the motifs are clustered together into a 'sequence-cluster'.

4 Discussion

Cells respond to their changing environment by reprogramming expression of various genes throughout the genome. They achieve this through the transcriptional regulation of genes by the interaction of diverse regulatory proteins, both activators and repressors, with the genome via specific DNA binding sequences. One of the greatest challenges of today is in understanding and mapping regulatory networks (the combination of regulators and regulated genes and interactions between them).

Our ability to map gene regulatory networks has been enhanced greatly by the sequencing of genomes and the development of new tools to study genome expression. In recent years there has been an explosion of genome-wide expression data, and computational algorithms have been developed that identify potential regulatory sequences in promoter regions throughout the genome. This era of computational genomics has contributed vastly to our ability to probe, understand, and characterize transcriptional regulatory networks on a systems level.

Functional genomics provides bioinformatics with genome- and proteome-wide data with an unprecedented throughput. Yet, the optimal utilization of these data sources requires establishing efficient means to assess the extent of noise in the data, and potentially also to filter it out. It is desired that in parallel to technological improvements on the experimental side that will reduce the noise level, accompanying computational tools will be developed to provide noise-filtration. It is likely that such tools will have to involve integration of data from other sources (that themselves may be noisy as well). In the current work we achieve exactly that. By combining the location data with promoter sequence data and extensive information on mRNA expression, we have significantly improved the accuracy of the DNA-protein location data, a first step in deciphering key elements of the genetic architecture of the yeast transcriptional network. We have also created visualization tools that allow better understanding and use of the wealth of expression, sequence, and binding data available, and enable us to uncover interesting biological phenomena that may otherwise be overlooked. The various types of data

available today probe the system from many different angles; sequence, functionality, network topology, dynamically, structurally, evolutionarily, and at the level of DNA, RNA, and proteins. These visualization tools allow us to view the multi-dimensional data in a colorful, intuitive, integrative, and interactive manner.

Our analysis of the genome-wide location analysis dataset has provided a cleaner, more accurate dataset. It is important to note that while we have considerable confidence in transcription factor-gene assignments supported by at least one of the four filtration methods presented here, it is entirely possible that additional filters may be proposed that would support additional such assignments. Such filters may include functional annotations or genome-wide transcription response to deletion of transcription factors (TFs). In addition we stress that supporting evidence for assignments in this work are mainly proposed for cases in which the DNA-protein interaction data shows regulatory effect on gene expression. It is possible that some assignments represent true binding events that resulted in no detectable transcriptional effects.

Through the process of achieving a more accurate dataset of DNA-protein location data, we have gained new insights on gene network design principles. This study has allowed us to obtain answers to a variety of important basic questions. Some answers, of course, are left to future work. Several examples of such questions are:

How many genes, on average, does a transcription factor strongly control (i.e. cause coherent expression of)?

Do most transcription factors work alone, or in combinatorial interactions with other transcription factors?

How different can the gene expression patterns brought about by the same transcription factor be, e.g. when examined in different growth conditions?

When observing a single growth condition, do the majority of transcription factors dictate one distinct expression profile in the group of genes they regulate, or do most exert various transcriptional regulatory controls on their gene targets?

When various expression profiles are dictated by one TF, is this usually due to alternative regulatory partners of the TF?

In such alternative interaction cases, what determines the relative influence of each of the interacting TFs?

In what percent of the promoters of genes bound by a transcription factor can we detect an over-represented, specific sequence motif? In cases where such motifs are not found, does this teach us about the faults of our motif-finding algorithms? Or is their perhaps underlying biology that we have yet to understand in order to fully grasp how DNA binding proteins recognize their binding sites?

Can one TF recognize alternative motifs, and consequently, exert different modes of regulation on genes that have these different motifs, or variations of a motif?

In this study we used motif-finding algorithms for binding site predictions, and microarray expression data, both of which are noisy techniques which will be further refined in the future. However, as long as the situation of noisy genome-wide technologies prevails, the course of action must be cleaning of one noisy data by intersection with other, potentially noisy, data sources. This is of course legitimate only in cases where there is no correlation between noise in the different technologies, and there is no reason to assume that noise in expression, sequence, and location data should be correlated. Thus our final products are rigorously statistically prioritized observations for which support comes independently from multiple sources that each by itself may be noisy, yet their concurrence is unlikely by chance.

