Genome analysis

Method for identifying transcription factor binding sites in yeast

Huai-Kuang Tsai[1], Grace Tzu-Wei Huang[1], Meng-Yuan Chou[2], Henry Horng-Shing Lu[3] and Wen-Hsiung Li[1,4,*]

[1]Genomics Research Center and [2]Institute of Information Science, Academia Sinica, Taipei, 115 Taiwan, [3]Institute of Statistics, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30050, Taiwan and [4]Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

ABSTRACT

Motivation: Identifying transcription factor binding sites (TFBSs) is helpful for understanding the mechanism of transcriptional regulation. The abundance and the diversity of genomic data provide an excellent opportunity for identifying TFBSs. Developing methods to integrate various types of data has become a major trend in this pursuit.

Results: We develop a TFBS identification method, TFBSfinder, which utilizes several data sources, including DNA sequences, phylogenetic information, microarray data and ChIP-chip data. For a TF, TFBSfinder rigorously selects a set of reliable target genes and a set of non-target genes (as a background set) to find overrepresented and conserved motifs in target genes. A new metric for measuring the degree of conservation at a binding site across species and methods for clustering motifs and for inferring position weight matrices are proposed. For synthetic data and yeast cell cycle TFs, TFBSfinder identifies motifs that are highly similar to known consensuses. Moreover, TFBSfinder outperforms well-known methods.

Availability: http://cg1.iis.sinica.edu.tw/~TFBSfinder/

Contact: whli@uchicago.edu

Supplementary information: Supplementary data are available on *Bioinformatics* online.

1 INTRODUCTION

The transcription of genes is controlled by interaction between transcription factors (TFs) and their binding sites (TFBSs). Identifying and characterizing the binding sites of a TF can provide a better understanding of the function of the TF. Unfortunately, TFBSs are usually short (∼5–15 bp) and degenerate (Stormo, 2000), making it difficult to define TFBSs experimentally or computationally. A traditional computational approach starts with a collection of genes presumed to be bound by the same TF according to their biological functions, expression profiles (Spellman *et al.*, 1998) or protein–DNA binding assays (Lee *et al.*, 2002). The next step involves the identification of overrepresented sequence elements (Bailey and Elkan, 1995; Bannai *et al.*, 2004; Hertz *et al.*, 1990; Pizzi *et al.*, 2005; Roth *et al.*, 1998). Some methods further define a suitable background set as a control (Kato *et al.*, 2004; Liu *et al.*, 2002; Sinha, 2003; Sinha and Tompa, 2003; Zhu *et al.*, 2002). More recently, phylogenetic footprinting methods have been used to test the conservation of

TFBSs across species (Cliften *et al.*, 2003; Elemento and Tavazoie, 2005; Emberly *et al.*, 2003; Kellis *et al.*, 2003; Tanay *et al.*, 2005; Wang and Stormo, 2005). Two issues should be emphasized. First, a set of reliable target genes and a suitable background set for a TF can significantly help find motifs preferentially residing in the target set. Second, a change at a variable nucleotide position in a TFBS may cause only a small effect on the binding affinity, whereas a change at an invariant position may have a strong effect (Moses *et al.*, 2003). Thus, testing TFBS conservation requires a measure that can account for these features.

In this study, we propose a novel TFBS identification method, called TFBSfinder, that utilizes several data sources, including DNA sequences, phylogenetic information, microarray data and ChIP-chip data. In TFBSfinder, reliable target genes of a TF are selected from ChIP-chip data and are required to be co-expressed with each other or to have a temporal (time-shifted) relationship with the expression profile of the TF. The background set is selected from the non-target genes of the TF according to ChIP-chip data. Further, a new metric for measuring the degree of conservation at a binding site across species is proposed. Finally, methods for clustering conserved *k*-mers and for inferring the position weight matrix (PWM) are developed. Our scheme of *k*-mer indexing and subsequent clustering of *k*-mers is similar to that of Shalgi *et al.* (2005) for identifying 3′ UTR motifs that may affect stability or localization of mRNAs.

We test the ability of TFBSfinder to recover planted motifs in synthetic data and to identify TFBSs of cell cycle TFs in *Saccharomyces cerevisiae*. For synthetic data TFBSfinder recovers the planted motifs more accurately than well-known current methods. For the yeast cell cycle TFs studied, most of our predicted TFBSs are consistent with the consensus sequences in the literature. Moreover, compared with well-known methods, TFBSfinder recovers known binding sites with higher precision. Finally, we explore the effects of target gene selection and the conservation criterion on the accuracy of TFBS prediction.

