Nucleic Acids Research, 2006, Vol. 34, No. 3 ***917–927***
doi:10.1093/nar/gkj487

Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*

Xueping Yu[1], Jimmy Lin[1], Tomohiro Masuda[1,5], Noriko Esumi[1], Donald J. Zack[1,2,3,4] and Jiang Qian[1,*]

[1]Wilmer Institute, [2]Department of Molecular Biology and Genetics, [3]Department of Neuroscience and [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287 USA and [5]Department of Applied Biochemistry, Faculty of Agriculture, Utsunomiya University, Japan

Received November 21, 2005; Revised January 10, 2006; Accepted January 18, 2006

ABSTRACT

Combinatorial regulation by transcription factor complexes is an important feature of eukaryotic gene regulation. Here, we propose a new method for identification of interactions between transcription factors (TFs) that relies on the relationship of their binding sites, and we test it using *Saccharomyces cerevisiae* as a model system. The algorithm predicts interacting TF pairs based on the co-occurrence of their binding motifs and the distance between the motifs in promoter sequences. This allows investigation of interactions between TFs without known binding motifs or expression data. With this approach, 300 significant interactions involving 77 TFs were identified. These included more than 70% of the known protein–protein interactions. Approximately half of the detected interacting motif pairs showed strong preferences for particular distances and orientations in the promoter sequences. These one dimensional features may reflect constraints on allowable spatial arrangements for protein–protein interactions. Evidence for biological relevance of the observed characteristic distances is provided by the finding that target genes with the same characteristic distances show significantly higher co-expression than those without preferred distances. Furthermore, the observed interactions were dynamic: most of the TF pairs were not constitutively active, but rather showed variable activity depending on the physiological condition of the cells. Interestingly, some TF pairs active in multiple conditions showed preferences for different distances and orientations depending on the condition. Our prediction and characterization of TF interactions may help to understand the transcriptional regulatory networks in eukaryotic systems.

INTRODUCTION

Eukaryotic transcriptional regulation is multifaceted. Transcription factors (TFs), co-activators, chromatin structure, promoter elements and other proteins co-operate in the control of gene expression. With the availability of large datasets derived from high throughput experiments, the understanding of gene regulation is no longer confined to the study of single genes, TFs or regulatory elements. Instead, complex network relationships can now be explored.

Numerous technological advances make possible the understanding of combinatorial transcriptional control. Complete genome sequencing provides the DNA information of promoter regions (1,2); large-scale expression profiling provides the global perspective of gene expression (3,4); yeast two hybrid experiments provide interactions between proteins (5,6); and chromatin immunoprecipitation (ChIP) provides the protein–DNA interactions (7,8). With bioinformatics analysis, we can now integrate the many sources of large-scale biological data to gain more insight into the mechanisms of gene regulation.

Previously, various *in silico* approaches have been used to study combinatorial gene regulation. The main focus has been on examining the relationship between gene expression profiles and transcriptional control. Synergistic relationships between TFs are inferred when their common target genes show highly correlated expression patterns (9–11). Nonlinear models (12) and Bayesian approaches (13) have also been used to identify the relationship between gene expression and interacting motifs. In another approach, Nagamine *et al*.

*To whom correspondence should be addressed. Tel: +1 443 287 3882; Fax: +1 410 502 5382; Email: jiang.qian@jhmi.edu

(14) predicted cooperative TFs by using the information from protein–protein interaction networks, based on the hypothesis that proteins that are close to each other in the interaction networks are more likely to be co-regulated by the same set of TFs.

In this paper, we describe a new approach to understanding interactions between TFs, and we test the method using *Saccharomyces cerevisiae* as a model system. Instead of gene expression profiles, we focus on the sequence motifs in the upstream promoters. The strength of this algorithm lies in the fact that it is sequence-based; it can be applied to genes without expression data or previously determined binding motifs. By taking groups of genes whose upstream sequences are known to be bound by two TFs, we made *ab initio* predictions of their corresponding TF binding sites and examined the relationship between these two sites on the promoter sequences. The sequence relationships between the binding motifs were examined in terms of preferences in distance and orientation, reflecting possible spatial relationships between TFs. We further analyzed these predicted relationships using gene expression data and found that they are dynamic and condition-dependent.

MATERIALS AND METHODS

Identification of interacting motif pairs

Motif-PIE, a C++ program, was developed to identify interacting TF binding motif pairs. Interacting motif pairs are defined as those that have over-represented co-occurrence in the input promoters and the distances (in units of base pairs) between the two motifs are significantly different from random expectation. For a set of genes believed to be regulated by two TFs, Motif-PIE was designed to detect the interacting motif pairs that reside in the promoters of these genes. The promoter is operationally defined as the upstream sequence relative to the predicted translational start codon.

The program first calculates the most over-represented single motifs (5 to 7mers) in the input promoter sequences. It then enumerates all possible pair combinations between the top n motifs (e.g. $n = 10$). These motif pairs are ranked according to their P-values. If the most significant motif pair has a lower P-value than a threshold (see below), we predict that their binding TFs interact with each other. The program was performed on all TFs whose target genes have been determined by various experimental works (see Results). Although motif discovery methods via enumeration cannot detect motifs with sequence variation, it is a widely used approach because of its efficient and easy implementation (15,16). Our preliminary analysis indicates that for most motifs a short core motif (5 to 7mers) is often conserved and thus can be detected by this method. Note that the main purpose of Motif-PIE is to detect interacting motif pairs, not to precisely predict motif sequences.

To evaluate the significance of motif pairs, we calculated the P-value for each possible motif pair. The P-value of a motif pair reflects two contributions—one is from motif pair co-occurrence and the other is from the distance constraint. The overall P-value is defined as:

$$P = P_{\mathrm{occ}} P_{\mathrm{d}}$$

where P_{occ} evaluates the over-representation of a motif pair occurrence (g) in the input promoters compared to its occurrence (G) in all promoters in yeast genome, and P_{d} evaluates the deviation of the observed distance distribution from a random expectation.

The contribution of the occurrence over-representation, P_{occ}, is calculated according to

$$P_{\mathrm{occ}} = \sum_{k=g}^{\min(n,G)} \frac{\binom{n}{k}\binom{N-n}{G-k}}{\binom{N}{G}},$$

where n is the number of the input promoters; N is the total number of yeast promoters; g is the occurrence of the motif pair in the input promoters; and G is the overall occurrence of the motif pair in promoters of entire yeast genome. The equation is used to obtain the chance probability of observing the motif pair g or more times in the n input promoter, given that the motif pair occur G times in total N promoters.

