Connecting protein structure with predictions of regulatory sites

Alexandre V. Morozov[†] and Eric D. Siggia

Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021

Communicated by Barry H. Honig, Columbia University, New York, NY, February 21, 2007 (received for review December 6, 2006)

A common task posed by microarray experiments is to infer the binding site preferences for a known transcription factor from a collection of genes that it regulates and to ascertain whether the factor acts alone or in a complex. The converse problem can also be posed: Given a collection of binding sites, can the regulatory factor or complex of factors be inferred? Both tasks are substantially facilitated by using relatively simple homology models for protein–DNA interactions, as well as the rapidly expanding protein structure database. For budding yeast, we are able to construct reliable structural models for 67 transcription factors and with them redetermine factor binding sites by using a Bayesian Gibbs sampling algorithm and an extensive protein localization data set. For 49 factors in common with a prior analysis of this data set (based largely on phylogenetic conservation), we find that half of the previously predicted binding motifs are in need of some revision. We also solve the inverse problem of ascertaining the factors from the binding sites by assigning a correct protein fold to 25 of the 49 cases from a previous study. Our approach is easily extended to other organisms, including higher eukaryotes. Our study highlights the utility of enlarging current structural genomics projects that exhaustively sample fold structure space to include all factors with significantly different DNA-binding specificities.

protein–DNA interactions | homology models of transcription factors | weight matrix predictions

Transcription factors (TFs) are regulatory proteins used by the cell to activate or repress gene transcription. They interact with short nucleotide sequences, typically located upstream of a gene, by means of the DNA-binding domains that recognize their cognate binding sites. As a rule, regulation of gene transcription is analyzed by the bioinformatics methods designed to detect statistically overrepresented motifs in promoter sequences. Intergenic sequences bound by the TF can be identified by using DNA microarray technology, including chromatin immunoprecipitation (ChIP-chip) (1, 2), protein binding (3), and DNA immunoprecipitation (DIP-chip) arrays (4). Of special note is a recent genome-wide study that used ChIP-chip analysis to profile *in vivo* genomic occupancies for 203 DNA-binding transcriptional regulators in *Saccharomyces cerevisiae* (2). Using these data, the authors predicted binding specificities for 65 TFs by using the genomes of related species; a number that was later increased to 98 by MacIsaac *et al.* (5)

The DNA-binding domains of TFs can be classified into a limited number of structural families (6, 7). Structural studies of the protein–DNA complexes reveal that, within each family, the overall fold of the DNA-binding domain and its mode of interaction with the cognate binding site are remarkably conserved, resulting in a characteristic pattern of amino acid contacts with DNA bases. These interactions form the basis of the sequence-specific direct readout of nucleotide sequences by amino acids in the DNA-binding domain. Another contribution to the specificity of protein–DNA interactions is indirect and comes from the curvature imposed on the DNA by the contacts with the DNA phosphate backbone. Those DNA sequences that most readily adapt to the imposed shape will bind most favorably.

In some cases, DNA-binding proteins from the same family recognize sites with similar length, symmetry, and specificity. Alignment of such sites yields a binding profile that can be used to significantly enhance the signal-to-noise ratio in bioinformatics algorithms, allowing for the *ab initio* motif discovery in long metazoan promoter sequences (8–11). However, familial binding profiles are inappropriate when factors form dimers with varying distances between monomer-binding domains, associate with DNA in various orientations, or participate in different multimeric complexes. An averaged site in such cases has little meaning. Furthermore, sufficient binding site data are lacking in many cases. As a result of these complications, only 11 familial profiles are available in the JASPAR database (8), in contrast with 142 protein families that represent TF DNA-binding domains in the Pfam database (12).

Here we present an approach that uses structure-based biochemical constraints to guide motif discovery algorithms. The key observations are (*i*) that the binding site specificity is largely imparted by the contacts between DNA bases and amino acid side chains and (*ii*) that the number of atomic contacts with a base pair is strongly correlated with the degree of conservation of that base pair in the binding site (13). Therefore, a structure of the TF–DNA complex and a single consensus (highest affinity) sequence can be used to predict the position-specific weight matrix (PWM) for the TF (14, 15). We make PWM predictions for 67 *S. cerevisiae* TFs by using homology models (we use "homology" to imply similarity rather than evolutionary relationship). We use these predictions as the informative priors in the Bayesian Gibbs sampling algorithm (16, 17) applied to the intergenic sequences identified with the ChIP-chip experiments (2). We find genomic binding sites for TFs and TF complexes and consider whether the inferred binding specificities provide significant refinements of the PWM, based solely on the structural model. Our study extends the best current yeast regulatory map (2, 5), built around phylogenetic conservation (but employing no structural constraints), by predicting 18 additional TF specificities, correcting 16 previous predictions, and amending 10 others, mostly with respect to the length or the composition of the regulatory complex [see supporting information (SI) Fig. 4].

Our approach also makes it possible to solve the inverse problem of associating binding motifs of unknown origin (e.g., inferred from gene expression arrays by using standard bioinformatics methods) with TFs from the structural database, generating experimentally testable hypotheses about the identity of regulatory inputs. To demonstrate the utility of the inverse approach, we attempted to associate correct TFs with 49 PWMs from MacIsaac *et al.* (5) for which the homology models were available in the structural database. Using a database of structure-based PWMs, we were able to assign a correct fold in 25 cases (10 of which yielded our original structural template) and identified several clear instances of mis-

Author contributions: A.V.M. designed research; A.V.M. and E.D.S. performed research; and A.V.M. and E.D.S. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: DIP, DNA immunoprecipitation; PWM, position-specific weight matrix; TF, transcription factor.

[†]To whom correspondence should be addressed. E-mail: morozov@edsb.rockefeller.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701356104/DC1.

association (e.g., the GCN4 leucine zipper motif assigned to the ARG81 experiment, $P = 4.1 \times 10^{-6}$).

Finally, homology modeling of TFs on the genomic scale allows us to identify future targets for structural studies on the basis of the low similarity of their DNA-binding interfaces and those found in solved protein–DNA complexes. Thus, the goal of the structural genomics projects to sample all protein folds should be extended to include all factors with significantly different specificities from each DNA-binding family. One can then hope to find binding sites for the remaining ≈100 factors in yeast for which the protein localization data are available (2). Our method is computationally efficient, can be applied to any organism, including higher eukaryotes, and may be used independently of or in conjunction with phylogenetic footprinting. We make the structure-based modeling of DNA-binding proteins available through an interactive web site, Protein-DNA Explorer (http://protein-dna.rockefeller.edu).

Results

Overview. We have built a global map of transcriptional regulation in *S. cerevisiae* by using the constraints imposed by the structures of protein–DNA complexes as the informative priors in the Bayesian Gibbs sampling algorithm (see *Materials and Methods* for details). We have supplemented 10 native structures of yeast protein–DNA complexes with homology models for 92 TFs from 14 families (see SI Table 3). Using this data set, we were able to make reliable binding specificity predictions for 67 TFs, 57 of which are modeled by homology (see SI Table 4). Some models were discarded because of the low interface scores, others because it was impossible to infer the dimeric binding mode from the information available in the literature. In addition, we excluded all C2H2 zinc fingers because of the high protein–DNA interface variability in this family (7), which results in few good matches to multidomain zinc fingers in the structural database. Alternative knowledge-based methods for predicting C2H2 zinc finger binding specificities (18, 19) may be more suitable as input to the Gibbs sampling. The structure-based informative priors for the 67 TFs we chose to keep can already explain the ChIP-chip data reasonably well without further refinement (see SI Fig. 5). Knowledge of the structural features at the protein–DNA interface is clearly superior to knowledge of the consensus sequence alone and leads to more accurate PWM predictions (SI Fig. 5). Nonetheless, using the yeast genome to test and refine the initial models enables us to infer additional binding specificity caused by indirect readout and to correct the inaccuracies that are likely to occur as a result of our deliberately simple approach to structural modeling. Sometimes, we test several structural models and keep the one for which the most evidence is found in the genome, effectively using the genomic sequence to infer the structural features of the protein–DNA-binding interface. Furthermore, because we are able to reliably associate TFs with DNA sequence motifs, we can systematically determine the identity of various proteins that form multimeric regulatory complexes. In the remainder of this section, we analyze several representative cases in detail, with the idea of demonstrating the power and flexibility of our method and its utility for understanding transcriptional regulation in yeast and other organisms.

