BIOINFORMATICS

Vol. 1 no. 1 2002
Pages 1–10

Correlating Gene Promoters and Expression in Gene Disruption Experiments

Kimmo Palin[1], Esko Ukkonen[1], Alvis Brazma[2] and Jaak Vilo[2]

[1]University of Helsinki, Department of Computer Science, Helsinki, Finland and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: Finding putative transcription factor binding sites in the upstream sequences of similarly expressed genes has recently become a subject of intensive studies. In this paper we investigate how much gene expression regulation can be attributed to the presence of various binding sites in the gene promoters by correlating the binding sites and the changes in gene expression resulting from gene disruptions (e.g., knockouts).

Results: We have developed a data analysis method for comparing mRNA measurements of gene disruption experiments with information about gene promoters. The method was applied to a well–known dataset to uncover correlations between known transcription factor binding site motifs in the upstream regions of all *S. cerevisiae* genes and the gene expression changes in various gene disruption experiments. The possible explanations of the correlations were categorized and analyzed using e.g. expression cascades. Several correlations turned out to be consistent with existing biological knowledge while some new ones suggest themselves for further study.

Availability: The resulting tables are available at `http://www.cs.helsinki.fi/u/kpalin/CorrDisrupt/`.

Contact: `kimmo.palin@cs.helsinki.fi`

INTRODUCTION

Gene regulation in eukaryotic organisms is achieved through many complex mechanisms most of which are not well understood. Completely sequenced genomes together with the advances in microarray technology give the researchers a powerful tool to study gene regulation. For instance, one can take clusters of coexpressed genes and look into the genome sequences around these genes for a 'signal' that could potentially explain their coregulation.

A major role in eukaryotic gene regulation is played by specific proteins called *transcription factors* which can bind to relatively specific DNA regions called *binding sites* in the promoter regions of genes. Through binding to DNA and forming protein complexes with other proteins the transcription factors activate or repress the transcription of the genes. The specificity and exact context of the binding events are still largely unknown, as on the full genome scale there are many potential binding sites, which in fact are not functional (see Scherf *et al.* (2000)).

One difficulty in finding putative binding sites stems from the uncertainty about the location of the functional promoter regions of different genes. In the higher eukaryotes the promoter regions are sometimes found up to 30 000 base-pairs (bp) away from the gene. This is simpler in yeast, where it is believed that promoter regions of most genes are located in the vicinity of the gene translation start sites, predominantly within about 600 bp upstream of the gene or even closer.

Some correlation between the coexpression of genes and the presence of common sequence elements in their upstream regions have been demonstrated in several studies such as Brazma *et al.* (1998); Eisen *et al.* (1998); Vilo *et al.* (2000); Bussemaker *et al.* (2001); Jakt *et al.* (2001); Ge *et al.* (2001) reviewed for example in Zhang (1999); Vilo & Kivinen (2001). Still even if we narrow down the putative promoter regions to 600 bp, we keep finding many matches that do not significantly correlate with the gene expression profiles. Vice versa, for some clusters of tightly coexpressed genes it has not been possible to find common sequence patterns in their upstream regions (Spellman *et al.* (1998)).

There are several reasons why the correlation between the presence of binding sites and gene expression profiles are far from perfect. One possible reason often referred to is that the chromatin structure makes some regions of the genome inaccessible for the transcription factors, i.e., many potential binding sites are actually hidden from the influences of the transcription factors.

Second it must be noted that typically the transcription is not directly regulated by several proteins binding to individual binding sites but by the proteins forming a complex which binds to its own site. This way a single transcription factor can affect many target genes through several protein complexes without directly binding to any upstream DNA sequences. Also it should be remembered that microarrays measure only the mRNA expression while it is known that the mRNA levels do not linearly correlate with protein levels (including transcription factors). Apparently considerable part of gene regulation is happening in the translation stage and through post–translational modifications such as phosphorylation of transcription factors.

It has been shown that the correlation between co-expression and the presence of particular sequence elements in the gene upstream regions can be improved if one looks for combinations of transcription factor binding sites (Pilpel *et al.* (2001)). Still many questions why particular binding sites are not active, and why some genes are coexpressed without obvious coregulatory mechanisms, remain unanswered.

In this paper we study the correlation between the presence of binding sites and gene expression by comparing the changes of gene expression in gene disruption experiments and the presence of putative binding sites in the upstream regions of various genes. In a gene disruption experiment the expression of a certain gene is artificially repressed. When the mutated organism is grown few generations and the expression of its genes is measured, all differences in the expressions compared to wild–type expressions is due to the disrupted gene. By using these gene disruption experiments we can equate the expression measurements with the genes that were disrupted in that experiment.

