Finding transcription factor binding motifs for coregulated genes by combining sequence overrepresentation with cross-species conservation

Jia, Hui.; Li, Jinming.

2012

Jia, H., & Li, J. (2012). Finding Transcription Factor Binding Motifs for Coregulated Genes by Combining Sequence Overrepresentation with Cross-Species Conservation. Journal of Probability and Statistics, 2012,1-18.

https://hdl.handle.net/10356/99457

https://doi.org/10.1155/2012/830575

Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 830575, 18 pages
doi:10.1155/2012/830575

Research Article

Finding Transcription Factor Binding Motifs for Coregulated Genes by Combining Sequence Overrepresentation with Cross-Species Conservation

Hui Jia[1] and Jinming Li[1,2]

[1] *School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551*
[2] *Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou Dadao Bei1838, Guangzhou 510515, China*

Correspondence should be addressed to Jinming Li, jmli@smu.edu.cn

Received 1 March 2012; Accepted 29 April 2012

Academic Editor: Xiaohua Douglas Zhang

Novel computational methods for finding transcription factor binding motifs have long been sought due to tedious work of experimentally identifying them. However, the current prevailing methods yield a large number of false positive predictions due to the short, variable nature of transcriptional factor binding sites (TFBSs). We proposed here a method that combines sequence overrepresentation and cross-species sequence conservation to detect TFBSs in upstream regions of a given set of coregulated genes. We applied the method to 35 *S. cerevisiae* transcriptional factors with known DNA binding motifs (with the support of orthologous sequences from genomes of *S. mikatae*, *S. bayanus*, and *S. paradoxus*), and the proposed method outperformed the single-genome-based motif finding methods *MEME* and *AlignACE* as well as the multiple-genome-based methods *PHYME* and *Footprinter* for the majority of these transcriptional factors. Compared with the prevailing motif finding software, our method has some advantages in finding transcriptional factor binding motifs for potential coregulated genes if the gene upstream sequences of multiple closely related species are available. Although we used yeast genomes to assess our method in this study, it might also be applied to other organisms if suitable related species are available and the upstream sequences of coregulated genes can be obtained for the multiple closely related species.

1. Introduction

To understand the mechanisms that regulate the gene expression in eukaryotes is a major challenge in modern molecular biology. Gene regulation is accomplished by a number of regulatory proteins called transcriptional factors (TFs), which bind to specific DNA motifs in the promoter region of the target gene. TFs and their binding motifs interact with each other

and help cells to respond to diverse stimuli. Identifying TFBSs in the upstream region of coregulated genes (genes regulated by a common TF) is crucial for inferring gene regulatory networks, since these motifs might be the building blocks of the regulatory network structures [1]. Most DNA binding motifs contain 6–25 bps and have a range of variability. Regulatory systems can take advantage of the variability in the binding sites to better control transcription [2]. Classical computational motif finding methods can be classified into two major categories: (1) enumerative methods, which explore all possible motifs up to a certain length; (2) local search algorithms using statistic approaches such as EM [3–6] and Gibbs sampling [7–9]. Under the second category, *MEME* [10–12] and *AlignACE* [13, 14] are two computer programs used frequently in finding motifs in a group of related DNA sequences. Recently, comparative genomics approaches such as phylogenetic footprinting have been developed for identifying TFBSs based on the premise that selective pressure causes functional elements to evolve at a slower rate than that of nonfunctional sequences [15]. Phylogenetic footprinting is mostly applied to finding well-conserved regions in a set of orthologous sequences from multiple species [15–18]. Although substantial progresses have been made in developing computational methods for predicting transcription factor binding motifs, currently available motif finding tools still yield many false positives due to relatively short length and high variability of DNA binding motifs. These motif finding tools with standard parameter settings usually report one putative TFBS out of 500 to 5000 bps, whereas only 0.1% of the predictions is likely to be functional [19]. Recently, gene expression profile analysis using microarray data and statistical clustering has resulted in numerous sets of potential co-regulated genes. Furthermore, the complete sequencing of more and more eukaryotic genomes makes it easier to obtain the upstream sequences of these co-regulated genes. Hence the development of a novel method with improved specificity in predicting transcription factor binding motifs for co-regulated genes becomes necessary and feasible.

We proposed here a method of finding TF binding motifs by considering both sequence overrepresentation in promoter regions and their conservation across closely related species. DNA binding motifs are believed to appear more frequently in the upstream regions of the genes being regulated, and these motifs are usually conserved across multiple closely related species. We use the degree of sequence conservation among multiple species as an additional constraint to reduce the false positive predictions. For a given set of co-regulated genes from a certain organism, we collect orthologous sequences from multiple closely related species and align them using multiple alignment programs such as *ClustalW* [20]. The statistically overrepresented sequences will be firstly selected as initial motif candidates, and then we evaluate their conservation in the alignments of orthologous upstream sequences of coexpressed genes. A statistic procedure based on the above principles was designed to scan for potential motifs, and a Perl script was written to conduct the procedure. To evaluate the proposed method, we collected 35 yeast TFs with known DNA binding motifs, and for genes co-regulated by each of these TFs, we searched the upstream regions for potential binding motifs. We compared our method with single-genome-based motif finding methods such as *MEME* and *AlignACE*, as well as with the multiple-genome based methods such as *PHYME* and *Footprinter*; the results suggested that the rank of the known binding motifs among the predictions of our method are generally higher than that using other methods.

2. Results

We used 35 well-studied yeast transcription factors (see Table 1) to evaluate the proposed method. The criteria for selecting the TFs are (1) their true DNA binding motifs are known;

(2) the orthologous genes are available in all the four yeast species, and the upstream sequences of these genes are also available. For each TF, we built two sets of genes, namely, the positive set (*PS*) and the negative set (*NS*). The *PS* consisted of all the genes that are known to be co-regulated by the TF (see Table 1), whereas the *NS* consisted of randomly selected genes from the *S. cerevisiae* genome. The *NS* was used to introduce the background information and serve as a control in our motif finding process. For each gene in both *PS* and *NS,* we extracted its promoter region sequences from the genomes of four yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus,* and *S. paradoxus*. We took *S. cerevisiae* as the principal species in our study.

The method was implemented using a PERL script to find potential binding motifs in the upstream sequences of the genes co-regulated by a given TF (see Table 1 for the 35 TFs considered in this study). We found the known binding motifs for 25 out of the 35 TFs. In Table 2, we listed the known DNA binding motif and the motif found using our method for each TF.

