doi: 10.1016/S0022-2836(02)00026-8 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 318, 71–81

JMB

Computational Identification of Transcription Factor Binding Sites *via* a Transcription-factor-centric Clustering (TFCC) Algorithm

Zhou Zhu†, Yitzhak Pilpel† and George M. Church*

Department of Genetics and Lippar Center for Computing and Genetics Harvard Medical School, 200 Longwood Avenue Boston, MA 02115, USA

*Corresponding author

While microarray-based expression profiling has facilitated the use of computational methods to find potential *cis*-regulatory promoter elements, few current *in silico* approaches explicitly link regulatory motifs with the transcription factors that bind them. We have thus developed a TF-centric clustering (TFCC) algorithm that may provide such missing information through incorporation of biological knowledge about TFs. TFCC is a semi-supervised clustering algorithm which relies on the assumption that the expression profiles of some TFs may be related to those of the genes under their control. We examined this premise and found the vicinities of TFs in expression space are often enriched with the genes they regulate. So, instead of clustering genes based on the mutual similarity of their expression profiles to each other, we used TFs as seeds to group together genes whose expression patterns correlate with that of a particular TF. Then a Gibbs sampling algorithm was applied to search for shared *cis*-regulatory elements in promoters of clustered genes. Our working hypothesis was that if a TF-centric cluster indeed contains many targets of the seeding TF, at least one of the discovered motifs would be the site bound by the very same TF. We tested the TFCC approach on eight cell cycle and sporulation regulating TFs whose binding sites have been previously characterized in *Saccharomyces cerevisiae*, and correctly identified binding site motifs for half of them. In addition, we also made *de novo* predictions for some unknown TF binding sites.

Keywords: computational biology; transcription factor; clustering; DNA regulatory motif; expression profile

Introduction

Understanding how the expression levels of thousands of genes are regulated at all times in the life of a cell remains one of the greatest challenges of molecular biology. A major component of gene regulation occurs at the level of transcription. Central to this mechanism are transcription factors (TFs), proteins that typically bind to specific, short DNA sequence motifs of ~5–25 bp in the *cis*-regulatory region (promoter, enhancer) of a gene and activate or repress its transcription. The identification of relevant TFs and their binding sites is an important step in elucidating the mechanism of transcriptional regulation of a particular gene.

Traditionally, TF binding sites have been characterized by a variety of different experimental approaches.[1] The recent advancement in high-throughput gene expression monitoring technology and availability of complete genome sequences enable the use of computational methods to find potential TF binding sites. For instance, genes can be grouped into disjoint clusters on the basis of similarity in their expression profiles or functional annotations.[2–6] Genes in the same cluster are thought to be transcriptionally co-regulated, and their regulatory regions can be analyzed for the presence of shared sequence motifs.[6–8] But few of these approaches attempt to explicitly link computationally discovered

† These authors made an equal contribution to this work.

Abbreviations used: TF, transcription factor; TFCC, transcription-factor-centric clustering; ORF, open reading frame.

E-mail address of the corresponding author: church@arep.med.harvard.edu

Table 1. Enrichment of target genes in the vicinity of TFs in expression space

TF	Average[a]	Right-side tail[b]
Ndt80	<0.001	<0.001
Met32	<0.001	<0.001
Fkh1	<0.001	<0.001
Mbp1	0.97	0.02
Mbp1[c]	<0.001	<0.001
Swi4	<0.001	<0.001
Fkh2	<0.001	<0.001
Swi5	0.33	0.956
Mcm1	0.133	1

[a] *P*-values on the hypothesis that the average correlation coefficient between a TF and its target genes is equal or lower than the average correlation coefficient between it and the same number of randomly sampled genes.
[b] *P*-values on the hypothesis that the fraction of target genes whose correlation with a TF is above 0.95 (for sporulation data set) or 0.7 (for cell cycle data set) is equal or less than the fraction of randomly sampled genes whose correlation with the TF is above that same threshold.
[c] The correlation coefficients between Mbp1 and its target genes or randomly sampled genes were calculated with a time delay of ten minutes and negative correlation type.

regulatory motifs with the transcription factors that bind them. We have developed a TF-centric clustering (TFCC) algorithm that may provide such missing information.

TFCC is built upon the assumption that the expression profiles of some TFs may be related to those of the genes under their control. Previous experimental observations suggest that in at least a few instances, the mRNA levels of TFs and some of their targets appear to be correlated.[9–11] We also notice that in expression space, genes often reside in the vicinity of the TFs that regulate them (see below). Since TFCC groups together genes that share a common expression pattern with a particular TF, we suspect some of them may contain the regulatory motif bound by this factor. AlignACE,[3,5] a Gibbs sampling-based motif finding algorithm, is applied to search for such shared *cis*-regulatory elements in promoters of clustered genes. Our working hypothesis is that if a TF-centric cluster indeed contains many targets of the seeding TF, one would expect at least one of these discovered motifs to be the site bound by the very same TF (Figure 1(a)). Here we not only tested the approach on eight TFs whose binding sites have been previously characterized in *Saccharomyces cerevisiae*, but also proposed novel predictions for some TFs with unknown binding sites.