For each disrupted gene we find the set of genes that change their expression levels significantly. On the other hand we take previously known and verified binding sites and find the corresponding sets of genes having these binding sites in their upstream regions. Then for each pair of a disrupted gene and a binding site we look how much the respective sets intersect. If the intersection is significantly larger than we would expect by random chance we infer that the binding site and the disrupted gene are somehow related.

We will develop an analysis method along these lines and apply it to correlate the gene disruption data of Hughes *et al.* (2000) for the yeast with the putative binding sites from Pilpel *et al.* (2001).

THE APPROACH

Let us denote by $\mathbb{G}$ the set of all genes of an organism. Consider a gene $t \in \mathbb{G}$ whose protein product is a transcription factor that binds to the sequence motif $m(t)$. Let $\mathbb{R}_t \subset \mathbb{G}$ be the set of all genes that have the motif $m(t)$ in their promoter region. In this case we say that the promoter region has the binding site t. We call $\mathbb{R}_t$ the *regulation set of the gene* t. Note that $\mathbb{R}_t$ is defined only for genes t, that code for DNA binding transcription factors. We denote the set of the genes encoding transcription factors by T. Assuming that all binding sites present in the gene upstream regions are functional, we can hypothesize that the set $\mathbb{R}_t$ is the set of genes directly regulated by the transcription factor coded by gene t. Note that this hypothesis ignores the importance of the context of the binding site.

Consider then experiments where a gene has been disrupted (e.g., knocked out) and the changes in the expression level of all other genes have been measured in that particular experiment. For every gene d that has been disrupted in one of the experiments — we denote this set of disrupted genes by D — we can find the subset $\mathbb{E}_d$ of the genes, whose mRNA expression is significantly altered when the gene d has been disrupted. We call $\mathbb{E}_d$ the *effectual set of gene* d.

We would assume that $\mathbb{R}_g$ is the set of genes whose expression depends directly on the expression of the gene g. This means that their expression should be affected when the gene g is disrupted. Hence we expect $\mathbb{R}_g$ and $\mathbb{E}_g$ to have large intersection. There are several reasons why this may not always be the case. First, if gene g is such that it has not been significantly expressed in the control population, the disruption of gene g may not have noticeable effect on the concentration of the transcription factor. The second reason comes from the way the expression measurements are performed — because the expression is measured when the cell population has divided after the mutation it is impossible to distinguish the direct from the indirect regulatory effects of the disruption.

In this paper we consider all pairs of genes (t, d), where $t \in T$ and $d \in D$, and look whether the sets $\mathbb{R}_t$ and $\mathbb{E}_d$ intersect significantly more than expected by chance, in which case we say that $\mathbb{R}_t$ *correlates* with $\mathbb{E}_d$ or, simply, the gene t *correlates* with the site d. Evidently we would expect that $\mathbb{R}_t$ correlates with $\mathbb{E}_d$ when $t = d$, as we should assume that the expression of genes that are directly regulated by a certain transcription factor depend on the gene coding for that transcription factor.

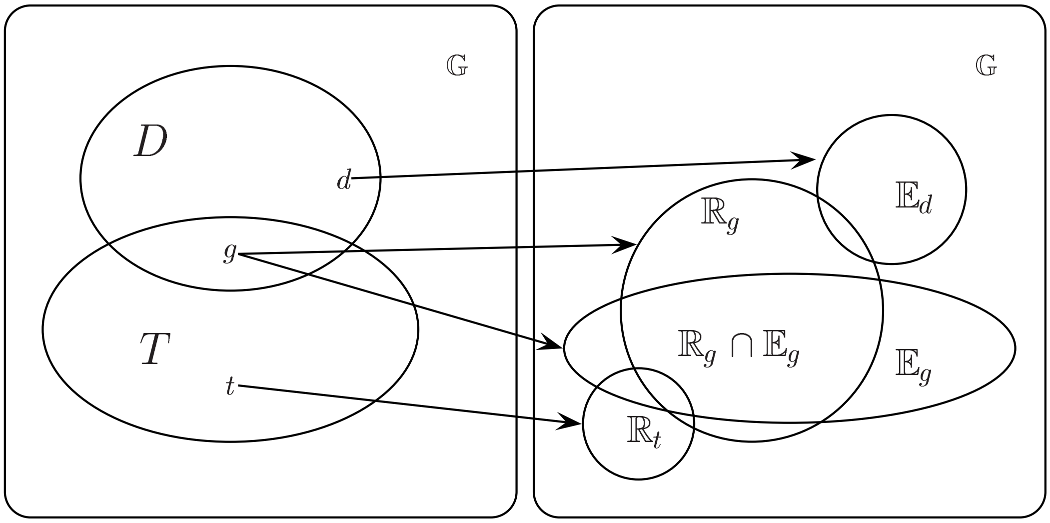

Figure 1. Disrupted gene d maps to its effectual set $\mathbb{E}_d$ and transcription factor t maps to its regulation set $\mathbb{R}_t$. Gene g that belongs to $D \cap T$ has both sets $\mathbb{R}_g$ and $\mathbb{E}_g$.