We compared our method with the single-genome-based motif finding methods *MEME* and *AlignACE*, as well as with the multiple-genome-based methods *PHYME* and *Footprinter* for the majority of these transcription factors. We used the upstream sequences of *S. cerevisiae* genes in the *PS* of each TF as the input of *MEME* and *AlignACE*. All the parameters were set to default when we used *AlignACE* to find motifs. To apply *MEME* to the motif finding, we set the minimum length of the potential motif to 6, and we set the number of motifs expected to be found to the same as the number of motifs predicted by our method. The results are listed in Tables 3 and 4, respectively. Since our method takes into account the conservation of candidate motif sequences among multiple species, the number of predicted motifs found for each TF is in general less than that found by *AlignACE* (Table 3) or *MEME* (Table 4). Tables 3 and 4 showed that our method is more efficient in finding the true motifs than *AlignACE* or *MEME*, in the sense that it returned less predicted motifs, and the ranks of the known motifs are also generally higher than those in the output of *AlignACE* or *MEME*. For example, there were 11 potential motifs found by *AlignACE* for *STE12*, and the known motif of *STE12* ranked second in the output; however, using our method only one motif was found, and it was the known motif. The results for other TFs showed the same tendency. *AlignACE* and *MEME* could only find the known binding motifs for 14 and 12 TFs, respectively, out of the 25 TFs whose known binding motifs were found using our method. Our method could not find the known binding motifs for 10 TFs among the 35 (Table 5) with any of the three parameter threshold settings. Out of these 10 TFs, using *AlignACE* and *MEME* we can find known binding motifs for 5 and 3 TFs, respectively.

Unlike single-genome-based motif finding methods such as *AlignACE* and *MEME*, our method uses the sequence information from multigenomes, so it is more reasonable to compare it with *PHYME* and *Footprinter,* which are two popular multiple-genome-based methods. For a given TF, we found that *Footprinter* usually yields overwhelming number of predictions, and this makes it difficult to do a comparison. To apply *PHYME* to the motif finding, we set the motif length limit to 17, which is the maximum length of all known binding motifs of the 35 TFs. For each regulon, the number of motifs predicted was set to 10 and the motifs were searched on both strands. The results were listed in Table 6. Using *PHYME* we found known motifs for 23 TFs, among them there were 6 TFs whose known binding motifs were not found using our method. From Table 6, we can see that our method and *PHYME* nearly have the same power in motif finding; however, the ranks of the known motifs found using our method are generally higher than those found by *PHYME*. Table 7 gives a list of the TFs whose known binding motifs could be found using our method but could not be found by *MEME*, *AlignACE,* and *PHYME*. Our method could not find the known binding motifs for

Table 1: Transcription factors and the genes being regulated.

TF	Number of co-regulated genes	Genes regulated by the TF
Ste12	9	YBR083, YCL055W, YFL027C, YJL170C, YLR452C, YML047C, YMR232W, YNL279W, YPL156C
Gal4	10	GAL2, GAL3, GAL1/10, GAL7, MTH1, FUR4, PCL10, GAL80, PGM2, GCY1
MET31	8	YEL015W, YEL016C, STR3, MET16, NUT2, SSN8, YJL060W, YEL072W
Mbp1	18	YEL018W, MMS21, YCK2, MCD1, MCM2, RPS9A, MOT1, OPY2, CLB5, YER071C, VTC1, YJL045W, MSH6, YNR009W, HXT10, YER087C-A, TOF1, YNL274C
Leu3	10	YDR279W, LEU1, OAC1, YOR271C, YDL228C, YHR209W, YHR207C, BAT1, ILV2, RRP6
Cbf1	16	YAL026C, YBR089C-A, YBR225W, YDR438W, YIL074C, YIL126W, YIL127C, YJL167W, YJL168C, YJL209W, YJR010W, YKL191W, YKL192C, YNL094W, YNL095C, YNL282W
Ace2	1	YLR286C
Gcn4	6	YBL103C, YDL170W, YKL015W, YLR451W, YML099C, YNL103W
Abf1	15	YAL038W, YBR248C, YCR012W, YFL038C, YFR031C, YGL234W, YGR059W, YHR174W, YIL160C, YJL166W, YKL112W, YLR203C, YLR204W, YOR116C, YPR110C
Hap1	4	YEL039C, YJR048W, YML054C, YOR065W
Ino4	6	YDR050C, YER026C, YGR157W, YHR123W, YMR084W, YNR016C
Mcb	6	YDL102W, YDL164C, YJL194W, YMR199W, YNL102W, YOR074C
Mse	1	YGR059W
Nbf	1	YJL153C
Pdr3	2	YBL005W, YGR281W
Pho4	2	YDR481C, YGR233C
Put3	1	YHR037W
Rap1	8	YFL014W, YFR031C, YGL123W, YKL062W, YLR399C, YNL216W, YOL082W, YPR102C
Swi5	2	YDL227C, YNL327W
Uasino	1	YJL153C
Uasrad	2	YCR066W, YGL058W
Adr1	2	YDR256C, YMR303C
Mig1	7	YBR019C, YBR020W, YDR009W, YDR146C, YIL162W, YKL109W, YPL248C
T4c	2	YJL106W, YJL153C
Uasphr	14	YBR114W, YDL200C, YDR217C, YEL037C, YER095W, YER142C, YGL058W, YIL066C, YJL026W, YJR035W, YJR052W, YML032C, YNL250W, YPL022W
Ap-1	1	YGR209C
Bas2	2	YCL030C, YGL234W
Csre	2	YER065C, YNL117W
Mac1	11	YDR058C, YDR075W, YER145C, YER146W, YGR136W, YJR049C, YJR050W, YNL250W, YNL251C, YPR110C, YPR111W
Gcr1	2	YAL038W, YGR215W
Mcm1	17	YAL040C, YBR160W, YBR202W, YDR146C, YDR403W, YER111C, YFL026W, YGL008C, YGR108W, YJL159W, YJL194W, YKL178C, YKL209C, YKR066C, YNL277W, YPR113W, YPR119W

Table 1: Continued.

