BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1576–1584

Comparing expression profiles of genes with similar promoter regions

Peter J. Park, Atul J. Butte and Isaac S. Kohane*

Informatics Program and Division of Endocrinology, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA

Received on March 1, 2001; revised on January 17, 2002; May 16, 2002; accepted on May 21, 2002

ABSTRACT

Motivation: Gene regulatory elements are often predicted by seeking common sequences in the promoter regions of genes that are clustered together based on their expression profiles. We consider the problem in the opposite direction: we seek to find the genes that have similar promoter regions and determine the extent to which these genes have similar expression profiles.

Results: We use the data sets from experiments on *Saccharomyces cerevisiae*. Our similarity measure for the promoter regions is based on the set of common mapped or putative transcription factor binding sites and other regulatory elements in the upstream region of the genes, as contained in the *Saccharomyces cerevisiae* Promoter Database. We pair up the genes with high similarity scores and compare their expression levels in time-course experiment data. We find that genes with similar promoter regions on the average have significantly higher correlation, but it can vary widely depending on the genes. This confirms that the presence of similar regulatory elements often does not correspond to similarity in expression profiles and indicates that finding transcription factor binding sites or other regulatory elements starting with the expression patterns may be limited in many cases. Regardless of the correlation, the degree to which the profiles agree under different experimental conditions can be examined to derive hypotheses concerning the role of common regulatory elements. Overall, we find that considering the relationship between the promoter regions and the expression profiles starting with the regulatory elements is a difficult but useful process that can provide valuable insights.

Contact: peter_park@harvard.edu

INTRODUCTION

Gene regulation has been studied intensely for decades, but the current knowledge of transcriptional control mechanisms is still fragmented. One of the major challenges in this context is to relate DNA sequences to their gene regulatory functions. For example, given a sequenced control region of a gene, we would like to make predictions about its expression patterns. Technological advances in rapid genome sequencing have provided entire genome sequences of many organisms, allowing researchers to apply various techniques to search for DNA sequences related to transcription control signals. However, the prediction of these regulatory elements is a difficult problem, and current methods are not accurate enough to be useful in the automatic annotation of a genome.

A major development in recent years is the use of high-throughput DNA microarrays. This allows monitoring of gene expression levels for thousands of genes simultaneously by quantifying the number of mRNA copies in transcription. Some of the earliest studies using microarrays were done on the budding yeast *Saccharomyces cerevisiae*, and it continues to be the focus of much attention (Wodicka *et al.*, 1997; Johnston, 2000; Simon *et al.*, 2001).

Some recent work in transcriptional regulation has taken advantage of both sequence and microarray data. By clustering the genes according to their expression profiles in one or more experiments, one is able to find co-expressed genes. Then, one hypothesizes that these genes are also co-regulated. The idea is then to extract the upstream sequences of these genes and apply an algorithm that searches for shared patterns over-represented in the sequences.

A primary consideration in these searches is in identifying transcription factor binding sites. Transcription factors (TFs) are proteins that can bind to a particular DNA sequence, called transcription factor binding site, usually of 5–25 base pairs in length. TFs are the elementary units of transcription regulatory mechanism. For a positively regulated gene, for example, a transcription activator protein might bind with an upstream sequence to prepare a gene for transcription. Therefore, identifying the TF binding sites is an important step in the characterization of a promoter region.

Detecting the binding sites for TFs in a genome sequence, however, is complicated by many factors,

*To whom correspondence should be addressed.

particularly for eukaryotes. For example, the consensus sequences recognized by some transcription factors are short and variable, and it has been known that sequence motifs of interest may contain too little diagnostic information for easy identification (Bucher, 1999). In general, the function of a regulatory region is complex, involving multiprotein complexes interacting with TFs bound to neighboring DNA sites. Therefore, in order to identify the correct sites, they need to be considered in the context of their interactions with other TFs (Wagner, 1999; Pilpel *et al.*, 2001).

There have been many methods for finding the common regulatory motifs. Inferring the motif sequence from a set of sequences known to contain the TF binding site, one can directly search for the specific patterns of nucleotides (Stormo and Hartzell, 1989; van Helden *et al.*, 2000; Lawrence *et al.*, 1993). With the availability of microarray data, these searching techniques could be made more efficient, by applying them to the upstream sequences of a group of co-expressed genes rather than to the sequences of all genes. It was recognized early (DeRisi *et al.*, 1997; Spellman *et al.*, 1998) that some groups of genes with similar expression patterns belong to similar regulatory pathways and that they also contain the binding sites for relevant transcription factors. In Brazma *et al.* (1998), 300 bp upstream regions for genes in the same expression clusters were compared against all other upstream regions and a large number of over-represented patterns were identified, many of which turn out to have matches to substrings in known TF binding sites.