In the present analysis subsets of co-expressed genes assigned to a TF are considered true positives even if they display expression profiles completely un-correlated with that of the TF itself. This reflects the fact that not all TFs vary at the mRNA expression level (e.g. post-translational modifications), that TF-gene interaction may include negative effects (Zhu et al. 2002; Segal et al. 2003), and that various logical interactions may be used to combine multiple regulators such that the expression profiles of target genes are a function of the combined expression profiles of several regulators (Pilpel et al. 2001; Setty et al. 2003). While the basic building blocks of transcription regulatory networks are the transcription factors and the regulatory motifs they bind, this work also provides the next level in gene network deciphering, namely TF combinations. We

provided here two largely independent methods analyzing interactions of pairs of TFs, and were encouraged to find that a very significant number of predictions are in the intersection of the two otherwise unrelated methods. This is a very strong attest to the strength of the combination of filters.

Combinatorial interactions among multiple regulators provide organisms with exponentially growing computational capacity, as well as the potential to respond to their multi-dimensional complex environment. These responses control the level of activity of the genes in the genome. In the present analysis TF combinatorics plays a dual role. On the technical level it serves to clean the location data. On the biological level, the discovery (and rediscovery) of combinatorial interactions constitute a crucial step towards full deciphering of the architecture of gene regulatory networks. Might it be that in addition to the role of TF combinatorics in representation of the multi-dimensional cellular environment, it is also employed by biology itself for the task of noise-filtering? Since the DNA-binding sites of most TFs are relatively short (5-20 base pairs (Stormo 2000)), their specificity towards their actual sites, which reduces sharply with increased genome size, is very low even for small genomes such as yeast's. A potential solution could be perhaps to increase the size of the individual DNA binding sites of transcription factors, but this would probably require a complete redesign of their protein folds. The obvious alternative is to employ simple AND-gated combinatorics, of homo- or hetero-TF combinations, in order to filter out genes that are bound by individual TFs but should not be regulated by them.

4.1 Future Directions

The combination of numerous data types in order to extract regulatory signal from datasets such as the genome-wide location analysis dataset is not limited to those data types and methods which we have used. As stated above, one future research direction may be to include functional data as a filtration method.

It will be very interesting to continue the analysis of the new location dataset by Harbison et al. (Harbison et al. 2004), and in addition to uncovering new

biology, to understand what new biology can be learned from the experiments done on those TFs that were not studied by Lee et al., and were not discussed in the present work.

The combinatorial nature of transcriptional regulation is a fascinating field, which we now are better prepared to delve into. Comparative genomics may greatly help to understand which transcription factors have been maintained as regulatory partners throughout evolution. This information can be used to understand the regulatory maps we have created using synergistic effects and co-localization of transcription factors in common promoters of regulated genes. One can also use functional annotation to aid in understanding these networks.

We have also begun to incorporate into our model the chromosomal location of the gene targets of transcription factors. The logic underlying this investigation is the notion that if a transcription factor regulates a number of genes, perhaps the simplest mechanism of action is the binding of the transcription factor to regions of close physical proximity. This may mean binding to close locations on a single chromosome, in which case we will find a cluster of gene targets of the TF near a single chromosomal location, or multiple chromosomal regions, or even different chromosomes which are physically proximal in the nucleus. Perhaps distance from the telomere is another factor influencing the regulatory control of transcription factors on their gene targets. It remains to be established whether we have enough data to make testable hypotheses about the effects of chromosomal location on gene regulation.

5 References

Aach, J., W. Rindone, et al. (2000). "Systematic management and analysis of yeast gene expression data." Genome Res **10**(4): 431-45.

Banerjee, N. and M. Q. Zhang (2003). "Identifying cooperativity among transcription factors controlling the cell cycle in yeast." Nucleic Acids Res **31**(23): 7024-31.

Bar-Joseph, Z., G. K. Gerber, et al. (2003). "Computational discovery of gene modules and regulatory networks." Nat Biotechnol **21**(11): 1337-42.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell **117**(2): 185-98.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." J. Roy Stat Soc **57**: 289-300.

Bulyk, M. L., E. Gentalen, et al. (1999). "Quantifying DNA-protein interactions by double-stranded DNA arrays." Nat Biotechnol **17**(6): 573-7.

Bulyk, M. L., X. Huang, et al. (2001). "Exploring the DNA-binding specificities of zinc fingers with DNA microarrays." Proc Natl Acad Sci U S A **98**(13): 7158-63.

Bussemaker, H. J., H. Li, et al. (2001). "Regulatory element detection using correlation with expression." Nat Genet **27**(2): 167-71.

Causton, H. C., B. Ren, et al. (2001). "Remodeling of yeast genome expression in response to environmental changes." Mol Biol Cell **12**(2): 323-37.

Cawley, S., S. Bekiranov, et al. (2004). "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs." Cell **116**(4): 499-509.