Results

Enrichment of target genes around TFs in expression space

We first set to examine the basic premise of our method, namely genes regulated by some TFs are clustered around them in expression space. Testing such a hypothesis is confounded by our lack of accurate and complete knowledge about the sets of genes regulated by many TFs. As an estimate, we used recent genome-wide chromatin immunoprecipitation and mutational analyses,[12–14] and obtained a list of genes most likely to be controlled by the TFs studied here. They were classified as targets for the purpose of this study. We then assessed the extent to which TFs are surrounded by their target genes in expression space from the evaluation of the pairwise Pearson correlation coefficients between each TF and its known targets. We derived a *P*-value on the null hypothesis that there is no enrichment of target genes around a TF by comparing the above correlation coefficient values to those obtained using a random sample of genes and the same TF. As seen in Table 1, five (Ndt80, Fkh1, Swi4, Fkh2 and Met32) of the eight TFs we studied have expression profiles significantly ($P < 0.001$) more highly correlated with their respective targets than with genes randomly selected from a pool of 3000 open reading frames (ORFs) that vary most in expression.

Table 2. Sizes of the six types of clusters generated for each TF

TF	No time delay, +[b]	Time delay,[a] +[b]	No time delay, −[b]	Time delay,[a] −[b]	No time delay, +/−[b]	Time delay,[a] +/−[b]
Fkh1	30	11	5	6	35	17
Fkh2	23	1	17	4	40	5
Mcm1	4	0	27	2	31	2
Met32	9	0	0	1	9	1
Swi5	56	48	15	8	71	56
Mbp1	17	17	1	22	18	39
Swi4	59	12	1	0	60	12
Ndt80	198	NA[c]	3	NA[c]	201	NA[c]

Cutoffs of 0.8 and 0.95 were set on the correlation coefficient scores for the cell cycle and sporulation data, respectively. We used a higher cutoff for the sporulation data because this response is characterized by a lower number of different types of profiles (data not shown).
[a] Time delay indicates ORF expression was delayed from TF expression by one time point during clustering.
[b] (+) Positive correlation only; (−) negative correlation only; (+/−) positive and negative correlation.
[c] Only clusters without time delay were produced for Ndt80 because a much longer and uneven time interval was used in the sporulation experiment.[12]