TF	Number of co-regulated genes	Genes regulated by the TF
Reb1	12	YCR012W, YDL164C, YDR007W, YDR050C, YDR146C, YER086W, YFL039C, YGL026C, YNL216W, YOL004W, YOL006C, YPL231W
Rox1	2	YDR044W, YPR065W
Scb	2	YDL227C, YMR199W
Sff	3	YDR146C, YGR108W, YPR119W-

10 TFs out of the 35 (Table 5). For these 10 TFs, *AlignACE* and *MEME* can find known binding motifs for 5 and 3 TFs, respectively. With *PHYME*, we can find known binding motifs for 6 TFs out of these 10 (Table 6).

3. Discussion

Transcription factors and their DNA binding sites are two of the most important functional elements in eukaryotic genomes. A thorough study of the interactions of the two is important for mapping the regulatory pathways and understanding the potential function of the genes regulated by the TFs [21]. In the past decade, clustering of gene expression profiles obtained from large-scale DNA microarray experiments has been successfully used in identifying coexpressed genes [22, 23], and we believed that these coexpressed genes may share common regulators that bind to their upstream regions. Finding the TF binding motifs of these potentially co-regulated genes becomes critical for understanding the interaction of the genes and their regulators [24–27]. So far the binding specificities have been well characterized only for a small number of TFs [19, 21]. TFBSs are usually quite short (around 6–25 bp) and degenerate, which leads to the difficulties in finding them reliably using current motif finding tools. Even though the *ab initio* motif finding tools have been used successfully in many cases, their performances are far from satisfying. The major drawback of these tools is that they produce many false positive predictions. Under default parameter settings, they yield usually tens or hundreds of putative motifs, and it is difficult to judge which candidate motifs out of them are functional [19]. Phylogenetic footprinting methods have been proposed recently [15–18], by which the interspecies comparative sequence information is used for helping to signal the presence of TF binding sites that might not have been predicted using sequences from a single genome. For example, binding sites found in human sequences that are also found in orthologous mouse or other mammalian sequences are far more likely to be functional than those found only in human [28]. We refer to these short orthologous sequences that are conserved over 6 bp or more as phylogenetic footprints.

Our method proposed here considers both overrepresentation and cross-species conservation of potential binding motifs. We used binomial test to determine the statistically overrepresented candidate sequences, and the average relative entropy of the aligned sequence block was used to measure the cross-species conservation of these candidates. The relative entropy is a popular measure of the degree of conservation at a site in a DNA or protein sequence alignment [29]. In our method, the input data are the upstream sequences of two groups of genes, namely, the co-regulated genes of a TF (*PS*) and the control genes (*NS*) selected randomly from the genome of the principal species under study, as well as the

Table 2: Comparison between the known motifs and the motifs found using our method. Different parameter threshold settings are used in our motif finding. (a) P-value in a magnitude of 10^{-6} (after Bonferroni adjustment), $\text{ARE}_p = 1.0$, and Z-value = 2.0; (b) P-value = 0.01 (without Bonferroni adjustment), $\text{ARE}_p = 1.0$, and Z-value = 2.0; (c) P-value = 0.01 (without Bonferroni adjustment), $\text{ARE}_p = 0.8$, and Z-value = 2.0.

TF	Genes in PS	Known motif	Motif found	P-value	Z-value	ARE_p
Ste12[(a)]	9	TGAAACA	TGAAACA	$5.6e{-}12$	3.68	1.00
Gal4[(a)]	10	CGGNNNNNNNNNNCCG	CGGNNNNNNNNNNCCG	$3.7e{-}12$	2.43	1.22
Mbp1[(a)]	18	ACGCGTNA	ACGCGT	$3.0e-7$	3.37	1.35
Leu3[(a)]	10	CCGGNNCCGG	CGGNNNCGG	$7.0e{-}13$	3.15	1.27
Cbf1[(a)]	16	RTCACRTG	CACGTG	$7.7e{-}13$	2.98	1.19
MET31[(a)]	8	CTGTGGC	TGTGGC	$6.7e-7$	3.39	1.06
Abf1[(b)]	15	TCRNNNNNNACG	TCANNNNNNACG	$1.3e-3$	3.73	1.26
Ace2[(b)]	1	GCTGGT	TGCTGGT	$1.4e-3$	6.07	1.55
Gcn4[(b)]	6	TGANTN	ATGACT	$8.7e-4$	4.45	1.10
Hap1[(b)]	4	CGGNNNTANCGG	TGCCGNNNNNNNCGG	$2.3e-4$	6.09	1.64
Ino4[(b)]	6	CATGTGAAAT	CATGTT	$2.9e-4$	5.60	1.31
Mcb[(b)]	6	WCGCGW	CGCNTCG	$4.1e-4$	4.66	1.36
Mse[(b)]	1	CRCAAAW	GACNCAA	$8.3e-3$	4.05	1.19
Nbf[(b)]	1	ATGYGRAWW	CATGTG	$5.9e-3$	5.85	1.36
Pdr3[(b)]	2	TCCGYGGA	TCCNNGGA	$4.3e-4$	2.88	1.03
Pho4[(b)]	2	CACGTK	GCGCGT	$1.8e-3$	3.55	1.20
Put3[(b)]	1	CGGNNNNNNNNNNCCG	TCGNNNNNNNNNNNCG	$2.6e-4$	4.65	1.51
Rap1[(b)]	8	RMACCCA	GTCNNNNNCCCAT	$8.8e-3$	3.16	1.01
Swi5[(b)]	2	KGCTGR	TGCTGG	$6.5e-4$	4.45	1.19
Uasino[(b)]	1	ATCTGAAWW	CATGTG	$5.9e-3$	5.83	1.36
Uasrad[(b)]	2	WTTTCCCGS	TCCNGCT	$1.1e-3$	4.42	1.24
Adr1[(c)]	2	TCTCC	CTCCNNNNNTCC	$1.6e-3$	2.18	0.88
Mig1[(c)]	7	CCCCRNNWWWWW	ACCCCA	$7.2e-3$	2.18	0.82
Uasphr[(c)]	14	CTTCCT	TCTNNNNNNNNNNTCCT	$2.2e-3$	2.38	0.93
T4c[(c)]	2	TTTTCTYCG	TTTTCNNNNNTCC	$1.2e-3$	2.69	0.96

orthologous sequences from other species, which are closely related to the principal species. Usually the co-regulated genes are collected through wet lab experiments or predicted through gene expression profile analysis using microarray data. The upstream sequences of genes in *PS* and *NS* could be extracted from the genome of the principal species, and the corresponding upstream sequences from other species could be obtained by doing BLAST [30] or by downloading from the publicly available databases.