Cho, R. J., M. J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell **2**(1): 65-73.

Cho, R. J., M. Huang, et al. (2001). "Transcriptional regulation and function during the human cell cycle." Nat Genet **27**(1): 48-54.

Chu, S., J. DeRisi, et al. (1998). "The transcriptional program of sporulation in budding yeast." Science **282**(5389): 699-705.

Costanzo, M. C., J. D. Hogan, et al. (2000). "The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information." Nucleic Acids Res **28**(1): 73-6.

De Smet, F., J. Mathys, et al. (2002). "Adaptive quality-based clustering of gene expression profiles." Bioinformatics **18**(5): 735-46.

Dequard-Chablat, M., M. Riva, et al. (1991). "RPC19, the gene for a subunit common to yeast RNA polymerases A (I) and C (III)." J Biol Chem **266**(23): 15300-7.

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

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

Famulok, M. and J. W. Szostak (1992). "In Vitro Selection of Specific Ligand Binding Nucleic Acids." Angew. Chem. Int. Ed. Engl. **31**: 979-988.

Gao, F., B. C. Foat, et al. (2004). "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data." BMC Bioinformatics **5**(1): 31.
Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Mol Biol Cell **11**(12): 4241-57.
Goldberg, D. S. and F. P. Roth (2003). "Assessing experimentally derived interactions in a small world." Proc Natl Acad Sci U S A **100**(8): 4372-6.
Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.
Heyer, L. J., S. Kruglyak, et al. (1999). "Exploring expression data: identification and analysis of coexpressed genes." Genome Res **9**(11): 1106-15.
Hughes, J. D., P. W. Estep, et al. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae." J Mol Biol **296**(5): 1205-14.
Ihmels, J., G. Friedlander, et al. (2002). "Revealing modular organization in the yeast transcriptional network." Nat Genet **31**(4): 370-7.
Iyer, V. R., M. B. Eisen, et al. (1999). "The transcriptional program in the response of human fibroblasts to serum." Science **283**(5398): 83-7.
Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**(6819): 533-8.
Jelinsky, S. A., P. Estep, et al. (2000). "Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: rpn4 links base excision repair with proteasomes." Mol Cell Biol **20**(21): 8157-67.
Johnson, J. M., S. Edwards, et al. (2005). "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments." Trends Genet **21**(2): 93-102.
Kadonaga, J. T. (2004). "Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors." Cell **116**(2): 247-57.
Kim, S. K., J. Lund, et al. (2001). "A gene expression map for Caenorhabditis elegans." Science **293**(5537): 2087-92.
Lapidot, M. and Y. Pilpel (2003). "Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription." Nucleic Acids Res **31**(13): 3824-8.
Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.
Lieb, J. D., X. Liu, et al. (2001). "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." Nat Genet **28**(4): 327-34.
Liu, X., D. M. Noll, et al. (2005). "DIP-chip: Rapid and accurate determination of DNA-binding specificity." Genome Res.
Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.
Mukherjee, S., M. F. Berger, et al. (2004). "Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays." Nat Genet **36**(12): 1331-9.
Orlando, V. (2000). "Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation." Trends Biochem Sci **25**(3): 99-104.

Orlando, V. and R. Paro (1993). "Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin." Cell **75**(6): 1187-98.
Pilpel, Y., P. Sudarsanam, et al. (2001). "Identifying regulatory networks by combinatorial analysis of promoter elements." Nat Genet **29**(2): 153-9.
Ptashne, M. and A. Gann (1997). "Transcriptional activation by recruitment." Nature **386**(6625): 569-77.
Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.
Roberts, C. J., B. Nelson, et al. (2000). "Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles." Science **287**(5454): 873-80.
Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.
Segal, E., Y. Barash, et al. (2002). "From Promoter Sequence to Expression: A Probabilistic Framework." Proc. 6th Inter. Conf. on Research in Computational Molecular Biology (RECOMB), Washington, DC.
Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-76.
Segal, E., R. Yelensky, et al. (2003). "Genome-wide discovery of transcriptional modules from DNA sequence and gene expression." Bioinformatics **19 Suppl 1**: i273-82.
Setty, Y., A. E. Mayo, et al. (2003). "Detailed map of a cis-regulatory input function." Proc Natl Acad Sci U S A **100**(13): 7702-7.
Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-8.
Simon, I., J. Barnett, et al. (2001). "Serial regulation of transcriptional regulators in the yeast cell cycle." Cell **106**(6): 697-708.
Solomon, M. J., P. L. Larsen, et al. (1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene." Cell **53**(6): 937-47.
Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.
Springer, C., M. Kunzler, et al. (1996). "Amino acid and adenine cross-pathway regulation act through the same 5'-TGACTC-3' motif in the yeast HIS7 promoter." J Biol Chem **271**(47): 29637-43.
Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.
Sudarsanam, P., Y. Pilpel, et al. (2002). "Genome-wide Co-occurrence of Promoter Elements Reveals a cis-Regulatory Cassette of rRNA Transcription Motifs in Saccharomyces cerevisiae." Genome Res **12**(11): 1723-31.
Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.
Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-5.