The contribution of the distance constraint between two motifs in promoter sequences, P_d, is calculated by comparing the observed distance distribution with a background distribution using the Kolmogorov–Smirnov (KS) test. The background distance distribution is considered to be from motif pairs that do not interact with each other. In other words, if we simultaneously throw two random motifs on promoters and measure the distances between them (in unit of bp), the obtained distance distribution is our background. Given the length of one promoter sequence (L) and motif pair distance (d), the number of all possible arrangements for the motif pair is $L - w_{\mathrm{f}} - w_{\mathrm{b}} - d + 1$, where w_{f} and w_{b} are the widths of the two motifs. The chance of observing distance d is proportional to the number of arrangements for a given d, and can be normalized as

$$f_d(L) = \frac{L - w_{\mathrm{f}} - w_{\mathrm{b}} - d + 1}{\sum_{i=1}^{L-w_{\mathrm{f}}-w_{\mathrm{b}}+1} (L - w_{\mathrm{f}} - w_{\mathrm{b}} - i + 1)}.$$

Given the length distribution $F(L)$ of promoter sequences in yeast, the random distribution of the motif distances is $f_{\mathrm{d}} = \sum_L F(L) f_{\mathrm{d}}(L)$. P_{d} is calculated by comparing the observed distance distribution and f_{d}.

Threshold for significant motif pairs

To determine if a motif pair is significant enough (i.e. the two motifs interact with each other), we derived a threshold from a background simulation. We constructed 8000 background groups of randomly selected promoter sequences from yeast genome in which we expect no meaningful motif pairs. The sizes of random groups were set to be the same as input groups. The same procedure was applied to the random gene groups and the resultant distribution for the most significant P-values from each random group was obtained (Supplementary Figure 1). The P-value at the 95th percentile of the random distribution is defined as the threshold for a significant motif pair. The final thresholds for the heterotypic and homotypic TF pair groups were defined as $-\log(P) = 8.0$ and 3.6, respectively. The Motif-PIE program is available from the authors upon request.

Comparison between known motif and detected motif sequences

We compared the detected motif and known motif sequences. The sequence similarity between two motif pairs, $A1{:}A2$ and $B1{:}B2$, is defined as

$$S(A1{:}A2, B1{:}B2) = \max\left[\sqrt{S(A1,B1)S(A2,B2)}, \sqrt{S(A1,B2)S(A1,B1)}\right].$$

where $S(A1,B1)$ is defined as the matching percentage of two single motifs. We then compared the known motif pair of a TF pair with the top-10 motif pairs derived by Motif-PIE, and we used the maximal similarity to measure the ability of Motif-PIE to recover known motif pairs.

Note that if the binding motif sequences of two input TFs are known, we can compare the predicted and known sequences as described above. However, in cases for which the two input TFs have no documented binding sequences, we cannot compare predicted and known sequences. In such cases, we are unable to link the predicted motif sequences to the input TFs.

Characteristic distance between two motifs

The distances between two motifs are not always uniformly distributed. If two motifs show a strong preference for particular distances, we defined these distances as characteristic distances. We obtained the characteristic distance by calculating the probability that a random system has a frequency at this distance more than or equal to the observed one. This probability can be expressed as

$$P_c(q \geqslant q_d) = \sum_{q \geqslant q_d} \binom{Q}{q} \quad f_d^q (1 - f_d)^{Q-q},$$

where q_d is the observed frequency at this distance; Q is the total frequency at all distances (i.e. the size of the gene group targeted by the motif pair); p is the expected probability at this distance as discussed above. For an extremely small P_d, the probability that a random system has a frequency at any distance more than or equal to q_d is approximately $P_d N_d$, where N_d is the number of possible distances. We set this probability to be 0.01, and with Bonferroni correction the corresponding P_d is $0.01/N_d$. The average N_d is 850, so the threshold for $-\log(P_d)$ is 4.93.

Effect of TF pairs on the expression of their target genes

For a TF pair, we obtained its target genes by searching the motif pair in promoter sequences and calculated expression correlations of all pairs of target genes. To get the background gene expression correlation, we calculated the correlations of any gene pairs in the entire yeast genome. By comparing the correlation distribution of the target genes with the background using the KS test, we obtain a P-value indicating the deviation of the observed correlation from the random expectation. We assumed that the deviations are due to the effects of the TF pairs. Thus, the degree of deviation corresponds to the effect of the TF pair on the expression of its target genes.

RESULTS

Detecting interacting motif pairs

The purpose of our approach is to predict interacting TF pairs based on the co-occurrence of their binding motif pairs in a set of promoter sequences. Given a pair of TFs, we searched significant motif pairs in the promoters of their common target genes. If the P-value of the detected motif pair is lower than a threshold, we inferred that the two TFs interact with each other (Figure 1).

We collected an initial set of known target genes of 152 TFs by integrating currently available chromatin immunoprecipitation on microarray (ChIP-chip) experiments and traditional genetic and biochemical results (7,8,17–19). For each of the pair combinations of the TFs ($152 * 151/2 = 11\,476$), Motif-PIE discovered their binding motifs in the promoters of their common target genes (Figure 1). We then compared the most significant P-values from each TF pair with a predetermined threshold (see Materials and Methods). From the

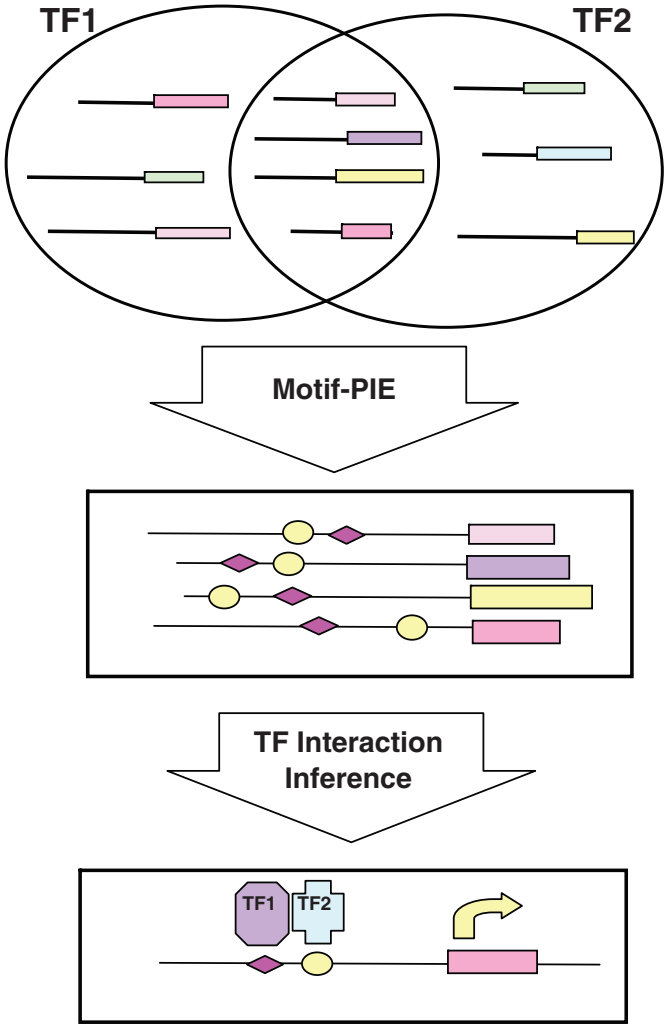

Figure 1. Prediction of interacting motif pairs. The figure shows a schematic description of the algorithm used for identifying heterotypic TF interactions. Based on known TF target genes, we searched significant motif pairs in the promoters of the common target genes of two TFs. We inferred TF interaction if we found a significant motif pair.