2 METHODS

Figure 1 shows the flowchart of TFBSfinder (for a more detailed figure, see our website). For a TF, we first define its target genes and non-target genes using ChIP-chip data. Then, these target genes are filtered as follows: they must either be co-expressed or have a temporal relationship (synchronous or time-shifted) with the TF expression profile. Next, we collect a pool of *k*-mers (6–9 bp) that occur more frequently and are

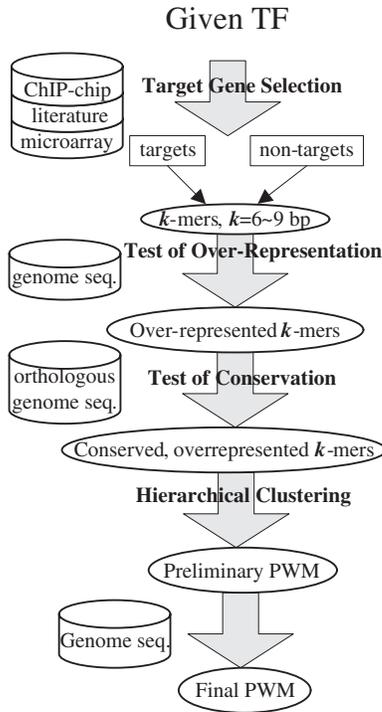*To whom correspondence should be addressed.

Given TF

Fig. 1. Flowchart of the TFBSfinder method.

more evolutionarily conserved in the promoters of target genes than in the promoters of non-target genes using sequence data from related yeast species. These candidate motifs are clustered and aligned to generate a preliminary PWM. Finally, we use this PWM to select core sequences from the promoter regions of target genes and use their respective flanking regions from the genome to find the best motif to represent the TFBS.

2.1 Defining target and non-target genes for a TF

For TF α, a set of target genes (G^α) and a set of non-target genes ($G^{-\alpha}$) are defined using the ChIP-chip data of Harbison *et al.* (2004). A gene belongs to G^α if the *p*-value in the TF α ChIP-chip experiment is smaller than a certain low threshold (e.g. $p_c < 0.0001$), but belongs to $G^{-\alpha}$ if the *p*-value exceeds a certain high threshold (e.g. $p_c > 0.9$). Since the *in vivo* DNA binding of a TF indicated by ChIP-chip data does not necessarily imply regulation, we use gene expression data to select reliable target genes from G^α as follows.

We first use the temporal relationship identification algorithm (W.S. Wu, W.H. Li and B.S. Chen, submitted for publication) to identify genes in G^α whose expression patterns are significantly correlated with that of TF α, possibly with a time lag. Let $\vec{x} = (x_1, \ldots, x_N)$ be the regulatory profile of TF α (derived from a sigmoid transformation of its expression profile) and $\vec{y} = (y_1, \ldots, y_N)$ be the expression profile of gene *y*. We define a correlation coefficient $r(j)$ for each pair of $\vec{x}$ and $\vec{y}$, where *j* is the time lag of *y* behind *x* that results in their maximum correlation. Note that *j* should be substantially smaller than *N*; e.g. in the case of the cell cycle data of Spellman *et al.* (1988), $N = 18$ (covering two cell cycles) and we used $j \leq 8$, so that the possible time lag was shorter than one cell cycle. Then we test the null hypothesis H_0: $r(j) = 0$ against the alternative H_1: $r(j) \neq 0$. Those genes with a *p*-value smaller than a cutoff are considered to be the target genes of TF α with a temporal relationship.

In addition to genes selected by the above procedure, we use the co-expressed genes in G^α. For this selection procedure, we use only the time points where TF α is functional; for example, in the application to cell cycle, the time points of TF α are selected following the procedure demonstrated for the cell cycle in our previous study (Tsai *et al.*, 2005). For the

selected time points, we calculate a threshold *T*, determined as the 95th percentile correlation coefficient value of all the pairwise correlation coefficients between 1000 gene pairs randomly chosen from the *S.cerevisiae* genome. We quantify the expression profiles within a set of genes using the EC score (expression correlation score), which is defined as the fraction of gene pairs in the set with a correlation higher than *T* (Pilpel *et al.*, 2001; Banerjee and Zhang, 2003). A gene is discarded from G^α if elimination of the gene yields a higher maximum EC score for the remaining genes. This process is continued until the correlation of each pair in the remaining genes becomes larger than *T*.

Additional target genes are recruited into G^α if there is experimental evidence from any of the four TF databases (Cherry *et al.*, 1998; Mewes *et al.*, 1999; Zhu and Zhang, 1999; Wingender *et al.*, 2001).

2.2 Identifying overrepresented *k*-mers in target gene promoters

For a *k*-mer sequence *S*, let f_α and $f_{-\alpha}$ be, respectively, the proportions of genes in G^α and $G^{-\alpha}$ with *S* occurring in their 500 bp upstream non-coding regions. A *k*-mer *S* is considered overrepresented in G^α if f_α is significantly greater than $f_{-\alpha}$. For each *S*, we test H_0: $f_\alpha - f_{-\alpha} = 0$ against H_1: $f_\alpha - f_{-\alpha} > 0$ using the one-sided two-sample proportion test. We reject the null hypothesis if