In Figure 1 we illustrate the case where effectual and regulation sets of a gene $g \in T \cap D$ intersect significantly. The disruption of a gene $d \in D \setminus T$ on the other hand does not affect more than expected number of genes having the binding site for the non-related transcription factor g. The same idea works also the other way around: the effectual set of gene g contains only relatively few genes having the binding site for $t \in T \setminus D$.

In the case where the promoter region of a gene contains the binding site of g but the gene is not affected by the disruption of g, there is a possibility that the particular binding site in the particular promoter is not functional due to the context. It is also possible to have genes that do get affected but do not have the binding site. If the gene g regulates some transcription factor, one can expect the targets of this intermediate transcription factor to be affected even though their promoter regions do not have the binding site of g. This unavoidable noise forces us to analyze the statistical significance of the intersection of the sets $\mathbb{R}_g$ and $\mathbb{E}_g$.

We can divide the genes g in $T \cap D$ into two classes according to their regulation and effectual sets: the ones for whom the regulation and effectual sets correlate, and the ones for whom they do not. The second group should be explained in some biologically meaningful way.

Next we want to identify pairs (t, d), such that $t \neq d$, but the regulation set $\mathbb{R}_t$ does correlate with the effectual set $\mathbb{E}_d$. There are number of reasons why such unexpected spurious looking correlation can happen. One way of analyzing this is by what we call *gene expression cascade*.

Disruption networks and gene expression cascades

The potential regulatory relations discovered by a set of gene disruption experiments can conveniently be summarized as a graph structure which we call a *disruption network*. Such a network has a node for each gene and there is a directed edge from a gene x to a gene y if and only if the disruption of x has a significant effect on the expression of y. Similar networks have recently been used e.g. by Wagner (2001) and Rung *et al.* (2002).

We say that a gene b is in the *gene expression cascade* of a gene g (the *cascade* of g for short), if disrupting b significantly affects the expression of g. Hence the cascade of g consists of all genes from which there is an edge to g in the disruption network. The cascade of g is denoted by $\mathbb{C}_g$.

Figure 2 gives a hypothetical example, in which the cascade of G is $\mathbb{C}_G = \{B, D, C\}$. In the situation of Figure 2, if we have found that the sets $\mathbb{E}_B$ and $\mathbb{R}_G$ correlate, we can explain it by saying that gene B regulates gene G which defines the set $\mathbb{R}_G$. This seems sound as B belongs to $\mathbb{C}_G$. One has to be careful though not to stretch this type of explanation to genes outside $\mathbb{C}_G$, such as A. If we have found that sets $\mathbb{E}_A$ and $\mathbb{R}_G$ correlate, we might be tempted to say that "Disruption of A affects C affects G which defines $\mathbb{R}_G$". This is unfortunately doubtful to say the least. This can be seen as follows.

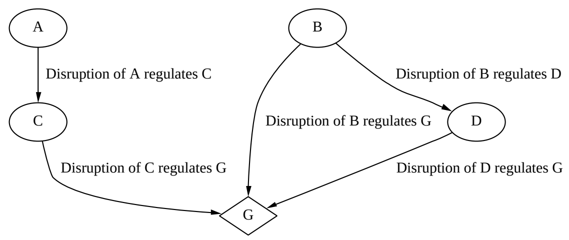

Figure 2. A disruption network. Disruption of source–gene affects the target-gene.

Because the mRNA measurements are made in steady state one could expect (by ignoring the effects that compensate each other) that all of the regulation paths of the network in Figure 2 should in fact have collapsed. This could be the case of gene B whose effect on the gene G may be mediated by D but, because of the collapse due to steady state, we see also the direct edge from B to G. On the other hand, the gene A, although having a regulation path to G, was not observed to regulate G for some reason, as indicated by the fact that a direct edge from A to G is missing in Figure 2. This regulation is lacking possibly because disruption of A affects also some other genes which nullify the effect of gene C on gene G.

Enough said, the gene expression cascades are useful for our purposes of explaining surprising correlations. For some transcription factors g almost all of the genes in its cascade $\mathbb{C}_g$ turns out to correlate with the regulation set $\mathbb{R}_g$ of the transcription factor.

All of the spurious correlations can not be explained by the cascades derived from just the mRNA measurements. If the production of the transcription factor protein is regulated by translational or post–translational mechanisms all gene disruptions affecting these mechanisms correlate also with the regulatory set of the given transcription factor.