Table 1. Representative results of predicted motif pairs

Input TF pair	Known binding motifs	Predicted top motif pair	Known TF–TF interaction	$-\log(p)$
Mbp1:Swi4	ACGCGTnA:TTTTCGCG	ACGCG:CGCGa	1	25.6
Fkh1:Fkh2	TTGTTTACST:TTGTTTACST	TTGTTTA:TGTTTA	1	17.0
Msn4:Swi5	CCCCT:KGCTGR	CCCTg:caGCcGc		16.7
Fkh2:Swi6	TTGTTTACST:ACGCGT	TGTTTAC:ACGCG		16.2
Fkh2:Mbp1	TTGTTTACST:ACGCGTnA	TTGTTTA:gACGCG		16.1
Ste12:Swi6	RTGAAACA:ACGCGT	cAcGAAA:ACGCG		14.2
Pdr1:Rap1	CCGCGG:ACACCCATACATTT	GCGG:ACCCA		14.0
Cup9:Yap6	N.A.:N.A.	GGCAC:TGCAGG		13.7
Rlm1:Swi6	CTAWWWWTAG:ACGCGT	AAACTA:aACGCG		13.6
Fkh1:Swi6	TTGTTTACST:ACGCGT	TTGTTTA:ACGCG		13.5
Mcm1:Swi6	TTWCCCnWWWRGGAAA:ACGCGT	ATCGGGA:ACGCG		12.9
Fkh2:Mcm1	TTGTTTACST:TTWCCCnWWWRGGAAA	TGTTTAC:TAGGA	1	11.3
Ace2:Fkh1	GCTGGT:TTGTTTACST	GCTGGTt:TcTTTAg		11.2
Ino2:Ino4	CATGTGAAAT:CATGTGAAAT	CATGTGA:caCATGC	1	10.4
Msn4:Rap1	CCCCT:ACACCCATACATTT	CACCCA:CCCCCA		9.4
Mcm1:Ste12	TTWCCCnWWWRGGAAA:RTGAAACA	GGAAAtt:TGAAACA	1	9.2

11 476 possible TF pairs, we found that 300 of them have statistically significant interactions, which involve 77 TFs (Supplementary Figure 1).

To examine possible interaction of a single TF with itself, we examined potential 'homotypic TF–TF interactions' with an altered Motif-PIE setting so that only homotypic motif pairs were under consideration (Supplementary Figure 2). With a threshold setting of 3.6 (95th percentile of the random distribution), 45% (69/152) of the TFs were predicted to have homotypic interactions. Note that in some cases, the predicted 'homotypic' interactions may actually represent interaction between two distinct TFs that share the same binding motif.

Table 1 lists some representative predictions. Mbp1 and Swi4 (denoted as Mbp1:Swi4) are known to form a complex to regulate cell cycle process (20). Motif-PIE detected a significant motif pair from the promoters of their known target genes, which corresponds to known binding sites of these two factors. Motif-PIE also predicted many previously unknown TF interactions based on significant motif pairs. In most cases, the predicted motif sequences are similar to known binding motifs. The average similarity between predicted and known motif sequences for the heterotypic TF pairs is 0.763. Also, the average similarity for the homotypic TF pairs is 0.874 (see Materials and Methods for definition of similarity). These high similarities indicate that Motif-PIE performs well in discovering significant motif pairs, and the detected motifs are highly likely to correspond to the input TFs. However, for two factors without known binding motifs (e.g. Cup9:Yap6 in Table 1), we cannot make an association between predicted motif sequences and their binding TFs. The assignment of first and second motifs in the Table (column 3) is arbitrary. For example, the first motif (GGCAC) is not necessary to be the binding motif of Cup9. The main purpose of Motif-PIE is not to predict single TF binding motif. Instead, it is used to detect possible interacting motif pairs and to infer TF interactions based on the motif pairs. The entire list of our predictions can be found in Supplementary Table 1.

Distance constraint for interacting motif pairs

For interacting TF proteins, it is conceivable that they must satisfy certain spatial requirements to have a functional interaction. Consequently, their corresponding binding motifs may demonstrate characteristic distance relationships in promoter sequences. To explore this possibility, we calculated the distance distribution of motif pairs in their target promoters.

Figure 2 shows some typical distance distributions for interacting motif pairs. The analysis revealed that several interacting motif pairs demonstrate a preference for specific separation distances. For example, the distances between motifs for Dig1 and Ste12 (Dig1:Ste12) have a predominant peak at 34 bp, whereas the peak for Gat3:Yap5 is at 49 bp. Compared with the random expectation for the distance distribution (dashed lines, see Materials and Methods), these peaks are highly significant. A similar observation was true for homotypic interactions, such as Mcm1:Mcm1 with a distance peak at 43 bp (Figure 2C).

We defined a threshold for significance, and the peaks above the threshold are called characteristic distances for the motif pair. The threshold was corrected for multiple hypothesis testing and defined as $-\log(p) = 4.93$ (see Materials and Methods for details). According to this threshold, 154 of 300 detected motif pairs have one or more characteristic distances, which is remarkably larger than the expected number from random events ($300 \times 1\% = 3$). It indicates that our detected motif pairs are much more likely to have distance constraints than are random motif pairs.

From the distribution of characteristic distances, we found that 75% of the characteristic distances are smaller than 166 bp, with 25% sporadically distributed in the broad region ranging from 166 to 2536 bp (Figure 2E). The occurrence of short genomic distances between motif pairs is suggestive of possible physical interaction between their respective TFs. The finding that some pairs have large characteristic distances may reflect secondary structure DNA looping or indirect interaction through complex formation.

We further explored the biological relevance of the characteristic distances by looking at the co-expression of their target genes. The degree of co-expression of gene groups targeted by a motif pair with a characteristic distance is significantly higher than that by a motif pair without characteristic distances (Figure 2F). Thus, characteristic distances provide an additional constraint for TF interaction.