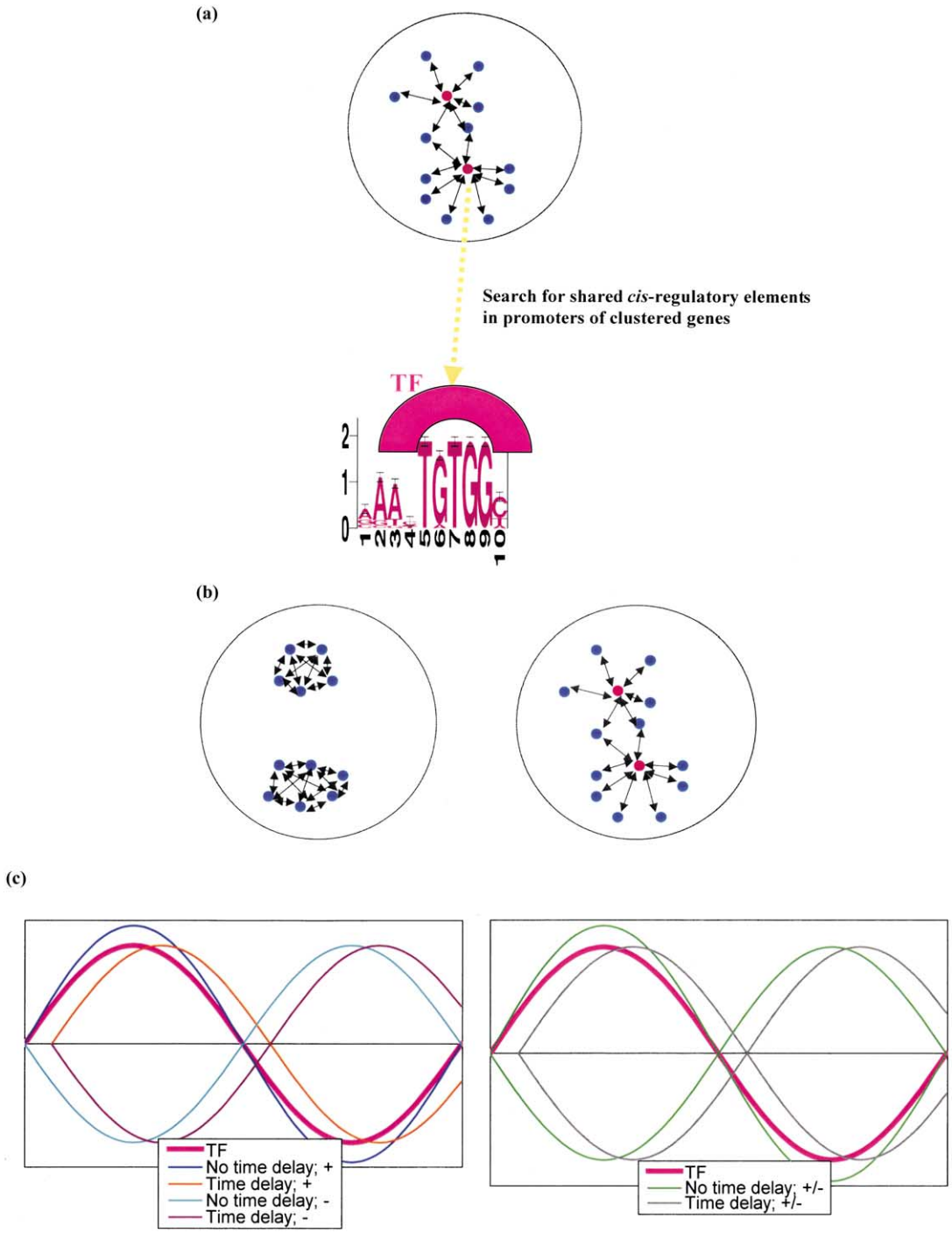

Figure 1. (a) Strategy for identifying TF binding sites *via* TFCC. Genes (blue) that share an expression pattern analogous to that of a particular TF (red) were grouped together. Then a motif finding algorithm was applied to search for shared *cis*-regulatory elements in the promoters of clustered genes. Our working hypothesis was that if a TF-centric cluster indeed contains many targets of the seeding TF, at least one of the discovered motifs would be the site bound by the TF that was used to seed the cluster. (b) The customary (left) *versus* TFCC (right) clustering schemes. In the customary clustering scheme, genes (blue) are clustered based on their correlation to each other. In the TFCC clustering scheme, genes (blue) are clustered based on their correlation to a TF (red), which serves as the center/seed of the cluster. (c) Types of TF-centric clusters. Time delay indicates ORF expression is delayed from TF expression during clustering by one data point. (+) Positive correlation only; (−) negative correlation only; (+/−) positive and negative correlation.

Benchmark testing of TFCC

Next we tested TFCC's capability of discovering transcription factor binding sites by generating clusters with these eight TFs as seeds. For every TF, we computed the Pearson correlation coefficients between it and each of the 3000 most variable ORFs according to their variance-normalized