Zhu, Z., Y. Pilpel, et al. (2002). "Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm." J Mol Biol **318**(1): 71-81.

6 Appendices

6.1 Appendix A

List of TFs in Lee et al. study (Lee et al. 2002):

1 - ABF1	41 - HIR2	81 - RTG1
2 - ACE2	42 - HMS1	82 - RTG3
3 - ADR1	43 - HSF1	83 - RTS2
4 - ARG80	44 - IME4	84 - SFL1
5 - ARG81	45 - INO2	85 - SFP1
6 - ARO80	46 - INO4	86 - SIG1
7 - ASH1	47 - IXR1	87 - SIP4
8 - AZF1	48 - LEU3	88 - SKN7
9 - BAS1	49 - MAC1	89 - SKO1
10 - CAD1	50 - MAL13	90 - SMP1
11 - CBF1	51 - MAL33	91 - SOK2
12 - CHA4	52 - MATa1	92 - SRD1
13 - CIN5	53 - MBP1	93 - STB1
14 - CRZ1	54 - MCM1	94 - STE12
15 - CUP9	55 - MET31	95 - STP1
16 - DAL81	56 - MET4	96 - STP2
17 - DAL82	57 - MIG1	97 - SUM1
18 - DIG1	58 - MOT3	98 - SWI4
19 - DOT6	59 - MSN1	99 - SWI5
20 - ECM22	60 - MSN2	100 - SWI6
21 - FHL1	61 - MSN4	101 - THI2
22 - FKH1	62 - MSS11	102 - UGA3
23 - FKH2	63 - MTH1	103 - USV1
24 - FZF1	64 - NDD1	104 - YAP1
25 - GAL4	65 - NRG1	105 - YAP3
26 - GAT1	66 - PDR1	106 - YAP5
27 - GAT3	67 - PHD1	107 - YAP6
28 - GCN4	68 - PHO4	108 - YAP7
29 - GCR1	69 - PUT3	109 - YBR267W
30 - GCR2	70 - RAP1	110 - YFL044C
31 - GLN3	71 - RCS1	111 - YJL206C
32 - GRF10(Pho2)	72 - REB1	112 - ZAP1
33 - GTS1	73 - RFX1	113 - ZMS1
34 - HAA1	74 - RGM1	
35 - HAL9	75 - RGT1	
36 - HAP2	76 - RIM101	
37 - HAP3	77 - RLM1	
38 - HAP4	78 - RME1	
39 - HAP5	79 - ROX1	
40 - HIR1	80 - RPH1	

List of 40 conditions:

Below are short descriptions of the **40 conditions** from which expression data was gathered. These datasets were downloaded from ExpressDB (Aach et al. 2000).

1 - Cho-cell cycle
2 - Chu-sporulation
3 - Environmental response-Acid
4 - Environmental response-Alkali
5 - Environmental response-Heat
6 - Environmental response-NaCl
7 - Environmental response-Peroxide
8 - Environmental response-Sorbitol
9 - Eisen - exposure to cold
10 - Gasch environmental response- diauxic shift
11 - Eisen – exposure to dtt
12 - Eisen - exposure to heat
13 - Jelinsky - exposure to DNA Damage
14 - Gasch environmental response- 37-25 shock
15 - Gasch environmental response- Amino Acid starvation
16 - Gasch environmental response- diamide
17 - Gasch environmental response- dtt1
18 - Gasch environmental response- dtt2
19 - Gasch environmental response- heat shock 1
20 - Gasch environmental response- hs 29-33 1m sorbitol
21 - Gasch environmental response- hs 29-33
22 - Gasch environmental response- hs 29-33 No sorbitol
23 - Gasch environmental response- hs2(3 time zero)
24 - Gasch environmental response- constant h2o2
25 - Gasch environmental response- Menadione
26 - Gasch environmental response- Hypo-osmotic
27 - Gasch environmental response- Nitrogen Depletion
28 - Gasch environmental response- sorbitol
29 - Gasch environmental response- hs various temp to 37c
30 - Gasch environmental response- various temp growth
31 - Gasch environmental response- var temp steady state
32 - Gasch environmental response- x media vrs car1
33 - Gasch environmental response- YPD1
34 - Gasch environmental response- YPD2
35 - Gasch environmental response- YPx media vrs car2
36 - MapK
37 - Spellman cell-cycle alpha
38 - Spellman cell-cycle cdc15
39 - Spellman cell-cycle cdc28
40 - Spellman cell-cycle eluteration