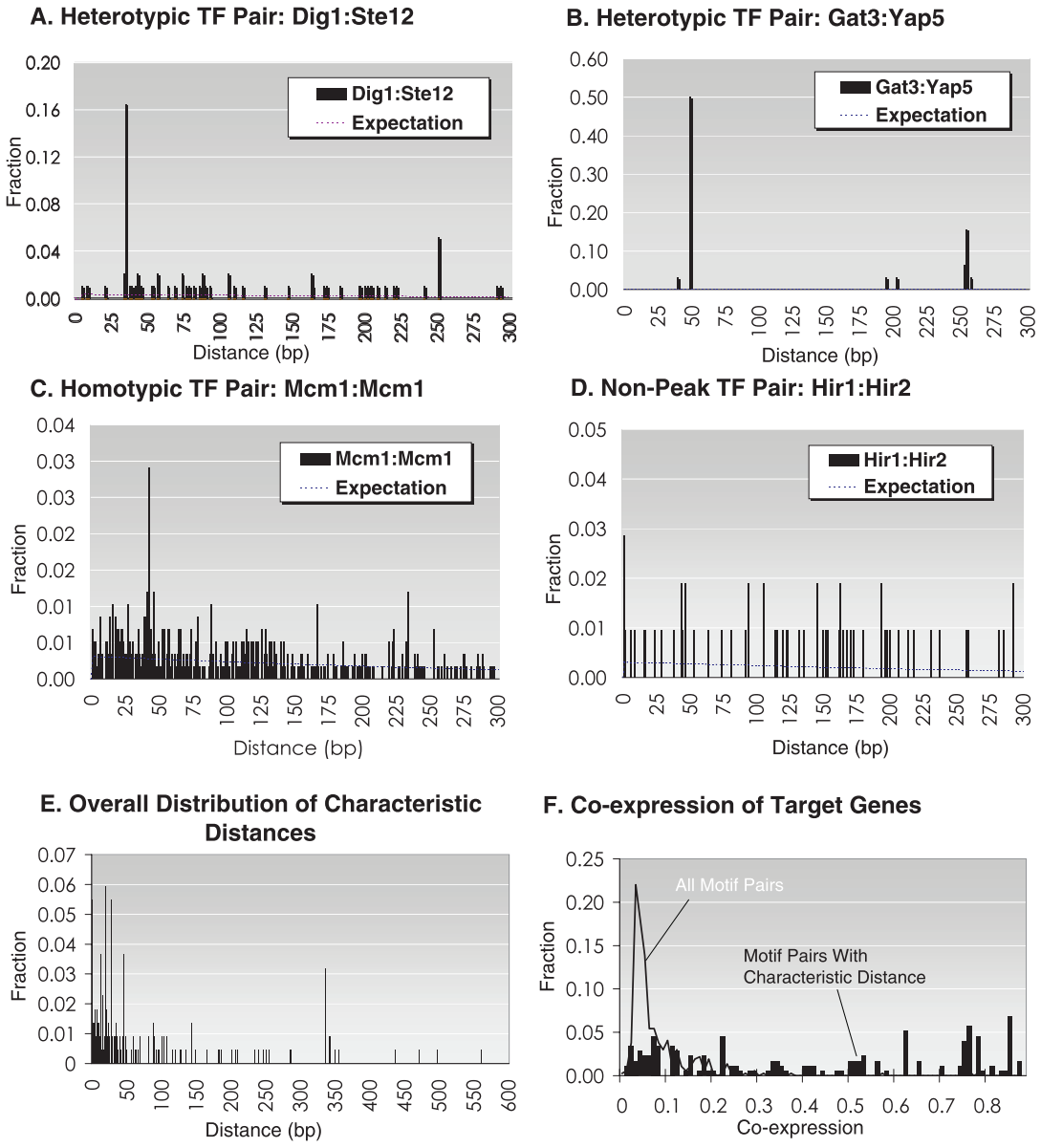

Figure 2. Distance features of interacting motif pairs. The distribution of the distances for each motif pair was plotted to reveal significant characteristic distances. Examples for both heterotypic (**A** and **B**) and homotypic interactions (**C**) are shown. In these three examples, significant characteristic peaks can be seen. (**D**) Shows an example of a motif pair without characteristic peaks. (**E**) The distribution of characteristic distances for all interacting motif pairs. Most of the characteristic distances are smaller than 100 bp. (**F**) The co-expression distribution for target genes of motif pairs. The target genes with characteristic distances are more co-expressed than those without distance peaks.

We have therefore designed Motif-PIE so that motifs pairs with characteristic distances are given a higher significance than those without. However, motif pairs without distance preference are not excluded because some known interacting TF pairs, such as Hir1:Hir2 (Figure 2D), do not show evidence of a characteristic distance.

Orientation constraint for interacting motif pairs

In addition to characteristic distances, we also explored the orientation of TF DNA binding sites among motif pairs. Orientation was defined as the relative directions of a motif pair on the genome sequence. Four possible relative orientations were examined for each motif pair: one divergent (opposite directions away), one convergent (opposite directions towards) and two tandem orientations (Note that they have different interaction interfaces) (Figure 3).

To determine if there is any orientation preference for an interacting motif pair, we defined a quantity P_o as the fraction of the most dominant orientation. In an even distribution of the four orientations, there would be no orientation preference and $P_o = 0.25$ (see random distribution in Figure 3A). A large P_o indicates strong orientation preference (i.e. one orientation is significantly over-represented).

Figure 3A shows the orientation distribution for the pairs with and without characteristic distances. One can see that for those motif pairs without characteristic distances, the distribution is only slightly shifted toward larger values as