Three parameters are considered in our method: (1) P value, which is used to evaluate the overrepresentation of a candidate sequence, (2) average relative entropy ARE_P of S_{OP}, which gives the degree of conservation of a candidate motif, (3) Z-value, which is used to assess the statistical significance of the conservation. In order to have a balanced consideration of the sensitivity and the specificity and to cope with different situations, we applied three different parameter threshold settings to scan for candidate motifs, and they are (a) P-value in a magnitude of 10^{-6} (after Bonferroni correction),

Table 3: Comparison to *AlignACE*. For each TF, we listed the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	*AlignACE*		Our method	
	The number of motifs found	The rank of the known motif	The number of motifs found	The rank of the known motif
Ste12[(a)]	11	2	1	1
Gal4[(a)]	9	2	1	1
Leu3[(a)]	22	3	1	1
Mbp1[(a)]	20	7	2	1
Cbf1[(a)]	29	1	4	2
Met31[(a)]	20	15	1	1
Abf1[(b)]	10	4	5	3
Ace2[(b)]	7	Not found	4	2
Gcn4[(b)]	11	4	39	6
Hap1[(b)]	6	Not found	23	1
Ino4[(b)]	13	Not found	22	4
Mcb[(b)]	18	2	11	1
Mse[(b)]	6	Not found	5	5
Nbf[(b)]	6	1	16	15
Pdr3[(b)]	12	Not found	10	2
Pho4[(b)]	8	Not found	9	1
Put3[(b)]	2	Not found	4	1
Rap1[(b)]	13	6	14	11
Swi5[(b)]	9	Not found	3	1
Uasino[(b)]	4	Not found	14	13
Uasrad[(b)]	4	Not found	20	3
Adr1[(c)]	4	2	14	2
Mig1[(c)]	30	2	32	30
Uasphr[(c)]	14	Not found	47	13
T4c[(c)]	9	1	28	6

ARE_P = 1.0, and Z-value = 2.0; (b) P-value = 0.01 (without Bonferroni correction), ARE_P = 1.0, and Z-value = 2.0; (c) P-value = 0.01 (without Bonferroni correction), ARE_P = 0.8, and Z-value = 2.0. Theoretically, we can find most of the known motifs as long as we make the criteria for overrepresentation and conservation loose enough, but the less strict criteria may result in numerous putative motifs that are actually false positives. Considering the high cost of verifying a predicted motif through lab experiment, we used firstly a strict criterion for candidate motif screening, so parameter setting (a) was set as default in our method. Using this strict parameter threshold setting we may miss some true TF binding motifs (see Tables 3 and 4), especially those without very high-level statistical significance of overrepresentation, and the method may not be able to return any predictions. We loosen the criteria by using setting (b) or setting (c) in actual motif finding process, if using the default threshold setting, we can find no hit at all. Setting (b) has a moderate criterion for overrepresentation, so it allows more candidate motif to pass the screening. With setting (c), we loosen the criterion

Table 4: Comparison to *MEME*. *MEME* requests a predetermined number of predicted motifs as its input, and we let it be the number of motifs predicted using our method. For each TF, we listed the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	*Meme*		Our Method	
	The number of motifs found	The rank of the known motif	The number of motifs found	The rank of the known motif
Ste12[(a)]	1	Not found	1	1
Gal4[(a)]	1	1	1	1
Leu3[(a)]	1	1	1	1
Mbp1[(a)]	2	1	2	1
Cbf1[(a)]	4	1	4	2
Met31[(a)]	1	Not found	1	1
Abf1[(b)]	5	Not found	5	3
Ace2[(b)]	4	Not found	4	2
Gcn4[(b)]	39	Not found	39	6
Hap1[(b)]	9	Not found	23	1
Ino4[(b)]	22	10	22	4
Mcb[(b)]	11	1	11	1
Mse[(b)]	5	Not found	5	5
Nbf[(b)]	16	3	16	15
Pdr3[(b)]	10	1	10	1
Pho4[(b)]	9	4	9	1
Put3[(b)]	4	Not found	4	1
Rap1[(b)]	14	Not found	14	11
Swi5[(b)]	3	Not found	3	1
Uasino[(b)]	14	Not found	14	13
Uasrad[(b)]	20	1	20	3
Adr1[(c)]	10	Not found	14	2
Mig1[(c)]	32	2	32	30
Uasphr[(c)]	47	Not found	47	13
T4c[(c)]	28	23	28	6

of the degree of conservation, since there do exist some known TF binding motifs with ARE_P less than 1.0 (see Table 2).

The method proposed here is, nevertheless, not a replacement of the prevailing motif tools such as *MEME* and *AlignACE*. The major limitation of our method is its strong prerequisite. Multiple closely related species and the upstream sequences of each co-regulated gene for all species under study are requested, and in many cases these prerequisites may not be satisfied, so the method is, therefore, not generally applicable. Another problem is how to choose the appropriate species to evaluate the cross-species conservation. In principle, the species selected in the study should be close enough so that the conservation of motif sequences could be detected in a multiple alignment, in the meanwhile their evolutionary distances should not be too close, so that the signals could be distinguished from the noises [31]. The number of species used in the method is also a factor that may need

Table 5: The TFs whose known binding sites cannot be found using our method. The expected number of motifs predicted by *MEME* was set at 10.

TF	Known motif	Using *AlignACE*	Rank of the known motif/total predictions	Using *MEME*	Rank of the known motif/total predictions
Ap-1	TTANTAA	Not found		TTAGTAA	3/10
Bas2	TAATRA,TAANTAA	Not found		Not found	
Csre	YCGGAYRRAWGG	Not found		GTCCGGAC	8/10
Mac1	GAGCAAA	GGAAGCAAA	17/33	Not found	
Gcr1	CWTCC	ATTGTTTTCC	5/5	Not found	
Mcm1	CCNNNWWRGG	TTACCNNNTAGGAAA	2/11	TTTCCTAATTAGGAAA	1/10
Reb1	YYACCCG	TTACCCGCACGGC	3/8	Not found	
Rox1	YYNATTGTTY	Not found		Not found	
Scb	CNCGAAA	AAGCCACGAAAA	1/13	Not found	
Sff	GTMAACAA	Not found		Not found	

to be considered. We recommended three or four, since using too many species may bring up strong noise and reduce the detection power of the method.