$$f_\alpha - f_{-\alpha} > z_{1-0.01} \text{ s.d.}(f_\alpha - f_{-\alpha}) \cong z_{1-0.01}\sqrt{f(1-f)\left(\frac{1}{|G^\alpha|} + \frac{1}{|G^{-\alpha}|}\right)},$$

where $f = (|G^\alpha|f_\alpha + |G^{-\alpha}|f_{-\alpha})/(|G^\alpha| + |G^{-\alpha}|)$ is the pooled estimator of the population proportion of a binomial distribution under the null hypothesis and $z_{1-0.01}$ is the *z*-score at the 0.01 critical value from the standard normal distribution using the asymptotic normal approximation. Before we proceed, we impose the following constraint on the upper threshold of $f_{-\alpha}$ to eliminate simple repetitive elements such as the TATA-box:

$$f_{-\alpha} + z_{1-0.01} \text{ s.d.}(f_{-\alpha}) \cong f_{-\alpha} + z_{1-0.01}\sqrt{f_{-\alpha}(1 - f_{-\alpha})/|G^{-\alpha}|} \leq \delta,$$

where $f_{-\alpha}$ is assumed to come from a binomial distribution and δ is set to 0.16. (The 0.16 cutoff was empirically determined from our experiments with 27 known cell cycle TFs.)

2.3 Identifying conserved *k*-mers in target genes

We develop a method to test whether a *k*-mer *S* has a higher degree of conservation across related species in G^α than in $G^{-\alpha}$. For this purpose we calculate a conservation matrix as follows. For genes in *S.cerevisiae* whose upstream regions contain *S*, we perform an alignment (ClustralW) on the promoter regions (500 bp, intergenic regions only) of its orthologues in *Saccharomyces paradoxus*, *Saccharomyces kudriavzevii*, *Saccharomyces mikatae* and *Saccharomyces bayanus* and collect the orthologous *k*-mers most similar to *S* in the vicinity of *S*'s positions in *S.cerevisiae* (a region within 3*k* bp). We create a conservation matrix for *S* from all these potentially orthologous *k*-mers by computing the frequency of each nucleotide at each position and correcting for background frequencies, similar to a PWM. The value for base *b* at position *i* in the conservation matrix is

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)},$$

where $p(b)$ is the background probability of base *b* and $p(b,i)$ is the corrected probability of base *b* at position *i*, calculated as

$$p(b,i) = \frac{f_{b,i} + p(b)}{N + 1},$$

where *N* is the number of orthologous *k*-mers and $f_{b,i}$ is the counts of base *b* at position *i*. This procedure is to eliminate null values before log-conversion and to correct for a small sample size (Wasserman and Sandelin, 2004). Each orthologous *k*-mer can be scored with the

conservation matrix by summing the values that correspond to the observed nucleotide at each position:

$$\sum_{i=1}^{k} W_{l_i, i,}$$

where l_i is the nucleotide at position i in the orthologous k-mer. For each gene, we collect all the orthologous k-mer scores as the conservation scores of S. If there are multiple occurrences of S within the same gene, the ones with the highest sum across species are selected. The conservation matrices of all k-mers are calculated in advance to speed up the computation.

Subsequently, we perform the one-sided Kolmogorov–Smirnov (KS) test to select k-mers with significantly higher conservation scores in G^α than in $G^{-\alpha}$. The KS test is a non-parametric test to determine if two distributions differ significantly. For each k-mer S, we test $H_0 : F_{G^\alpha} = F_{G^{-\alpha}}$ against $H_1 : F_{G^\alpha} <_{st} F_{G^{-\alpha}}$ using the one-sided KS test, where F denotes the cumulative distribution function of the conservation scores of a gene group. If H_0 is rejected, $F_{G^\alpha} <_{st} F_{G^{-\alpha}}$, which means that the conservation scores in group G^α are 'stochastically greater' than those in group $G^{-\alpha}$.

2.4 Constructing the position weight matrix

Since the binding motif can be variable at several positions, we group the candidate patterns (the overrepresented and conserved k-mers) together based on overlapping genomic positions and their similarity. First, we record all incidences of these candidate patterns in the 500 bp upstream regions of the target genes. If the occurrences of different k-mers overlap in the genome, the longest continuous pattern spanning the candidate patterns is added into the candidate pool to eliminate redundancy. Then, we perform a hierarchical clustering algorithm to group and align similar candidate patterns. At the start, the algorithm assigns each candidate pattern to its own group, and the similarity between two groups is calculated from their optimal alignment without gaps. At every step, the two groups with the greatest similarity are merged into a new group according to the optimal alignment. The similarities between the new group and the other groups are then updated. The process is iterated until no pair of clusters shares a similarity above 0.6. (This cutoff point and the following one were empirically determined from our experiments with 27 known cell cycle TFs. However, different cutoff thresholds between 50 and 65% do not have strong influences on the performance of our method. The only pronounced effect it makes is on the number of motif groups, as shown in Supplementary Table 1.)