List of TFs in Harbison et al. study (Harbison et al. 2004):

1 - A1	42 - Gcn4	83 - Met4	124 - Rpi1	165 - Uga3
2 - Abf1	43 - Gcr1	84 - Mga1	125 - Rpn4	166 - Ume6
3 - Abt1	44 - Gcr2	85 - Mig1	126 - Rtg1	167 - Upc2
4 - Aca1	45 - Gln3	86 - Mig2	127 - Rtg3	168 - Usv1
5 - Ace2	46 - Gts1	87 - Mig3	128 - Rts2	169 - War1
6 - Adr1	47 - Gzf3	88 - Mot3	129 - Sfl1	170 - Wtm1
7 - Aft2	48 - Haa1	89 - Msn1	130 - Sfp1	171 - Wtm2
8 - Arg80	49 - Hac1	90 - Msn2	131 - Sig1	172 - Xbp1
9 - Arg81	50 - Hal9	91 - Msn4	132 - Sip3	173 - Yap1
10 - Aro80	51 - Hap1	92 - Mss11	133 - Sip4	174 - Yap3
11 - Arr1	52 - Hap2	93 - Mth1	134 - Skn7	175 - Yap5
12 - Ash1	53 - Hap3	94 - Ndd1	135 - Sko1	176 - Yap6
13 - Ask10	54 - Hap4	95 - Ndt80	136 - Smk1	177 - Yap7
14 - Azf1	55 - Hap5	96 - Nnf2	137 - Smp1	178 - YBL054W
15 - Bas1	56 - Hir1	97 - Nrg1	138 - Snf1	179 - YBR239C
16 - Bye1	57 - Hir2	98 - Oaf1	139 - Snt2	180 - YBR267W
17 - Cad1	58 - Hir3	99 - Opi1	140 - Sok2	181 - YDR026C
18 - Cbf1	59 - Hms1	100 - Pdc2	141 - Spt10	182 - YDR049W
19 - Cha4	60 - Hms2	101 - Pdr1	142 - Spt2	183 - YDR266C
20 - Cin5	61 - Hog1	102 - Pdr3	143 - Spt23	184 - YDR520C
21 - Crz1	62 - Hsf1	103 - Phd1	144 - Srd1	185 - YER051W
22 - Cst6	63 - Ifh1	104 - Pho2	145 - Stb1	186 - YER130C
23 - Cup9	64 - Ime1	105 - Pho4	146 - Stb2	187 - YER184C
24 - Dal80	65 - Ime4	106 - Pip2	147 - Stb4	188 - YFL044C
25 - Dal81	66 - Ino2	107 - Ppr1	148 - Stb5	189 - YFL052W
26 - Dal82	67 - Ino4	108 - Put3	149 - Stb6	190 - YGR067C
27 - Dat1	68 - Ixr1	109 - Rap1	150 - Ste12	191 - Yhp1
28 - Dig1	69 - Kre33	110 - Rco1	151 - Stp1	192 - YJL206C
29 - Dot6	70 - Kss1	111 - Rcs1	152 - Stp2	193 - YKL222C
30 - Ecm22	71 - Leu3	112 - Rdr1	153 - Stp4	194 - YKR064W
31 - Eds1	72 - Mac1	113 - Rds1	154 - Sum1	195 - YLR278C
32 - Fap7	73 - Mal13	114 - Reb1	155 - Sut1	196 - YML081W
33 - Fhl1	74 - Mal33	115 - Rfx1	156 - Sut2	197 - YNR063W
34 - Fkh1	75 - Mbf1	116 - Rgm1	157 - Swi4	198 - Yox1
35 - Fkh2	76 - Mbp1	117 - Rgt1	158 - Swi5	199 - YPR022C
36 - Fzf1	77 - Mcm1	118 - Rim101	159 - Swi6	200 - YPR196W
37 - Gal3	78 - Mds3	119 - Rlm1	160 - Tbs1	201 - Yrr1
38 - Gal4	79 - Met18	120 - Rlr1	161 - Tec1	202 - Zap1
39 - Gal80	80 - Met28	121 - Rme1	162 - Thi2	203 - Zms1
40 - Gat1	81 - Met31	122 - Rox1	163 - Tos8	
41 - Gat3	82 - Met32	123 - Rph1	164 - Tye7	