Transcriptional Regulation in the G1-S Cell Cycle Stage in Fungi: Insights through Computational Analysis

Viktor Martyanov[1] and Robert H. Gross[2,*]

[1]*Department of Genetics, Dartmouth Medical School, Hanover, NH 03755, USA*

[2]*Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA*

Abstract: The transcription factor complexes Mlu1-box binding factor (MBF) and Swi4/6 cell cycle box binding factor (SBF) regulate the cell cycle in *Saccharomyces cerevisiae.* They activate hundreds of genes and are responsible for normal cell cycle progression from G1 to S phase. We investigated the conservation of MBF and SBF binding sites during fungal evolution. Orthologs of *S. cerevisiae* targets of these transcription factors were identified in 37 fungal species and their upstream regions were analyzed for putative transcription factor binding sites. Both groups displayed enrichment in specific putative regulatory DNA sequences in their upstream regions and showed different preferred upstream motif locations, variable patterns of evolutionary conservation of the motifs and enrichment in unique biological functions for the regulated genes. The results indicate that despite high sequence similarity of upstream DNA motifs putatively associated with G1-S transcriptional regulation by MBF and SBF transcription factors, there are important upstream sequence feature differences that may help differentiate the two seemingly similar regulatory modes. The incorporation of upstream motif sequence comparison, positional distribution and evolutionary variability of the motif can complement functional information about roles of the respective gene products and help elucidate transcriptional regulatory pathways and functions.

Keywords: G1-S transition, cell-cycle, fungal evolution, motif finding, transcriptional regulation, regulons, TFBS, gene regulation.

INTRODUCTION

Eukaryotic cells commit to the cell cycle late in G1 at a G1-S phase called Start. Transcription of hundreds of genes is induced as part of the G1-S transition, including those responsible for DNA synthesis, budding and spindle pole body duplication [1]. Much of this transcriptional program depends on MBF and SBF transcription factors [2].

MBF and SBF are heterodimeric complexes sharing a common trans-activating or regulatory subunit, Swi6 [3]. Their DNA-binding components (Mbp1 for MBF and Swi4 for SBF) are related proteins with different target DNA sequences, ACGCG (MCB, or Mlu1 cell cycle box) for Mbp1 [2] and CRCGAAA (SCB, or Swi4/6 cell cycle box) for Swi4 [4].

There is a considerable amount of information known about MBF and SBF and their effect on G1-S-regulated transcription. They are generally known as G1-S transcriptional activators. While single deletion mutants of either *Mbp1 or Swi4* are viable in *S. cerevisiae*, the double deletion mutant is lethal with arrest occurring in G1 [3]. It has been shown that in some cases the removal of either transcription factor minimally influences the transcription rate of putative target genes [5]. It has also been reported that the DNA-binding subunits, Mbp1 and Swi4, may be functionally redundanteither because some promoters have instances of both binding sites [6] or because of cross-binding [7]. Finally, a number of genes show increased expression in the absence of MBF and/or SBF suggesting their role as repressors or repressor activators for some genes [3].

It is assumed that cell cycle regulatory components are conserved among eukaryotes, reflecting the importance of this process [8]. Nevertheless, there have been only two studies that mentioned the question of conservation of G1-S transcriptional regulation across fungal evolution [9, 10]. It should be noted, however, that the main focus of those papers was the conservation and evolution of multiple other regulatory networks. Additionally, in both cases the research focused on a more limited selection of fungal species (14 [9] and 17 [10]), seven of which were members of the *Saccharomyces* genus. In this study, we analyzed 38 fungal species that are evenly distributed across fungal phylogeny representing three major phyla, Zygomycota, Basidiomycota and Ascomycota and spanning across hundreds of millions of years of evolution.

These studies recognized the potential of computational methods by combining biologically generated data with the output of a motif-finding algorithm [9]. Motif finding is the process of computational identification of sequence patterns that are important both for transcription regulation analysis and protein function prediction [11]. Different motif finders employ various search strategies and use different DNA motif representations [12]. However, a majority of existing tools

*Address correspondence to this author at the Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA;
Tel: 603-646-2059; Fax: 603-646-1347;
E-mail: robert.h.gross@dartmouth.edu

1875-0362/12

try to find only a specific type of motif. Because of that, they exhibit similar performance over multiple datasets from multiple species. In fact, in a benchmark study, 13 widely used motif finders displayed low absolute measures of correctness [13].

Combining several motif finding approaches can improve the accuracy of prediction. Our method, Suite for Computational identification Of Promoter Elements (SCOPE) uses an ensemble approach to combine three search strategies by looking for three kinds of motifs: short and non-degenerate (e.g. ACGCG), short and degenerate (e.g. WCGYG) and long and bipartite (e.g. AYGNNNNNNCRT) [14]. The individual algorithms are all run in parallel and then combined to return the best scoring motifs. SCOPE outperforms numerous existing motif finders and is highly robust to the presence of extraneous sequences in the input gene set [14].

In this paper, we analyzed the evolutionary conservation of putative upstream regulatory motifs responsible for G1-S transcriptional regulation. Starting with well-studied sets of *S. cerevisiae* MBF and SBF gene targets, we generated a list of orthologs in other fungal species. Orthologs from each species were then analyzed by SCOPE. High-scoring candidate motifs from each run were compared to the *S. cerevisiae* computationally predicted binding sites. Sequence logos of these motifs were then displayed on the fungal tree in order to gain additional insights into the relationship between fungal evolution and consensus sequence of the putative transcription factor binding sites. We have also analyzed combinations of computationally predicted motifs in the upstream regions of corresponding gene sets. Finally, MBF- and SBF-regulated genes from *S. cerevisiae* and sets of orthologous genes from other fungal species were analyzed for functional enrichment and functional specificity. Using this experimental approach, we showed overall conservation of G1-S transcriptional regulons across fungal evolution in terms of upstream regulatory motifs, their patterns and functional annotations of the respective genes.

MATERIALS AND METHODOLOGY

An overview of our approach is shown in Fig. (**1**) and described in the following sections.

S. cerevisiae Gene Set Source

Lists of MBF and SBF gene targets were generated as follows. The initial transcriptional regulatory map of *S. cerevisiae* [15] was constructed by combining genome-wide chromatin immunoprecipitation data, phylogenetic conservation and prior knowledge. It was refined by applying conservation-based motif discovery algorithms, PhyloCon [16] and Converge [17]. Following the approach undertaken by Harbison *et al.* [15], we restricted the analysis to genes from MacIsaac *et al*. [17] containing highly significant motifs (bound by the corresponding factor at a p-value $\leq 10^{-3}$ denoting a high confidence of the interaction between a regulator and promoter regions based on both genome-wide location data and several motif discovery methods) conserved across three out of four related yeast species (*S. cerevisiae*, *S. mikatae*, *S. kudriavzevii* and *S. bayanus*). Using these criteria and removing gene targets for both MBF and SBF, we generated two unique sets of 68 MBF and 40 SBF gene targets.