For example, if the transcription factor t needs to have a certain phosphorylation state to function properly and a gene d regulating the phosphorylation is disrupted then the effectual set of d and the regulatory set of t correlate even though the mRNA levels of t are not affected. For some genes also the translation of the mRNA to the protein requires some specific proteins. According to Ideker *et al.* (2001) the mRNA and the protein abundance have only about 0.6 correlation in yeast disruption experiments.

RESULTS

When investigating the presence and absence of correlations between the sets $\mathbb{R}_g$ and $\mathbb{E}_d$ we can potentially have the following categories 1–4.

1. The transcription factor and the disrupted genes are the same, i.e., $g = d$ and the sets $\mathbb{R}_g$ and $\mathbb{E}_d$ correlate as expected.
2. For the given g, the set $\mathbb{R}_g$ correlates with set(s) $\mathbb{E}_d$, for $d \neq g$, such that the correlation can be explained with existing knowledge.
3. For the given g, the set $\mathbb{R}_g$ correlates with set(s) $\mathbb{E}_d$ for $d \neq g$, such that there is no obvious reasons why it should happen.
4. For the given g, there is no d such that $\mathbb{R}_g$ and $\mathbb{E}_d$ correlate.

The first case we consider is that when the sets correlate as expected. The second case can often be explained by the product of gene d being a part of a protein complex regulating gene g. The cases in which the sets do not correlate or have unexpected correlation are the ones that need to be investigated.

We analyzed the expression data from 287 gene disruptions in yeast that have been reported in Hughes *et al.* (2000). In the measurements the expression levels in cells with the disrupted gene were compared with the control population of wild type yeast cells. This public dataset is hence an adequate source of information for our effectual sets.

As regulation sets $\mathbb{R}_t$ we used 356 sets defined by yeast binding sites in Pilpel *et al.* (2001). Of these binding site motifs, 37 are previously known and the rest are putative sites generated from MIPS families. It should be noted that even the computationally generated motifs have their information from the curated MIPS database (Mewes *et al.* (2002)) and not from the typical expression similarity clustering.

We analyzed a subset of 263 expression measurements against all the 356 motifs. Our results were highly variable both across the binding sites and the disruption experiments. A summary of the results for the 37 known binding site motifs is shown in Figure 3, and a more detailed view of some of the sites is given in Table 1. The rows of Figure 3 stand for disrupted genes and the columns stand for the motifs. Computationally generated motifs and disruptions without any correlation or cascade relations are omitted. See the web supplement at `http://www.cs.helsinki.fi/u/kpalin/CorrDisrupt/` for correlations of the computationally generated binding sites.