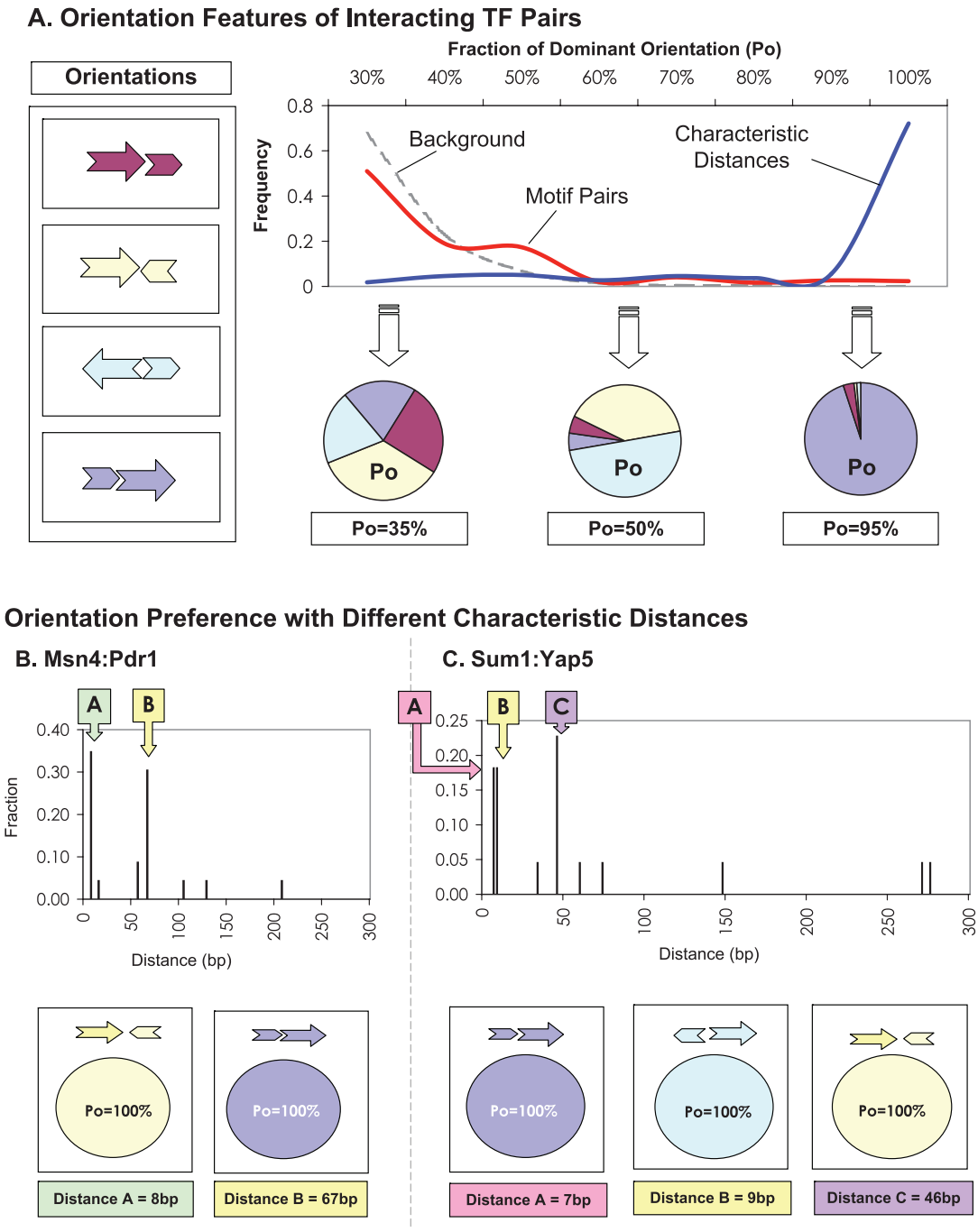

Figure 3. Orientation features of interacting motif pairs. TF pairs can be oriented with respect to each other in four configurations: one towards, one away and two in the same directions, as shown in the left panel of (**A**). To examine the orientation preference of a TF pair, we calculated the fraction of the orientation in all instances that is most over-represented for a given motif pair, which we named P_o. As shown by the schematic pie charts below the *x*-axis, a large P_o (e.g. 95%) means that the dominant orientation represents 95% of the total orientation, whereas a small P_o (e.g. 35%) shows almost an even distribution between the four different distributions. The plot in (**A**) shows the distributions of P_o for all predicted interacting motif pairs, for the subset with characteristic distances, and the background control group. As shown, those with characteristic distances have a strong preference for one type of orientation. Over 70% of the motif pairs with characteristic distances have $P_o = 100\%$, meaning that all instances for a motif pair exhibit the same orientation. In contrast, the distribution for overall motif pairs is almost the same as that for background. In (**B**) and (**C**), two examples of TF pairs are shown with multiple characteristic distances. At specific distances, labeled A, B or C, each of the distribution of the orientations are shown below. Interestingly, for both Msn:Pdr1 and Sum1:Yap5, at the characteristic distances, all the motifs share the same orientation, but at different distances the orientations are different.

compared with a random distribution. If we only consider the target genes with characteristic distances, the orientation preference is dramatically increased. On closer examination, we see that in most cases (>70%), motif pairs adopt only one orientation (i.e. $P_o = 1$). These strong orientation preferences of motif pairs presumably reflect requirements for particular spatial arrangements for appropriate TF interactions.

Characteristic distances are associated with orientations

Further analysis of the 154 motif pairs with characteristic distances revealed that 35.1% have more than one characteristic distance. When we examined the orientation distribution at these distances, we found that different distances often correspond to different preferred orientations. These multiple characteristic distances may reflect various interaction configurations. For example, the Msn4 and Pdr1 pair has two characteristic distances: one at 8 bp and the other at 67 bp (Figure 3B). Interestingly, at a distance of 8 bp, all eight instances share one orientation, whereas at a distance of 67 bp, all seven instances share a different orientation. Figure 3C provides another example: Sum1:Yap5 has three characteristic distances, 7, 9 and 46 bp. For each of the characteristic distances, all the sites with that distance share the same relative orientation, yet the common orientation at each distance is distinct from that at the other distances. These results reveal the potential spatial nature of TF physical interactions, with different spatial alignments and orientations associated with characteristic distances.

Evaluation of prediction by known protein–protein interactions

The genomic features of interacting motif pairs strongly suggest possible TF–TF interactions. Although our predicted 'interactions' between motifs could include both direct physical interactions between the factors themselves and indirect interactions mediated by co-activators or other mechanisms, in order to have a stringent evaluation scheme, we compared our predictions with available data on direct protein–protein interactions. We collected a set of known TF–TF interactions from databases of general protein–protein interactions. We utilized information from the DIP (21), TRANSFAC (18) and MIPS (22) databases, but excluded data from high throughput experiments such as yeast two hybrid (Y2H) screen. One reason for excluding the datasets from high throughput experiments is a concern about the quality of the data. Another reason is that Y2H may not be suitable for detecting interactions between TFs because of the problem of auto-activation (23).

The database survey identified 20 heterotypic and 21 homotypic interactions between the 152 TFs. Motif-PIE correctly predicted 70% (14/20) of the known heterotypic interactions and 81% (17/21) of the known homotypic interactions (Table 1). One possible explanation for at least some of the false negative predictions is that one of the interacting TFs may not directly bind to the promoter. One such example is Hap3:Hap4, where Hap4 does not directly bind to DNA but still has a physical interaction with Hap3 (24).

Motif-PIE also predicted a number of interactions between TFs that were not present in the available databases. Interestingly, some of these interactions have been reported in the literature. Examples include Fhl1:Rap1, which was recently suggested experimentally (25,26) and Yap5:Yap5, which was confirmed by two independent high throughput experiments (27,28).