Dimeric Structural Variability and Binding Specificities in the Zn2-Cys6 Family. The Zn2-Cys6 binuclear cluster family is the most common in yeast and other fungi. TFs in this family bind DNA as homodimers and recognize sites of various lengths because of a flexible linker region that joins DNA-binding and dimerization domains (20). This structural flexibility results in the variable spacing between the monomeric 5′-CGG-3′ half-sites and in several possible orientations of the monomeric half-sites with respect to one another (Fig. 1). The observed variability of the binding sites makes it impossible to create a familial binding profile (8) for Zn2-Cys6 TFs. Besides the linker flexibility, additional discrimination between target sites is attributed to the protein–DNA interactions outside the canonical CGG half-sites, often caused by the asymmetric binding of the monomeric subunits (e.g., HAP1 and PUT3 in Fig. 1) (20–22). Gibbs sampling enables us to refine the initial structure-based PWM predictions with sequence data: for example, the HAP1-DNA structure was solved by using the CGC half-sites (21). However, in the Gibbs sampling PWM the half-sites become a mixture of the CGG and CGC triplets, in accordance with prior knowledge about HAP1 binding sites (21).

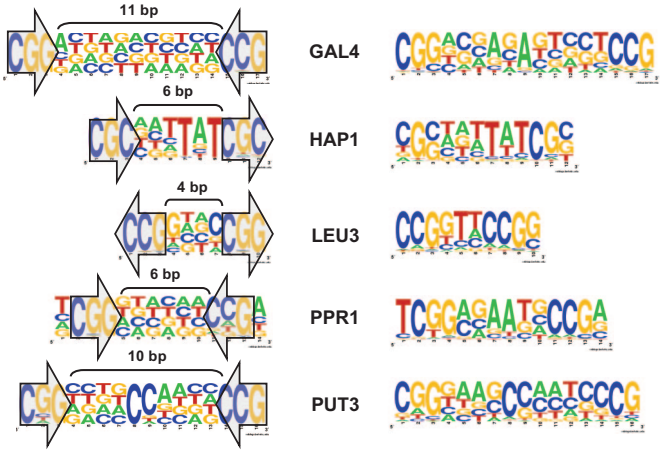


Fig. 1. PWM predictions for five TFs in the Zn2-Cys6 binuclear cluster family, with co-crystal structures showing extensive spacing and orientation variability. (*Left*) Structure-based priors. (*Right*) PWMs refined with Gibbs sampling (see *Materials and Methods*). Arrows show the relative orientation of two monomeric half-sites in the dimeric site from the crystal structure. The monomeric half-sites can be arranged in direct (tail-to-head; HAP1), inverted (head-to-head; GAL4, PPR1, PUT3), and everted (tail-to-tail; LEU3) orientations.

The base in the middle of the 17-bp GAL4 PWM is strongly conserved in the Gibbs prediction, despite the total absence of direct amino acid contacts with DNA bases (Fig. 1) but in accordance with the observed DNA conformational change in the center of the binding site (23). The PUT3-DNA crystal structure has extensive contacts between a β strand from one asymmetrically bound protein subunit and DNA minor groove, resulting in a DNA kink in the middle of the binding site (22). The additional specificity resulting from these contacts is evident from the Gibbs prediction (Fig. 1) and was previously discovered by Siddharthan *et al.* (17) using a phylogenetic approach. PPR1 activates the transcription of genes involved in regulation of pyrimidine levels; it binds extended $TCGGN_6CCGA$ sites: van der Waals contacts are made to the bases flanking the canonical CGG triplets (24). Surprisingly, we could not find many PPR1 sites in the ChIP-chip intergenic regions for which the binding was reported. There is no overlap between ChIP-chip promoter sequences and those of nine genes likely to be involved in the pyrimidine pathway (*URA1–URA8*, *URA10*). Using the latter set, we found canonical sites in seven of these sequences ($S_{track} = 0.68$; SI Table 4). Finally, LEU3 is the only Zn2-Cys6 TF with a native structure to bind an everted repeat: $CCGN_4CGG$. The Gibbs search refines the structure-based model to reveal preference for TT in the middle of the binding site (Fig. 1).

Homology Modeling of the ARG80–ARG81–MCM1 Regulatory Complex. We predicted 14 additional Zn2-Cys6 PWMs by homology (see SI Table 3). For most of these proteins, the homolog has a different spacing between its monomeric half-sites, and the structure-based prior has to be modified accordingly (see SI Table 5). We either obtain information about the binding site length from the literature or explore a range of half-site arrangements. We illustrate our procedure with ARG81, which coordinates the expression of arginine metabolic genes and binds DNA as an ARG80–ARG81–MCM1 complex with the MADS box proteins ARG80 and MCM1

BIOPHYSICS

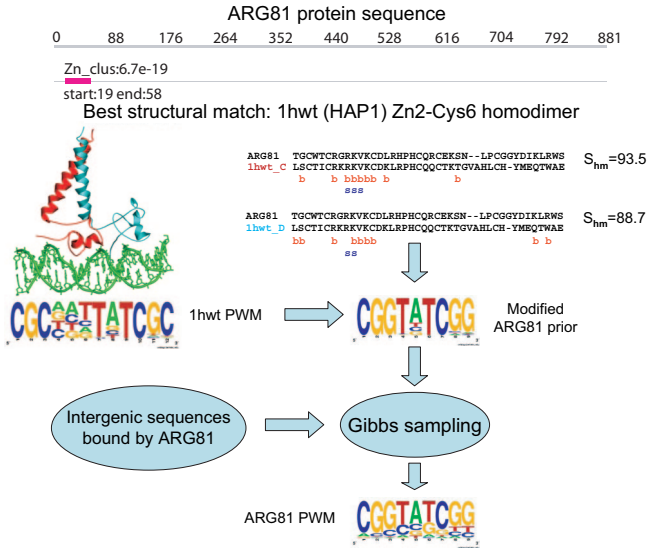


Fig. 2. Illustration of how structural and sequence data are mined in the case of ARG81. A DNA-binding domain of the Zn2-Cys6 binuclear cluster type is found in the ARG81 protein sequence. The HAP1 homodimer (PDB code 1hwt) is identified as the homolog with the highest interface scores S_{hm} (93.5 for chain C, 88.7 for chain D). The interface scores reflect the similarity of the HAP1 and ARG81 DNA-binding interfaces on the basis of their protein sequence alignments. Interface amino acids are labeled ''b'' for the DNA phosphate backbone contacts and ''s'' for the DNA base contacts. Observed amino acid mutations at the interface are sufficiently conservative and thus are assumed not to change the binding specificity significantly. However, to approximate previously characterized ARG81 binding sites (26), columns 4–6 are removed from the HAP1 PWM, and the CGC half-sites are replaced by the more common CGG half-sites. The 1hwt-based PWM modified in this way is used as the informative prior for the Gibbs sampling algorithm, which is run on the intergenic sequences known to be bound by ARG81 from the ChIP-chip experiment (2). After the ARG81 sites are identified, their alignment is used to compile the ARG81 PWM. Each site in the alignment is weighted by its posterior probability $p(s, c)$ (>0.05).