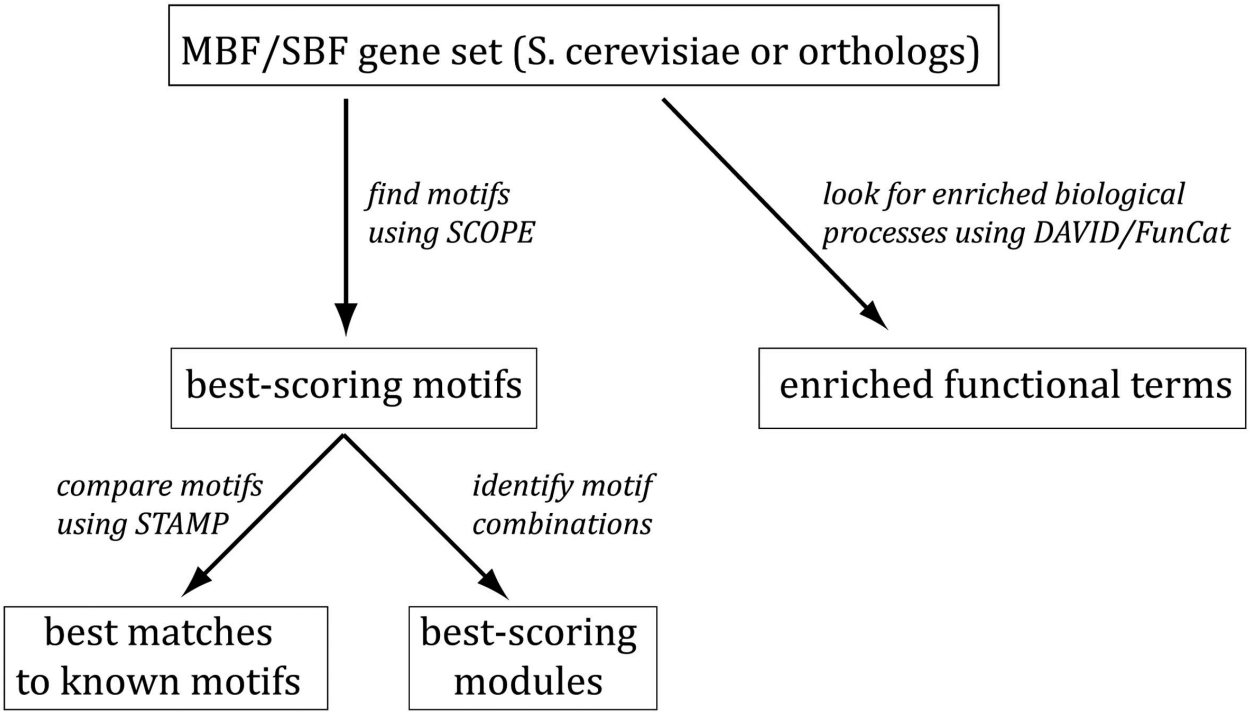

Fig. (1). Flowchart of the experimental pipeline.

A set of genes regulated by MBF or SBF transcription factors from S. cerevisiae (or their orthologs from other fungal species) are analyzed *via* SCOPE to identify best-scoring upstream DNA motifs. These motifs are compared to the known regulatory motifs *via* STAMP and their combinations are assessed *via* a module analysis approach. Starting gene sets are analyzed *via* DAVID and FunCat enrichment in functional annotations.

Fungal Ortholog Identification

We used a maximum likelihood fungal phylogeny that was constructed from a concatenated alignment of 153 universal orthologs in 42 fungal genomes [18]. Thirty-eight of these species were analyzed in this study (Table **S1**).

Fungal orthologs for *S. cerevisiae* targets of MBF and SBF were identified by mining PhylomeDB [19], a public repository of complete collections of gene phylogenies. This database utilizes combination of multiple approaches for phylogeny reconstruction, including maximum likelihood or Bayesian tree inference, alignment trimming and evolutionary model testing. When the internal identifiers were not easily convertible to standard gene names, we used the RSAT suite and looked for mutually highest scoring ortholog pairs [20].

SCOPE Runs

For each run, standard gene names were used as input. A fixed upstream sequence length of 800 bps was used since it corresponds to most frequently used upstream region size for fungal analyses [20].

Motif Comparison

STAMP [21] allows users to query defined motifs against databases of known motifs or against user-provided datasets. For each input motif, an alignment between the motif and the known transcription factor binding site or the user-specified motif is performed and the p-value of an alignment (that is a relative measure of motif similarity) is calculated. The p-values are calculated based on the methods from Sandelin and Wasserman [22] where 10,000 simulated matrix modules were constructed for the analysis of score significance given the lengths of aligned matrices. This extensive analysis enables the assignment of empirical p-values to the alignment scores.

We first used STAMP to calculate the similarity between *S. cerevisiae* computationally predicted motifs (from SCOPE) and biologically verified transcription factor binding sites from SGD [23] in order to identify the best matches. Then we calculated similarities between these best-matching SCOPE motifs for *S. cerevisiae* and highest-scoring motifs from each SCOPE run of corresponding orthologs in other fungal species. The *S. cerevisiae* motif served as an input motif and was compared to a user-defined dataset of SCOPE motifs from each fungal ortholog run. For each motif predicted by SCOPE, its PWM (position weight matrix) was converted into Jaspar/Consite input format supported by STAMP in which motif representation is preceded by a >-containing line which lists the motif name. The actual motif representation has 4 rows of characters that begin with the DNA letter represented by the frequencies in this row[1]. STAMP analysis was done with the following default parameters: Pearson Correlation Coefficient for column comparison metric, ungapped Smith-Waterman for alignment method, iterative refinement for multiple alignment strategy and UPGMA for tree-building algorithm.

[1] The authors wish to thank Piotr Teterwak for help in converting the SCOPE output data into an appropriate input format for STAMP.

Module (Motif Pattern) Analysis

In addition to identifying individual upstream sequences, we were also interested in analyzing patterns of motifs (or modules). We define a module as the arrangement of two motifs separated by a conserved distance. A module may consist of two of the same motifs (homotypic module) or two different motifs (heterotypic module)..

In order to distinguish between different modules, we have developed a comprehensive approach that calculates a module score reflecting the overall quality of a module[2]. In its current implementation, the module score includes five components that are assigned a weight from 0 to 1 and are combined to generate the overall module score from 0 (denoting a low quality module) to 1 (denoting a high quality module). The five components are as follows:

1) Sig value, which is the sum of Sig values for two motifs comprising the module. The Sig value of a motif is a significance value as calculated by SCOPE that is an indication of the overrepresentation and position distribution of a motif in the gene set compared to the rest of the genome. For detailed description of Sig value, see [24];

2) ABC score, a measure of how non-random the intermotif distance distribution is for the entire set of instances of the module. ABC stands for Area Between Curves and indicates how non-random the intermotif distribution for the actual modules is in comparison to randomly generated modules;

3) Module position, a measure of how non-random the position distribution of the module is for the entire set of module instances;

4) Coverage which is equal to the fraction of genes in the set that contain at least one instance of a given module;

5) Motif orientation within modules, a measure of the asymmetry of the occurrences of a module with both motifs in the same orientation vs. motifs in different orientations.