It is important to note that the underlying hypothesis in these studies is that co-expression of genes implies common regulatory mechanism, and in particular, the presence of common TF binding sites. However, it is well known that the correspondence between gene cluster and common motifs is imprecise in both directions, i.e., many genes in the cluster do not contain the motif and those with the motif often are not expressed (Bussemaker *et al.*, 2000). Sometimes different transcriptional mechanisms may result in similar expression patterns and sometimes the same mechanism may have different effect on the expression. Therefore, the success of identifying regulatory elements guided by expression similarity relies in part on the extent to which this 'controversial' hypothesis (Altman and Raychaudhuri, 2001) holds true for the given set of expression profiles.

In this paper, we examine this hypothesis by posing the question in the opposite direction. Rather than starting with the expression profiles and asking what regulatory elements they have in common, we ask whether the genes with similar promoter regions, i.e., genes that have common TFs or other regulatory elements, are in fact co-expressed according to the microarray data. Understanding the complex relationship between the promoter region and the expression is a great challenge, and we believe it is important to consider the problem in both directions.

The extent of disagreement between the similarities of the promoter regions and the expression profiles helps to evaluate the effectiveness of the methods in which one tries to extract regulatory sites in the promoter regions of co-expressed genes. In addition, whether or not the correlation between profiles is high, we may be able to gain valuable insights by examining the profiles. By seeing how the profiles are correlated for each of the experimental data sets, we may be able to hypothesize the conditions under which the promoters become relevant. This would be a way to elucidate the role of some common promoter elements.

METHODS

We would like to define a reasonable measure of similarity among different promoter regions, pick out the most similar pairs, and examine their correlations across an expression space. We make use of the regulatory elements to describe promoter regions.

SCPD database

S. cerevisiae Promoter Database (SCPD) at Cold Springs Harbor Laboratory contains a list of genes with their TF binding sites, collected from the literature (Zhu and Zhang, 1999). This database has been used for several studies (Bussemaker *et al.*, 2000; Hughes *et al.*, 2000; Cohen *et al.*, 2000), often to verify the performance of an algorithm for identifying regulatory sites. For a region of specified length upstream from a gene, one can retrieve the TFs and their binding sites as well as some other regulatory factors, either mapped or putative.

We use this database to extract TF binding sites in the 600 base pairs upstream of the genes. (Among the regulatory sites examined by Roth *et al.* (1998), about 85% were within 600 bases upstream of translation start.) In SCPD, there are over two hundred genes with at least one mapped site. There are total of 455 mapped sites; among them, 203 sites are present once, 69 in two copies, 19 in three, nine in four, three in five and one in six (Zhu and Zhang, 1999).

While it is clear that information collected in SCPD is far from complete, the mapped sites in the database provide a fairly reliable measure. The database also contains a much larger number of putative sites. These have been obtained using a version of the expectation–maximization (EM) algorithm which has been shown to work well in finding the sites as well as the weight matrices for representing them (Stormo, 2000). However, the variability in the binding site sequences and many other factors make an accurate description of putative sites difficult and their prediction even more difficult. As a

result, there are sometimes up to 20 times more putative elements than mapped elements for a gene, but only a small fraction of them has been demonstrated as true positives. In our comparisons, we first use mapped and putative sites separately and then use them together.

Similarity measure for promoter regions

Unfortunately, how to define similarity between the promoter regions is not clear. It is complicated, for example, by the fact that often there are many TFs involved in the regulation of each gene, possibly with multiple binding sites for each TF. It is known that contextual information is important, as transcriptional control frequently involves interactions of many TFs, but it is difficult to formalize this to come up with a measure that should apply in all cases. Some other considerations include how to weigh the mapped sites against the putative ones, how to weigh multiple matches of the binding sites for the same TF, how to weigh the binding sites that are present for only one gene, and whether the prevalence of a TF should diminish its weight.