(25). Our structure-based approach allows us to differentiate among regulatory inputs from the different TFs involved in the complex.

We modeled ARG81 by using the HAP1 structure but had to make the dimer spacing consistent with the known literature sites (26) (Fig. 2). Because the reported binding sites were not delineated clearly enough to deduce the spacing unambiguously, we used alternative models with 3, 4, and 5 bp between the half-sites. The prior with 3 bp was chosen because it yielded most sites with Gibbs sampling, although in principle multiple binding modes are possible. The best model for ARG80 is in fact MCM1, which was crystallized in the MATα2–MCM1–MATα2 complex with DNA [Protein Data Bank (PDB; www.rcsb.org/pdb/home/home.do) code 1mnm]. Although the binding interface is almost completely conserved between MCM1 and ARG80, ARG80 has weaker *in vitro* binding affinity for the canonical MCM1 P-box site, CC(A/T)$_6$GG (27). This is attributed to the I21Q and K40R mutations in ARG80, which we classify as phosphate backbone contacts. Because of this interface similarity, we cannot differentiate between ARG80 and MCM1 binding sites. Indicative of the ARG80–ARG81–MCM1 complex formation, ARG80 and ARG81 bind some of the same intergenic regions (Table 1). Consequently, we are able to discover ARG80/MCM1 and ARG81 sites in both sets of intergenic sequences (Table 2). Furthermore, several composite sites are spaced similarly to known ARG80–ARG81–MCM1 sites (two P-boxes with the ARG81 site in between) (25).

For the set of intergenic sequences bound by ARG81, MacIsaac *et al.* (5) predict GCN4 binding specificity (Table 2). This is reasonable biologically, given that GCN4 is a master regulator of amino acid biosynthesis (25, 26), but unreasonable structurally because GCN4 is a leucine zipper homodimer that binds symmetric AP-1 sites [ATGA(C/G)TCAT] *in vivo* (28). GCN4 binding specificity is inconsistent with the prior for ARG81, and the length of the AP-1 site is very different from what is expected for a protein complex. We find GCN4 sites in both ARG80- and ARG81-bound promoter sequences by using a prior based on the structure of GCN4 bound to the AP-1 site (PDB code 1ysa) (Table 2).

Table 1. Partial overlaps between two sets of intergenic regions bound by the members of multiprotein complexes

Experiment 1	Experiment 2	N_1†	N_2†	N_{12}‡	$\langle N_{1*}\rangle$§	$\langle N_{*2}\rangle$¶
ARG80 (YPD)	ARG81 (SM)	29	27	6	0.81	0.48
	GCN4 (SM)	29	143	5	0.81	1.69
INO2 (YPD)	INO4 (YPD)	35	32	14	0.40	0.48
CBF1 (SM)	MET28 (YPD)	195	16	0	1.88	0.14
	MET4 (SM)	195	37	8	1.88	0.70
YOX1 (YPD)	YHP1 (YPD)	61	18	2	2.64	0.15

All intergenic regions are classified by Harbison *et al.* (2) as bound at the $P <$ 0.001 confidence level.
†Number of intergenic regions bound by factor 1 (factor 2) in the ChIP-chip experiment.
‡Number of intergenic regions bound by both factors.
§Average number of intergenic regions shared between factor 1 and all other factors.
¶Average number of intergenic regions shared between factor 2 and all other factors.

Identical Binding Specificities of TF Heterodimers. TFs known to function as heterodimers should bind the same intergenic regions and exhibit identical specificities. For example, the helix–loop–helix proteins INO2 and INO4 form a heterodimer involved in derepression of phospholipid biosynthesis genes in response to inositol deprivation (29). Consistent with the heterodimer formation, we find that 14 of the 35 intergenic regions bound by INO2 are also bound by INO4, many more than expected by chance (Table 1). The INO2–INO4 complex is homologous to the Myc–Max complex bound to the E-box (CACGTG) site (30). Using the Myc–Max-based homology model, we find many E-box sites in the sequences bound by either INO2 or INO4, in agreement with previous studies (2, 5) (see Table 2 and SI Table 4).

Table 2. Summary of DNA-binding specificity predictions for the multiprotein complexes in Table 1

Prior†	Experiment‡	S_{track}§	p_{MF}¶
ARG80	ARG80 (YPD)	0.41	0.90
	ARG81 (SM)	0.22	0.80
ARG81	ARG80 (YPD)	0.65	0.54
	ARG81 (SM)	0.66	0.63
GCN4	ARG80 (YPD)	0.36	0.028
	ARG81 (SM)	0.30	9.1×10^{-6}
INO2-INO4	INO2 (YPD)	0.71	7.7×10^{-8}
	INO4(YPD)	0.80	1.7×10^{-7}
CBF1	CBF1 (SM)	0.98	1.1×10^{-13}
	MET28 (YPD)	0.29	0.29
	MET4 (SM)	0.91	0.65
YOX1-MCM1-YOX1	YOX1 (YPD)	0.74	2.5×10^{-4}
YHP1-MCM1-YHP1	YHP1 (YPD)	0.45	0.39

†TF for which the structure-based informative prior was constructed.
‡Set of intergenic regions (identified by the bound factor and the environmental condition) for which Gibbs sampling with the informative prior was carried out.
§Motif quality as measured by the PhyloGibbs posterior probability $p(s, c)$ averaged over the top 10 sites (all sites if <10 sites were tracked; S_{track} close to 1.0 indicates well defined motifs).
¶Probability that the optimal overlap between our PWM and that of MacIsaac *et al.* (5) is due to chance (see *Materials and Methods*).

Dual Mechanism of Gene Regulation by a Helix–Loop–Helix TF. The helix–loop–helix protein TYE7 is highly homologous to the human sterol regulatory element (StRE; ATCACCCCAC) binding protein (SREBP-1a; PDB code 1am9). The SREBP-1a binding specificity is attributed to the asymmetric homodimer conformation and to the ARG→TYR mutation at the binding interface (31). However, because SREBPs also bind E-boxes *in vitro* with comparable affinity (32), it is conceivable that their *in vivo* function is mediated by both types of sites. The E-box motif is dominant in TYE7 sequences: it is found by using either the vertebrate Max homodimer bound to the E-box (PDB code 1an2) or the SREBP-1a bound to StRE (PDB code 1am9) as the informative prior ($P = 1.2 \times 10^{-5}$). This is not surprising because 28 of the 56 intergenic sequences bound by TYE7 are also bound by another helix–loop–helix factor, CBF1. The CBF1 homodimer is a homolog of Max (1an2) and thus can be expected to bind E-boxes. However, using both 1an2 and 1am9 informative priors in a single Gibbs sampling run reveals a secondary StRE motif ($S_{track} = 0.80$). Thus, TYE7 may bind E-box sites in complex with CBF1 and may bind both E-box and StRE sites as a homodimer.