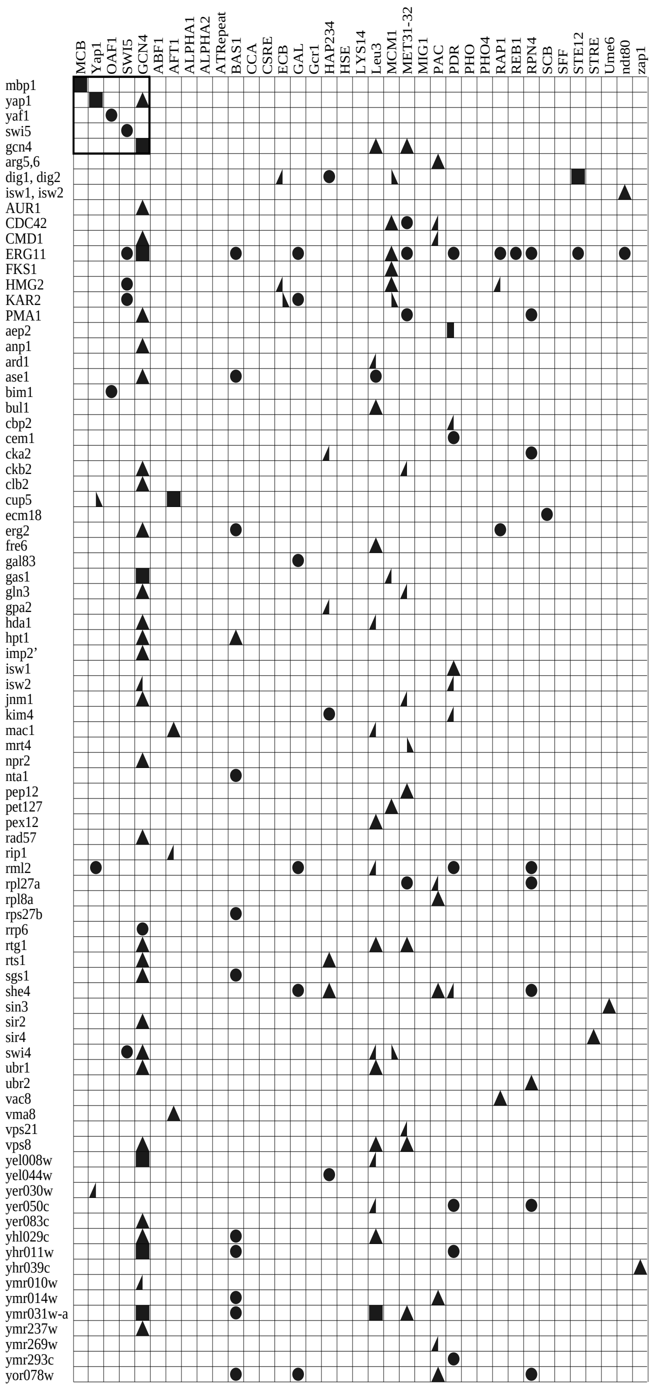

Figure 3. Correlations between the disruptions on the left and binding sites on the top. Only left or right part of the triangle or rectangle is shown if the correlation occurs only with weaker or stronger (left and right half of an icon) definition of significantly altered expression. A triangle ▲ for correlation (6 strong, 29 weak, 61 both), a circle • (54) for the disruption in the expression cascade of the transcription factor. A rectangle ■ is for a correlation explained by the cascade (1 weak, 11 both).

Unfortunately there are only five genes that are disrupted and whose binding site is known in the dataset (i.e., the intersection of T and D consist of only five elements). These five genes are located in the upper left hand corner of Figure 3.

An entry in Figure 3 is marked with a triangle if the effectual set of the gene on the left and the regulation set of the motif above have significant correlation. The entry contains a circle if the gene on the left is in the expression cascade of the transcription factor on top. If both of these conditions are true the entry contains a rectangle. If the correlation is detected only with stronger or weaker definition of altered expression (See Methods) the triangle or rectangle is only the left or right half of the original icon, respectively.

Next we explain the results in more detail. The biological information in the following sections is obtained from the Yeast Protein Database Costanzo *et al.* (2001) unless otherwise cited.

Disruption specific motifs

For a few transcription factor genes our test gave exactly the expected results. For example the *mbp1* disruption correlated with the MCB site which it is known to bind to as a part of the MBF complex. Also *yap1* disruption correlates best with its known binding site. These two genes together with *gcn4* make the whole category 1 of our results.

Positively there are some binding sites in the category 2, whose respective gene was not disrupted in any of the experiments, but the binding site still correlates well with some disruptions of genes related to their own. For example *sin3* disruption correlates with Ume6 site and their proteins are known to interact. *dig11* gene is known to repress gene *STE12*, so it is not surprising to see *dig1, dig2* double disruption mutant to correlate with STE12 binding site.

The binding sites of zap1 protein correlate most significantly with the disruption of gene *msc7* (ORF code *yhr039c*) which is understandable as the *zap1* is known to regulate aldehyde dehydrogenase genes to whom *msc7* has strong similarity. The BAS1 protein regulates adenine biosynthesis genes but the repression activity is bypassed in *hpt1* mutant, thus BAS1 site correlates with *hpt1* disruption.

The ndt80 binding site correlates well with *isw1, isw2* double mutant which is as expected since ndt80 is not required for the activation of meiotic genes in an *isw2* mutant. The binding site for protein RPN4 correlates with *ubr2* disruption, and these both proteins are related to protein degradation. *RAP1* gene is related to vacuole

(Foury (1990)) as is gene *vac8* whose disruption correlates well with RAP1 binding site.

General motifs

About 7 binding sites correlate with distinctively high number of disruptions. In category 3 the sites Leu3, MET31-32 and GCN4, whose genes are all in the amino acid pathway, correlate with a wide range of disruptions. Also others, including a coregulatory protein MCM1, early cell-cycle box ECB and PDR site that is related to drug resistance, correlated with several disruptions.

In the easily explainable category 2 the AFT1 site has somewhat interesting correlations. It correlates strongly with three disruptions *cup5*, *mac1* and *vma8* all of which are related to small molecule transport and iron uptake.

The usefulness of the gene expression cascades can be seen in in the case of GCN4 binding site which resides in categories two and three. The gene expression cascade for *gcn4* is illustrated as a graph in Figure 4. In the graph the nodes are genes whose disruption affects the gene *gcn4*. An edge is solid if the disruption of the source down regulates the target and dashed if the the regulation is upwards. A node is a rectangle if its effectual set correlates with $\mathbb{R}_{GCN4}$ and otherwise a circle.