Because of these ambiguities in defining biological similarity, whichever measure one employs will be arbitrary to some extent. We have tried to capture the main effects and performed sensitivity analysis for the parameters, which showed that the similarity measure we introduce is fairly robust to parameter values. For each pair of genes, we used the following score S:

$$S = \sum_{j=1}^{2} \gamma_j \sum_i \left(f_i^j \right)^{-1/2} \left[\left(\frac{2}{N_{1i}^j + N_{2i}^j} + \alpha \right) C_i^j - \beta \, (N_{1i}^j + N_{2i}^j) \mathrm{I}_{\{C_i^j = 0\}} \right],$$

where $j = 1, 2$ correspond to the mapped and putative sites, respectively, and i denotes the distinct binding sites; γ_j are the relative weights between mapped and putative sites; omitting the superscript j for convenience, N_{1i} and N_{2i} are the number of binding sites for the two genes in a pair; and C_i is the number of common sites between the two genes. The term $2\,C_i/(N_{1i} + N_{2i})$ is the proportion of sites in common; the next term αC_i is to give preference to the pairs with more matches when the proportion of the matches as given by the first term is the same. The term containing β is the penalty term for the cases in which a binding site appears for only one gene, specified by indicator function $\mathrm{I}_{\{C_i=0\}}$, which is equal to 1 if $C_i = 0$ and 0 otherwise. Factor $(f_i)^{-1/2}$ is the weighting of the number of common occurrences by the inverse square root of its frequency in the database.

In defining the similarity score S, we have tried to satisfy few basic properties. First, we normalize the number of matched sites between the genes by the number of sites in both genes. This avoids automatically penalizing genes with fewer binding sites. However, we do give some preference to having extra matches, as specified by α. (With each letter representing a TF binding site: AAA/AAA should be preferable to A/A, but not by a factor of 3.) By including the penalty term with β, we account for those sites that do not have matches. (A/A should be preferred over AB/A) We also include the frequency information (A/A should be preferred over B/B if A occurs few times in the database while B occurs hundreds of times). When we simply consider the number of shared elements without accounting for their overall frequency, we find that those genes with few commonplace TFs dominate the list of high-scoring pairs. Dividing by the frequency, on the other hand, de-emphasized the common regulatory elements too much. The choice of square root in the denominator is a compromise, giving decreasing weight to successive occurrences. If there are some other effects that we want to include in the similarity score, it can be done easily. We can, for example, prefer the ordered matches (ABC/ABC over ABC/ACB); however, this effect should be relatively small and the complication it would bring with an additional parameter seems to outweigh the benefits.

If the similarity measure is to be based only on the counts of different motifs, one can use the Poisson-based similarity metric (van Helden, 2002). This models the probability of an occurrence for a motif by Poisson distribution, with the average number of the motif in all genes as the mean. When this measure was applied to our data, however, we observed that the average of the correlation coefficients for the expression profiles is much lower. We suspect that this is related to the sparsity of the mapped sites in the database. For other cases, the Poisson-based measure and the heuristic measure we propose give a very high concordance.

Scope of the expression space

After we decide on the similarity measure for the promoter regions, we must still decide on the appropriate way to measure their expression space similarity. The difficulty arises due to the inevitable lack of exact correspondence between the promoter elements found on a pair of genes and the presence of the experiments that capture the effect of those elements. On the one hand, if the expression space we compare is too small, there may be regulatory elements present in the two genes but without the profiles that would reflect their presence; on the other hand, if the expression space is too large, there may be profiles that are not related to any of the promoters found. In both cases, the profile similarity would not reflect the promoter similarity exactly.

For the purpose of measuring correspondence between promoter regions and the expression profiles, the compromise we adopt is a moderate size expression profile space from extensively studied regulatory mechanisms. Since the mapped binding sites would most likely be related to the well-studied mechanisms, the use of these profiles seems appropriate. The compilation of 79 profiles contains data from several experiments such as mitotic cell division cycle, sporulation, the diauxic shift, and shock responses (Eisen *et al.*, 1998). Of the 236 genes that have mapped sites in the database, 183 are included in the 2467 genes used in the experiments. We also chose this particular data set because it has been demonstrated already that clustering with this data set results in groups of genes with the same functional categories (Eisen *et al.*, 1998) and that a set of clusters based on this data set can be used to recover many regulatory elements (Vilo *et al.*, 2000).

RESULTS

Using mapped and putative sites separately

We first compare the promoter regions using only the mapped sites. We do this by setting $(\gamma_1, \gamma_2) = (1, 0)$ in the similarity measure. The top pairs giving the highest Pearson correlations are listed in Table 1, and the promoter regions of the top five pairs are displayed in Figure 1 with each symbol representing a mapped TF binding site. The first pair consists of ADE5,7 and HIS4. They share both copies of BAS1 and one copy of BAS2. GCN4 is only contained in the HIS4 promoter but it appears relatively often and hence carries a smaller weight; BAS2, on the other hand, appears in no other promoter regions and carries a large weight. Few of the pairs on the list share exactly the same set of TFs. For example, the second pair, COX6 and QCR8, both have ABF1 and HAP2;HAP3;HAP4 genes.