For each alignment of the remaining clusters, a nucleotide position is excluded if <50% of the patterns selected have a nucleotide at that position. We construct a preliminary PWM for the remaining positions. Those aligned sequences are scored according to the preliminary PWM and the third quartile of these scores is used as the cutoff to eliminate noisy patterns. Using the preliminary PWM with the cutoff, we scan the promoter sequences of target genes. Those patterns with a score larger than the cutoff are retained. These patterns are then padded with 50 bp of their flanking sequences in the genome. Next, we calculate the entropy at each nucleotide position. Starting from the position of the minimum entropy, we define a core region and extend it in both directions until the entropy rises above the cutoff, which is defined as the 95th percentile of an empirical entropy distribution generated by scrambling the padded alignments 1000 times. The core region is represented as a PWM with the cutoff.

3 RESULTS

We studied the performance of TFBSfinder using simulation and known PWMs. Further, we compared the performance of TFBSfinder with those of three well-known methods: AlignACE (Roth *et al.*, 1998), MEME (Bailey and Elkan, 1995) and MDscan (Liu *et al.*, 2002). AlignACE and MEME are designed to identify degenerate motifs that are overrepresented in the input set with respect to a background frequency model. AlignACE uses an iterative masking approach to identify multiple overrepresented motifs, based on the Gibbs sampling algorithm. MEME uses the expectation maximization (EM) technique to identify motifs by optimizing an E-value of a statistic that is the product of the p-values of position information contents. MDscan adopts a word numeration strategy to find enriched motifs in the input and employs a heuristic method to update the motif model, with the third-order Markov model or a given background set. We used the default values for parameter setting. Motif width is set to be 6–9 bp in MEME and 10 bp in MDscan. The versions used are AlignACE v4.0, MEME v3.5.2 and MDscan v1.0.

3.1 Synthetic data simulation

In the simulation studies, the four methods compared were applied to 50 datasets each of which contained both a set of 'target' genes (between 25 and 50 genes) and a set of 'non-target' genes (between 500 and 800 genes) randomly selected from the yeast genome. Ideally, a method should not detect motifs from a random set of genes. TFBSfinder reported motifs only for 7 out of the 50 datasets, while AlignACE, MDscan and MEME each reported a number of motifs for all 50 datasets. While some of these motifs may be biologically meaningful, most of them seem to be random motifs (simple repetitive elements) or common *cis*-elements such as TATA box.

In addition, we investigated the recovery rate of planted motifs for each method. From those 43 sets of synthetic data where TFBSfinder reported no motif, we randomly selected 10 sets for motif planting. For each set, a binding motif with known consensus was randomly selected to be planted and designated as the 'answer'. The abundance of each binding motif planted was determined using the degree of conservation and degeneracy similar to its occurrences in the yeast genome. The motifs were inserted into the promoter regions of the target genes and their orthologous sites, where the inserted positions were selected at random between 20 and 480 bp upstream of the start site. The probability of each target gene containing at least one occurrence was randomly chosen between 0.2 and 0.5, and the probability of the motif occurring one, two or three times in the same promoter region of a target gene was 70, 20 and 10%, respectively. As shown in Supplementary Table 2, TFBSfinder produced motifs most similar to the planted motifs in 8 out of 10 runs (the similarity is defined in Section 3.4). Overall, it also achieved a higher sensitivity [TP/(TP+FN)] and specificity [TP/(TP+FP)] in motif recovery (Table 1). Note that our method produced only one motif for each of the synthetic runs, whereas the other methods gave multiple motifs and the motif identified by AlignACE or/and MEME that matched best to the answer often failed to rank number one in its outputs. For more details, see Supplementary Table 2.

3.2 Identifying the binding sites of yeast cell cycle TFs

For each of the 50 cell cycle TFs identified in Tsai *et al.* (2005), the binding threshold (p_b) is set to 0.0001 or 0.001 and the non-binding threshold (p_{nb}) to 0.8 or 0.9 to select a sufficient number of target and non-target genes. Twelve TFs (Arg80, Ask10, Haa1, Hal9, Hir2, Hir3, Met18, Phd1, Rcs1, Rme1, Skn7, and Spt23) are found to have fewer than 10 potential target genes and are excluded from further analysis. The cutoff p-values for the tests of overrepresentation (one-tailed two-sample proportion test) and conservation (KS test) are set to 0.01 and 0.005, respectively.

Table 1. Average performances of TFBSfinder, AlignACE, MDscan and MEME on 10 synthetic datasets

	TFBSfinder	AlignACE	MDscan	MEME
Similarity (S)	**0.93/0.93**	0.68/**0.93**	0.62/0.72	0.78/0.82
Sensitivity (Sn)	**0.86/0.86**	0.30/0.72	0.29/0.52	0.60/0.54
Specificity (Sp)	**0.92/0.92**	0.26/0.79	0.24/0.46	0.53/0.52

The four methods are evaluated using the similarity between an inferred motif and the planted motif (S) and the ability to recover the planted TFBS [sensitivity (Sn) and specificity (Sp)]. Within each entry, the value on the left denotes the average similarity, sensitivity or specificity of a given method, with only the highest ranking motif considered, while the value on the right denotes the average similarity, sensitivity or specificity of a given method for the motif that matches best to the answer, regardless of its ranking.