After comparing with the motif finding software such as *MEME*, *AlignACE*, and *PHYME*, we can reach the following conclusions: (1) Our method screens for candidate motifs in terms of both overrepresentation and conservation, therefore, it gives relatively less predicted motifs for a group of co-regulated genes (Tables 3 and 4), hence it is helpful for reducing false positive predictions; (2) The rank of known motif in the output of our method is in general higher (Tables 3 and 4), and this is of practical importance, since we usually focus only on putative binding motifs with high ranks despite the large number of predicted motifs; (3) unlike the most common motif finding tools, our method requests no prior inputs such as the length of the motifs or the number of predictions. Although we used yeast genomes to assess our method, it could also be applied to other organisms if suitable related species are available and the upstream sequences of co-regulated genes could be obtained for the multiple species.

4. Materials and Methods

4.1. Materials

In this study, we considered gene promoter regions of four yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus*. All these four are members of the *Saccharomyces sensu stricto* group. The last three are believed to be separated from *S. cerevisiae* by an estimated 5–20 million years of evolution and are found to have sufficient sequence similarity to *S. cerevisiae* such that orthologous regions can be aligned reliably [32].

We obtained the information about gene regulation network of *S. cerevisiae* from the database SCPD (The Promoter Database of Saccharomyces cerevisiae) [33], which

Table 6: Comparison to *PHYME*. For each TF, we listed the number of predicted motifs and the rank of the known motif in the predictions. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	Known motif	Found by our method	Rank	Found by *PHYME*	Rank
Ste12[(a)]	TGAAACA	TGAAACA	1	TGAAACA	3
Gal4[(a)]	CGGNNNNNNNNNNCCG	CGGNNNNNNNNNNCCG	1	CCGAATAGTCTGCCCCG	8
Mbp1[(a)]	ACGCGTNA	ACGCGT	1	ACGCGTCA	3
Leu3[(a)]	CCGGNNCCGG	CGGNNNCGG	1	CCGGTACCGG	3
Cbf1[(a)]	RTCACRTG	CACGTG	2	GTCACGTG	2
MET31[(a)]	CTGTGGC	TGTGGC	1	Not found	
Abf1[(b)]	TCRNNNNNNACG	TCANNNNNNACG	3	Not found	
Ace2[(b)]	GCTGGT	TGCTGGT	2	Not found	
Gcn4[(b)]	TGANTN	ATGACT	6	TGAGTC	6
Hap1[(b)]	CGGNNNTANCGG	TGCCGNNNNNNNCGG	1	Not found	
Ino4[(b)]	CATGTGAAAT	CATGTT	4	Not found	
Mcb[(b)]	WCGCGW	CGCNTCG	1	ACGCGT	1
Mse[(b)]	CRCAAAW	GACNCAA	5	CACAAAA	3
Nbf[(b)]	ATGYGRAWW	CATGTG	15	ATGTGAAAT	1
Pdr3[(b)]	TCCGYGGA	TCCNNGGA	1	TCCGCGGA	2
Pho4[(b)]	CACGTK	GCGCGT	1	Not found	
Put3[(b)]	CGGNNNNNNNNNNCCG	TCGNNNNNNNNNNNCG	1	Not found	
Rap1[(b)]	RMACCCA	GTCNNNNNCCCAT	11	AAACCGA	4
Swi5[(b)]	KGCTGR	TGCTGG	1	TGCTGAATG	1
Uasino[(b)]	ATCTGAAWW	CATGTG	13	Not found	
Uasrad[(b)]	WTTTCCCGS	TCCNGCT	3	TTTCCCAC	4
Adr1[(c)]	TCTCC	CTCCNNNNNTCC	2	ACTCC	4
Mig1[(c)]	CCCCRNNWWWWW	ACCCCA	30	CCCCGCCCC	4
Uasphr[(c)]	CTTCCT	TCTNNNNNNNNNNTCCT	13	GCTTTCTTT	8
T4c[(c)]	TTTTCTYCG	TTTTCNNNNNTCC	6	TTTTTCTTTT	1
Ap-1	TTANTAA	Not found		Not found	
Bas2	TAATRA,TAANTAA	Not found		TAATAG	8
Csre	YCGGAYRRAWGG	Not found		Not found	
Mac1	GAGCAAA	Not found		GAGAAAA	3
Gcr1	CWTCC	Not found		Not found	
Mcm1	CCNNNWWRGG	Not found		CCGTTTTGGG	5
Reb1	YYACCCG	Not found		CTACCCG	5
Rox1	YYNATTGTTY	Not found		Not found	
Scb	CNCGAAA	Not found		CACGAAA	1
Sff	GTMAACAA	Not found		GTTAACAA	6

Table 7: The TFs whose known binding motifs cannot be found by *MEME*/*AlignACE*/*PHYME*, but can be found using our method. Three different parameter threshold settings, namely, (a), (b), and (c), are used in our method as given in Table 2.

TF	Genes in *PS*	Our method	Known motif
Ace2[(b)]	1	GCTGGT	TGCTGGT
Hap1[(b)]	4	TGCCGNNNNNNNCGG	CGGNNNTANCGG
Mse[(b)]	1	GNCACAA	CRCAAAW
Put3[(b)]	1	TCGNNNNNNNNNNNCG	CGGNNNNNNNNNNCCG
Swi5[(b)]	2	TGCTGG	KGCTGR
Uasino[(b)]	1	CATGTG	ATCTGAAWW
Uasphr[(c)]	14	TCTNNNNNNNNNNTCCT	CTTCCT

contained TFs and genes co-regulated by them. The upstream region sequences of the co-regulated genes of each TF for all the four yeast species were downloaded from http://www.broad.mit.edu/.

The genes known to be co-regulated by specific TFs such as *STE12* and *GAL4* were used to evaluate the method. We let *PS* (positive set) denote the collection of *S. cerevisiae* genes co-regulated by a common TF, and we built an *NS* (negative set) by randomly selected *S. cerevisiae* genes. For each gene in both *PS* and *NS,* we extracted the promoter region sequences for all the four species and aligned them using multiple sequence alignment program *ClustalW*.