Related Binding Specificities in the Leucine Zipper Family. We made PWM predictions for 14 TFs in the leucine zipper family. Our analysis shows that many of the binding specificities in this family are similar, and thus only four structural templates are required to model all leucine zippers in yeast (see SI Table 4). As for the Zn2-Cys6 binuclear cluster family, the spacing between the monomeric half-sites is variable. Inasmuch as GCN4 is known to have comparable *in vitro* binding affinity for AP-1 (ATGA(C/G)TCAT) and ATF/CREB (ATGACGTCAT) sites (33), it is likely that the other leucine zippers are also capable of binding both types of sites. For example, all transcriptional regulators in the YAP family (CAD1, CIN5, YAP1, YAP3, YAP5, YAP6, YAP7, and ARR1) (34) are homologous to the PAP1 TF from *Schizosaccharomyces pombe* complexed with the GTTACGTAAC PAP1 site (PDB code 1gd2) (35). Similarly to GCN4, PAP1 homodimers are known to recognize shorter [GTTA(C/G)TAAC] and longer (GTTACGTAAC) sites *in vitro* (34, 35). Using both types of priors, we find more shorter sites for CAD1, YAP1, YAP3, YAP5, and ARR1 and more longer sites for CIN5 and YAP6. For YAP7, there is comparable evidence for both types of sites (SI Table 4), of which MacIsaac *et al.* (5) find only the shorter. Interestingly, only half of the palindromic site is strongly conserved in YAP5, suggesting a contribution from monomeric binding or cofactors. This notion is supported by the significant overlap of the YAP5-bound intergenic regions and those bound by PDR1 and GAT3 (data not shown).

Indirect Recruitment of MET28 and MET4. Two leucine zipper proteins, MET28 and MET4, form regulatory complexes with either the helix–loop–helix protein CBF1 or the highly related zinc finger proteins MET31 and MET32 (36, 37). In particular, the CBF1–MET28–MET4 complex acts as a transcriptional activator in the sulfur–amino acid metabolism and biosynthesis pathway. The MET4 sequence is very diverged (with an e-value of 0.0017 for a match to the $bZIP_2$ family and no homologous structures), whereas the MET28 DNA-binding interface is reasonably similar to GCN4. Nonetheless, it is believed that neither MET28 nor MET4 interacts with DNA directly (37, 38), being recruited instead through association with CBF1, MET31, or MET32. MacIsaac *et al.* (5) assigned MET31/32 binding specificity to MET4 sequences (even though MET4 is a leucine zipper); besides the MET31/32 motif, we also find CBF1 E-box sites in MET4 sequences (Tables 1 and 2).

Surprisingly (and contrary to previous studies), we could not find strong evidence supporting CBF1 and MET31/32 binding in MET28 sequences (data not shown and Tables 1 and 2). MacIsaac *et al.* (5) report a MET31/32 site from the literature for MET28. It is possible that MET28 participates in pathways other than sulfur metabolism by forming complexes with other factors.

Cooperative Binding of the Homeodomain TFs. Another example of synergistic TF action involves proteins in the homeodomain family, which often increase their specificity by forming homo- and heterodimers and interacting with cofactors (39). Our results suggest that all yeast homeodomains (MATα2, CUP9, PHO2, YHP1, and YOX1) employ this strategy to some extent. In particular, CUP9 is homologous to the extradenticle (Exd) homeodomain from the Ubx–Exd–DNA complex in *Drosophila melanogaster* (40). Even though CUP9 shares lower interface similarity with Ubx (see SI Table 4), using the whole complex yields a PWM corresponding to the CUP9 homodimeric binding mode. Similarly, PHO2 is homologous to the homeodomain homodimer from the *Drosophila* protein paired (Pax class) (41). Consistent with this observation and previous footprinting studies (42), we find PHO2 dimeric sites in the intergenic sequences (SI Table 4). Neither complex was found in the previous study (5).

YOX1 and YHP1 are the transcriptional repressors that, together with MCM1, bind early cell cycle boxes found in the promoters of genes expressed during the M/G_1 phase of the cell cycle (43). We used the MATα2–MCM1–MATα2 complex as a prior for both YOX1 and YHP1 bound sequences and found matches in more than half the genes. A previously found YOX1 site (5) corresponds to just the YOX1–MCM1 part of the larger complex, whereas for YHP1, a standard TAAT monomeric homeodomain site from the literature is reported (5).

Inferring the Identity of Regulatory Inputs with the Structure Database. Our method is also useful for associating TFs with independently discovered sequence motifs. By correlating 49 PWMs from MacIsaac *et al.* (5) for which we could build homology models, and 252 structure-based TF binding specificity predictions, we were able to assign the correct protein fold in 25 cases [we considered a fold to be predicted correctly if the *P* value for the correlation between the MacIsaac *et al.* (5) PWM and at least one PWM from the correct fold was among the lowest three]. The correct fold had the lowest *P* value in 16 of 25 instances. In 10 cases, the actual structural template used in homology modeling was among the top three (this measure is less objective because in many cases several homologs are equally plausible). Some of the failures are clearly due to the misassociation of TFs and sets of bound intergenic regions, such as the GCN4 motif reported for ARG81 or the MCM1 motif reported for YOX1 (5). Other motifs are missed because dimers with correct spacings are absent in the structural database. Indeed, by adding 24 PWMs with adjusted spacings (see SI Table 5) we were able to place 28 of 49 motifs into the correct fold and to identify 14 structural templates correctly. Thus, we have developed a computational equivalent of the yeast one-hybrid assay, a useful tool for postprocessing motifs found in microarray data analyses.

Discussion

We have developed a computational approach that uses structure-based TF binding specificity predictions as priors in the Bayesian Gibbs sampling algorithm (a related method was recently described in refs. 44 and 45). Structure-based PWM predictions are based on the observations (*i*) that for most TFs, amino acid conservation at the DNA-binding interface leads to similar specificities and (*ii*) that the number of atomic contacts with the consensus base is a good predictor of the degree of its conservation in the binding site (13). Thus, a simple probabilistic model based on the number of protein–DNA atomic contacts can be used to make PWM predictions by homology. The structural predictions are subsequently refined with sequence data from the ChIP-chip experiments. The sequence-structure approach is computationally inexpensive and thus should be used as a standard bioinformatics tool. If sequences from several species are available, the structural priors can be easily combined with phylogenetic footprinting. Our approach is not limited to the relatively simple structure-based models used in this work: any prior specificity information, including more sophisti-

cated structural predictions and experimentally inferred PWMs, can be used as input to the Gibbs sampling algorithm.

We have observed limited diversity of TF binding specificities: DNA-binding domains with 30–50% overall sequence identity routinely bind similar sites and share high interface homology. Thus, to obtain comprehensive structural coverage for the analysis of transcriptional regulation in any organism, it is sufficient to have just one representative protein–DNA structure for each distinct binding specificity subclass in every TF family. This notion extends to protein–DNA complexes the basic principle of structural genomics projects: at least one structure should be solved for every protein fold.

SI Fig. 4 summarizes our predictions for 67 *S. cerevisiae* TFs and contrasts them with prior work employing phylogenetic footprinting (2, 5). Eighteen of our predictions were not found in the previous study (5), and 26 of the 49 remaining predictions disagree with it either partially (e.g., significantly differing in length) or completely (SI Fig. 4 and SI Table 4). Phylogenetic conservation alone cannot provide a direct link between the physicochemical properties of the TF–DNA complex and the sequence-derived motifs, resulting in cases of "mistaken identity" (e.g., GCN4 PWM assigned to ARG81 and ARG80; Table 2) and incorrectly assigned specificity (YOX1–MCM1 PWM reported as YOX1; ARO80 PWM given as three monomeric sites rather than a dimer).