Table 2 displays the 38 predicted TFBSs as 38 sequence logos. Our inferred PWMs agree well with 10 known PWMs, 13 known binding consensus sequences and four (Dig1, Fhl1, Met4 and Stb1) predicted consensus sequences (Harbison *et al.*, 2004). For the remaining 11 motifs, several observations can be made. Although the binding motif of Mig2 is unknown, Mig2 is structurally homologous to Mig1, and the TFBSs predicted for the two TFs are highly similar. Ndd1 does not bind to DNA directly but is recruited by Fkh1/Fkh2 to regulate genes. Indeed, for Ndd1 we inferred a TFBS that greatly resembles the binding site of Fkh1/Fkh2. YOX1 is found to co-occur with MCM1 and the putative binding sites are thought to be TAATTR (Pramila *et al.*, 2002). TFBSfinder reveals a conserved motif TTAGGAAW as the binding site of Yox1 with a slightly more degenerate motif YAATTA nearby. TFBSfinder predicts TTMGCR as the binding motif for Hir1, for which no known consensus is available. Yap5 is an interesting example in that a majority of Yap5 target genes reside in the subtelomeric regions of 12 yeast chromosomes and lack orthologues in the other four species. These targets are likely duplicated genes, because their coding regions and promoters reveal a high degree of similarity, possibly due to subtelomeric segmental duplications. The highly similar promoters make it difficult for TFBSfinder to pick out the TFBSs in these genes. Similar situations are found for Dat1, Gat3 and Pdr1.

No candidate is found for Smp1, and the results for Gal4 and Sut1 fail to match literature evidence, in part due to an insufficient number of target genes (16, 12 and 18 for Gal4, Sut1 and Smp1, respectively). The failure to identify the correct binding site for Gal4, specifically, can be attributed to the fact that the number of variable positions (11 Ns) in Gal4 (CGGN$_{11}$CCG) far exceeds the capacity of our current program. While Sut1 has a sequence located in the C-terminal half that is similar to the Zn(II)2Cys6 binuclear cluster DNA-binding domain shared by many TFs such as Gal4, it does not appear to bind to DNA directly. Rather, it physically interacts with Cyc8 to prevent the Cyc8-Tup1 complex from repressing hypoxic genes through their association with Rox1. This added layer of binding complexity limits our ability to identify the TFBS.

From the results, we deduce groups of TFs whose TFBSs are highly similar, including Msn2-Msn4, Fkh1-Fkh2-Ndd1, Swi4-Swi6-Mbp1-Stb1, Met31-Met32-Met4, Dig1-Ste12 and Ace2-Swi5. This is consistent with experimental evidence that either

they are proteins with homologous DNA-binding domains or one is the piggy-back binding TF of the other. For example, Msn4 is a structural homolog of Msn2. Msn2 is mostly responsible for the binding of STRE (STress Response Element) and Msn4 can weakly interact with STRE and can partially compensate for the absence of MSN2. Msn2 might recognize promoter sequences of the SUC2 gene that are normally bound by the Mig1 repressor (CCCCC/CCCCG), which is similar to STRE. In fact, four TFs (Msn2, Msn4, Mig1 and Mig2) have the same C2H2 zinc finger domain. Swi4 and Swi6 form the SBF complex and Mbp1 and Swi6 form the MBF complex. We were able to correctly identify the TFBSs of Swi4 and Mbp1 and to discover both motifs for Swi6. For Stb1, which plays a role in the regulation of MBF-specific transcription, TFBSfinder also yielded a motif highly similar to that of the MBF complex.

3.3 Importance of target gene selection and test of motif conservation

To explore the effects of target gene selection and the test of motif conservation, we evaluate the performance of TFBSfinder with and without these two procedures for 27 TFs with known PWMs, known consensuses or predicted consensuses. We use the consensus sequences as the correct answers and call a candidate motif pattern a 'hit' if there is 80% or higher similarity in their overlapping regions, which are required to be at least 4 bp in length. Highly degenerate positions (B, D, H, V and N) in the consensus are ignored. We then define the hit ratio as the proportion of 'hits' within a set of candidate patterns. Note that there are two major groups of candidates for SWI6 that represent the motifs of SCB (CNCGAAA) and MCB (ACGCGT), respectively. Candidate motif patterns that match either one are considered hits.