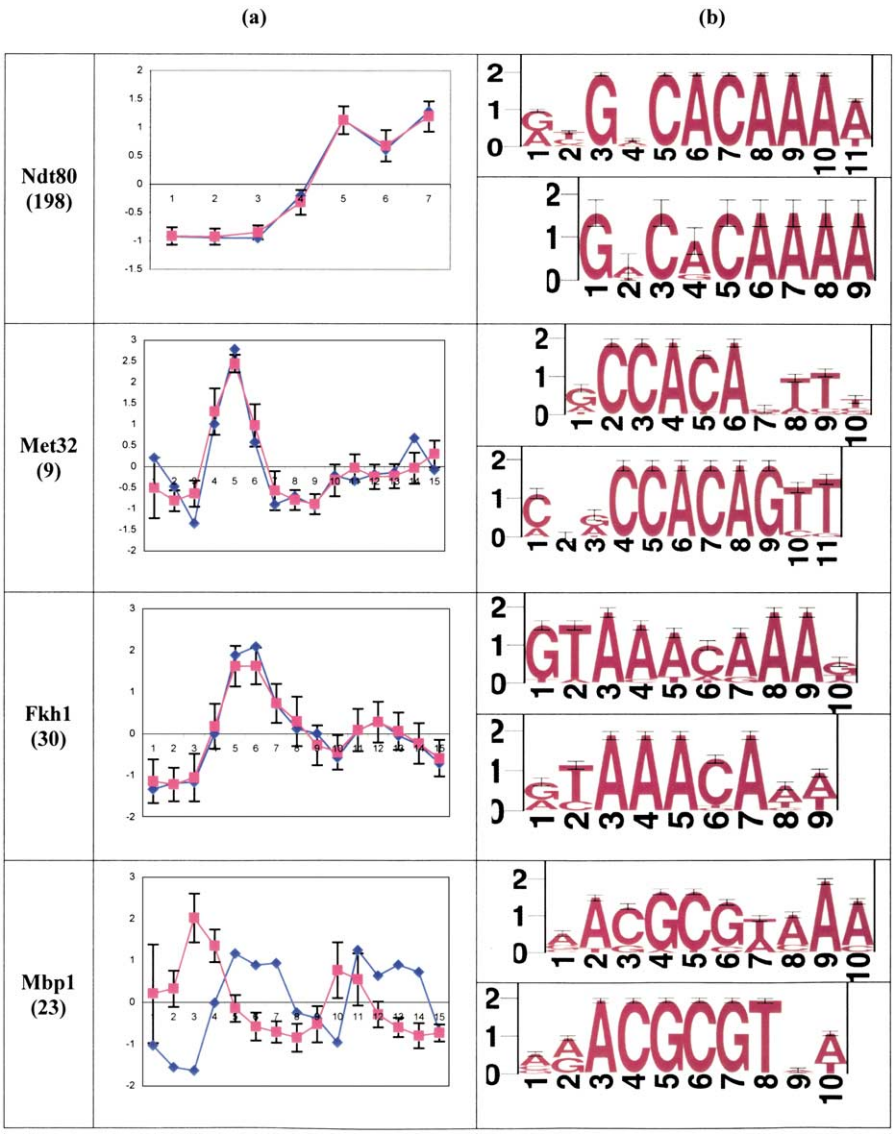

Figure 2. Clusters that generate motifs corresponding to the correct binding site of the seeding TF (CompareACE score ≥ 0.7). Two motifs are usually considered similar with a CompareACE score of 0.7 and above.[5,17,25] (a) The temporal profile of a cluster, named according to the seeding TF, with the size of the cluster in parenthesis. It is represented by a plot of variance-normalized expression pattern of the seeding TF (blue) and the average of all the genes within the cluster (red) (x-axis: Normalised expression level; y-axis: Time point.). (b) Sequence logo[27] representation of the motif discovered from the cluster (top) that is most similar to the known binding site of the seeding TF (bottom). CompareACE scores are 0.98, 0.96, 0.92 and 0.86 for Ndt80, Met32, Fkh1 and Mbp1, respectively. The overall height of the stack at each position signifies the information content of the sequence at that position (0–2 bits). The size of each base is determined by multiplying the frequency of that base by the total information content at that position. The bases are sorted with the most frequent one on top.

expression profiles. Genes that are closely correlated with the TF were grouped together (Figure 1(b)). In order to allow for various regulatory modes, multiple types of clusters were generated for each TF (Figure 1(c)). The effect of TFs on the transcription of the genes they regulate may not be immediate. For instance, it was recently reported that cell cycle transcriptional activators that function during one stage of the cell cycle regulate transcriptional activators that function during the next stage.[13] To account for such scenarios, we introduced a time delay between TF and gene expressions by calculating correlation coefficient at time points i of the TF profile and $i+1$ of the gene profile. Additionally, TFs can be activators, repressors or both. For example, Mig1, a well-studied DNA-binding zinc finger protein involved in glucose repression, is a transcriptional repressor;[15] Rap1, on the other hand, can function as either an activator or repressor of transcription, depending upon the context of its binding site.[16] Therefore, we not only clustered genes whose expression profiles are positively correlated, but also those whose expression profiles are negatively

Table 3. The significant motif with the highest CompareACE score for each TF

TF	Time delay[a]	Correlation type[b]	CompareACE score[c]	MAP score	Group specificity score, $-\log_{10}$	Rank in respective cluster: By MAP score	Rank in respective cluster: By group specificity score
Ndt80	0	+	0.98	66	26	3 (out of 8)	1 (out of 8)
Met32	0	+	0.96	31	10	1 (out of 16)	2 (out of 16)
Fkh1	0	+	0.92	9	11	5 (out of 11)	1 (out of 11)
Mbp1	1	−	0.86	36	12	1 (out of 13)	1 (out of 13)
Swi4	0	+/−	0.58	92	34	1 (out of 5)	1 (out of 5)
Fkh2	0	−	0.34	6	14	7 (out of 16)	4 (out of 16)
Swi5	0	−	0.33	5	10	15 (out of 24)	1 (out of 24)
Mcm1[d]							

[a] Time delay indicates by how many time points ORF expression was delayed from TF expression during clustering.
[b] (+) Positive correlation only; (−) negative correlation only; (+/−) positive and negative correlation.
[c] A CompareACE score of below 0.7 suggests the correct binding site was not found.
[d] No significant motifs were derived for Mcm1.

correlated, with that of the TF. This resulted in six clusters per TF (see Materials and Methods for details). Table 2 lists, for every TF, the number of genes in each cluster type. The temporal profile of a cluster is represented by a plot of variance-normalized expression patterns of the seeding TF and the average of all the genes within the cluster (Figure 2(a)).

Then we conducted a blind and systematic search for the upstream DNA sequence elements shared by members of each cluster. This was done with the program AlignACE,[3,5] which identifies motifs that are over-represented in a set of unaligned input sequences. As a total from all six types of clusters, AlignACE generated an average of 48($\pm$16) motifs per cell cycle TF (23 motifs for Ndt80, only three clusters were produced in consideration of a much longer and uneven time interval used in the sporulation experiment).[12]

Because a large number of motifs were generated for each TF, we selected the most statistically significant ones based on two measures used by AlignACE, namely MAP and group specificity scores[3,5] (see Materials and Methods for details about these scores). Previous studies have shown that most real motifs have a specificity score of $\leq 10^{-10}$ and MAP score of ≥ 5.[5,17] For seven out of eight TFs (except for Mcm1), we found motifs that pass both these thresholds. In this way we reduced the AlignACE output to a list of about four candidate motifs per TF, a reasonable number to test with conventional experimental approaches.

We then compared all the significant motifs discovered for each TF with its known binding site matrix. The motif most similar to the known site for each TF, calculated by CompareACE,[5] is reported in Table 3. Our approach correctly identified the binding sites for four (Mbp1, Fkh1, Met32 and Ndt80) out of the seven TFs for which we obtained significant motifs (Figure 2(b)). We noticed that the motifs corresponding to the correct binding site usually rank quite high in their respective clusters in terms of MAP and/or group specificity scores (Table 3), suggesting the combination of these scores may serve to prioritize candidate motifs prior to experimental testing of *de novo* predictions.