Functional Enrichment Analysis

Functional annotation tools DAVID [25, 26] and FunCat [27] were used to calculate functional enrichment in MBF/SBF-regulated gene sets and their orthologs. DAVID functional enrichment calculation was done for the *S. cerevisiae* and *Schizosaccharomyces pombe*. FunCat was used for *Neurospora crassa*, *Ustilago maydis* and *Fusarium graminearum*. Different tools were used for different species because only a limited selection of species could be analyzed by a single tool.

RESULTS

SCOPE Analyses of MBF and SBF Target Gene Sets

Sets of MBF-regulated *S. cerevisiae* genes and their orthologs in other fungal species were analyzed *via* SCOPE. According to SGD [23], the sequence of the binding site for Mbp1, the DNA-binding subunit of MBF is ACGCGT. From the *S. cerevisiae* SCOPE run, we identified the ACGCGTH motif as the best match to ACGCGT, with a STAMP p-value of 8.3×10^{-10}. We compared SCOPE-predicted motifs from

[2] Manuscript in preparation.

Table 1. Putative Fungal MBF Regulatory Motifs Identified by SCOPE

Species	Motif	Sig Value
Ashbya gossypii	DA**CRCG**W	77.1
Aspergillus fumigatus	DA**CGCG**Y	41.5
Aspergillus nidulans	A**CGCG**TB	82.4
Aspergillus terreus	A**CGCG**	73.2
Botrytis cinerea	**CGCG**WNH	78.1
Candida albicans	A**CGCG**	65.5
Candida dubliniensis	DD**CGCG**W	92.4
Candida glabrata	A**CGCG**TH	51.0
Candida guilliermondii	**CGCG**NH	62.0
Candida lusitaniae	A**CGCG**TH	71.3
Candida tropicalis	A**CGCG**	90.4
Chaetomium globosum	D**CGCG**HY	74.3
Coccidioides immitis	DW**CGCG**W	106.0
Coprinopsis cinerea	A**CGCG**	37.8
Cryptococcus neoformans	RW**CGCG**W	66.6
Debaryomyces hansenii	**CGCG**NH	78.8
Fusarium graminearum	D**CGCG**HH	96.1
Fusarium verticillioides	D**CGCG**NY	55.0
Histoplasma capsulatum	**CGCG**TB	43.1
Kluyveromyces lactis	DHDW**CGCG**W	94.4
Kluyveromyces waltii	A**CGCG**TNH	126.6
Magnaporthe grisea	**CGCG**WH	79.7
Neurospora crassa	D**CGCG**HH	70.3
Rhizopus oryzae	D**CRCG**HH	42.8
Saccharomyces bayanus	HNA**CGCG**W	334.7
Saccharomyces castellii	A**CGCG**WH	243.0
Saccharomyces cerevisiae	A**CGCG**TH	374.8
Saccharomyces kluyveri	DNW**CGCG**W	126.5
Saccharomyces kudriavzevii	A**CGCG**TND	262.5
Saccharomyces mikatae	A**CGCG**	247.1
Saccharomyces paradoxus	D**CGCG**TB	331.8
Schizosaccharomyces pombe	DW**CGCG**W	54.8
Sclerotinia sclerotiorum	D**CGCG**HH	111.7
Stagonospora nodorum	A**CGCG**	39.4
Trichoderma reesii	D**CGCG**HH	68.7
Uncinocarpus reesii	A**CGCG**TB	43.6
Ustilago maydis	A**CGCG**	29.7
Yarrowia lipolytica	**CGCG**NH	24.7

The CGCG core is marked in **bold**. Sig value is the SCOPE measure of the statistical significance of a predicted motif.

other fungal runs to ACGCGTH *via* STAMP [21]. All the best STAMP hits were high-scoring motifs with a CGCG tetranucleotide core (Table **1**). The median Sig value for all fungal runs was 75.7 which is approximately equivalent to the p-value of 1.6×10^{-23} ($\sim 1/2^{75.7}$). These motifs displayed strong sequence similarity to *S. cerevisiae* SCOPE-predicted motifs (Table **S2**). STAMP p-values of alignments between the *S. cerevisiae* SCOPE motif for MBF (ACGCGTH) and similar motifs from other fungi were in the range of 10^{-5} – 10^{-12}, with the median p-value of 5.5×10^{-9}. These data imply that the MBF motif has been highly conserved during fungal evolution.

The same analyses were done for the sets of SBF-regulated *S. cerevisiae* genes and their fungal orthologs. In terms of the binding site, we used the sequence CRCSAAA which is a composite of two SGD motifs (CACGAAAA and CGCSAAA) for Swi4, the DNA-binding component of SBF. Among *S. cerevisiae* motifs predicted by SCOPE, CRCGARAD had the highest similarity to CRCSAAA, with a STAMP p-value of 5.9×10^{-11}. We used STAMP to compare motifs from all other fungal SCOPE runs to the CRCGARAD sequence. The list of the best STAMP matches was used to assemble heterogeneous motifs as shown in Table **2**. The median Sig value in this case was much lower (18.9) than for MBF motifs, but still highly significant with an approximate p-value of 2.0×10^{-6} ($\sim 1/2^{18.9}$). Similarly, STAMP p-values were less significant, being in the range of 10^{-1} – 10^{-14}, with the median p-value of 3.7×10^{-3} (Table **S3**).

Analysis of Position Distributions of Putative MBF and SBF Binding Sites

We investigated the position distributions of putative MBF-binding sites in the upstream regions of the analyzed fungi (Fig. **2**). It is apparent that the CGCG-containing motif typically occupies a preferred location in the first 200 nucleotides upstream of the gene start and does not display a uniform distribution across the upstream region as a whole. On average across all species, 51% of all occurrences of CGCG-containing motifs were in the 0-200 upstream region (Table **3**). Thirty-one out of 38 species had a majority of MBF motif occurrences in the 0-200 quartile (7 remaining species had a majority of MBF motif instances in the 201-400 quartile).

Analysis of positional data for the putative SBF-binding sites showed a less distinct distribution in the upstream regions of the fungal gene sets (Fig. **3**). In this case, a plurality of motif instances in all species (32%) was found in the 201-400 upstream region (Table **4**). Only 15 out of 38 species had a majority of SBF motif occurrences in the 201-400 quartile. Of the remaining 23 species, 7 had most SBF instances in the 0-200 quartile, 7 in the 401-600 quartile, and 8 in the 601-800 quartile.