4.2. Methods

The method proposed here requests promoter region sequences from multiple closely related ortholog species. Usually we are interested in motif finding for only one of the species, namely, the principal species, whereas the sequences from other species are helpful for the reduction of false positives. For a given TF, we need two sets of genes, namely, positive set (*PS*) and negative set (*NS*). *PS* consists of the genes co-regulated by the TF, whereas *NS* consists of genes randomly collected from the genome of the principal species.

4.3. Finding Overrepresented Sequences

We only consider the principal species for finding overrepresented sequences. We first search the promoter regions of the genes in *NS* for each possible sequence pattern of length M ($6 \leq M \leq 17$) that satisfying the following constraints: the first three nucleotides in the left flank and the last three nucleotides in the right flank are the core elements and fixed, between the two core elements there might be M_0 nucleotides ($M_0 = 0, 1, 2, \ldots, 11$) and each of them could be any of the nucleotides A, C, T, and G. So within the L-bp upstream region of a gene, there are $L - M + 1$ possible locations that can be occupied by a sequence pattern of length M. We call the fraction as the background probability of the given sequence pattern

$$p = \frac{c_n}{c}, \tag{4.1}$$

where c_n is the number of total occurrences of the pattern in the promoter regions of genes in NS, and c is the total number of possible locations for an M-bp sequence in promoter regions of the genes in NS. In the same way, we can obtain the number of the pattern occurrences K and the total number (N) of possible M-bp locations in the promoter regions of the genes in PS. Using binomial distribution, we can calculate the probability of the pattern occurring more than K times as following [27, 34]:

$$P = \sum_{k=K}^{N} \frac{N!}{(N-k)!k!} p^k (1-p)^{N-k}, \tag{4.2}$$

where p is the background probability. We choose the sequence patterns with P less than a threshold p^* (usually in magnitude of 10^{-6} after Bonferroni adjustment) for further analysis. If the overlap of two sequences is longer than 80% of one of the two, we eliminate the sequence with larger P-value from the collection of overrepresented sequences. Both DNA strands are considered when we calculate the number of occurrences for a given sequence in the upstream region. If the sequence is a palindrome, we just use the count in one strand as the total occurrence.

4.4. Bonferroni Adjustment

We used the Bonferroni adjustment to the multiple statistical tests for determining overrepresented sequences, so that it was more "difficult" for any single test to be significant. The adjustment was accomplished by setting the P-value threshold at the common significant level (usually 0.05 or 0.01) divided by the number of tests being performed. In our case, the p^* was set as 0.05 divided by the number of all possible sequences in the form of $NNNnn\ldots nnNNN$, where NNN stands for three fixed nucleotides and $nn\ldots nn$ stands for unfixed number (from 0 to 11) of nucleotides.

4.4.1. Relative Entropy and Conservation Criteria

Let α be the background nucleotide distribution and β the nucleotide distribution at a given position in a multiple sequence alignment. For the two probability distributions α and β, the relative entropy (also known as Kullback-Leibler "distance") is defined by [29, 35]

$$H(\beta||\alpha) = \sum_{i=1}^{4} \beta_i \log \frac{\beta_i}{\alpha_i}. \tag{4.3}$$

We can prove that relative entropy is always a nonnegative value, and it reflects the extent of the deviation of actual nucleotide distribution from background distribution at a given site in the alignment. The larger the value, the greater the deviation between the actual distribution and the background distribution at that site [29].

Given an overrepresented sequence O, we search for its occurrences in the alignment of upstream sequences from the four species for each gene in PS and NS, respectively. If we find an occurrence in the alignment of a gene in PS, we extract the corresponding sequence block from the alignment and put the four segments that form this block to a sequence set S_{OP}. Similarly, we also build a sequence set S_{ON} for genes in NS.

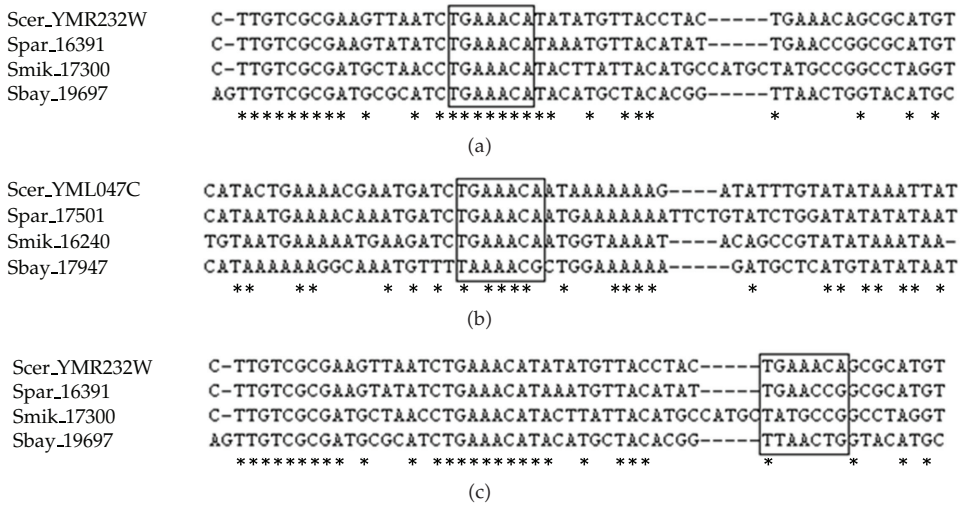

Figure 1: DNA binding motif *TGAAACA* of transcription factor *STE12*. The known DNA binding motif *TGAAACA* of transcription factor *STE12* for *S. cerevisiae* genes YML047C (a), YLR452C (b), and YCL055W (c) is conserved in the alignment of orthologous gene promoter regions of closely related yeast species, namely, *S. cerevisiae*, *S. mikatae*, *S. bayanus*, and *S. paradoxus*.

We further align all the sequences in S_{OP} and S_{ON}, respectively. These two alignments are used to evaluate the degree of conservation of O across closely related species. We define the average relative entropy (ARE_P) of S_{OP} as

$$\mathrm{ARE}_P = \frac{\sum_{i=1}^{M} EP_i}{M}, \tag{4.4}$$

where EP_i is the relative entropy at the position i of the alignment of the sequences in S_{OP}, and M is the length of O. If O is not found in the alignment of upstream sequences for any gene in NS, then we deposit O to the collectionof candidate motifs for further consideration. Otherwise, we could also calculate the average relative entropy ARE_N for the sequences in nonempty set S_{ON}. We define a Z-score as

$$Z = \frac{\mathrm{ARE}_P - \mathrm{ARE}_N}{\sqrt{s_N^2/M}}, \tag{4.5}$$

where s_N is the standard deviation of the relative entropies at different positions of the multiple upstream sequence (across multiple species) alignments of genes in NS.