It is interesting to note that sometimes we do not find sites with expected specificity in the ChIP-chip intergenic sequences, even if our confidence in the informative prior is high (e.g., for PPR1 and NDT80, for which co-crystal structures are available). This can happen if a protein associates with DNA through cofactors (e.g., MET4 in the CBF1–MET28–MET4 complex) or if the binding is dominated by relatively weak, nonspecific interactions. Finally, we note that in all likelihood only a fraction of sites reported here are functional *in vivo*. Additional filters based on phylogenetic conservation (2, 5) and the proximity to the coding regions and cofactor sites can easily be used in specific situations.

Materials and Methods

Structure-Based PWM Predictions. We have built a database of 515 protein structures bound to double-stranded DNA, including 252 transcription factors from 40 families. We predicted PWMs for all protein–DNA complexes in this database by using a simple model that exploits the structure of the protein–DNA complex but does not require any detailed predictions of the protein–DNA energetics (14). The performance of this model was previously found to be only slightly inferior to that of more sophisticated but less computationally efficient all-atom models (14, 15). We construct a PWM by using the consensus DNA sequence and the number of atom–atom contacts, N_i, between all protein side chains and the DNA bases in the base pair i (we use a 4.5 Å distance cutoff; hydrogen atoms are excluded from the counts). We assume that the three nonconsensus bases occur with equal probabilities and that the consensus base is favored over a nonconsensus base by $N_i/N_{\max}$ ($0 \le N_i \le N_{\max}$):

$$w_i^{\alpha}(N_i) = \begin{cases} \frac{1}{4}\,(1 - N_i/N_{\max}) & (\alpha \neq \text{consensus}) \\ \frac{1}{4}\,(1 + 3N_i/N_{\max}) & (\alpha = \text{consensus}) \end{cases}, \qquad [1]$$

where $w_i^{\alpha}(N_i)$ is the probability of base $\alpha = \{A,C,G,T\}$ in the PWM column i, N_i is the number of protein atoms in contact with the base pair i, and $N_{\max}$ is the number of contacts at which the native base pair becomes absolutely conserved: $w_i^{\text{consensus}}(N_i) = 1$, $N_i \ge N_{\max}$. Note that if $N_i = 0$, all four bases are equally likely. $N_{\max}$ is a free parameter of the model; its optimum value was found to be 20 by fits to experimental data (14). Fig. 3 illustrates our method, using PHO4 as an example.

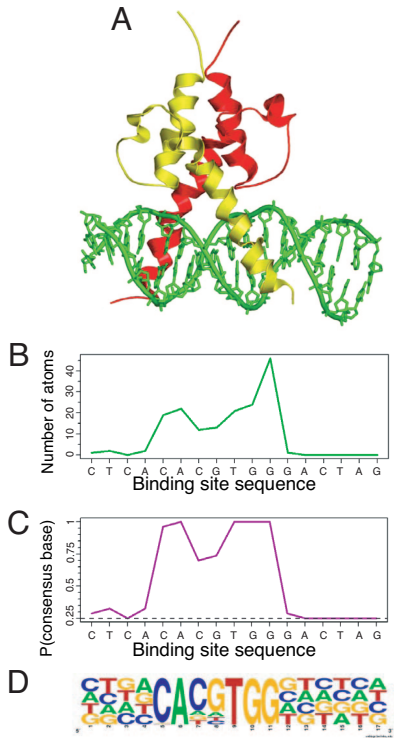


Fig. 3. Prediction of the informative prior for the phosphatase system regulator PHO4. (*A*) Crystal structure of the PHO4 helix–loop–helix dimer bound to its consensus site (PDB code 1a0a). (*B*) Atomic profile: the number of heavy atoms, N_i, within 4.5 Å of base pair i in the binding site. (*C*) Consensus base probability profile: the probability $w_i^{\alpha}(N_i)$ of the consensus base α at position i in the binding site (*cf.* Eq. **1**). (*D*) Structure-based PWM prediction.

TF Homology Modeling. For each structure, we found matches to the protein families in the Pfam database (see *SI Materials and Methods*) and identified amino acids at the DNA-binding interface by using a distance cutoff of 4.5 Å to a DNA base pair. We classified all contacts as DNA base and/or DNA phosphate backbone/sugar ring (a given amino acid can make both types of contacts with one or several base pairs). We then searched all *S. cerevisiae* proteins for matches to those Pfam families with at least one hit in the structural database. This procedure yielded both a Pfam classification of yeast protein factors and the alignments of their sequences with the putative structural homologs, providing information about the amino acids in contact with DNA.

For each sequence-structure protein alignment, we computed the amino acid substitution score S_{hm} at the DNA-binding interface. Given an alignment of the query protein sequence with the target sequence from the protein–DNA structural database, the protein–DNA interface score is given by $S_{\text{hm}} = 0.5\,S_{\text{base}} + 0.5\,S_{\text{bb}}$, where S_{base} and S_{bb} are the DNA base and DNA backbone substitution scores, respectively. These scores are defined as:

$$S_{\text{base/bb}} = \frac{1}{N_{\text{cont}}} \sum_{\text{contacts}} s(\text{aa}_1, \text{aa}_2), \qquad [2]$$

where the sum is over N_{cont} amino acid–DNA contacts (base contacts for S_{base}, backbone contacts for S_{bb}), aa_1 is the amino acid in the query sequence, aa_2 is the amino acid from the target sequence aligned with aa_1 and in contact with DNA, and $s(\text{aa}_i, \text{aa}_j)$ is the PET91 amino acid substitution score (46). Note that the amino acids contacting multiple bases will make a proportionately greater contribution to Eq. **2**. The protein–DNA interface score defined in this way has a range 0–100, with 100 assigned when both

DNA base- and DNA backbone-contacting amino acids are fully conserved in the two-sequence alignment.

In most cases, the structure with the highest interface score was chosen as the structural template. If several structures had comparable homology scores, we chose either the most accurate one (using measures such as resolution of x-ray diffraction) or the one most relevant in the biological context (using information about cofactors and the dimerization state). Computing the score only from amino acids that contact DNA, rather than from entire aligned sequences, assumes that amino acid–DNA interactions are local: if the amino acids at the DNA-binding interface are conserved between two protein–DNA complexes, they will adopt similar geometric arrangements with respect to DNA, regardless of the rest of the protein (7, 47, 48). For example, a comparison of the engrailed and $\alpha 2$ homeodomain–DNA complexes revealed an extensive set of conserved contacts with DNA, even though the amino acid sequences were only 27% identical (7). A more recent study (48) identified a number of cases in which the local interface geometry was conserved, even if DNA conformational change was required in order to accommodate it.

Informative Priors. In a majority of cases, we modeled only those yeast factors for which a protein–DNA complex with an interface homology score S_{hm} exceeding the empirical cutoff of 80 could be found (in multimeric complexes, S_{hm} was averaged over all protein chains). Typically, the corresponding structure-based PWM was used to create the informative prior without further modification. We discarded all cases in which nonconservative amino acid mutations lowered our confidence in the homology template. More structures could be modeled if we were able to predict the new specificity with accurate descriptions of protein–DNA energetics. We also avoided updating the structure-based PWMs by using the information about interface mutations, because predicting the number of atoms in contact with DNA bases would require explicit modeling of mutated side chains. In some cases, for TFs that bind as dimers the spacing and relative orientation of the monomeric subsites were adjusted based on the information about previously characterized factor binding sites. Finally, the structure-based PWMs were multiplied by the total number of pseudocounts $\bar{n}$ and used as the informative priors in the Bayesian Gibbs sampling algorithm (17). We found empirically that setting $\bar{n} = n/2$ (where n is the number of input intergenic sequences that approximates the expected number of binding sites) biases the search fairly strongly toward the expected binding sites but at the same time is weak enough so that the final PWM prediction can be completely different if the genomic sites disagree with the prior. Fig. 2 demonstrates our method for predicting TF binding specificities with sequence and structural data in the case of ARG81.