MBF and SBF Motif Sequence Conservation across Fungal Evolution

We compared MBF motif sequences across all species studied (Fig. **4**). There appears to be a trend towards conservation of the MBF binding site in general and its CGCG core in particular throughout fungal evolution.

Similar analysis of SBF motifs across fungal evolution showed several interesting patterns (Fig. **5**). A majority of

Table 2. Putative Fungal SBF Regulatory Motifs Identified by SCOPE

Species	Motif	Sig Value
Ashbya gossypii	RCACRCGAAA	15.7
Aspergillus fumigatus	TTCTT	28.6
Aspergillus nidulans	TTTCC	16.9
Aspergillus terreus	TCTNNNCCC	22.1
Botrytis cinerea	MCGCACGA	18.2
Candida albicans	AGANNNNNAAA	20.6
Candida dubliniensis	HNYWCRTT	40.0
Candida glabrata	TTTSGCR	11.9
Candida guilliermondii	AGAAAA	12.4
Candida lusitaniae	GCGANNNNNNCCC	13.3
Candida tropicalis	SRVRAGAGA	106.8
Chaetomium globosum	DCRCGWY	29.2
Coccidioides immitis	TCGAHAW	22.2
Coprinopsis cinerea	ACGCG	13.6
Cryptococcus neoformans	GADAATANA	17.2
Debaryomyces hansenii	CHTCTC	14.3
Fusarium graminearum	TTTCTT	18.5
Fusarium verticillioides	GAANNAAA	14.1
Histoplasma capsulatum	AAYYAAA	24.5
Kluyveromyces lactis	GAAAAA	25.4
Kluyveromyces waltii	CVSGAAD	22.4
Magnaporthe grisea	TTYGCSTC	11.1
Neurospora crassa	CGTCGCC	14.4
Rhizopus oryzae	CKCGAAAA	16.1
Saccharomyces bayanus	CRCGARA	75.2
Saccharomyces castellii	CSCGAMA	27.3
Saccharomyces cerevisiae	CRCGARAD	85.1
Saccharomyces kluyverii	GCGRRM	18.2
Saccharomyces kudriavzevii	CRCGAAA	87.0
Saccharomyces mikatae	CRCGAAA	86.4
Saccharomyces paradoxus	CRCGARAH	93.9
Schizosaccharomyces pombe	TTTNNNNNYTC	19.2
Sclerotinia sclerotiorum	CNHTTH	48.6
Stagonospora nodorum	AGCGCG	26.1
Trichoderma reesii	CGAGCA	12.4
Uncinocarpus reesii	TYTWYCRCS	16.1
Ustilago maydis	CTCNNNTTC	13.2
Yarrowia lipolytica	CCANNACC	14.8

fungi, from *Rhizopus oryzae* to *Candida albicans*, did not display a highly specific conserved sequence. Instead, 16 out

Analysis of Motif Patterns in MBF and SBF Gene Sets

In order to evaluate potential conserved patterns of two motifs (which we call modules) for MBF and SBF gene sets, we analyzed homotypic modules formed by two adjacent instances of MBF/SBF binding site in each species. We then calculated module scores and other relevant module statistics such as number of occurrences per gene, median distance between motifs within a module, coverage, median upstream position and ABC score (see Materials and Methods). Complete data for MBF and SBF gene sets are shown in Tables **S4** and **S5**, respectively. According to median statistics, a typical MBF module was consistently different from a typical SBF module in terms of higher module score, smaller inter-motif median distance, higher coverage, closer location to the transcription start site and higher ABC score. For example, a typical MBF module was formed by two instances of MBF binding sites separated on average by 38.6 ± 31.4 bps, present in 46% of the genes in the gene set and located at 218 bps upstream. A typical SBF module had an average inter-motif distance of 78.6 ± 63.2 bps, was found in 38% of the genes in the gene set and was located at 386 bps upstream.

We have also investigated the behavior of the module statistics across four major fungal groups derived from the phylogeny in [18]: 1) Scer (*S. cerevisiae*)-like fungi comprising primarily *Saccharomyces* species; 2) Ncra (*Neurospora crassa*)-like fungi combining *Sordariomycetes* and *Leotiomycetes* representatives; 3) Afum (*Aspergillus fumigatus*)-like species comprising *Eurotiomycetes* species including members of *Aspergillus* genus; and 4) Calb (*Candida albicans*)-like fungi combining *Candida* species. For each module statistic, we subdivided all fungal data into top and bottom halves and looked at the distribution of the species from a given fungal group according to their values. This way, we were able to associate each fungal group with a unique pattern of module criteria. While some module metrics may behave similarly in different fungal groups, their combinations result in unique profiles distinguishing different sets of closely related species from each other. For MBF, module median position was the least informative (Table **5**), whereas for SBF that was true for module score (Table **6**).

Functional Enrichment Analysis of MBF and SBF Gene Sets

We analyzed the original MBF input genes from *S. cerevisiae* in terms of their functional annotations. According to DAVID [25, 26] and FunCat [27] analyses, *S. cerevisiae* targets of MBF were generally enriched in various cell cycle and DNA-related processes, such as DNA metabolism, processing, recombination and repair. Functional annotation analyses of the orthologs of *S. cerevisiae* MBF-regulated genes in other species showed enrichment in similar functional terms (Table **S6**).

S. cerevisiae genes regulated by SBF were generally responsible for cell wall organization and biogenesis. The analysis of orthologs of these genes in other fungi showed that, together with cell cycle and DNA processing functional

categories, they were also involved in biogenesis of cellular components, budding and cell polarity (Table **S7**).

For MBF analysis, we separated the starting *S. cerevisiae* MBF-regulated gene set into a subset with modules and a subset without modules. We compared these subsets to each other and to the original (undivided) gene set in terms of functional enrichment. Most of the functionality of the starting gene set could be accounted for by looking only at the genes with modules which had multiple functional terms with better statistics than in the original gene set (Table **7**). The subset of starting genes with modules was very similar to the original gene set (in terms of the enrichment in cell cycle- and DNA metabolism-related processes) and while the