Dynamic effect of TF pair interactions on target genes

Having studied the yeast TF interaction network under static conditions, we next attempted to explore the possible dynamic behavior of interacting TF pairs. We examined the effects of TF pairs on the expression of their downstream target genes under various physiological conditions. For each of the TF pairs, we performed a whole-genomic search in the upstream promoter regions of the genes, and the target genes were defined as those genes whose promoter sequences contain the corresponding motif pair.

Researchers have previously used co-expression of common target genes to predict interacting TF pairs (9–11). Similarly, we assume that if we observe significant co-expression of their target genes under one condition, we can infer that the TF pair is active under this condition; otherwise, the TF pair is likely to be inactive. This definition of 'active' is operational and based upon the overall expression behavior of the downstream targets of the TF pair under consideration. It does not necessarily mean that none of target genes are activated by the TF pair, nor does it exclude the possibility that the TFs could be physically interacting despite not leading to co-expression. The effect of TF pairs on target gene expression was calculated by checking the deviation of expression correlation of target genes from a random expectation (see Materials and Methods).

The combined database of expression data that was used consisted of 82 experiments and six conditions (3,4,29). Those conditions were cell cycle, elutriation, heat shock, DNA damage, sporulation and drug treatment. Figure 4A presents the degree of effect of TF pairs on their target genes under the six conditions. From this plot, one can see that some TF pairs show effect under all conditions, but most TF pairs are active only under certain specific conditions. More than 40% of the TF pairs are active only under a single condition. In contrast, 16% of the TF pairs are constitutively active under all six conditions (Figure 4A). This finding suggests that, although a few are constitutively active, most TF pair interactions are transient and affect downstream target gene expression only under certain conditions.

Under different conditions, TFs interact with different partners. In Figure 4B and C, we show two TFs, Ndd1 and Ste12, that have such varying interactions. Figure 4B shows the interaction of Ndd1 with other TFs. Ndd1 cooperates with eight TFs (Swi4, Mcm1, Fkh2, Swi5, Rlm1, Swi6, Ace2 and Mbp1) to affect expression of target genes under all six conditions. However, when Ndd1 interacts with Fhl1 or Smp1, there is no cooperative effect under the condition of sporulation. More strikingly, TF pairs Ndd1:Ume6 and Ndd1:Gcr2 are active only in sporulation. Even though Ndd1 alone is described as a TF involved in the mitotic cell cycle (30), when we study it in the context of TF interactions, it demonstrates different activity patterns across the various conditions. The function of a TF and its activity may be better described in relation to its interacting partners than to one factor alone.

Figure 4C shows the different interactions of Ste12, which is known to be activated by a MAP kinase cascade and to activate genes involved in mating or invasive growth pathways (31,32). Consistent with its being activated only under specific biological circumstances, the interactions between Ste12 and its partners do not show constitutive function, and each pair is activated under a small number of experimental conditions. The interactions of Ste12 with its different partners show different patterns of activity.

We then investigated why some TF pairs are constitutively active whereas the others are only active under certain

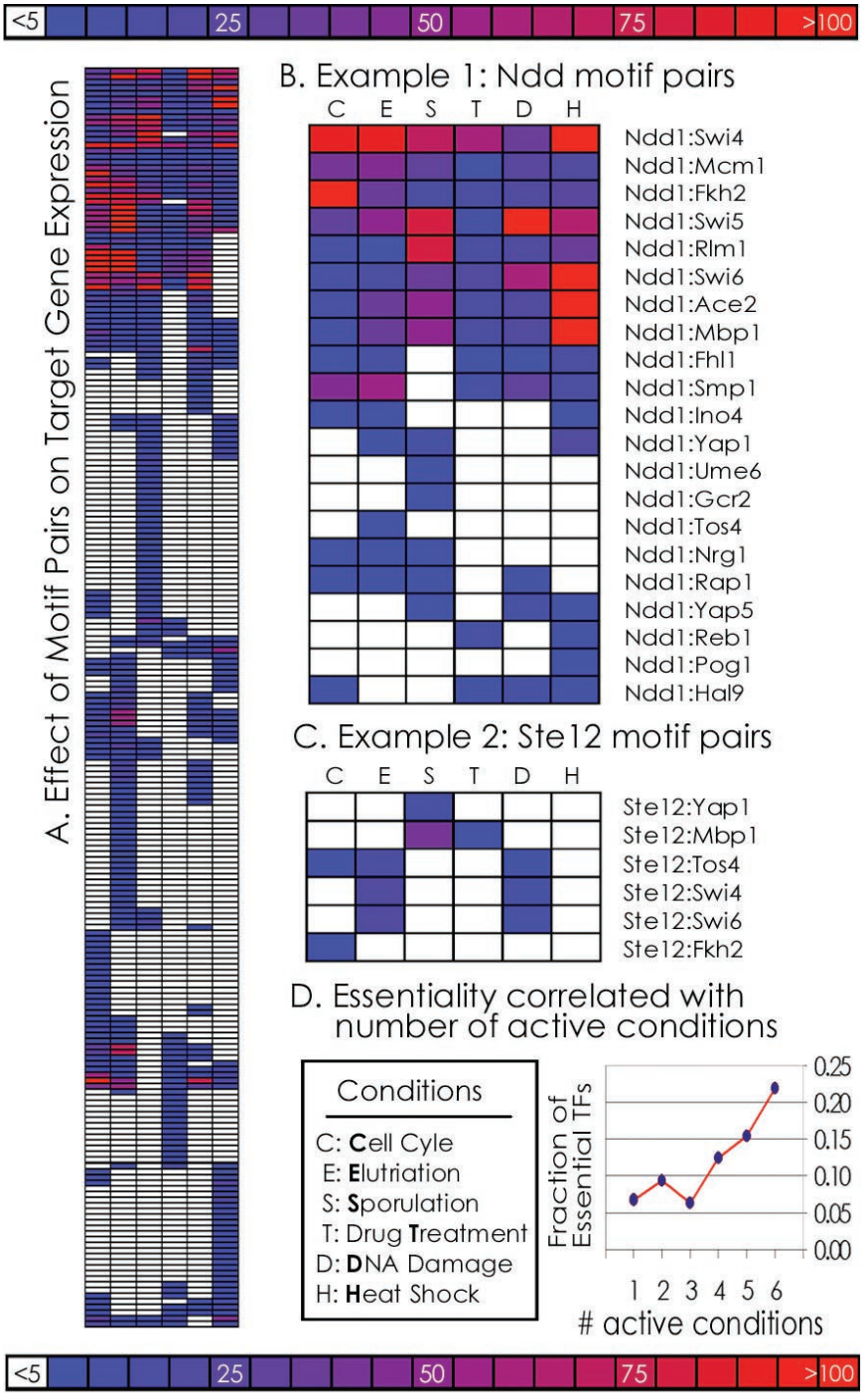

Figure 4. Dynamic effect of TF pair interactions on target genes. The effect of different TF pairs can be seen by measuring the downstream expression of target genes. (**A**) provides a summary view of the different effects of the TF pairs. Each row represents a TF pair; and each column is one physiological condition. The cells show the degree of co-expression of the target genes of TF pairs under various conditions. The values shown in the color bar are $-\log(p)$. Not only is there overall variation, but also individual TFs may interact with different TF partners under different conditions. (**B** and **C**) show two expanded views of (A). They are two examples of TFs interacting with a number of possible partners, depicting expression correlations under six sets of experimental conditions (noted in the box at the bottom of the figure). (**D**) shows how the essentiality of a TF factor is related to the number of conditions under which a TF is active. TFs active in higher numbers of conditions are more likely to be essential.