We find 19 out of 27 TFs to achieve a higher hit ratio when target genes are pre-selected (Fig. 2). In particular, target gene selection greatly increases the hit ratios for Ace2, Mig1 and Swi5. For Dig1 and Rlm1, the candidate patterns identified without target gene selection were so well conserved and overrepresented that the hit ratio was already 100%, which could not be further improved by a target gene selection procedure. The omission of target gene selection does not alter the candidate hit ratios for Tec1, Met4 and Msn2 because the numbers of target genes eliminated by target gene selection are too few to make an impact on the selection of candidate patterns (detailed results can be found on our website). For the remaining three TFs, Msn4, Met31 and Stb1, the hit ratios without target gene selection are somewhat higher than those with pre-target gene selection; note that the consensuses of Stb1 was predicted, not experimentally verified. The average hit ratios of TFBSfinder with and without target gene selection are 0.76 and 0.61, respectively, indicating that target gene selection indeed reduces noise in the target gene set. However, the procedure for testing the correlation between the expression profiles of the target gene and the TF is not recommended if the experimental time points are far apart or not evenly distributed.

Figure 2 shows an ∼25% reduction in hit ratio when the test of motif conservation is omitted, suggesting that the test is effective in helping eliminate overrepresented but false candidate motifs. If TFBSfinder is carried out when neither the test of motif conservation nor target gene selection is included, the average hit ratio falls to only 39%. The results indicate that both the target gene selection

Table 2. Comparison of predicted binding specificities to known PWMs or consensus sequences

TF	Results (PWM)	known PWM or consensus	TF	Results (PWM)	known PWM or consensus
ABF1*	*(logo)*	TCACTATACACG	MIG1*	*(logo)*	AAAAATGTGGGG
ACE2*	*(logo)*	ACCAGC	MIG2	*(logo)*	*Unknown*
BAS1*	*(logo)*	TGACTC	MSN2*	*(logo)*	AAGGGG
DAT1	duplicate genes in subtelomeres	*unknown*	MSN4*	*(logo)*	AAGGGG
DIG1*	*(logo)*	TGTTTCA[#]	NDD1	*(logo)*	*unknown*
FHL1*	*(logo)*	CAYCCRTACA[#]	PDR1	duplicate genes in subtelomeres	*(logo)*
FKH1*	*(logo)*	TTGTTTACC	RAP1*	*(logo)*	*(logo)*
FKH2*	*(logo)*	GGTAAACAA	REB1*	*(logo)*	*(logo)*
GAL4	*(logo)*	*(logo)*	RLM1*	*(logo)*	*(logo)*
GAT3	duplicate genes in subtelomeres	*unknown*	SMP1	*not available*	*(logo)*
HAP1*	*(logo)*	CGGNNNTANCGG	STB1*	*(logo)*	TTSGCGTYY[#]
HAP4*	*(logo)*	YCNNCCAATNANM	STE12*	*(logo)*	*(logo)*
HIR1	*(logo)*	*unknown*	SUT1	*(logo)*	GCSGSGNNSG[#]
MAC1*	*(logo)*	GAGCAAA	SWI4*	*(logo)*	*(logo)*
MBP1*	*(logo)*	*(logo)*	SWI5*	*(logo)*	*(logo)*
MCM1*	*(logo)*	*(logo)*	SWI6*	*(logo)*	TTTCGNG
MET31*	*(logo)*	AAACTGTGG	TEC1*	*(logo)*	RGAATG
MET32*	*(logo)*	AAACTGTGG	YAP5	duplicate genes in subtelomeres	TTATCAA
MET4*	*(logo)*	GSCRCSMCASWTKKY[#]	YOX1	*(logo)*	TAATTR[#]

Degenerate codes: R: A or G, Y: C or T, S: G or C, W: A or T, M: A or C, K: G or T, B: C, G, or T, D: A, G, or T, H: A, C, or T, V: A, C, or G, N: A, C, G, or T. A consensus of a TF marked with # indicates that the consensus is obtained from computational prediction. TFs marked with * are used in further analysis.
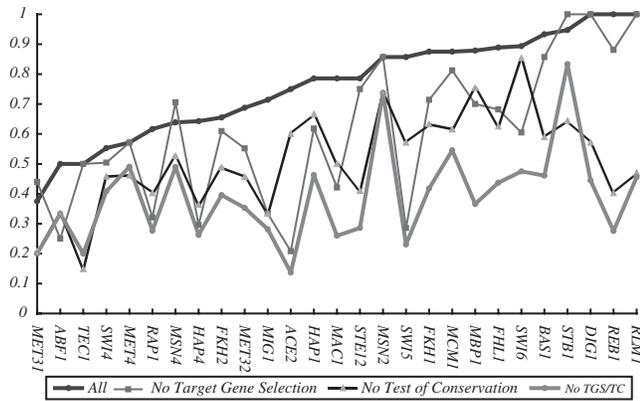
Fig. 2. Importance of target gene selection and test of motif conservation illustrated by the hit ratios of candidate motif patterns of the following four methods: 'All' stands for TFBSfinder; 'No Target Gene Selection' stands for TFBSfinder without target gene selection; 'No test of Conservation' stands for TFBSfinder without conservation filtering and 'No TGS/TC' stands for TFBSfinder without target gene selection and test of motif conservation. The *y*-axis denotes the hit ratio.

procedure and the test of motif conservation improve the accuracy of TFBS detection.