When we examine the correlation coefficients for the expression profiles of many of these pairs, we find that they are not as high as one would expect. We see in Table 1 that it may be as high as 0.955 (ENO2 and PGK1), but there are also few negative numbers (four out of 40). The average and the standard deviation of these numbers are shown in Table 2. Using mapped sites only, the mean of the correlation coefficients for the top 40 genes is 0.404 (the standard deviation is 0.307). To put the correlations in context, we also computed correlations between randomly selected genes. A random sample of 10^6 pairs shows the average to be 0.04 (the standard deviation is 0.26). It is definitely the case that the selected pairs of genes have higher correlations on average than those obtained from random pairs of genes. But the correlations are not uniformly high and there is a large variability.

One possible reason for the lack of corresponding agreement in expression profiles is that not enough sites have been mapped, at least as contained in SCPD. If this is the case, we may be able to obtain more uniformly high correlations between the expression profiles by including the putative sites. The number of putative sites is usually several times larger than that of the mapped sites for each gene in SCPD. However, when we use the putative sites alone, the mean decreases to 0.157 (standard deviation of 0.283). This seems to indicate that the putative sites have not been predicted accurately, and that they do not provide a reliable description of the promoter region. This is not surprising since the number of predicted sites far exceeds the expected number of sites and there are many false positives. This clearly indicates the limitations of the current computational approaches for identifying putative sites.

Using mapped and putative sites together

Though not as accurate as mapped sites, putative sites may provide some additional information when combined with the mapped sites. We examine this by choosing the measure of relative influence through γ_i accordingly. We have found that values in the neighborhood of $(\gamma_1, \gamma_2) = (0.8, 0.2)$ give the best result, defined as the average correlation of the top 20 or 40 pairs. We find that the overall average of the correlation in fact increases when putative sites are incorporated, as shown in Table 2. This shows that while putative sites by themselves are a poor measure compared to mapped sites, they should be used in addition to the mapped sites. The top 20 pairs are listed in Table 1. The list is quite similar to the one obtained by mapped sites alone: the first two are the same; 18 of the top 20 are shared by both lists.

While there are some parameters in this formulation, the list of pairs with high similarity score S are fairly robust to the perturbations in the parameters chosen, particularly to α and β. To determine the parameters, we searched through the parameter space exhaustively in our simulations, looking for those values that give the highest average correlations in the top pairs. We found that $(\alpha, \beta, \gamma_1, \gamma_2) = (1, 0.1, 0.8, 0.2)$ are near optimal and stable values. Perturbations in α and β resulted in only small changes and the lists of top pairs were almost identical; different values of γ_1 and γ_2 resulted in more substantial changes. Some other variations on the similarity measure, such as different ways of incorporating the frequency information, seem to result in gene pairs with similar or lower average correlations.

Identifying role of regulatory elements

Once we have the gene pairs with similar promoter regions, we can examine the extent of agreement in their profiles under different experimental conditions. By seeing what common regulatory elements confer similar

Table 1. Gene pairs with the highest similarity score S. The list on the left was obtained using mapped sites only $(\alpha, \beta, \gamma_1, \gamma_2) = (1, 0, 1, 0)$; the one on the right was obtained using both mapped and putative sites $(\alpha, \beta, \gamma_1, \gamma_2) = (1, 0.1, 0.8, 0.2)$. In order to see how similar the expression profiles of these gene pairs are, we compute the Pearson correlation coefficients over the aggregate data. A summary of the coefficients is presented in Table 2