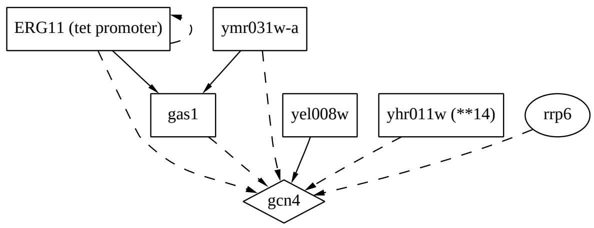

Figure 4. The expression cascade of gene *gcn4*. The effectual set $\mathbb{E}$ of a gene in a rectangle correlates with the regulation set $\mathbb{R}_{GCN4}$. Dashed edge is for up regulation and solid for down regulation.

Remarkably 5 out of the 6 disruptions which affect *gcn4* also correlate with the regulation set of *gcn4*. The figure also illustrates the collapsed transitive paths where the gene *gas1* possibly mediates the effect of *ERG11* and *ymr031w-a* on *gcn4*.

The gene expression cascade explains only few of the excess correlations for the GCN4 binding site as seen on the fifth column of Figure 3. The other correlations might be due to the translation level regulation of GCN4 protein (Hinnebusch (1984); Albrecht *et al.* (1998)). Because of the translation level regulation, it is not possible to infer the regulatory network of GCN4 protein just from microarray experiments which only measure mRNA abundances.

Also two of the putative motifs had this kind of widely correlating behavior. The motifs found from MIPS classes *rRNA processing* and *rRNA transcription* correlate well with about 9 disruptions, including ribosomal genes *rpl27a* and *rpl8a*. The correlation was very strong with disruptions of the little known genes *she4*, *bud21* and *bud22*. This observation suggest that the role of these genes in protein synthesis deserves further studies.

Indecisive motifs

Most of the analyzed transcription factor binding sites did not show significant correlation with any of the disruptions. Out of the 356 analyzed motifs only 102 had at least one correlating disruption at 0.01 confidence level. Positively out of the 37 known motifs 20 had at least one correlation.

Out of the 319 generated motifs only 85 had a correlating disruption. In general, for most of these putative motifs the correlations are rare and weak. This is not very surprising because the common MIPS classification used in the motif generation does not imply common activity or regulation.

The non-correlating sites include also SWI5 and OAF1 whose respective genes were disrupted in our dataset. This might seem bad at first but these lacking correlations can be explained by environmental conditions as follows.

The SWI5 protein is in the cell nucleus only during the G1 phase of the cell cycle, so its DNA binding activity can only take place under haploid conditions. Because we used only the expression measurements conducted with diploid state yeast, it is understandable that *swi5* disruption did not have an effect on its targets. The *yaf1* (alias *OAF1*) gene is activated with oleate, and in the oleate free environment the disruption of *yaf1* did not have significant effect on any of the measured genes.

Most of the motifs generated from the MIPS classes and the following 17 known motifs did not have significant correlation: ABF1, ALPHA1, ALPHA2, ATRepeat, CCA, CSRE, GAL, Gcr1, HSE, LYS14, MIG1, OAF1, PHO, PHO4 REB1, SCB, SFF, SWI5. Many of these binding site definitions are of insufficient specificity. For example the binding sites ALPHA1, SFF and SCB occurred in the upstreams of 1400–3200 genes when most of the disruptions affected only less than 500 genes.

It also seems probable that the disruptions in the otherwise homogeneous experimental conditions

$\lvert\mathbb{R}_b\rvert$	Site	$\lvert\mathbb{E}_g\rvert$	Disruption	$\lvert\mathbb{R}_b \cap \mathbb{E}_g\rvert$	Description
184	MCB	8	*mbp1*	5	Part of a DNA binding complex.
78	YAP1	55	*yap1*	6	Binding site of *yap1* factor
116	Ume6	346	*sin3*	20	Interacting proteins.
210	zap1	3	*msc7*	3	Relation via hydrogenases.
243	STE12	437	*dig11,dig2*	36	*dig1* represses *STE12*.
153	ndt80	151	*isw1,isw2*	13	Genetic interaction with *isw2*.
180	RPN4	33	*ubr2*	17	Similar cellular role.
257	RAP1	121	*vac8*	23	Weak link through vacuole
480	BAS1	23	*hpt1*	11	Adenine response.
149	STRESS	126	*sir4*	13	Unexplained.
116	HAP234		4 disruptions		Unexplained.
151	GCN4		34 disruptions		Central biosynthesis regulator.
89	Leu3		20 disruptions		In biosynthesis pathway.
58	MET31-32		16 disruptions		In biosynthesis pathway.
188	AFT1		*cup5 mac1 vma8*		Small molecule transport, iron uptake.
907	rRNA proc.		9 disruptions		Ribosomal activity.
514	PAC		9 disruptions		Ribosomal activity.
356	ECB		5 Weak disruptions		Early Cell-Cycle box
371	PDR		11 Weak disruptions		Unexplained
410	MCM1		10 disruptions		*MCM1* needs coregulators.