Binding motifs tend to be conserved in the orthologous species (see Figures 1, 2, and 3), so we remove the sequences that are overrepresented but not conserved from our collection of candidate sequences. We set the Z-score threshold as 2, such that the sequences with $Z > 2$ are kept as the candidate sequences for further consideration.

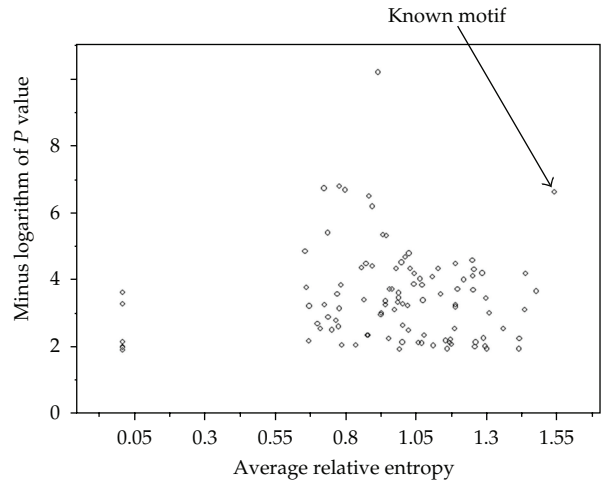

Figure 2: Average relative entropy of the motifs. We choose the 100 sequences found in the upstream of Ace2-regulated *S. cerevisiae* genes with the smallest overrepresentation *P*-values and compute their average relative entropies in the multiple alignments of orthologous upstream sequences of the four related species. The results were displayed with a scatter plot of *P*-value versus average relative entropy. The arrow points to the known binding motif *GCTGGT* of Ace2 in *S. cerevisiae*. The average relative entropy of the known binding motif is greater than that of most other sequences.

4.5. Building a Profile for a Candidate Sequence

Each candidate sequence will be searched for in the alignment of upstream sequences (from the multiple species) of each gene in *PS*. If an instance is found in any of the species, we extract the corresponding alignment block for further consideration. We use e_B to denote the average of the relative entropies at M different positions of an alignment block of length M. For each block, we set $h_P = \mu_P + 2(\sigma_P/\sqrt{M})$ as our cutoff value for block selection, where μ_P and σ_P are the mean and the standard deviation of the relative entropies, respectively, at different positions in the alignments of upstream sequences (from multiple species) of the genes in *PS*. For a given candidate sequence, we use all the blocks with e_B greater than h_P to build a profile to represent the candidate motif. For example, we search for a candidate sequence *GTTTCA* in the alignments of upstream sequences of genes in *PS*. If we can find it in any species in the alignments, we extract the corresponding alignment block, calculate e_B, and compare it with h_P to decide whether we keep this block for profile building. Using all the blocks selected, we calculate the base frequencies at each position and create thereafter the profile to represent the initial candidate motif. Both strands are considered when we build the profile.

4.6. Species-Specific PSSM Building

The profile obtained above represents the initial candidate motif derived from all the ortholog species. Usually we are only interested in the motif finding for one species, which is named as principal species in our analysis, and it is necessary to build a species-specific *PSSM* (Position Specific Score Matrix) for the candidate motif [36]. For the genes in *PS,* which are from the principal species, we search for the candidate motif in their upstream sequences in terms

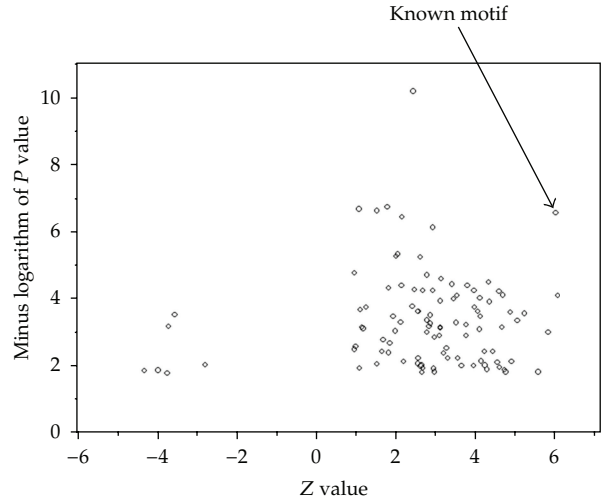

Figure 3: Z-value of the motifs. We choose the 100 sequences found in the upstream of *Ace2*-regulated *S. cerevisiae* genes with the smallest overrepresentation P-values and compute their Z-values in the multiple alignments of orthologous upstream sequences of the four related yeast species. The arrow points to the known motif *GCTGGT* of Ace2 in *S. cerevisiae*. The Z-value of the known binding motif is greater than that of the most other sequences.

of the initial motif profile, and all the significant hits found are used in building the final species-specific *PSSM*. The profile search is performed as follows. For each M-bp segment of upstream sequences of the genes in PS, we calculate a score

$$Sc = \prod_{i=1}^{M} q_i, \tag{4.6}$$

where q_i is the probability of observing the ith nucleotide of the segment, which is defined by the position-specific nucleotide distribution in the initial profile of the candidate motif. To determine the significance criterion, we calculate Scs for all the possible M-bp segments of the upstream sequences (for principal species only) of genes in NS and rank these scores in the descending order. We use the 0.001-quantile of these ranked scores, denoted as Sc^*, as the threshold value to determine whether a match is significant in the profile search. For example, if there are 1000 genes in the NS and the length of each promoter region is L-bp, then there are totally $1000^*(L-M+1)$ possible segments, so we have $1000^*(L-M+1)$ scores. We sort the scores in the descending order and set the nth value as the cutoff score Sc^* with $n = L-M+1$. We calculate Sc for each possible segment in the upstream sequences (principal species only) of the genes in PS. If $Sc \geq Sc^*$, we deposit the segment into I, which is the set of the incidences of the candidate motif.