	Mapped only			Mapped and putative		
Rank	Gene pair	Score	Pearson corr	Gene pair	Score	Pearson corr
1	ADE5,7, HIS4	2.096	0.600	ADE5,7, HIS4	1.691	0.600
2	COX6, QCR8	1.816	0.712	COX6, QCR8	1.549	0.712
3	ENO2, PGK1	1.699	0.955	LEU1, LEU2	1.469	−0.097
4	PGK1, TPI1	1.651	0.906	ANB1, HEM13	1.379	0.111
5	PHO5, PHO84	1.579	0.481	PGK1, TPI1	1.320	0.906
6	CLN2, IME1	1.414	0.040	ENO2, PGK1	1.317	0.955
7	LEU1, LEU2	1.414	−0.097	FBP1, ICL1	1.316	0.751
8	PHR1, URA3	1.414	0.030	FAS1, RPO21	1.276	0.374
9	ANB1, HEM13	1.363	0.111	CLB1, SWI5	1.262	0.458
10	CLB1, SWI5	1.359	0.458	PHO5, PHO84	1.207	0.481
11	BAR1, STE2	1.358	0.137	BAR1, STE2	1.159	0.137
12	ARG1, ARG8	1.289	0.277	ARG1, ARG8	1.116	0.277
13	ADE5,7, HIS7	1.262	0.677	CDC21, CDC6	1.060	0.299
14	FBP1, ICL1	1.224	0.751	PDR5, SNQ2	1.056	0.477
15	FAS1, RPO21	1.189	0.374	CDC21, CDC9	1.053	0.726
16	CYT1, MET16	1.154	0.111	ADE5,7, HIS7	1.050	0.677
17	CYT1, PGK1	1.154	−0.038	PHR1, URA3	1.013	0.030
18	MET16, PGK1	1.154	−0.236	CYT1, MET16	1.013	0.111
19	HIS4, HIS7	1.152	0.598	CDC6, CDC9	1.002	0.397
20	ADE2, ADE5,7	1.133	0.667	CLB1, CLB2	0.994	0.428

Table 2. Correlation coefficients. We compute the mean and standard deviation of the top 40 gene pairs generated by the similarity measure with different γ_i values. Using only the mapped sites $(\gamma_1, \gamma_2) = (1, 0)$ is more efficient than using only the putative sites $(\gamma_1, \gamma_2) = (0, 1)$, but combining them gives the best result. The first column is for a large number (500 000) of gene pairs chosen at random. The pairs chosen according to S have substantially higher correlations. If the same parameters $(\alpha, \beta) = (1, 0.1)$ are used for all three cases, the means (standard deviations) are 0.383 (0.327) and 0.125 (0.344) for the mapped and putative only cases, respectively

	All	Top 40 pairs $(\gamma_1, \gamma_2) = (1, 0)$ $(\alpha, \beta) = (1, 0)$	$(\gamma_1, \gamma_2) = (0, 1)$ $(\alpha, \beta) = (0.6, 0.5)$	$(\gamma_1, \gamma_2) = (0.8, 0.2)$ $(\alpha, \beta) = (1, 0.1)$
Mean	0.041	0.404	0.157	0.436
s.d.	0.26	0.307	0.283	0.316

expression in which experiments, we may gain more understanding of the transcriptional regulations involved.

We plot the profiles of two pairs in Figure 2, along with the correlation coefficients for each subset of the data listed under the name of the experiment. The genes in the top pair, ADE5,7 and HIS4, have been studied extensively in the context of biosynthesis. In particular, HIS4 gene is one of the best characterized yeast genes related to amino acid biosynthesis. These two genes achieve a high similarity score due to the presence of the binding sites for the transcriptional activators BAS1 and BAS2. It turns out that the role of BAS1 and BAS2 in ADE5,7 and HIS4 have been characterized separately in general (Rolfes *et al.*, 1997; Arndt *et al.*, 1987); it also has been noted that BAS1 and BAS2 are cross-pathway regulators of both the histidine and purine biosynthetic pathway (Daignan-Fornier and Fink, 1992).

In the expression space, ADE5,7 and HIS4 have an overall correlation coefficient of 0.600, but we see that the coefficients vary widely depending on the experiment. There is a strong agreement six of the eight experiments, with coefficients all greater than 0.731. However, in ELU (centrifugal elutriation), the coefficient is low (0.118); in CDC15 (temperature-sensitive cdc15 mutant), the coefficient is negative (-0.396). From this we may suspect that the mechanism involved in the two experiments ELU and CDC15 may be different from the others that involve BAS1 and BAS2, and that they may be related