Table 1. Disruption – Binding Site correlations: Columns for size of the regulation set, name of the binding site motif, size of the effectual set of the best correlating disruption, name of the disruption, size of the intersection of the two sets and description of the result. Size of the effectual set and the intersection is only given for patterns that had one clearly best correlation.

did not disturb most of the regulatory pathways. Therefore it is understandable that only few of the binding sites had significant correlations with the expression profiles.

METHODS

Our microarray data consists of 263 disruption experiments selected from the compendium of expression profiles Hughes *et al.* (2000). From the 300 experiments in the compendium we chose not to include drug treatments or experiments conducted with yeast under haploid state. This way we wish to have more standard conditions over different expression measurements.

For each of the 263 disruptions we selected two sets $\mathbb{E}_g$ with different thresholds for the affected genes. The affected genes are selected according to the p-value statistic reported in Hughes *et al.* (2000). It takes into account the uncertainties due to low intensity spots and the inherent gene specific fluctuations in the transcript abundances. One of the sets $\mathbb{E}_g$ contains the genes having p-value less than 0.025 and the other one uses a more lenient cutoff value of 0.05. The gene expression cascades in Figures 3 and 4 have the regulators whose effect has p-value less than 0.0075.

Out of the 6312 ORFs whose mRNA levels were measured, 5390 had significantly altered expression at least in one experiment with 0.05 significance. 90 percent of the ORFs were significantly altered in less than 12 disruptions. The most sensitive ORF was affected in 52 experiments.

We used binding site motifs and site locations from Pilpel *et al.* (2001). The motif database contains 356 binding sites from which 37 are previously known from the literature and rest are generated with AlignACE of Roth *et al.* (1998) from gene upstream regions grouped by the MIPS classes. The 37 known binding sites correlate with 73 different disruptions. The most influential disruptions correlated with three different motifs.

Statistical measures

To measure the similarity between the two sets $\mathbb{R}_g$ and $\mathbb{E}_d$ we tried several statistics. First we computed a ratio between the proportion of altered genes within the genes having binding site versus the proportion of altered genes of all the genes. Formally our ratio statistic is $\frac{|\mathbb{E}_d \cap \mathbb{R}_g| / |\mathbb{R}_g|}{|\mathbb{E}_d| / |\mathbb{G}|}$. Unfortunately this ratio is not comparable between disruption experiments with different number of altered genes. Especially when the denominator is small even a very high ratio statistic might be due to random noise.

To statistically assess the similarity between the two sets we computed for each pair (d, g) the probability of obtaining an intersection $\mathbb{R}_g \cap \mathbb{E}_d$ this large or more, given the two sets are independent. Our null hypothesis is that there are $|\mathbb{G}|$ genes of which $|\mathbb{E}_d|$ are marked and we have randomly picked $|\mathbb{R}_g|$ of all the genes. With these assumptions our experiment is classical sampling without replacement and the size of the intersection $X = |\mathbb{R}_g \cap \mathbb{E}_d|$ is distributed according to the hypergeometric distribution. While we know the distribution of X we can compute the p-value $\hat{p}$ for our null hypothesis as the probability of observing an intersection this large or more given that the two sets were picked independently. The p-value is easily computed by formula

$$\hat{p} = \mathrm{P}(X \geq k) = 1 - \sum_{i=0}^{k} \frac{\binom{|\mathbb{E}_d|}{i}\binom{|\mathbb{G}|-|\mathbb{E}_d|}{|\mathbb{R}_g|-i}}{\binom{|\mathbb{G}|}{|\mathbb{R}_g|}}.$$

Because we compare each binding site with multiple disruptions, we need to adjust our p-values to represent the worst case binding sitewise probabilities. Without this correction we would seriously underestimate the probability of finding a correlation under the null hypothesis and we would find lots of seemingly significant correlations due to just the random noise. For the adjustment we use the sequential Holm's correction Holm (1979). The correction works by first sorting the p-values from n tests to increasing order $\hat{p}_1 \leq \hat{p}_2 \leq \ldots \leq \hat{p}_n$ and the corrected p-value is obtained by multiplying $p_i = (n + 1 - i)\hat{p}_i$. In our case the p-values are reported with respect to each binding site with two expression cutoff values, so our $n = 2 \cdot 263$. With this n we bound the p–value for each binding site. A more stringent observer might want to limit the p–value also across the binding sites by using $n = 2 \cdot 263 \cdot 356$ but since we look at the results one binding site at a time, we feel that the correction for one binding site is enough. If the binding sitewise corrected p-value p_i is smaller than our threshold 0.01, we reject the null hypothesis and say that $\mathbb{R}_g$ and $\mathbb{E}_d$ are not independent of each other.