To assess the significance of the above results, we designed three types of negative controls. First, we assigned to seeding TFs random gene sets with sizes (i.e. number of ORFs in a cluster) identical to the real clusters we obtained from TFCC. Since a large number of motifs ($\sim$48.5 for cell cycle TFs, and 23 for Ndt80) were found for each TF, it is conceivable that some of them may pass the MAP and group specificity score cutoffs and match the

Table 4. Identification of TF binding sites by TFCC from clusters that are correlated with the seeding TF by various levels of expression similarity

TF[a]	0.9–0.8[b]	0.8–0.7[b]	0.7–0.6[b]	0.6–0.5[b] and lower[c]
Met32	NF[d]	0.94[e]	NF[d]	NF[d]
Mbp1	0.84[e]	0.96[e]	0.93[e]	NF[d]
Swi4	NF[d]	0.72[e]	0.81[e]	NF[d]
Fkh1	NF[d]	NF[d]	NF[d]	NF[d]

[a] For Fkh2, Swi5 and Mcm1, the three cell cycle TFs with which TFCC failed, no significant motifs corresponding to their known binding sites were found from clusters in any of the above correlation coefficient ranges.
[b] Pearson correlation coefficient range.
[c] The results of lower ranges (i.e. 0.5–0.4, 0.4–0.3, 0.3–0.2, 0.2–0.1) are exactly the same as those obtained for 0.6–0.5.
[d] No significant motif corresponding to the known TF binding site was found.
[e] The CompareACE score of the significant motif that is most similar to the known TF binding site.

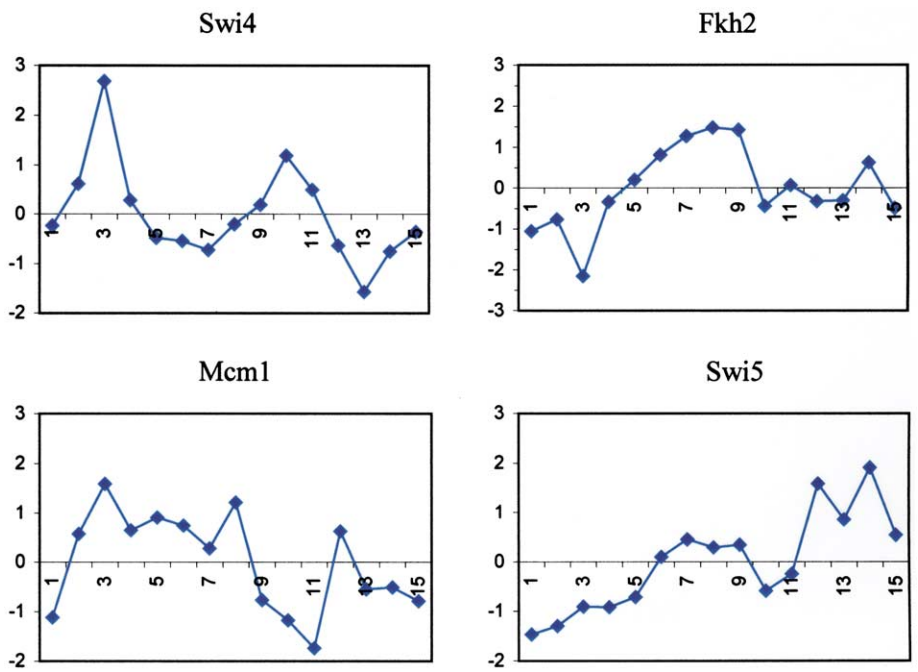

Figure 3. The expression profiles of the four TFs that TFCC fails to derive any significant motifs corresponding to their documented binding sites as for Figure 2. (*x*-axis: Normalised expression level; *y*-axis: Time point.).

known binding site of the seeding TF by chance. The random control produced a comparable number of motifs (~55 per cell cycle TF and 34 for Ndt80) but was incapable of deriving correct binding sites for any of the eight TFs. Thus, our results cannot be simply explained by the large number of motifs produced.

Second, we collected genes that surround each TF at different levels of expression similarity (i.e. 0.9–0.8, 0.8–0.7, 0.7–0.6, ..., 0.3–0.2, 0.2–0.1). Because those in the lower end of the correlation coefficient spectrum are less related to the seeding TFs than those in the higher end, they serve as a negative control for the TFCC approach. The fact that correct binding sites were only derived from genes that are located in the vicinity of a few TFs (>0.6), but not those distant from them (Table 4), demonstrates the necessity of a high enough similarity between the expression profiles of the seeding TF and the remaining genes in the cluster for TFCC to succeed.

Third, we introduced a reverse time shift by calculating the correlation coefficient at time points i of the TF profile and $i - 1$ of the ORF profile. This way genes are still tightly clustered, but probably not related to the seeding TF in a biologically meaningful manner, since the expression profiles of TFs are less likely to be delayed behind their gene targets as the time intervals between measurements increase. Therefore, it enables us to investigate whether clustering alone is sufficient to produce the results obtained by TFCC. Although the exact same motif search and selection procedures were carried out, the reverse control failed to derive correct binding sites for any of the eight TFs, suggesting that the link between a motif and the TF that was used to seed the cluster it came from is not random.