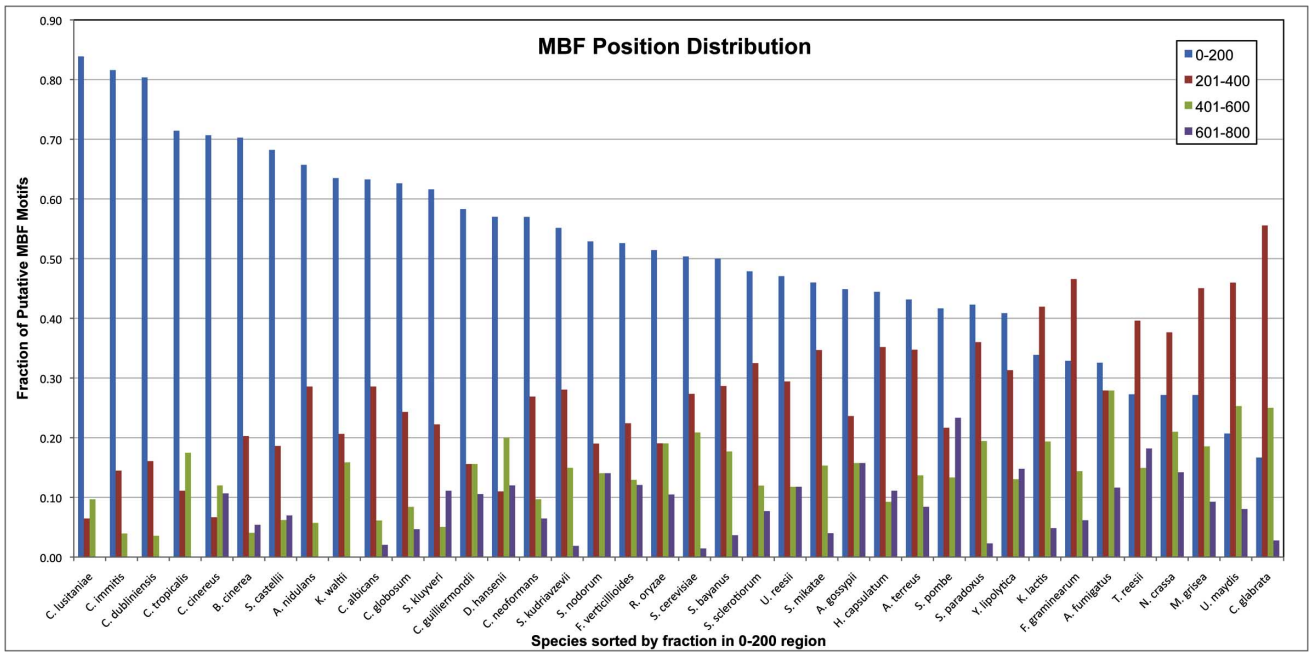

Fig. (2). Position distribution of putative MBF binding sites predicted by SCOPE for upstream regions of 38 fungal species.
X-axis shows fungal species and y-axis shows fraction of putative MBF motifs in each of four different upstream quartiles for each species.

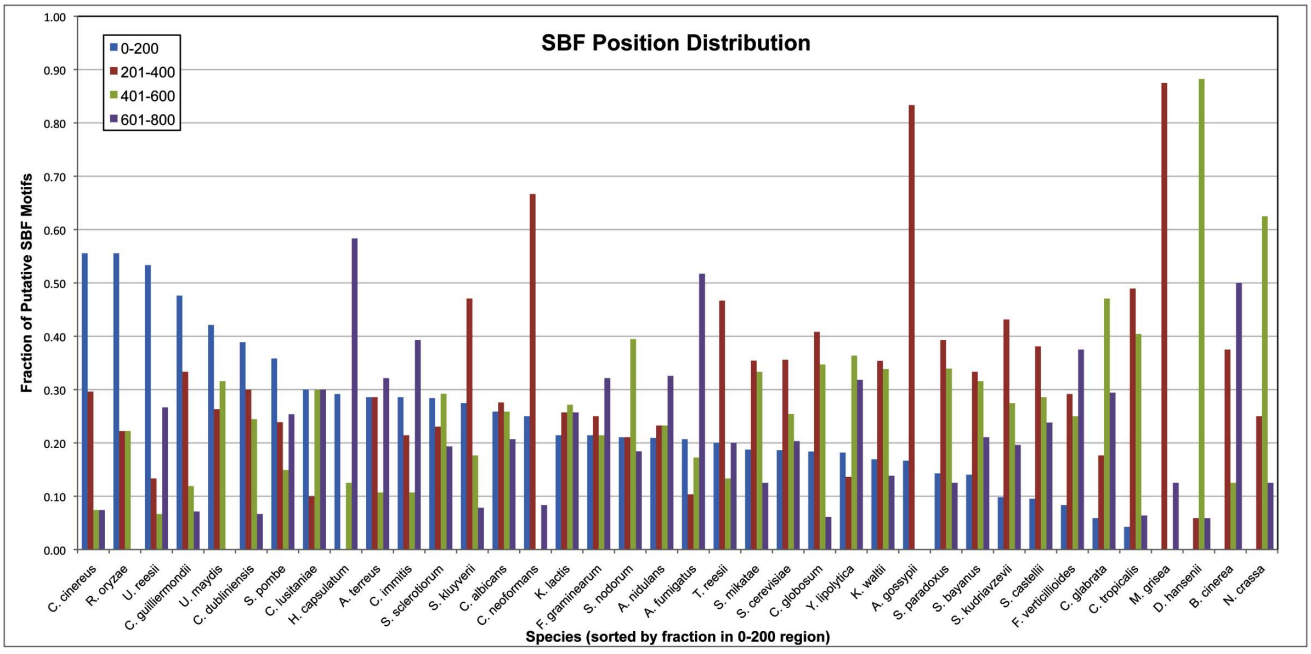

Fig. (3). Position distribution of putative SBF binding sites predicted by SCOPE for upstream regions of 38 fungal species.
X-axis shows fungal species and y-axis shows fraction of putative MBF motifs in each of four different upstream quartiles for each species.

Table 3. Distribution of Occurrences of MBF-like Motifs for four Upstream Quartiles of Orthologs of *S. cerevisiae* MBF-regulated Genes

Motif Quartiles	Mean	Standard Deviation	Standard Error
0-200	0.51	0.17	0.03
201-400	0.27	0.11	0.02
401-600	0.14	0.06	0.01
601-800	0.08	0.06	0.01

Mean value represents average fraction of occurrences for all fungal species for a given quartile region.

Table 4. Distribution of Occurences of SBF-like Motifs for Four Upstream Quartiles of Orthologs of *S. cerevisiae* SBF-regulated Genes

Motif Quartiles	Mean	Standard Deviation	Standard Error
0-200	0.22	0.15	0.02
201-400	0.32	0.18	0.03
401-600	0.25	0.17	0.03
601-800	0.21	0.14	0.02

Mean value represents average fraction of occurrences for all fungal species for a given quartile region.