4.7. Optimal Motif Length

Let k be the number of sequence segments in I. In order to determine the optimal length of the potential motif, we extend 0 to 5 bp in both flanks of each M-bp segment in I according to its mother sequence in the gene upstream region. So we have totally 36 possible combinations

(left flank extended by $M_L = 0, 1, 2, 3, 4,$ or 5 bp; right flank extended by M_R = 0, 1, 2, 3, 4, or 5 bp). For each possible combination (M_L, M_R), we put the newly added flanks into a block with k rows and $M_L + M_R$ columns. We calculate the average relative entropies of all 36 blocks and choose the combination (M_{L^*}, M_{R^*}) that delivers the maximum average relative entropy e_{B^*} for further consideration. In the meanwhile, we randomly generate 1000 sequence blocks, each with k rows and $M_{L^*} + M_{R^*}$ columns, in terms of the background nucleotide distribution α. We calculate the mean e_{rand} and the standard deviation s_{rand} of the average relative entropies of these 1000 blocks. If e_{B^*} is greater than $e_{\text{rand}} + 2s_{\text{rand}}$, then we accept the extension (M_{L^*}, M_{R^*}) and set the final motif length at $M + M_{L^*} + M_{R^*}$; otherwise, we still keep the original motif length M. The extended sequences (M_{L^*} bp in left flank and M_{R^*} bp in right flank) of the segments in I form a new sequence set I_e, which is the set of the incidences of the extended motif. Using all the sequences in I_e, we build the *PSSM* for a general representation of the final motif.

4.8. Implementation

We used a PERL script to implement the method. The script and the example of input data are available upon request.

Abbreviations

TFBS: Transcription factor binding sites
TF: Transcription factor
bp: Base pair
EM: Expectation maximization.

Conflict of Interests

The authors have declared that no competing interests exist.

Acknowledgments

This work is financially supported by the Nanyang Technological University Research Grant RG64/06 (to JMLI) and a start-up grant from Southern Medical University and Guangdong Province (to JMLI).

References

[1] T. I. Lee, N. J. Rinaldi, F. Robert et al., "Transcriptional regulatory networks in *Saccharomyces cerevisiae*," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.
[2] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.
[3] Elkan TLBaC, *Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers*, AAAI Press, Menlo Park, Calif, USA, 1994.
[4] C. E. Lawrence and A. A. Reilly, "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, no. 1, pp. 41–51, 1990.

[5] L. R. Cardon and G. D. Stormo, "Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments," *Journal of Molecular Biology*, vol. 223, no. 1, pp. 159–170, 1992.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] G. Thijs, K. Marchal, M. Lescot et al., "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes," *Journal of Computational Biology*, vol. 9, no. 2, pp. 447–464, 2002.

[8] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats," *Protein Science*, vol. 4, no. 8, pp. 1618–1632, 1995.

[9] S. Sinha and M. Tompa, "Discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Research*, vol. 30, no. 24, pp. 5549–5560, 2002.

[10] T. L. Bailey and C. Elkan, "The value of prior knowledge in discovering motifs with MEME," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 3, pp. 21–29, 1995.

[11] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994.

[12] T. L. Bailey and M. Gribskov, "Combining evidence using *P*-values: application to sequence homology searches," *Bioinformatics*, vol. 14, no. 1, pp. 48–54, 1998.

[13] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, no. 10, pp. 939–945, 1998.

[14] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*," *Journal of Molecular Biology*, vol. 296, no. 5, pp. 1205–1214, 2000.

[15] M. Blanchette and M. Tompa, "Discovery of regulatory elements by a computational method for phylogenetic footprinting," *Genome Research*, vol. 12, no. 5, pp. 739–748, 2002.

[16] L. A. McCue, W. Thompson, C. S. Carmack et al., "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes," *Nucleic Acids Research*, vol. 29, no. 3, pp. 774–782, 2001.

[17] M. Blanchette, B. Schwikowski, and M. Tompa, "Algorithms for phylogenetic footprinting," *Journal of Computational Biology*, vol. 9, no. 2, pp. 211–223, 2002.

[18] S. Sinha, M. Blanchette, and M. Tompa, "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences," *BMC Bioinformatics*, vol. 5, article 170, 2004.

[19] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[20] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[21] P. Qiu, "Recent advances in computational promoter analysis in understanding the transcriptional regulatory network," *Biochemical and Biophysical Research Communications*, vol. 309, no. 3, pp. 495–501, 2003.

[22] L. Ma, J. Li, L. Qu et al., "Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways," *Plant Cell*, vol. 13, no. 12, pp. 2589–2607, 2001.

[23] G. B. Fogel, D. G. Weekes, G. Varga et al., "Discovery of sequence motifs related to coexpression of genes using evolutionary computation," *Nucleic Acids Research*, vol. 32, no. 13, pp. 3826–3835, 2004.

[24] M. Caselle, F. Di Cunto, and P. Provero, "Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes," *BMC Bioinformatics*, vol. 3, article 7, 2002.

[25] L. Mao, C. Mackenzie, J. H. Roh, J. M. Eraso, S. Kaplan, and H. Resat, "Combining microarray and genomic data to predict DNA binding motifs," *Microbiology*, vol. 151, no. 10, pp. 3197–3213, 2005.

[26] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3339–3344, 2003.

[27] P. M. Haverty, U. Hansen, and Z. Weng, "Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification," *Nucleic Acids Research*, vol. 32, no. 1, pp. 179–188, 2004.

[28] A. Prakash and M. Tompa, "Discovery of regulatory elements in vertebrates through comparative genomics," *Nature Biotechnology*, vol. 23, no. 10, pp. 1249–1256, 2005.
[29] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1998.
[30] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
[31] S. W. Doniger, J. Huh, and J. C. Fay, "Identification of functional transcription factor binding sites using closely related Saccharomyces species," *Genome Research*, vol. 15, no. 5, pp. 701–709, 2005.
[32] E. Herrero, "Evolutionary relationships between *Saccharomyces cerevisiae* and other fungal species as determined from genome comparisons," *Revista Iberoamericana de Micologia*, vol. 22, no. 4, pp. 217–222, 2005.
[33] J. Zhu and M. Q. Zhang, "SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*," *Bioinformatics*, vol. 15, no. 7-8, pp. 607–611, 1999.
[34] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor, "Toucan: deciphering the cis-regulatory logic of coregulated genes," *Nucleic Acids Research*, vol. 31, no. 6, pp. 1753–1764, 2003.
[35] S. Kullback, *Information Theory and Statistics*, John Wiley & Sons, New York, NY, USA, 1959.
[36] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2004.