While we correctly identified the binding sites for Fkh1, Mbp1, Met32 and Ndt80, we did not derive any significant motif similar enough to the known binding sites of the other four TFs we tested, namely Swi4, Fkh2, Swi5 and Mcm1 (Figure 3). It may be noted, however, that "correct" motifs were found for Swi4 and Fkh2 (CompareACE scores[5] 0.73 and 0.70, respectively), yet they did not pass the double significance thresholds defined by MAP and group specificity scores.[5,17] The lack of success with Swi5 may be explained in part by a previous report that Swi5 does not have a highly conserved binding sequence.[18] In the case of Mcm1, although we failed to identify its binding site, we managed to avoid making any false prediction as no significant motif was found from Mcm1-seeded clusters. Comparing expression profiles of the TFs for which we succeeded (Figure 2(a)) and of those for which we failed (Figure 3) shows no simple correlation between rate of success and expression-related parameters such as mean expression level and variance.

Our results confirm that the success of the TFCC approach is dependent on whether genes are clustered near their regulating TFs in expression space. Among the four TFs (Mbp1, Fkh1, Met32 and Ndt80) that TFCC managed to derive correct binding sites for, the vicinities of three of them (Fkh1, Met32 and Ndt80) are significantly enriched with the genes they regulate (Table 1). The fact that the motif for Mbp1 was only discovered with a time shift of ten minutes and negative correlation may be explained by the observation that many of its experimentally defined targets do not cluster

Table 5. Predictions made by TFCC for TFs with unknown binding sites

TF	Prediction of binding site	Time delay[a]	Correlation type[b]	MAP	Group specificity score, $-\log_{10}$
Zds1		0	−	6	11
Rsc3		0	+/−	6	11
		0	+	7	12
		1	−	23	9
		1	−	5	10
Swi1		0	−	5	12
		1	−	9	10
		1	−	6	10
		1	−	5	10
		1	+/−	6	11
Ndd1		0	+/−	10	10
		0	+	12	10

[a] Time delay indicates ORF expression was delayed from TF expression by one time point during clustering.
[b] (+) Positive correlation only; (−) negative correlation only; (+/−) positive and negative correlation.

with it directly (Table 1). A close examination of Mbp1 indicates that the expression profiles of the TF and its known targets are significantly anti-correlated with a time shift of ten minutes ($P < 0.001$; Table 1). On the other hand, some of the cases where TFCC failed to derive the correct binding sites, e.g. Swi5 and Mcm1, correspond to TFs that do not cluster with the genes they are known to regulate (Table 1).

De novo prediction of TF binding sites

We also applied the TFCC approach to TFs whose binding sites have not been reported. A few of our *in silico* predictions can be found in Table 5 and the rest are on our supplementary website. For each TF, we list all the significant motifs (i.e. group specificity score of $\leq 10^{-10}$ and MAP score of ≥ 5) derived by TFCC. In addition, we include the type of cluster they came from, which may suggest the mode of regulation (positive and/or negative; with or without time delay). We predict the binding site of the cell cycle regulator Zds1 to be CACGTG, which, interestingly, is palindromic (Table 5), and shows a significant preference towards a particular location in the promoter, 173 base-pairs from the transcription start site (P-value $= 3.2 \times 10^{-5}$ as measured by the positional bias score).[5]

Discussion

Current expression-based clustering methods can be effectively used to discover *cis*-regulatory elements. But they do not provide an immediate link between the computationally identified motifs and the TFs that bind them. An interesting study published recently attempted to establish such a link by decomposing *S. cerevisiae* promoter regions into 7-mers, followed by correlating the "composite" expression pattern of all genes containing each 7-mer with TF expression patterns.[11] Our TFCC algorithm may be viewed as a complementary method to this *cis*/TF approach for identifying TF binding sites *in silico*. From a methodological point of view, the two approaches are the converse of each other: TFCC begins with TF expression profiles and proceeds to identify motifs, while *cis*/TF begins with motif expression profiles and proceeds to identify TFs.

TFCC differs from customary expression-based clustering methods in several aspects (Figure 1(b)). First, it is a semi-supervised clustering method in that it utilizes the biological knowledge about which genes are TFs. Second, since TFs are used as seeds for clustering, it allows the grouping of genes that are less correlated with each other, provided they are sufficiently correlated with the same TF. Lastly, it permits genes to belong to more than one cluster, an important property for clustering studies aimed at exploring multifactorial gene controls.[19] The proposed method may not only predict the motif(s) bound by a TF, but also has the potential to imply which genes are under its regulation and the mode of such regulation (positive and/or negative; with or without time delay).

It should be noted that if several TFs share very similar expression profiles, they will seed similar cluster(s) and thus lead to similar motif(s). TFCC is limited in its capability to differentiate which of these TFs actually bind the motif(s). In such cases, however, TFCC may suggest an indirect relationship between these TFs and the motif(s) as co-expressed TFs are likely to be functionally related: a potential scenario is that these TFs form a complex over the motif and only one of them contacts the DNA physically. Of course, when unrelated TFs happen to have similar expression profiles, TFCC will produce wrong predictions for (at least) some of them.

The TFCC approach requires that the mRNA levels of seeding TFs fluctuate beyond "biological noise"[20] and should be applied to such transcripts only. However, intensive repetitive measurements and adequate (gene-specific) statistical models are needed to determine which genes (and in particular, which TFs) vary sufficiently. At least one attempt has been made in this direction.[20] We anticipate that when such data become more readily available, the TFCC approach will be applied more rigorously to TFs fulfilling the variance requirement. In the present study, we focused on a set of TFs that have been implicated in the regulation of either cell cycle[13] or sporulation, because their mRNA levels are likely to fluctuate beyond "biological noise" during the respective processes.