Gibbs Sampling. We use the PhyloGibbs implementation of the Gibbs sampling algorithm because its Bayesian formulation allows us to take the prior information into account in a straightforward way (17). PhyloGibbs assigns configurations C to the input sequence S; each configuration consists of the nonoverlapping TF binding sites and the background (modeled with a first-order Markov model) and can have multiple site instances for each TF. PhyloGibbs uses simulated annealing to find the configuration C^* with the highest posterior probability, $P(C^*|S)$ (see *SI Materials and Methods*). Once the optimal configuration C* is identified, the algorithm samples the distribution $P(C|S)$ without restrictions to estimate the probability $p(s, c)$ that a site s belongs to a TF c in C^* (17). After this so-called tracking phase, all sites s for which $p(s, c) \geq 0.05$ are reported for each TF in the configuration C^*. Tracking is used as a convenient means of summarizing the information contained in $P(C|S)$ and is able to both discover additional sites and remove spurious sites from the reference configuration C^*.

Sequence Data. We ran PhyloGibbs in the tracking mode on the intergenic sequences from the ChIP-chip experiments in yeast (factors bound with $P < 0.001$) (2) and from the literature (see *SI Materials and Methods*). Each run of the algorithm is used to infer a single PWM, using only *S. cerevisiae* intergenic sequences bound by a specific TF and the informative prior as inputs. In several cases, when none of the sites tracked sufficiently well, we reported sites from the simulated annealing configuration. In this case, the number of sites must be guessed in advance (we assume one site per promoter sequence). Gibbs sampling PWMs are inferred from the alignments of binding sites weighted by $p(s, c)$ (if available).

PWM Correlations. For each forward and reverse complement alignment between the two PWMs, we computed Pearson's correlation coefficients for the PWM log-probabilities (15). To avoid spurious short alignments, we set the minimum allowed PWM overlap to the length of the shorter PWM, minus a constant offset (2 for GATA factors, 4 for other PWMs). We reported the overlap with the lowest P value (Bonferroni-corrected for the number of tested alignments).

We thank Erik van Nimwegen, Harmen Bussemaker, Jonathan Widom, and Amos Tanay for comments on the manuscript and Michael Mwangi and Rahul Siddharthan for useful discussions. A.V.M. was funded by The Lehman Brothers Foundation through the Leukemia and Lymphoma Society; E.D.S. was supported by National Science Foundation Grant DMR-0129848.

1. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreider J, Hannett N, Kanin E, *et al.* (2000) *Science* 290:2306–2309.
2. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, MacIsaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, *et al.* (2004) *Nature* 431:99–104.
3. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML (2004) *Nat Genet* 36:1331–1339.
4. Liu X, Noll DM, Lieb JD, Clarke ND (2005) *Genome Res* 15:421–427.
5. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) *BMC Bioinformatics* 7:113.
6. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) *Genome Biol* 1:R001.
7. Pabo CO, Sauer RT (1992) *Annu Rev Biochem* 61:1053–1095.
8. Sandelin A, Wasserman WW (2004) *J Mol Biol* 338:207–215.
9. Xing EP, Karp RM (2004) *Proc Natl Acad Sci USA* 101:10523–10528.
10. Mahony S, Golden A, Smith TJ, Benos PV (2005) *Bioinformatics* 21(Suppl 1):i283–i291.
11. MacIsaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E (2006) *Bioinformatics* 22:423–429.
12. Kummerfeld SK, Teichmann SA (2006) *Nucleic Acids Res* 34:D74–D81.
13. Mirny LA, Gelfand MS (2002) *Nucleic Acids Res* 30:1704–1711.
14. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) *Nucleic Acids Res* 33:5781–5798.
15. Foat BC, Morozov AV, Bussemaker HJ (2006) *Bioinformatics* 22:e141–e149.
16. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) *Science* 262:208–214.
17. Siddharthan R, Siggia ED, van Nimwegen E (2005) *PLoS Comput Biol* 1:e67.
18. Benos PV, Lapedes AS, Stormo GD (2002) *J Mol Biol* 323:701–727.
19. Kaplan T, Friedman N, Margalit H (2005) *PLoS Comput Biol* 1:e1.
20. Schwabe JWR, Rhodes D (1997) *Nat Struct Biol* 4:680–683.
21. King DA, Zhang L, Guarente L, Marmorstein R (1999) *Nat Struct Biol* 6:64–71.
22. Swaminathan K, Flynn P, Reece RJ, Marmorstein R (1997) *Nat Struct Biol* 4:751–759.
23. Marmorstein R, Carey M, Ptashne M, Harrison SC (1992) *Nature* 356:408–414.
24. Marmorstein R, Harrison SC (1994) *Genes Dev* 8:2504–2512.
25. Messenguy F, Dubois E (2000) *Food Tech Biotechnol* 38:277–285.
26. De Rijcke M, Seneca S, Punyammalee B, Glansdorff N, Crabeel M (1992) *Mol Cell Biol* 12:68–81.
27. Jamai A, Dubois E, Vershon AK, Messenguy F (2002) *Mol Cell Biol* 22:5741–5752.
28. Garcia-Gimeno MA, Struhl K (2000) *Mol Cell Biol* 20:4340–4349.
29. Robinson KA, Lopes JM (2000) *Nucleic Acids Res* 28:1499–1505.
30. Nair SK, Burley SK (2003) *Cell* 112:193–205.
31. Párraga A, Bellsolell L, Ferré-D'Amaré AR, Burley SK (1998) *Structure (London)* 6:661–672.
32. Kim JB, Spotts GD, Shih H, Halvorsen Y, Ellenberger T, Towle HC, Spiegelman BM (1995) *Mol Cell Biol* 15:2582–2588.
33. Keller W, König P, Richmond TJ (1995) *J Mol Biol* 254:657–667.
34. Fernandes L, Rodrigues-Pousada C, Struhl K (1997) *Mol Cell Biol* 17:6982–6993.
35. Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T (2000) *Nat Struct Biol* 7:889–893.
36. Kuras L, Barbey R, Thomas D (1997) *EMBO J* 16:2441–2451.
37. Blaiseau P-L, Thomas D (1998) *EMBO J* 17:6327–6336.
38. Kuras L, Cherest H, Surdin-Kerjan Y, Thomas D (1996) *EMBO J* 15:2519–2529.
39. Wilson DS, Desplan C (1999) *Nat Struct Biol* 6:297–300.
40. Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK (1999) *Nature* 397:714–719.
41. Wilson DS, Guenther B, Desplan C, Kuriyan J (1995) *Cell* 82:709–719.
42. Vogel K, Hörz W, Hinnen A (1989) *Mol Cell Biol* 9:2050–2057.
43. Pramila T, Miles S, GuhaThakurta D, Jemiolo D, Breeden LL (2002) *Genes Dev* 16:3034–3045.
44. Narlikar L, Hartemink AJ (2006) *Bioinformatics* 22:157–163.
45. Narlikar L, Gordân R, Ohler U, Hartemink AJ (2006) *Bioinformatics* 22:e384–e392.
46. Jones S, van Heyningen P, Berman HM, Thornton JM (1999) *J Mol Biol* 287:877–896.
47. Pabo CO, Nekludova L (2000) *J Mol Biol* 301:597–624.
48. Siggers TW, Silkov A, Honig B (2005) *J Mol Biol* 345:1027–1045.