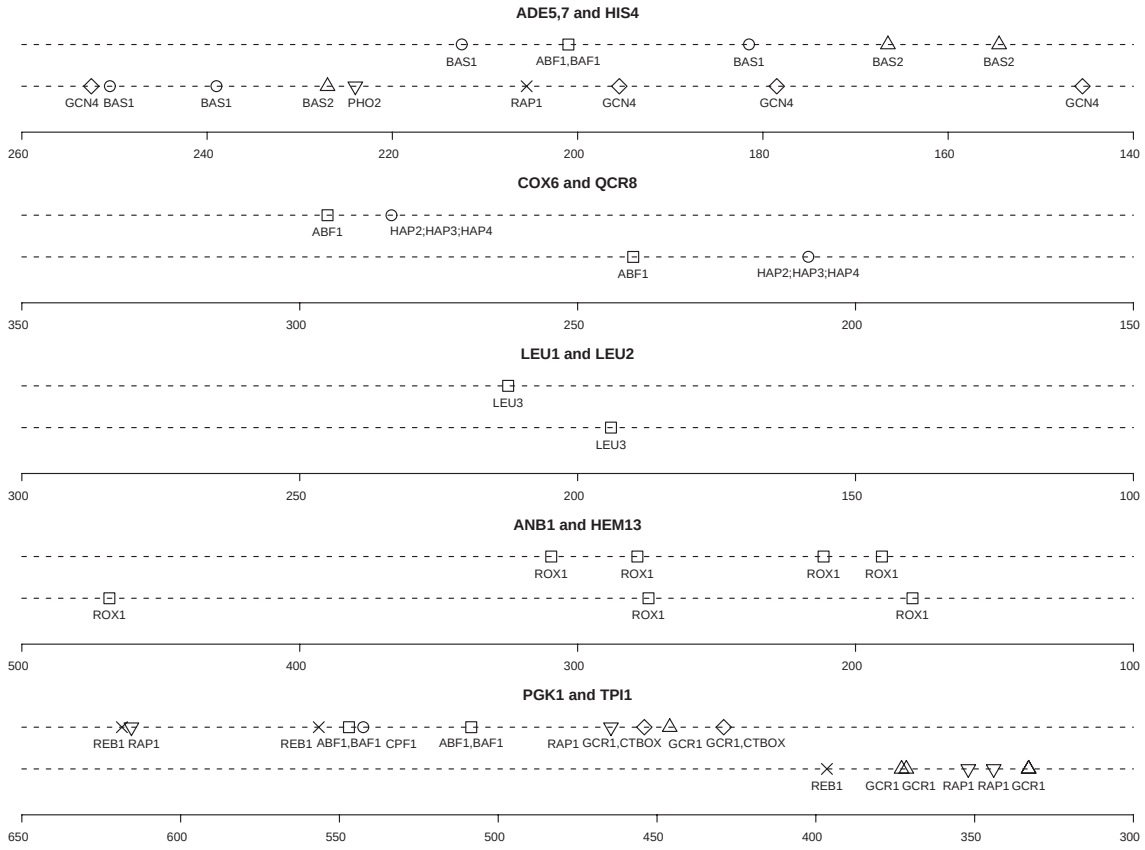

Fig. 1. Promoter regions of the five gene pairs with the highest similarity score *S*, when both mapped and putative sites are used. Only the mapped sites are shown here (a large number of putative sites makes it difficult to show them this way).

to the TFs that are not common in the two promoters, such as GCN4, RAP1, and ABF1;BAF1. This is a limited hypothesis, since it is based only on the mapped sites and a small data set. However, the fact that the profiles agree very well in the six experiments seems to indicate that two mapped sites are an important factor and suggest that the two remaining experiments may involve different transcriptional mechanisms.

For the third pair LEU1 and LEU2, the overall correlation is much lower (-0.097). (The second pair, COX6 and QCR8, is not interesting, as the correlation across all experimental conditions is high) The only common binding site is for LEU3, which is known to encode a factor for control of RNA levels of a group of leucine-specific genes. (Among the putative sites, they share GCN4, BAS2, and HSTF.) But based on the correlations coefficients shown in Figure 2, it seems likely that the presence of LEU3 by itself is not directly relevant in most experimental conditions. We note that the correlation coefficients do not take account of the noise level in the data, and small coefficients do not necessarily imply lack of correlation when the expression levels are comparable to the noise level.

Classification and Regression Trees (CART)

Another approach to better understand the relationship between the promoters and the expression profiles is a direct classification. The similarity measure we have used enumerates all the promoters in order to relate them to expression profiles. It is possible, however, that what determines the expression profile is a smaller subset of specific promoters that have dominant effects. In that case, there would exist a simple set of rules that can explain the profiles in terms of few particular promoters. The similarity measure we introduced earlier may not capture this effect as well as a direct classification scheme.

For this approach, we use a Classification And Regression Trees (CART) method, as implemented in the software package C5.0 (Quinlan, 1993). It is an algorithm that extracts informative patterns from data, with the aim of predicting the class of a sample based on its attributes. In our case, we would like to predict the type of expression profile given the promoter attributes.