A breakdown of the information provided by ChIP-chip, expression profiles and literature-derived databases can be found in Supplementary Figure 1. On average, target gene selection based on temporal correlation or co-expression generally retains 45% of the binding targets initially identified by ChIP-chip experiments. With the aid of expression profiles, ChIP-chip information contributes to selection of 60% of the final set of target genes. Evidence from the literature alone accounts for roughly 36%, while 4% of the target genes are selected based on both the literature and ChIP-chip/expression data.

3.4 Comparison to well-known methods

We evaluate the performance of TFBSfinder against AlignACE, MDscan and MEME on 27 yeast cell cycle TFs with known PWMs, known consensuses or predicted consensuses. This is done by measuring the similarity of the predicted motifs to the consensus sequences. For each of the other three methods, we restrict the numbers of output motifs to be the same as that for TFBSfinder. For those compared motifs available only in the consensus form, we generate the corresponding substitution-derived PWM (sdPWM) according to Doniger *et al.*'s (2005) model, which is constructed from all occurrences of the consensus sequence with 0 or 1 difference in the orthologous positions in the other four species. To calculate the similarity between a predicted PWM (*a*) and a substitution-derived PWM (*b*), they are aligned to maximize

$$1 - \frac{1}{w} \sum_{i=1}^{l} \frac{1}{\sqrt{2}} \sqrt{\sum_{L \in \{A, T, C, G\}} (a_{i,L} - b_{i,L})^2},$$

where *w* is the number of positions in the substitution-derived PWM, and $a_{i,L}$ and $b_{i,L}$ are the estimated probabilities of base *L* at position *i* in PWMs *a* and *b*, respectively. Both sides of *a* are padded with a sufficient number of bases (A:0.31, T:0.31, C:0.19 and G:0.19) to ensure that each position of the substitution-derived PWM has a corresponding position in *a*. This distance metric is a

Table 3. Performance comparisons of TFBSfinder, AlignACE, MDscan and MEME, using 27 yeast cell cycle TFs with known PWMs (with 'a'), known consensuses or predicted consensuses

TFs	TFBSfinder	AlignACE	MEME	MDscan
ABF1[a]	0.72	**0.87**	0.57	0.59
ACE2	**0.92**	0.59	0.56	0.35
BAS1	**0.95**	0.94	0.94	0.58
DIG1	**0.94**	**0.94**	0.93	0.51
FHL1	0.90	**0.91**	0.85	0.90
FKH1	0.86	0.81	**0.93**	0.88
FKH2	0.80	0.85	**0.87**	0.58
HAP1	**0.65**	0.50	0.51	0.47
HAP4	**0.78**	0.65	0.64	0.61
MAC1	0.82	**0.83**	0.82	0.80
MBP1[a]	**0.96**	0.95	0.93	0.95
MCM1[a]	0.85	0.86	0.46	**0.88**
MET31	0.83	0.52	**0.84**	0.43
MET32	**0.93**	0.84	**0.93**	0.42
MET4	0.76	**0.78**	0.73	0.46
MIG1[a]	**0.87**	0.51	0.83	0.51
MSN2	0.80	0.53	**0.83**	0.38
MSN4	0.85	0.82	**0.87**	0.38
RAP1[a]	0.80	**0.85**	0.82	0.84
REB1[a]	0.94	0.44	**0.95**	0.93
RLM1[a]	**0.69**	0.57	0.58	0.67
STB1	**0.87**	0.86	0.57	0.51
STE12[a]	0.82	0.65	**0.83**	0.77
SWI4[a]	0.82	**0.91**	0.83	0.60
SWI5[a]	**0.84**	0.69	0.80	0.58
SWI6	**0.88**	0.82	0.85	0.69
TEC1	**0.89**	0.70	0.61	0.60
Average	**0.84**	0.74	0.77	0.62

A substitution-derived PWM (sPWM) is generated for each known consensus according to Doniger *et al.*'s (2005) model. We modified the distance metric of Harbison *et al.* (2004) to calculate similarity between two PWMs. Each entry represents the similarity between the derived PWM and the sPWM or known PWM. For each TF (row), the entry in boldface indicates the derived PWM most similar to sPWM or known PWM.

modified version of that in Harbison *et al.* (2004). For each method and for each TF, we calculate the similarity between all generated motif groups and the substitution-derived PWM. The group with the maximum similarity is selected as the right answer. As shown in Table 3, among the four methods TFBSfinder identified motifs with the highest similarity in 13 out of 27 TFs, while AlignACE, MEME and MDscan performed best for 7, 8 and 1, respectively. Of those 14 TFs where TFBSfinder did not outperform all of the other three methods, the similarities of our predicted motifs to the consensuses were close to the highest. The average similarity for TFBSfinder for these 27 TFs (0.84) is significantly higher than the averages (0.74, 0.77 and 0.62) for the other methods.