Table 3. List of *S.cerevisiae* transcription factors with homologs in the structural database.

Family	Pfam[♭]	N[§]	$N_t^{‡}$	Transcription factors
Zn2-Cys6 binuclear cluster	Zn_clus	26	20	GAL4⋆, HAP1⋆, LEU3⋆, PPR1⋆, PUT3⋆, ARG81, ARO80, HAL9, OAF1, PDR1, PDR3, PIP2, RDS1, SIP4, STB4, SUT1, UGA3, UME6, YRR1, CHA4, DAL81, MAL33, STB5, THI2, UPC2, WAR1
Basic helix-loop-helix	HLH	8	16	PHO4⋆, CBF1, HMS1, INO2, INO4, RTG3, TYE7, RTG1
Basic region leucine zipper	$bZIP_1$ $bZIP_2$	15	42 38	GCN4⋆, CAD1, CIN5, CST6, ACA1, HAC1, SKO1, YAP1, YAP3, YAP5, YAP6, YAP7, ARR1, MET28, MET4
MADS box	SRF-TF	4	14	MCM1⋆, ARG80†, RLM1,† SMP1†
NDT80 type	NDT80_PhoG	1	11	NDT80⋆
Myb-like DNA-binding domain	Myb_DNA-binding	4	18	RAP1⋆, BAS1, REB1, SNT2
GATA type factor	GATA	7	6	DAL80†, GLN3†, GZF3†, GAT1†, GAT3, ASH1, SRD1
C2H2 zinc finger	zf-C2H2	21	76	ACE2, ADR1, AZF1, FZF1, MET31, MET32, MIG1, MIG2, MOT3, MSN2, MSN4, NRG1, RIM101, RME1, RPH1, RPN4, SFP1, STP1, STP2, SWI5, ZAP1
Heat shock factor type DNA-binding domain	HSF_DNA-bind	4	2	HSF1†, MGA1, HMS2, SKN7
Homeodomain	Homeobox	4	51	CUP9, PHO2, YHP1, YOX1
Fork head domain	Fork_head	3	3	FKH1, FKH2, FHL1
Histone-like transcription factor	CBFD_NFYB_HMF	2	8	HAP3, HAP5
RFX DNA-binding domain	RFX_DNA_binding	1	2	RFX1
High mobility group (HMG) box	HMG_box	2	10	ROX1, IXR1

$bZIP_1$ and $bZIP_2$ are grouped together because their domain matches overlap.

⋆ indicates a TF for which the structure of the protein-DNA complex is available.

† indicates a TF for which a homolog with no mutations at the DNA binding interface is available.

♭ Protein family identifier from the Pfam database; § Number of *S.cerevisiae* proteins in the Pfam family; ‡ Number of domain matches for this family in the structural database.

Table 4. List of PWMs derived using Gibbs sampling with the structure-based informative priors.

TF	PDB	S_{hm}	Env	L	L_{MF}	N_{seq}	$\langle N \rangle$	S_{track}	p_{MF}	C
Zn2-Cys6 binuclear cluster										
GAL4	1d66 (A,B)	100.0/100.0	RAFF	17	18	37	33.3	1.00	1.2e-15	
HAP1	1hwt (C,D)	100.0/100.0	YPD	12	10	116	96.1	0.97	2.7e-15	
LEU3	2ere (A,B)	100.0/100.0	SM	10	10	32	27.4	0.98	6.8e-12	
PPR1	1pyi (A,B)	100.0/100.0	SGD	14	-	9	7.6	0.68	-	4
PUT3	1zme (C,D)	100.0/100.0	SM	16	16	32	26.1	0.98	8.3e-07	
ARG81	1hwt‡ (C,D)	93.5/88.7	SM	9	8	27	19.2	0.66	0.63	1
ARO80	1hwt‡ (C,D)	93.3/86.9	SM	13	23	31	24.2	0.89	3.6e-08	2
OAF1	1pyi‡ (A,B)	75.2/72.7	YPD	24	-	51	36.8	0.81	-	4
PIP2	1pyi‡ (A,B)	78.2/75.9	YPD	24	-	31	24.2	0.75	-	4
PDR1	1pyi‡ (A,B)	93.0/92.3	YPD	8	11	68	50.6	0.82	0.017	2
PDR3	1pyi‡ (A,B)	95.2/94.3	YPD	8	8⋆	18	12.7	0.61	9.0e-05	3
STB5	1pyi‡ (A,B)	75.2/73.6	YPD	8	10	44	33.9	0.78	1.3e-09	2
YRR1	1pyi‡ (A,B)	80.5/77.5	YPD	8	12	29	15.4	0.36	0.84	1
RDS1	1pyi‡ (A,B)	72.7/70.7	H2O2Hi	8	6	49	40.8	0.98	1.1e-08	
SIP4	1pyi‡ (A,B)	97.5/95.0	SM	13	15	24	19.8	0.73	1.8e-10	
UGA3	2ere (A,B)	69.6/69.2	RAPA	10	10⋆	28	19.5	0.63	4.8e-04	3
CHA4	1pyi‡ (A,B)	68.9/69.8	SM	18	9	27	16.9	0.47	7.5e-03	2
THI2	1pyi‡ (A,B)	83.9/79.1	Thi-	20	16	49	35.0	0.74	0.51	1
WAR1	2ere‡ (A,B)	79.0/78.3	YPD	29	-	12	6.4	0.36	-	4

PDB, Protein Data Bank code of the template structure (with chain IDs in parentheses); $\mathbf{S_{hm}}$, homology template scores defined in the main text (one per protein chain); **Env**, the environmental condition from Harbison *et al.* (1) (except SGD, which stands for a gene list from the literature); **L** and $\mathbf{L_{MF}}$, predicted motif widths in this study and MacIsaac *et al.* (2), respectively; $\mathbf{N_{seq}}$, number of intergenic sequences; $\langle \mathbf{N} \rangle$, expected number of binding sites according to the posterior probability $p(s, c)$. Note that $\langle \mathbf{N} \rangle$ is not equal to the total number of found sites because sites with low posterior probabilities contribute less to the average. $\mathbf{S_{track}}$, average posterior probability of the top 10 sites (all sites if < 10 sites were tracked); $\mathbf{p_{MF}}$, P value for the comparison of PWMs predicted in this study with PWMs predicted by MacIsaac *et al*; **C**, categories in SI Fig. 4: 1, 16 predictions with $\mathbf{p_{MF}} > 0.05$; 2, 10 cases for which $\mathbf{p_{MF}} < 0.05$ but the MacIsaac *et al.* motif is classified as partially incorrect by manual inspection of the Weblogo images (available at http://protein-dna.rockefeller.edu); 3, 8 predictions for which there are sites in the literature; 4, 10 cases with no prior binding site information.

⋆ Motifs from the literature, as reported by MacIsaac *et al.*

† Sites that were reported from the highest probability cluster (without tracking).

‡ Indicates that changes were made to the structure-based informative prior based on the binding site information from the literature (PWM columns were added, removed, or modified; consensus sequences from the structure were altered: SI Table 5).

1. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., MacIsaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004) *Nature*, **431**, 99–104.
2. MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D. and Fraenkel, E. (2006) *BMC Bioinformatics*, **7**, 113.