Table 5. MBF Module Profiles for 4 Major Fungal Groups

Module	Fungal groups			
	Scer-like	Ncra-like	Afum-like	Calb-like
Module score	+	=	-	+
Occurrences/gene	-	+	-	-
Median distance	+	=	+	-
Coverage	=	+	-	-
Median position	=	=	=	+
ABC score	+	-	+	+

"+" means that most species in this group have a value of the statistic which is higher than median. "-" means that most species in this group have a value of the statistic which is lower than median. "=" means that equal number of species in this group has a value that is higher or lower than median. Scer - *Saccharomyces cerevisiae*, Ncra – *Neurospora crassa*, Afum – *Aspergillus fumigatus*, Calb – *Candida albicans.*

Table 6. SBF Module Profiles for 4 Major Fungal Groups

Module Statistic	Fungal Groups			
	Scer-like	Ncra-like	Afum-like	Calb-like
Module score	=	=	=	=
Occurrences/gene	-	=	=	+
Median distance	+	-	-	+
Coverage	+	-	+	-
Median position	=	=	-	+
ABC score	+	-	=	=

"+" means that most species in this group have a value of the statistic which is higher than median. "-" means that most species in this group have a value of the statistic which is lower than median. "=" means that equal number of species in this group has a value that is higher or lower than median. Group designations refer to shortened names of their representative members: Scer - *Saccharomyces cerevisiae*, Ncra – *Neurospora crassa*, Afum – *Aspergillus fumigatus*, Calb – *Candida albicans.*

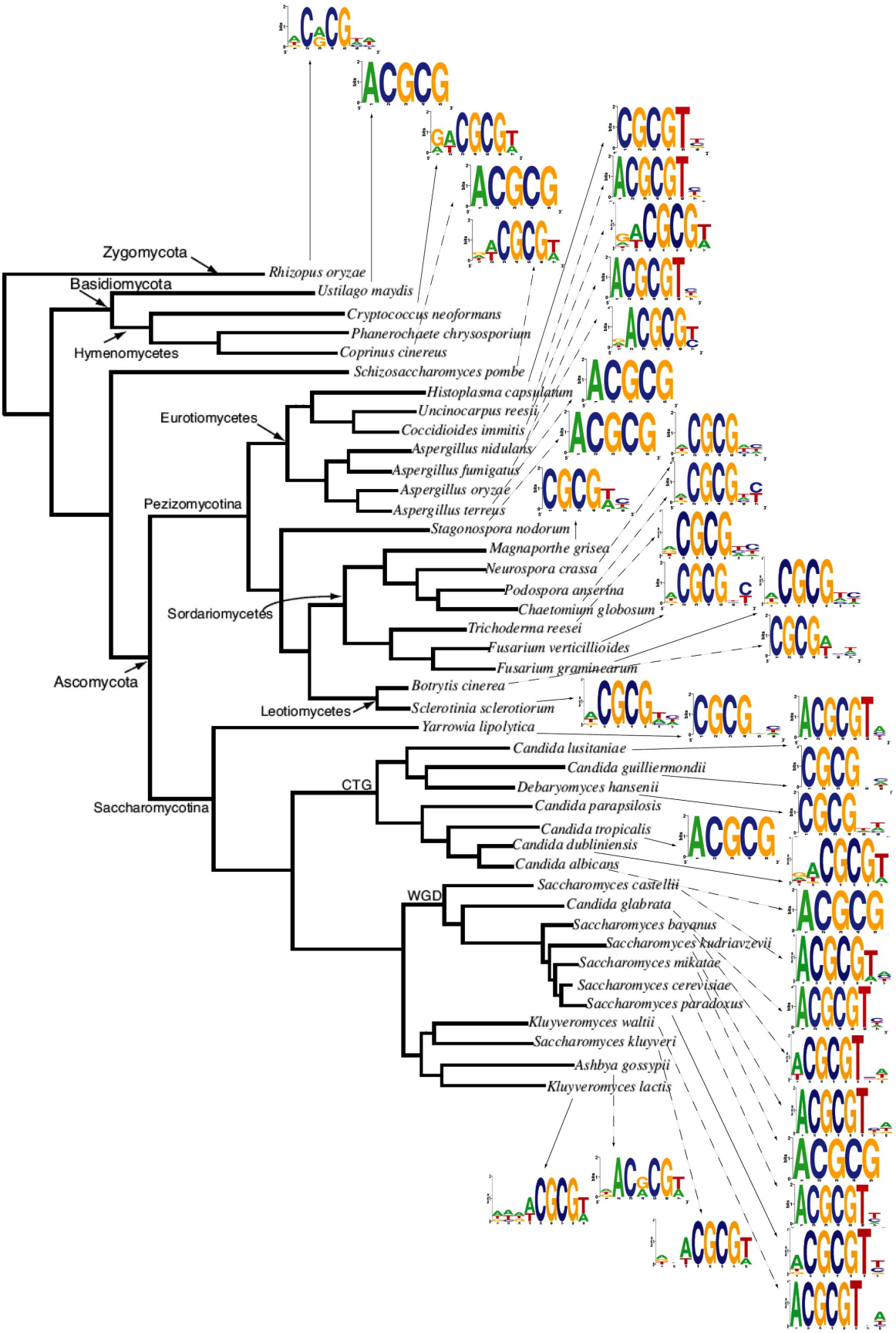

Fig. (4). Distribution of the best MBF STAMP matches for fungal phylogeny.
Sequence logos represent the best SCOPE matches to SGD MBF sequence for each species.