Next, we compare the accuracy of binding site predictions of the four methods to known TF binding sites. We collect the experimentally verified binding sites of these 27 TFs from the TRANSFAC and SCPD databases. For each TF, duplicate instances of the same binding site are removed, and those located more than 500 bp away from the start site are not considered. We also ignore binding sites when no motif can be found within the positions indicated. After the filtering, only 12 TFs with a total of 101 binding site records are retained. We measure the sensitivity and specificity

Table 4. Performance comparisons of TFBSfinder, AlignACE, MDscan and MEME, using 27 yeast cell cycle TFs: the sensitivity (Sn) and specificity (Sp) of four methods in recovering known TF binding sites

TFs	No. of sites	TFBSfinder		AlignACE		MEME		Mdscan	
		Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
ABF1	16	0.07	0.33	**0.60**	**0.91**	0.00	0.00	0.00	0.00
BAS1	6	**1.00**	0.70	0.67	**0.80**	0.83	0.50	0.00	0.00
HAP1	3	**0.25**	**0.20**	0.00	0.00	0.00	0.00	0.00	0.00
HAP4	4	**0.33**	**1.00**	**0.33**	**1.00**	0.00	0.00	0.00	0.00
MAC1	3	**1.00**	0.57	**1.00**	0.57	0.75	**0.60**	0.75	**0.60**
MCM1	21	0.78	0.61	**0.83**	**0.76**	0.17	0.08	0.78	0.66
MIG1	10	**0.80**	**0.62**	0.00	0.00	0.50	0.50	0.00	0.00
RAP1	9	**0.73**	0.41	**0.73**	0.89	**0.73**	1.00	**0.73**	0.77
REB1	11	0.73	0.89	0.00	0.00	**1.00**	**1.00**	0.82	0.82
STE12	2	**0.57**	**0.50**	0.00	0.00	0.29	0.29	0.43	0.38
SWI5	6	0.33	**1.00**	0.00	0.00	**0.50**	0.38	0.17	0.33
SWI6	1	**1.00**	**0.25**	0.00	0.00	**1.00**	**0.25**	0.00	0.00
Average		**0.63**	**0.59**	0.35	0.41	0.48	0.38	0.31	0.30

For each TF (row), the entry in boldface indicates the derived PWM most similar to sPWM or known PWM.

using the following criteria. A predicted binding site is counted as a hit if it overlaps with a true binding site by >50% of the length of the shorter one of the predicted or known binding sites. A binding site prediction is made based on the following rule. For each TF, we select the best motif from the output and assign a cutoff for its PWM, which is defined as the 5th percentile of the PWM scores of all sequences used to generate the PWM. We then scan the promoter sequences and make a prediction of a binding site if its PWM score is greater than the cutoff. As shown in Table 4, TFBSfinder has the highest sensitivity in 8 of the 12 TFs and the highest specificity in 6 TFs. The average sensitivity and specificity for TFBSfinder are 0.63 and 0.59, higher than those for the other three methods. These comparisons show the superiority of TFBSfinder over the three current methods.

Note that without test of conservation and target gene selection, our method will not be so robust. This is a limitation of our method. However, as long as more than one species are available, a test of conservation can be conducted. In principle, our approach can be applied to higher eukaryotes such as human but considerable modifications are needed—this will be explored in the future. In this paper, our target selection is intended for multiple-time-point data. For non-temporal data, the temporal correlation test cannot be done but the co-expression test can still be applied.

ACKNOWLEDGEMENTS

REFERENCES

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21–29.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Bannai,H. *et al.* (2004) Efficiently finding regulatory elements using correlation with gene expression. *J. Bioinform. Comput. Biol.*, **2**, 273–288.

Cherry,J.M. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Cliften,P. *et al.* (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Doniger,S.W. *et al.* (2005) Identification of functional transcription factor binding sites using closely related *Saccharomyces* species. *Genome Res.*, **15**, 701–709.

Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, **6**, R18.

Emberly,E. *et al.* (2003) Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics*, **4**, 57.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hertz,G.Z. *et al.* (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Kato,M. *et al.* (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.*, **5**, R56.

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Liu,X.S. *et al.* (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Mewes,H.W. *et al.* (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**, 44–48.

Moses,A.M. *et al.* (2003) Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.*, **3**, 19.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Pizzi,C. *et al.* (2005) Detecting seeded motifs in DNA sequences. *Nucleic Acids Res.*, **33**, e135.

Pramila,T. *et al.* (2002) Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes Dev.*, **16**, 3034–3045.

Roth,F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Shalgi,R. *et al.* (2005) A catalog of stability-associated sequence elements in 3′ UTRs of yeast mRNAs. *Genome Biol.*, **6**, R86.

Sinha,S. (2003) Discriminative motifs. *J. Comput. Biol.*, **10**, 599–615.

Sinha,S. and Tompa,M. (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **31**, 3586–3588.

Spellman,P.T. *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.

Tanay,A. *et al.* (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 7203–7208.

Tsai,H.K. *et al.* (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl Acad. Sci. USA*, **102**, 13532–13537.

Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Wingender,E. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

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

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