One difficulty is that the class that we wish to predict is a somewhat arbitrary grouping of a clustering result,

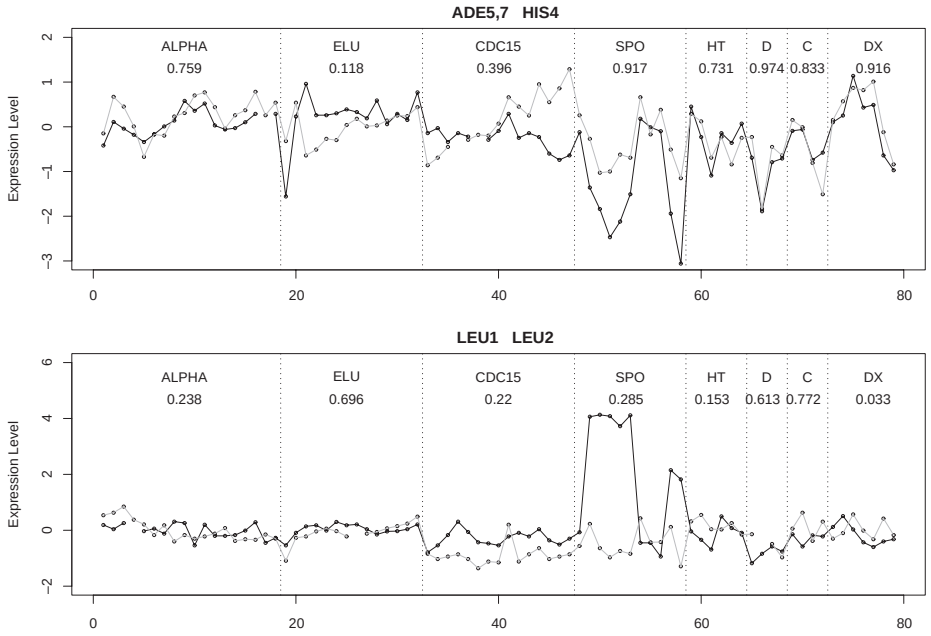

Fig. 2. Profiles of two pairs: ADE5,7 and HIS4; LEU1 and LEU2. Vertical lines divide the expression space into 8 experimental conditions that make up the data set: ALPHA (the cell division cycle after synchronization by alpha factor arrest, 18 points), ELU (centrifugal elutriation, 14 points), CDC15 (with a temperature-sensitive cdc15 mutant, 15 points), SPO (sporulation, 11 points), HT (shock by high temperature, 6 points), D (reducing agents, 4 time points), C (low temperature, 4 points), and DX (diauxic shift, 7 points). The break in expression lines is due to the missing values. The numbers below the experiment labels are the Pearson correlation coefficient for that segment. The overall correlations for the two pairs are 0.600 and −0.097, respectively.

and the result of classification may depend heavily on how one defines the classes of expression profiles. In the case of *S. cerevisiae*, we can avoid the problem of defining clusters of expression profiles by using the functional categories instead, as the expression profiles tend to be closely related to the function of the genes (Eisen *et al.*, 1998). MIPS database (Mewes *et al.*, 1999) contains the necessary functional category annotations.

We then seek to find a classification that can explain the functional categories of genes in terms of their promoters. The output of CART is a classification tree as well as optimal rules for determining the functional categories of all genes. If simple structures appear, then we may conclude that those few regulatory elements involved are the important ones to capture and that it is not necessary to account for all the elements in the promoter region.

Overall, we find that the presence or absence of various regulatory elements constitutes a poor description of the functional categories. The misclassification rates are 50.8% both for the decision tree and for the rules when only the mapped sites are used. When both mapped and putative sites are used, the rules become much more complicated, with only a slight improvement in correct classification. The error rate is little better (39.2%) for the decision tree and slightly worse (51.4%) for the rules.

DISCUSSION

The complex relationship between regulatory elements and gene expression has often been studied starting from the expression space. We considered the relationship in the opposite direction in the paper, trying to capture the promoter similarity with a simple measure. Based on the information contained in the SCPD database and the form of similarity measure we introduce here, we are able to find many pairs of genes whose expression levels have high correlation, but we also find some pairs with low or even negative correlation.

The fact that a high score on our similarity measure often does not result in a high correlation in the expression profile may be attributed to several things. Most importantly, as we discussed in the Methods section, the scope of the expression space may not match the promoter elements precisely. Although successful recovery of a large number of regulatory elements has been reported using the same expression space (e.g., Vilo *et al.* (2000)), the data set is not sufficiently large to represent the gene function in some cases, and it is too large and suppresses the effect of promoters in other cases. This aspect is critical but is often not addressed adequately.