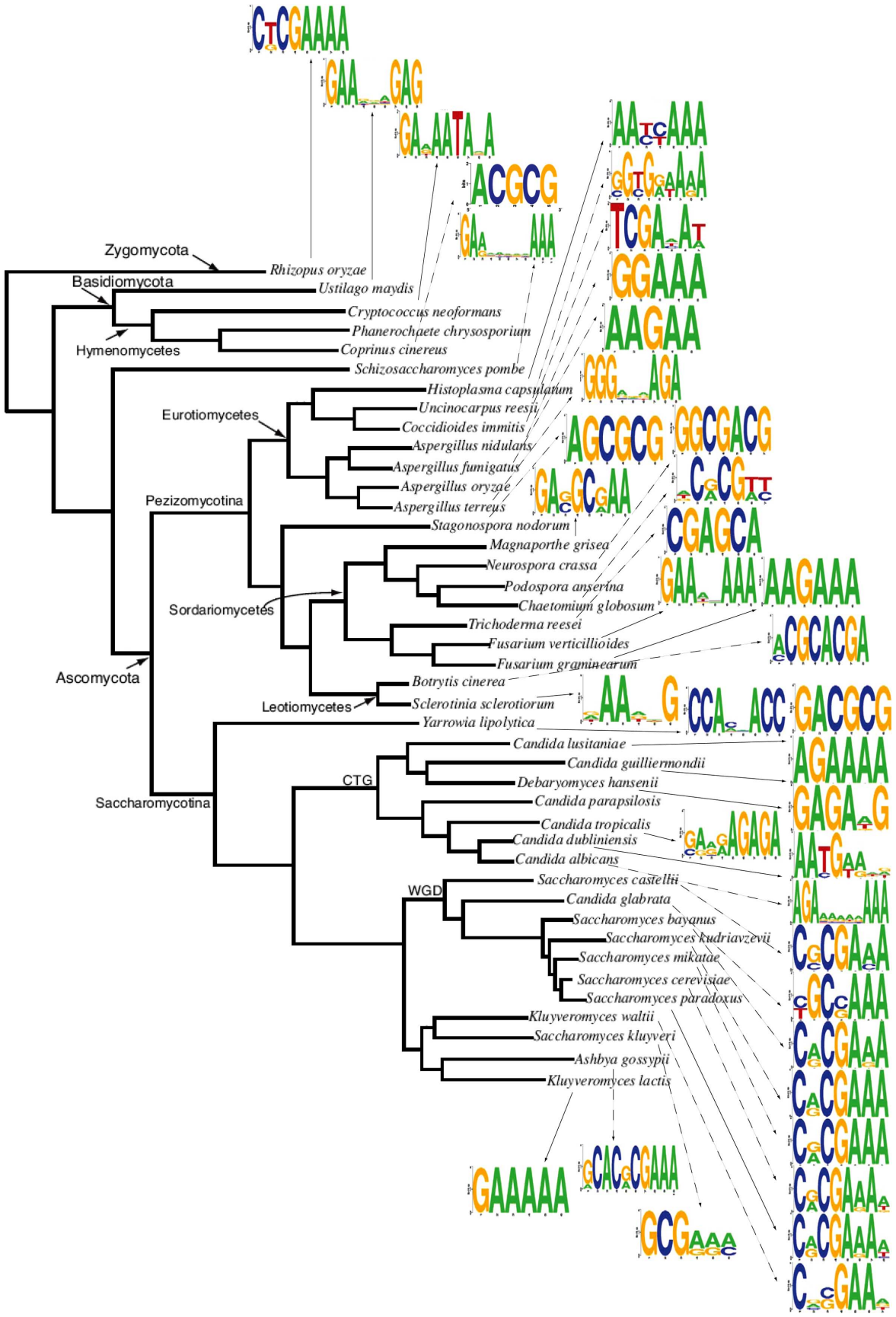

Fig. (5). Distribution of the best MBF STAMP matches for fungal phylogeny.
Sequence logos represent the best SCOPE matches to SGD MBF sequence for each species.

subset of genes without modules was enriched only in meiosis. We obtained similar results in several other fungi: subsets with modules were enriched in DNA-related processes and subsets without modules were enriched in a variety of functional terms including several processes related to protein metabolism (Table **S8**).

For SBF-regulated genes and their orthologs, we were unable to determine a specific relationship between module presence or absence and a set of functional terms for the corresponding subset. While for *S. cerevisiae* genes having modules did not show any functional enrichment, for *Ustilago maydis* genes with and without modules showed association with very different functional annotations (Table **S9**).

DISCUSSION

We have studied the conservation of transcriptionally regulated gene sets of two important *S. cerevisiae* transcription factors, MBF and SBF, responsible for the G1-S transition during the cell cycle. We have analyzed *S. cerevisiae* genes and their orthologs in 37 other fungal species. We have specifically examined significant upstream motifs, their position distribution, evolutionary conservation and functional enrichments.

We believe that these analyses enhance current understanding of the regulation of transcription by MBF and SBF. While the results indicate the existence of a certain similarity and overlap between the two sets of regulated genes, there are also important differences.

Both MBF and SBF datasets displayed enrichment in CGCG-containing motifs, although on average MBF motifs were more overrepresented in the upstream regions and much more similar to the known MBF binding site than their SBF counterparts. While the presence of the CGCG core seems to be important for the regulation of transcription of SBF gene targets or their orthologs, multiple STAMP matches did not have this particular submotif but instead contained a GAA core thus being more similar to the second part of the known SBF binding site, CRCSAAA. Despite relatively low significance and sequence alignment scores, upstream regions of SBF-regulated genes displayed numerous instances of multiple motifs similar to the known SBF regulatory sequence.

In terms of the positional analysis of the motifs, both MBF- and SBF-regulated gene sets showed preferential enrichment in the first (0-200) and second (201-400) upstream quartiles. However, the relative proportions were distinct: the clear majority of MBF motifs appeared in the first quartile, whereas the relative majority of SBF motifs were found in the second quartile. Additionally, for MBF the overwhelming majority of species had most motif sequences in the first quartile, whereas almost the same number of species had a majority of SBF occurrences in any one of three quartiles. It is likely that the position distributions of the MBF and SBF target sites are a distinguishing factor in gene regulation.

It is possible that the presence of multiple occurrences of the same motif within a relatively fixed distance from each other might be important for the successful regulation of the transcription. In *Neurospora crassa*, a motif that is necessary for mediating light induction of the clock gene has been shown to occur in numerous instances separated by a relative conserved distance in upstream regions of early light-responsive genes [28]. In our study, these combinations of the same motif generally displayed higher module scores for MBF than for SBF, which could be explained by simpler sequence composition of a typical MBF motif compared to its SBF counterpart. The modules displayed a particular behavior across different fungal groups, resulting in specific regulatory patterns different for each set of closely related fungi and different between MBF and SBF.

Finally, while sets of MBF- and SBF-regulated genes shared enrichment in cell cycle-related processes, they also displayed the presence of unique functional terms: DNA-related processes for MBF targets and cell wall and cellular component-related processes for SBF targets. Thus, the two transcription factors might have distinguishable roles in G1-S. In terms of the relationship between moduleness and functional annotation, subsets of MBF-regulated genes with

Table 7. Relationship between MBF Modules and Gene Function in *S. cerevisiae*

Gene Set	Unique and Enriched Categories	Note	p-value Fold Change
With modules	cell cycle	vs. no modules	3.70E+00
	cell cycle and DNA processing	vs. no modules	7.20E+03
	DNA binding	unique	
	DNA damage response	unique	
	DNA processing	vs. all	1.20E+01
	DNA recombination and repair	vs. all	6.70E+00
	DNA repair	vs. all	8.20E+00
	DNA synthesis and replication	vs. all	1.00E+01
	extension/polymerization activity	vs. all	1.80E+01
No modules	meiosis	vs. all	1.90E+00

P-value fold change means the improvement of p-value either compared to second subset with no modules or to the starting gene set.

modules were clearly enriched in DNA-related processes as opposed to gene subsets without modules. The pattern of these similarities and differences creates a coherent picture of transcriptional regulation for MBF and SBF targets.

On the basis of this study, MBF gene targets are enriched in ACGCG-like motifs mostly occurring in the 0-200 bps upstream region across the entire fungal tree, suggesting that MBF regulation is evolutionarily more ancient. This presence of an ACGCG-like motif seems to be correlated with functionalities related to cell cycle processes. These motifs tend to occur in numerous instances next to each other in the upstream regions of putative MBF gene targets. They tend to occur in genes specifically enriched in cell cycle and DNA-related biological processes.