Since TFs are usually lowly expressed and may be controlled post-transcriptionally, the feasibility and reliability of their mRNA abundance measurements from arrays have often been met with skepticism. Methods such as TFCC and *cis*/TF are built upon the expression patterns of TFs estimated by arrays and assume they are related to that of the genes under their control. The encouraging results obtained by TFCC and *cis*/TF, supported also by previous experimental observations,[9,10] suggest that microarray measurements are capable of capturing the variations of at least some TFs at the level of expression, and also that such changes can be correlated with the mRNA fluctuation of the genes they regulate.

Materials and Methods

Expression data

The raw expression data we used for our computational analysis came from Cho *et al.*[18] (cell cycle) and Chu *et al.*[12] (sporulation). There are 15 valid time points in the Cho data set, across two cell cycles (time points 90 and 100 minutes were excluded from our analysis due to the less efficient labeling of their mRNA during the original chip hybridizations);[6] the Chu data set

reports seven successive time points. We obtained both expression data sets from ExpressDB.[21] Based on the normalized dispersion in expression level of each gene across the time points (SD/mean), we chose the most variable 3000 ORFs.[6] Then each of their expression profiles was variance normalized by subtracting the mean across the time points, and dividing by the standard deviation across the time points:[6]

$$Y_{ij} = \frac{X_{ij} - \langle X_i \rangle}{\sqrt{\frac{1}{15}\sum_{j=1}^{15}(X_{ij} - \langle X_i \rangle)^2}}.$$

where X_{ij} represents the expression level of gene i at time point j, Y_{ij} represents the respective normalized value and $\langle X_i \rangle$ represents the mean expression level of gene i across all time points.

Benchmark TFs

As the Cho data came from a study of the mitotic cell cycle, we decided to focus on a recently defined set of TFs that are involved in cell cycle regulation:[13] Mbp1, Swi4, Mcm1, Fkh1, Fkh2 and Swi5 (Swi6, Ndd1 and Ace2, also well-known cell cycle regulators, were not covered in our analysis because either they do not bind DNA directly, or their binding site matrices are not available). Met32 was included as not only methionine biosynthesis is related to cell cycle,[4] but also the site bound by Met32 itself has appeared in a previous study utilizing cell cycle data.[6] To test the TFCC approach on another independent data set, we used Chu's sporulation data[12] and seeded with an established sporulation factor, Ndt80.

Cluster types

We generated six types of gene clusters for each cell cycle TF.

1. Positively correlated ORFs only and no time delay between TF and ORF expressions.
2. Positively correlated ORFs only and ORF expression is delayed from TF expression by one time point (corresponding to ten minutes in the Cho data).[18]
3. Negatively correlated ORFs only and no time delay between TF and ORF expressions.
4. Negatively correlated ORFs only and ORF expression is delayed from TF expression by one time point.
5. Positively and negatively correlated ORFs and no time delay between TF and ORF expressions.
6. Positively and negatively correlated ORFs and ORF expression is delayed from TF expression by one time point.

Only clusters without time delay were produced for Ndt80 because a much longer and uneven time interval was used in the sporulation experiment.[12]

Determining the statistical significance of the enrichment of target genes around TFs in expression space

The gene targets of Fkh1, Fkh2, Mbp1, Swi4, Swi5 and Mcm1 were obtained from a recent study.[13] For the sporulation factor Ndt80, we used the list of genes that are induced at least threefold when the factor is expressed ectopically as its potential targets.[12] As there is very limited literature on the genes regulated by Met32, we used the following two criteria to identify candidate target genes: (1) containing the experimentally verified binding site (5′AAACTGTGG3′)[14] in their promoters; (2) involved in amino acid metabolism as annotated in the Munich Information Center for Protein Sequences (MIPS) database.[22] A total of 11 genes satisfy both criteria. The list of target genes for each TF can be found on our supplementary website.

We calculated the average correlation coefficient between each TF and its target genes, as well as the fraction of target genes whose correlation coefficient with the TF is above 0.7 (for cell cycle data set) or 0.95 (for sporulation data set). A gene set with size identical to the number of known targets was randomly selected from the 3000 most variable ORFs in the genome. Their average correlation coefficient with the TF and the fraction over the same threshold were computed. We repeated the sampling process 1000 times, counting the number of these runs that achieved an average or fraction equal or higher than the values we obtained with the true gene targets of the seeding TF. A P-value of <0.001 indicates the outcome from none of the 1000 random runs reaches or surpasses that from the run using the true TF targets.