Second, it may be that the similarity measure we devised does not capture the main elements of the transcriptional

mechanism. It is certainly the case that this measure does not include all the factors that are known to affect the level of gene expression. One set of these factors is related to a more precise description of the TF binding sites, such as their location from the start of the transcription site and the order in which they are arranged. In developing a method to determine if a gene belongs to a particular class based on motif-based hidden Markov models of the promoter regions, Pavlidis *et al.* (2001) found that the relative locations of motifs as well as the number of their occurrences are important. They also suggested that the bendability determined by the base composition of the spacer regions between motifs should be included as it influences the binding of the TFs. There are also some post-transcriptional mechanisms, such as those controlling mRNA stability that affect the co-regulation for some genes.

There are also some sequence-independent effects, such as the distance of the gene from the centromere. It has been shown experimentally that the expression level can change substantially for the same gene placed on different location along the chromosome. Also not considered in this study are chromatin effects, such as the acetylation state of histones, which have been shown to play an important role in gene expression. In our analysis, we considered 600 bp upstream of the genes (larger segment seems to make little difference), and this does not account for long-range interactions that we know exist. While all these can affect transcription, we have tried to keep the model simple and capture the main effects.

Another reason for the discrepancy may simply be the incompleteness of the information contained in the database. The yeast genome is the most well studied of all eukaryotes and many genes have been characterized extensively through numerous experiments. However, only a fraction of sites has been mapped and the putative sites have not been predicted with great accuracy. With a more complete database, we would be able to draw stronger conclusions.

We note that this lack of correspondence between the similarity of the promoter regions, as described by regulatory elements, and the similarity of expression profiles has some implications in the common methods used to identify regulatory elements. While it can be effective in many cases, extracting common motifs from the upstream sequences of genes that cluster together based on expression data may be limited in its effectiveness in other cases. This approach is based on the assumption that co-expression in a set of experiments provides sufficient information to capture the promoter elements. But depending on the expression space, certain regulatory elements may or may not be shared by a large portion of those genes. Clearly, deeper understanding of the control mechanisms and more sophisticated computational methods will be needed to obtain a comprehensive set of regulatory elements. We also note the ineffectiveness of putative sites in our approach indicates that more work is needed in describing and identifying true positive sites.

Finally, we found that the fact that these profiles do not always correspond well can provide an opportunity to gain some understanding of the role of the promoters. We have examined the strength of correspondence under different experimental conditions, and we are able to make preliminary hypotheses on the involvement of the common promoters under various conditions. As more accurate information on the TF binding sites is gathered on the databases and more microarray data are accumulated, classification or clustering based on the properties of the promoter regions appears to be a promising approach.

ACKNOWLEDGMENTS

We thank the anonymous referees for their comments and Dr. Jacques van Helden for helpful discussions.

REFERENCES

Altman,R.B. and Raychaudhuri,S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, **11**, 340–347.

Arndt,K.T., Styles,C. and Fink,G.R. (1987) Multiple global regulators control HIS4 transcription in yeast. *Science*, **237**, 874–880.

Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, **11**, 1202–1215.

Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, **9**, 400–407.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.

Cohen,B.A., Mitra,R.D., Hughes,J.D. and Church,G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.*, **26**, 183–186.

Daignan-Fornier,B. and Fink,G.R. (1992) Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proc. Natl Acad. Sci. USA*, **89**, 6746–6750.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Johnston,M. (2000) The yeast genome: on the road to the golden age. *Curr. Opin. Genet. Develop.*, **10**, 617–623.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle

sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.*, **27**, 44–48.

Pavlidis,P., Furey,T.S., Liberto,M., Haussler,D. and Grundy,W.N. (2001) Promoter region-based classification of genes. *Pac. Symp. Biocomput.*, **6**, 151–164.

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

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Rolfes,R.J., Zhang,F. and Hinnebusch,A.G. (1997) The transcriptional activators BAS1, BAS2 and ABF1 bind positive regulatory sites as the critical elements for adenine regulation of ADE5,7. *J. Biol. Chem.*, **272**, 13343–13354.

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

Simon,I., Barnett,J., Hannett,N. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spellman,P.T., Sherlock,G., Zhang,M.Q. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. and Hartzell,G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.

van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **281**, 1808–1818.

van Helden,J. (2002) Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, in press.

Vilo,J., Brazma,A., Jonassen,I., Robinson,A. and Ukkonen,E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 384–394.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Wodicka,L., Dong,H., Mittmann,M., Ho,M.-H. and Lockhart,D.J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **15**, 1359–1367.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.