SBF gene targets are enriched in sequences having a G(A)n core mostly found in the 201-400 bps upstream region. More specifically, the *Saccharomyces* genus is enriched in CRCGAAA motifs possibly indicating that the current mode of SBF transcriptional regulation is a relatively recent phenomenon in fungal evolution. These sequences are also often found in modules. Partial functional overlap with MBF gene targets in terms of cell cycle-related processes is probably caused by a large number of CGCG-like sequences in the upstream regions of the SBF-regulated genes. However, enrichment in CRCGAAA/GAA-core sequences seems to be associated with functional roles in cell wall and cellular component biogenesis. It is interesting to hypothesize that a combination of CGCG tetrad and GAA core, both of which are found separately in more ancient fungi, allowed *Saccharomyces species* and their close relatives to fine-tune the transcriptional program and be able to respond simultaneously to divergent signals. It is possible that the presence of both sequence cores combined into a single DNA motif enabled *Saccharomyces* to simultaneously regulate cell cycle transition and cell wall biosynthesis thus optimizing the transcriptional regulation of cell cycle.

CONCLUSION

MBF and SBF are transcription factors important for the progression of the cell cycle in *S. cerevisiae*. We have analyzed their gene targets and that of orthologous genes in other fungal species. We were able to associate each set of gene targets with specific significant motifs in their upstream region and their combinations, their unique positional distributions, particular patterns of the evolutionary conservation and divergence of the regulatory motifs and unique functional processes. We believe that our approach of complementing gene functional information with upstream motif positional analysis and its evolutionary pattern has a great potential of elucidating otherwise hidden relationships within phylogenies in terms of interplay between regulatory sequence, biological function for which it is responsible and a history of species in which it occurs.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This research was supported by a grant to RHG from the National Science Foundation, DBI-0445967. The authors wish to thank Jason Moore and Michael Whitfield for thoughtful discussions of the research. We would also like to thank Piotr Teterwak for help with data manipulation.

REFERENCES

[1] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell*, vol. 9, pp. 3273-3297, Dec 1998.

[2] C. Koch and K. Nasmyth, "Cell cycle regulated transcription in yeast," *Curr. Opin. Cell Biol.*, vol. 6, pp. 451-459, Jun 1994.

[3] J. M. Bean, E. D. Siggia, and F. R. Cross, "High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in Saccharomyces cerevisiae," *Genetics*, vol. 171, pp. 49-61, Sep 2005.

[4] M. R. Taba, I. Muroff, D. Lydall, G. Tebb, and K. Nasmyth, "Changes in a SWI4,6-DNA-binding complex occur at the time of HO gene activation in yeast," *Genes Dev.*, vol. 5, pp. 2000-2013, Nov 1991.

[5] F. R. Cross, M. Hoek, J. D. McKinney, and A. H. Tinkelenberg, "Role of Swi4 in cell cycle regulation of CLN2 expression," *Mol. Cell Biol.*, vol. 14, pp. 4779-4787, Jul 1994.

[6] D. Stuart, and C. Wittenberg, "Cell cycle-dependent transcription of CLN2 is conferred by multiple distinct cis-acting regulatory elements," *Mol. Cell Biol.*, vol. 14, pp. 4788-4801, Jul 1994.

[7] I. A. Taylor, P. B. McIntosh, P. Pala, M. K. Treiber, S. Howell, A. N. Lane, and S. J. Smerdon, "Characterization of the DNA-binding domains from the yeast cell-cycle transcription factors Mbp1 and Swi 4," *Biochemistry*, vol. 39, pp. 3943-3954, Apr 11 2000.

[8] S. J. Elledge, "Cell cycle checkpoints: preventing an identity crisis," *Science*, vol. 274, pp. 1664-72, Dec 6 1996.

[9] A. P. Gasch, A. M. Moses, D. Y. Chiang, H. B. Fraser, M. Berardini, and M. B. Eisen, "Conservation and evolution of cis-regulatory systems in ascomycete fungi," *PLoS. Biol.*, vol. 2, p. e398, Dec 2004.

[10] A. Tanay, A. Regev, and R. Shamir, "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 7203-7208, May 17 2005.

[11] E. Zaslavsky and M. Singh, "A combinatorial optimization approach for diverse motif finding applications," *Algorithms Mol. Biol.*, vol. 1, p. 13, 2006.

[12] M. K. Das and H. K. Dai, "A survey of DNA motif finding algorithms," *BMC Bioinformatics*, vol. 8, Suppl. 7, p. S21, 2007.

[13] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites," *Nat. Biotechnol.*, vol. 23, pp. 137-144, Jan 2005.

[14] A. Chakravarty, J. M. Carlson, R. S. Khetani, and R. H. Gross, "A novel ensemble learning method for de novo computational identification of DNA binding sites," *BMC Bioinformatics*, vol. 8, p. 249, 2007.

[15] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99-104, Sep 2 2004.

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

[17] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel, "An improved map of conserved regulatory sites for Saccharomyces cerevisiae," *BMC Bioinformatics*, vol. 7, p. 113, 2006.

[18] D. A. Fitzpatrick, M. E. Logue, J. E. Stajich, and G. Butler, "A fungal phylogeny based on 42 complete genomes derived from

supertree and combined gene analysis," *BMC Evol. Biol.*, vol. 6, p. 99, 2006.

[19] J. Huerta-Cepas, A. Bueno, J. Dopazo, and T. Gabaldon, "PhylomeDB: a database for genome-wide collections of gene phylogenies," *Nucleic Acids Res*, vol. 36, pp. D491- D496, Jan 2008.

[20] J. van Helden, "Regulatory sequence analysis tools," *Nucleic Acids Res.*, vol. 31, pp. 3593 - 3596, 2003.

[21] S. Mahony, and P. V. Benos, "STAMP: a web tool for exploring DNA-binding motif similarities," *Nucleic Acids Res.*, vol. 35, pp. W253-258, Jul 2007.

[22] A. Sandelin and W. W. Wasserman, "Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics," *J. Mol. Biol.*, vol. 338, pp. 207-215, Apr 23 2004.

[23] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein, "SGD: Saccharomyces Genome Database," *Nucleic Acids Res.*, vol. 26, pp. 73-79, Jan 1998.

[24] J. M. Carlson, A. Chakravarty, and R. H. Gross, "BEAM: a beam search algorithm for the identification of cis-regulatory elements in groups of genes," *J. Comput. Biol.*, vol. 13, pp. 686-701, Apr 2006.

[25] G. Dennis, Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol.*, vol. 4, p. P3, 2003.

[26] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protoc.*, vol. 4, pp. 44-57, 2009.

[27] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, and H. W. Mewes, "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes," *Nucleic Acids Res.*, vol. 32, pp. 5539-5545, 2004.

[28] C. H. Chen, C. S. Ringelberg, R. H. Gross, J. C. Dunlap, and J. J. Loros, "Genome-wide analysis of light-inducible responses reveals hierarchical light signalling in Neurospora," *EMBO J*, vol. 28, pp. 1029-1042, Apr 22 2009.

Received: May 05, 2012 Revised: June 30, 2012 Accepted: July 05, 2012