conditions. We linked the dynamics of TFs to gene essentiality. The essential genes are those that render cells non-viable if they are knocked out. The gene essentiality in yeast has been examined on the genomic scale (33,34). Bioinformatics work has shown that the gene essentiality can be related to many of the topological characteristics (e.g. hubs) of protein–protein interaction networks (35). In terms of transcriptional regulatory networks, TFs with many targets are more likely to be essential than are other proteins (35). Beyond their relationships with the static regulatory networks, in this

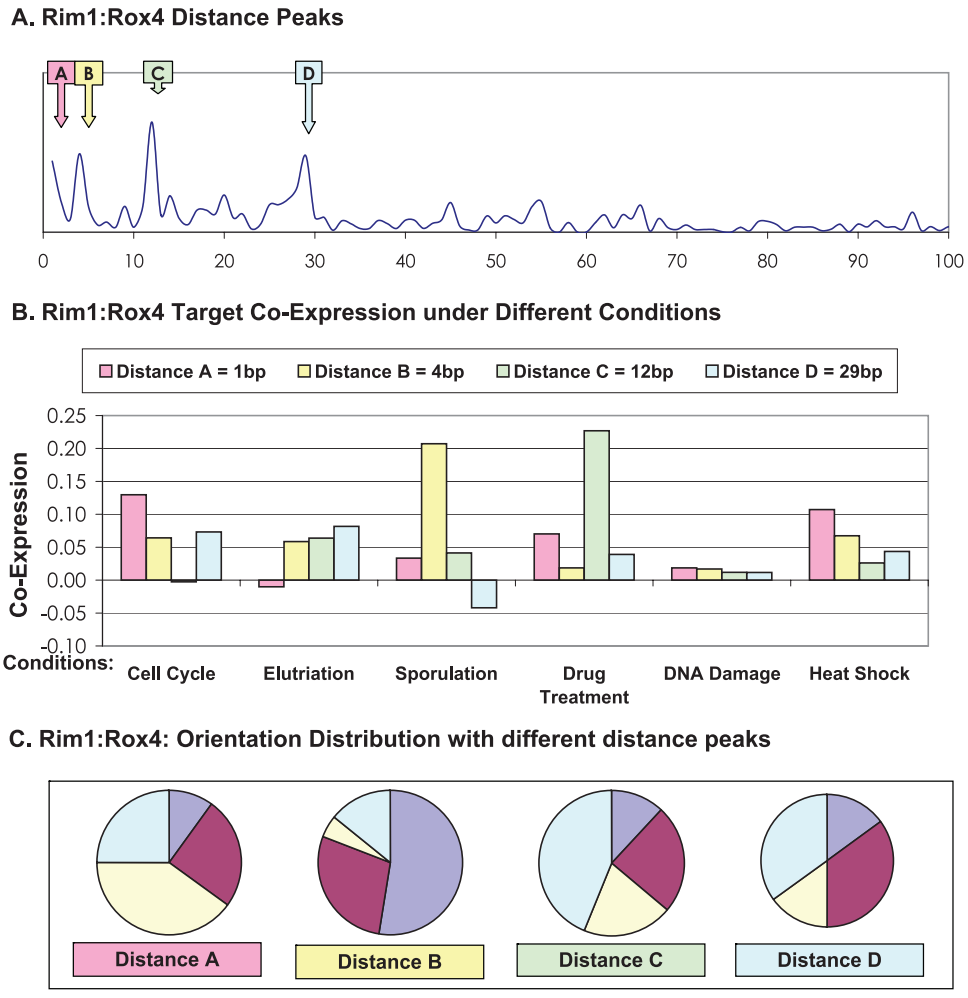

Figure 5. Integration of the spatial relationship and dynamic nature of the TF interactions. Under different conditions, there is a different preference for different distances. These different distances are shown clearly in (**A**). (**B**) shows co-expression of target genes with different characteristic distances under different conditions. With each of the distances, the associated orientation distribution is shown in (**C**). This example shows how under different conditions, there is a clear association with particular distances and orientations.

work we found that TF pairs involved in more conditions were more likely to include essential genes. The fraction of essential TFs shows an almost monotonic increase with number of active conditions. In other words, the TFs active under more conditions are more likely to be essential than those that are only active under specific conditions (Figure 4D).

Knowing the dynamic nature of these interactions as well as the relationship between characteristic distances and particular orientations, we explored the possible relationship between all three features. We found that different conditions preferred particular orientations as well as particular distances. For example, in Figure 5, the motif pair Rlm1:Rox1 is examined under all six conditions for all four characteristic distances and all four orientations. Interestingly, all six conditions show unique patterns of expression correlations for the subsets of target genes. Under the sporulation condition the targets with a characteristic distance of 4 bp show strong gene expression correlation, whereas under the drug treatment condition the expression correlation is highest for the genes with a 12 bp separation (Figure 5B). Furthermore, for these two characteristic distances, there were distinct orientation preferences as well (Figure 5C). This example suggests the dynamic nature and complexity of DNA binding and protein interaction as well as potential different spatial and physical arrangements of the TFs.

DISCUSSION

In this paper, we presented a novel sequence-based algorithm to identify transcription factor interactions. Using 1D genomic features, such as motif sequence and distance, this algorithm can be used more generally and in instances where gene expression experiments are not available. We hope that this new perspective will add to the existing methods of transcription factor analysis.

With Motif-PIE, the software suite we developed to implement our sequence-based algorithm, we were able not only to recover many known *S.cerevisiae* TF interactions, but also to make predictions about novel TF interactions and their targets. Using data from experimentally verified ChIP-chip experiments, we searched for statistically significant motif pairs over various combinations of single motifs—not an

insignificant computational problem. Based on stringent statistical criteria, we predicted 369 significant interactions, including both homotypic and heterotypic interactions.