Our hypothesis testing is quite powerful compared to the simple ratio test. For example the ratio statistic for disruption of *ERG11* and transcription factor binding site GCN4 is only 1.5 but the p-value is well below the threshold. The other end of the spectrum is the disruption of *arg80* and transcription factor binding site GCN4 whose ratio statistic is over 16, but due to the low number of genes affected by disruption of *arg80* our hypothesis test calls it a random chance.

DISCUSSION

At first it may seem that our result that only 102 of the 356 binding sites have correlation is discouraging. However, if we note that most of these binding site motifs are computationally predicted and only 37 of them are experimentally proven, our finding is in fact surprisingly positive. Also it should be taken into account that the 263 different disruptions are not necessarily covering all regulatory events of the yeast and that some binding sites may not be functional for any of the events studied.

In a way the most interesting case is to correlate the effects of the transcription factor gene disruptions with the set of genes having binding sites for that particular factor in their promoter regions (i.e., the case comparing $\mathbb{R}_g$ and $\mathbb{E}_d$ for $g = d$). Unfortunately only five genes in the dataset we used are in this category and only three of them correlate. The lack of correlation for the other two genes can be explained by the apparently poor quality of the binding site descriptions.

The finding that many binding sites correlate with disruptions of genes that are not transcription factors binding to these sites may first look surprising. We found 108 correlations for the 37 known motifs only 11 of which can be directly explained by the expression cascades. Possible explanations for the other correlations were briefly discussed in the introduction. A rather straightforward explanation of these cases is that these proteins may be vital parts of DNA binding complexes.

Another observation we make is that some binding sites are correlating with a wide range of disruptions. In these cases we can usually see that the binding site is a part of an important regulation pathway that can be affected by a wide range of events. We also uncovered some strong correlations for which we were unable to find a biological explanation.

An alternative way of viewing our problem is to see it as a graph comparison problem. One could understand a gene regulatory network simply as a representation of our knowledge of gene regulatory relationships in a form of a graph (i.e., the nodes of the graph are genes, and the edges are relationships between the genes). In this way we can obtain different graphs regarding on what knowledge we are representing. In this paper we considered two different sources of information: (1) the global gene expression data that tells which genes are affected by disruptions of other genes, and (2) the information about transcription factors and their binding sites in promoter sequences across the whole genome, telling essentially which genes are potentially regulated by each transcription factor. In effect we compared the regulatory networks defined by these two different data sources and demonstrated that they share many common features as well as contain also differences

ACKNOWLEDGEMENTS

The authors wish to thank Thomas Schlitt and Johan Rung for helpful discussions.

References

Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U. & Braus, G. H. (1998). Monitoring the gcn4 protein-mediated response in the yeast saccharomyces cerevisiae. *The Journal of biological chemistry*, **273**, 12696–12702.

Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome research*, **8**, 1202–1215.

Bussemaker, H. J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, **27**, 167–171.

Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., Lengieza, C., Lew-Smith, J. E., Tillberg, M. & Garrels, J. I. (2001). Ypd, pombepd and wormpd: model organism volumes of the bioknowledge library, an integrated resource for protein information. *Nucleic acids research*, **29**, 75–79.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Foury, F. (1990). The 31-kda polypeptide is an essential subunit of the vacuolar atpase in saccharomyces cerevisiae. *The Journal of biological chemistry*, **265**, 18554–18560.

Ge, H., Liu, Z., Church, G. M. & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nature Genetics*, **29**, 482–486.

Hinnebusch, A. G. (1984). Evidence for translational regulation of the activator of general amino acid control in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 6442–6446.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Jakt, L. M., Cao, L., Cheah, K. S. & Smith, D. K. (2001). Assessing clusters and motifs from gene expression data. *Genome research*, **11**, 112–123.

Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002). Mips: a database for genomes and protein sequences. *Nucleic acids research*, **30**, 31–34.

Pilpel, Y., Sudarsanam, P. & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, **29**, 153–159.

Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature biotechnology*, **16**, 939–945.

Rung, J., Schlitt, T., Brazma, A., Freivalds, K. & Vilo, J. (2002). Building and analysing genome–wide gene disruption networks. In *these proceedings*. Oxford University Press.

Scherf, M., Klingenhoff, A. & Werner, T. (2000). Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *Journal of molecular biology*, **297**, 599–606.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, **9**, 3273–3297.

Vilo, J., Brazma, A., Jonassen, I., Robinson, A. & Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB–2000)*. AAAI Press, pp. 384–394.

Vilo, J. & Kivinen, K. (2001). Regulatory sequence analysis: application to interpretation of gene expression. *European Neuropsychopharmacology*, **11**, 399–411.

Wagner, A. (2001). How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, **17**, 1183–1197.

Zhang, M. Q. (1999). Promoter analysis of co-regulated genes in the yeast genome. *Computers and Chemistry*, **23**, 233–250.