Searching for common upstream regulatory motifs and selection of significant ones

We used AlignACE (with default settings) to conduct a search for common DNA-sequence motifs in the upstream regions (800 bp)[23,24] of the ORFs within each TF-centric cluster. AlignACE is based on a Gibbs sampling algorithm and returns a series of motifs that are over-represented in the input set.[3,5] To select the most statistically significant motifs, we utilized MAP and group specificity scores, two parameters calculated by AlignACE and its accessory programs. MAP score measures the degree to which a motif is over-represented relative to the expected random occurrence of such a motif in the sequence under consideration; the group specificity score gauges how well a given motif targets the upstream regions of the genes used to find it relative to the upstream regions of all genes in the genome. We selected the most significant ones on the basis of a combined MAP and group specificity score thresholds of 5 and 10^{-10}, respectively, since it has been shown that most real motifs score higher than these cutoffs.[5,17]

Evaluation of discovered motifs against known TF binding sites

We compared the significant motifs discovered for each TF and its previously published binding site matrix with CompareACE,[5] a program that performs a pairwise comparison between the position-specific weight matrices of two motifs and returns a value between -1.0 and 1.0 for the best possible alignment. The value corresponds to the Pearson correlation coefficient between the base frequencies of the positions in the aligned portion of the motifs. Two motifs are considered similar with a CompareACE score of 0.7 and above.[5,17,25] We obtained the known binding site matrices from a previous study[26] and they are listed on our supplementary website.

Parameters and settings

As many other bioinformatics works, the TFCC procedure requires the determination of multiple settings and threshold values, including the correlation coefficient cutoff for expression profile clustering, the AlignACE running parameters, and motif significance thresholds (MAP and group specificity). While most of the settings chosen here are adopted from previous studies[5,6,17] and are somewhat arbitrary, a detailed "parameter landscape" analysis indicates the choice of threshold values from a variety of potential settings would have relatively little effect on the final results. As shown by sFigure 1 on our supplementary website, TFCC performance appears largely insensitive to a wide range of correlation coefficient cutoffs. This may be explained by AlignACE's capability of finding even slightly over-represented motifs. Note the above results do not contradict with our control no.2 where correct binding sites could not be derived from genes in the lower end of the correlation coefficient spectrum (Table 4). For more details, see our supplementary website sFigure1. The motif significance thresholds were adopted for two reasons. On the one hand, they decrease the number of candidate binding sites dramatically (approximately sevenfold with the group specificity score cutoff and twofold with the MAP score cutoff; sFigure 3). On the other hand, they should have minimal impact on true positives as previous studies have shown that most real motifs have a specificity score of $\leq 10^{-10}$ and a MAP score of ≥ 5.[5,17]

Supplementary website

A more complete list of our *in silico* predictions for unknown TF binding sites, the "parameter landscape" analysis, and the gene targets for each TF can be found on our supplementary website (http://genetics.med.harvard.edu/~zzhu/TFCC.html).

Acknowledgments

We thank Priya Sudarsanam for providing most of the known TF binding site matrices tested in this study. We are grateful to John Aach, Vasudeo Badarinarayana, Martha Bulyk, Barak Cohen, Patrik D'haeseleer, Aimee Dudley, Rob Mitra, Allegra Petti, Jay Shendure, Priya Sudarsanam and Rebecca Wingert for helpful comments and discussions. Z.Z. is a Howard Hughes Medical Institute predoctoral fellow. Y.P. is a postdoctoral scholar of the Fulbright program. This work was supported by DOE, NSF and the Lipper Foundation.

References

1. Carey, S. & Smale, S. (2000). Analysis and modeling of DNA–protein interactions. In *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*, pp. 448–462 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
3. 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 Biotechnol.* **16**, 939–945.
4. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B. *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
5. Hughes, J. D., Estep, P. W., Tavazoie, S. & 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.
6. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285.
7. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. (1998). Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.* **8**, 1202–1215.
8. Wolfsberg, T. G., Gabrielian, A. E., Campbell, M. J., Cho, R. J., Spouge, J. L. & Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.* **9**, 775–792.
9. Chu, S. & Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. *Mol. Cell*, **1**, 685–696.
10. DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
11. Birnbaum, K., Benfey, P. N. & Shasha, D. E. (2001). *cis* Element/transcription factor analysis (*cis*/TF): a method for discovering transcription factor/*cis* element relationships. *Genome Res.* 1583–1590.
12. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
13. Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L. *et al.* (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.
14. Blaiseau, P. L., Isnard, A. D., Surdin-Kerjan, Y. & Thomas, D. (1997). Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell Biol.* **17**, 3640–3648.
15. Carlson, M. (1999). Glucose repression in yeast. *Curr. Opin. Microbiol.* **2**, 202–207.
16. Shore, D. (1994). RAP1: a protean regulator in yeast. *Trends Genet.* **10**, 408–412.
17. McGuire, A., Hughes, J. D. & Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10**, 744–757.
18. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L. *et al.* (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
19. Bussemaker, H. J., Li, H. & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genet.* **27**, 167–171.
20. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D. *et al.* (2000). Functional discovery *via* a compendium of expression profiles. *Cell*, **102**, 109–126.

21. Aach, J., Rindone, W. & Church, G. M. (2000). Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**, 431–445.
22. Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A. *et al*. (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37–40.
23. Mellor, J. (1993). Multiple interactions control the expression of yeast genes. In *The Eukaryotic Genome, Organisation and Regulation* (Broda, S. G. O. P. & Sims, P. F. G., eds), pp. 275–320, Cambridge University Press, Cambridge, UK.
24. Zhu, J. & Zhang, M. Q. (1999). SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
25. McGuire, A. & Church, G. M. (2000). Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucl. Acids Res.* **28**, 4523–4530.
26. Pilpel, Y., Sudarsanam, P. & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.* **29**, 153–159.
27. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.

Edited by G. von Heijne

(Received 10 September 2001; received in revised form 29 January 2002; accepted 29 January 2002)