More detailed investigation revealed many interesting properties of the identified TF pair interactions. The analysis demonstrated that, in general, short distances between binding sites are preferred over longer distances. Of potentially greater significance, we found that there are specific and characteristic distances that are preferred for many TF motifs, and that these characteristic distances are TF-dependent. Moreover, these characteristic distances show strong preferences for particular orientations of the TF motifs, possibly reflecting physical constraints on the 3D interactions between TFs. Thus, TFs do not seem to interact arbitrarily at any distance, but to have specific preferences. Such information may be useful in aiding ongoing efforts to model the physical and spatial interactions between TFs.

We also explored the dynamics of TF interactions, i.e. how they vary according to a cell's physiological state. It is of course well known that as an organism goes through different stages of growth or is presented with various environmental conditions, TF activity is modulated to modify gene expression patterns (36,37). The dynamic nature of TF interactions was analyzed for six different conditions involving 82 microarray experiments. Not only did this analysis demonstrate that there are subsets of computationally determined TF interactions that show distinct activity patterns across multiple states, it also indicated that those TFs that are constitutively active are more likely to be encoded by essential genes.

Finally, we found that even the same TF pair may behave differently under different conditions, with different characteristic distances and different orientations. This finding suggests that interactions between the same two TFs can have distinct transcriptional effects depending on the relative orientation and distance between their respective binding sites. It should be noted that the occurrence of such spatially variable interactions is not rare—~30% of the interactions we identified fell into this class. The molecular basis for such spatially dependent activities is unclear, but potentially could be dependent upon interaction with distinct sets of co-factors.

Based on the results discussed above, we feel that Motif-PIE can be a useful addition to the armamentarium of methods currently available to study TF interactions. However, it certainly has a number of limitations, as well as ample room for further improvement and development. For instance, in this study we did not consider two overlapping binding sites; therefore, we would miss those cases in which two TFs compete for the same or overlapping binding sites. Also, we are aware that short inter-motif distances can be attributed to dimeric TFs, such as basic leucine zippers. Consequently, we might discover two segments of one TF binding sequence instead of two independent binding sites. We need to improve our algorithm to distinguish these two situations. One possible solution is to expand current vocabulary of motifs (5 to 7mers) by including motifs with spacer segments between conserved sites. An additional issues is that it is often not a trivial problem to associate a detected motif sequence with its respective binding TF, both because (i) in some situations the identified motif does not resemble a known TF binding site and (ii) even when the identified motif does resemble a known binding site, multiple TFs can bind to the same or similar DNA sequences. These issues are difficult to solve with a purely computational approach, and would likely be best addressed by combining programs such as Motif-PIE with laboratory-based experimental studies.

Having demonstrated that additional features of TF interactions, such as the distance between binding sites, orientation of sites and condition/time of activity, can be important, we can use these features for future predictions. We can potentially define 'super-motifs' that include not only the sequence of the binding sites, but also multiple sites with set distances between them with a particular orientation. This capability can greatly increase prediction specificity for genomic scans. In addition, this system can be applied to organisms, such as the human, and it can potentially help elucidate the mechanisms of gene regulation in diseases, such as cancer. With these additional features, we can better understand the gene regulatory networks in eukaryotic systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Drs Heng Zhu, Joel Bader and Shannath Merbs (Johns Hopkins University) for stimulating discussions. The authors also thank the two anonymous reviewers for their helpful comments. The research was supported in part by grants from the National Institute of Health (EY015684 to J.Q., EY009769 and EY001765 to D.J.Z.), unrestricted funds from Research to Prevent Blindness, Inc., and by generous gifts from the Guerrieri family and from Robert and Clarice Smith. T.M. is supported by JSPS Research Fellowships for Young Scientists. D.J.Z. is the Guerrieri Professor of Genetic Engineering and Molecular Ophthalmology.

Conflict of interest statement. None declared.

REFERENCES

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.
2. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
3. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
4. Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
5. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
6. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
7. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I.

et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
8. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
9. Yu,H., Luscombe,N.M., Qian,J. and Gerstein,M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet*., **19**, 422–427.
10. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet*., **29**, 153–159.
11. Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*., **31**, 7024–7031.
12. Das,D., Banerjee,N. and Zhang,M.Q. (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.
13. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
14. Nagamine,N., Kawada,Y. and Sakakibara,Y. (2005) Identifying cooperative transcriptional regulations using protein–protein interactions. *Nucleic Acids Res*., **33**, 4828–4837.
15. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol*., **281**, 827–842.
16. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′-UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
17. Horak,C.E., Luscombe,N.M., Qian,J., Bertone,P., Piccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes. Dev*., **16**, 3017–3033.
18. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
19. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*., **29**, 281–283.
20. Koch,C. and Nasmyth,K. (1994) Cell cycle regulated transcription in yeast. *Curr. Opin. Cell Biol*., **6**, 451–459.
21. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*., **30**, 303–305.
22. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*., **30**, 31–34.
23. Walhout,A.J. and Vidal,M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, **24**, 297–306.
24. Forsburg,S.L. and Guarente,L. (1989) Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes. Dev*., **3**, 1166–1178.
25. Wade,J.T., Hall,D.B. and Struhl,K. (2004) The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature*, **432**, 1054–1058.
26. Schawalder,S.B., Kabani,M., Howald,I., Choudhury,U., Werner,M. and Shore,D. (2004) Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. *Nature*, **432**, 1058–1061.
27. Newman,J.R. and Keating,A.E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science*, **300**, 2097–2101.
28. Marino-Ramirez,L. and Hu,J.C. (2002) Isolation and mapping of self-assembling protein domains encoded by the *Saccharomyces cerevisiae* genome using lambda repressor fusions. *Yeast*, **19**, 641–650.
29. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
30. Loy,C.J., Lydall,D. and Surana,U. (1999) NDD1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in *Saccharomyces cerevisiae*. *Mol. Cell Biol*., **19**, 3312–3327.
31. Madhani,H.D. and Fink,G.R. (1997) Combinatorial control required for the specificity of yeast MAPK signaling. *Science*, **275**, 1314–1317.
32. Roberts,R.L. and Fink,G.R. (1994) Elements of a single MAP kinase cascade in *Saccharomyces cerevisiae* mediate two developmental programs in the same cell type: mating and invasive growth. *Genes Dev*., **8**, 2974–2985.
33. Ross-Macdonald,P., Coelho,P.S., Roemer,T., Agarwal,S., Kumar,A., Jansen,R., Cheung,K.H., Sheehan,A., Symoniatis,D., Umansky,L. *et al*. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.
34. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H. *et al*. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
35. Yu,H., Greenbaum,D., Xin Lu,H., Zhu,X. and Gerstein,M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet*., **20**, 227–231.
36. Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
37. de Lichtenberg,U., Jensen,L.J., Brunak,S. and Bork,P. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.