TF	PDB	S_{hm}	Env	L	L_{MF}	N_{seq}	$\langle N \rangle$	S_{track}	p_{MF}	C
MADS box										
MCM1	1mnm (A,B)	100.0/100.0	Alpha	16	16	77	72.2	0.96	0.0e+00	
ARG80	1mnm (A,B)	100.0/96.7	YPD	16	6	29	14.9	0.41	0.90	1
RLM1	1egw (A,B)	100.0/100.0	YPD	13	10	57	31.7	0.45	3.5e-03	
SMP1	1egw (A,B)	100.0/100.0	YPD	13	14*	80	45.8	0.46	3.5e-07	3
NDT80 type										
NDT80	1mnn (A)	100.0	SGD	10	-	69	45.1	0.67	-	4
Myb-like										
RAP1	1ign (A)	100.0	YPD	15	12	109	105.5	1.00	0.0e+00	
REB1	1w0t (A)	74.6	YPD	8	8	99	88.9	0.90	1.1e-15	
GATA										
DAL80	4gat (A)	100.0	RAPA	6	5*	31	31†	-	0.049	3
GAT1	4gat (A)	100.0	RAPA	6	7	49	49†	-	0.13	1
GLN3	4gat (A)	100.0	RAPA	6	10	79	79†	-	0.23	1
GZF3	4gat (A)	100.0	RAPA	6	6*	42	42†	-	0.20	3
GAT3	6gat (A)	87.0	YPD	6	8	26	26†	-	0.55	1
ASH1	6gat (A)	76.5	BUT14	6	10	39	39†	-	0.16	1
Heat shock factor type										
HSF1	3hts (B,D)	100.0/100.0	H2O2Lo	8	14	74	60.3	0.91	4.9e-05	
MGA1	3hts (B,D)	93.2/93.2	YPD	8	-	62	42.2	0.70	-	4
SKN7	3hts (B,D)	93.2/93.2	H2O2Lo	8	8	148	128.5	0.62	1.00	1
Homeodomain										
CUP9	1b8i (A,B)	67.0/87.8	YPD	12	-	27	14.4	0.40	-	4
PHO2	1fjl (A,B)	87.4/87.8	YPD	13	6	32	11.8	0.27	0.27	1
YHP1	1mnm (A,B,C,D)	-	YPD	24	6*	18	10.9	0.45	0.39	3
YOX1	1mnm (A,B,C,D)	-	YPD	24	20	61	43.1	0.74	2.5e-04	2
Fork head										
FKH1	2c6y (A,B)	88.3/90.7	YPD	13	12	104	67.0	0.83	3.5e-10	2
FKH2	2c6y (A,B)	90.5/91.9	YPD	13	15	91	54.9	0.70	4.8e-06	2
FHL1	2c6y (A,B)	88.0/90.4	YPD	13	10	131	117.8	0.96	4.5e-10	2
RFX DNA-binding										
RFX1	1dp7 (B,P)	95.3/95.3	YPD	16	14	25	21.9	0.99	4.1e-10	
HMG box										
ROX1	1j47 (A)	89.7	YPD	9	10	67	30.3	0.43	0.59	1

TF	PDB	S_{hm}	Env	L	L_{MF}	N_{seq}	$\langle N \rangle$	S_{track}	p_{MF}	C
Helix-loop-helix										
PHO4	1a0a (A,B)	100.0/100.0	Pi-	12	7	24	20.4	0.76	5.6e-10	
CBF1	1an2 (A,D)	87.2/87.2	SM	8	8	195	186.6	0.98	1.1e-13	
INO2	1nkp (A)	84.6	YPD	10	9	35	28.2	0.71	7.7e-08	
INO4	1nkp (B)	82.8	YPD	10	9	32	26.4	0.80	1.7e-07	
RTG1	R-box	-	RAPA	8	-	42	15.9	0.28	-	4
RTG3	R-box	-	RAPA	8	8	55	27.4	0.56	0.022	2
TYE7	1am9 (C,D)	99.0/94.5	YPD	12	8	56	48.8	0.94	4.2e-07	
	1an2 (A,D)	79.1/79.1		8			49.0	0.96	3.7e-10	
HMS1	1am9 (C,D)	96.8/98.0	YPD	12	-	22	14.5	0.54	-	4
	1an2 (A,D)	80.2/80.2		8			14.3	0.43	-	
Basic region leucine zipper										
GCN4	1ysa (C,D)	100.0/100.0	SM	11	8	143	111.6	0.85	3.2e-09	
CAD1	1gd2‡ (F,E)	93.8/92.5	YPD	11	10	29	24.5	0.94	6.4e-13	
CIN5	1gd2 (F,E)	91.5/91.0	H2O2Lo	12	9	118	105.5	0.99	1.2-07	
YAP1	1gd2‡ (F,E)	96.2/95.5	H2O2Lo	11	11	37	30.6	0.89	3.4e-07	2
YAP3	1gd2‡ (F,E)	94.3/95.0	YPD	11	7⋆	24	14.5	0.40	1.0e-05	3
YAP5	1gd2‡ (F,E)	91.4/91.6	YPD	11	6	47	40.0	0.92	0.43	1
YAP6	1gd2 (F,E)	92.5/91.0	YPD	12	8	92	76.1	0.91	6.2e-07	
YAP7	1gd2‡ (F,E)	91.3/90.8	H2O2Hi	11	10	101	93.6	0.98	3.7e-13	
	1gd2 (F,E)			12			76.7	0.92	0.37	
ARR1	1gd2‡ (F,E)	82.2/82.6	YPD	11	8	28	15.8	0.40	0.49	1
ACA1	1dh3 (A,C)	89.9/90.6	YPD	8	-	5	0.5	0.05	-	4
CST6	1dh3 (A,C)	89.9/90.6	YPD	8	8	27	11.1	0.25	0.45	1
SKO1	1dh3 (A,C)	88.7/89.2	YPD	8	8	17	3.2	0.18	2.3e-05	
HAC1	1dh3 (A,C)	91.8/91.8	YPD	8	10	17	4.1	0.17	0.22	1
MET28	2dgc (A,D)	82.9/82.9	YPD	10	6⋆	16	7.7	0.29	0.75	3
MET4	1nwq (A,C)	38.0/46.7	SM	-	8	37	-	-	-	1

Table 5. List of TFs for which the structure-based priors were modified based on the previously characterized binding sites.

TF	Comments and references
ARG81	1hwt: CGG half-sites, columns 4-6 removed Rijcke *et al.* (1) ($CGG - N_{3-5} - CCG$)
ARO80	1hwt: CGG half-sites, one non-specific column inserted at position 4 MacIsaac *et al.* (2) ($CGG - N_7 - CCG$)
OAF1, PIP2	1pyi: 16 non-specific columns between the TCGG,CCGA half-sites Karpichev *et al.* (3) ($CGGNNNTNAN_{9-12}CCG$)
PDR1, PRD3, STB5, YRR1	1pyi: columns 5-10 removed between the TCCG,CGGA half-sites Akache *et al.*, (4) Le Crom *et al.* (5) (TCCGCGGA)
RDS1	1pyi: columns 5-10 removed MacIsaac *et al.* (2) (TCGGCCGA)
SIP4	1pyi: 5 non-specific bases inserted between the TCGG,CGGA half-sites MacIsaac *et al.* (2) ($TCGG - N_5 - CCGA$)
UGA3	2ere Nöel *et al.* (6) ($CCG - N_4 - CGG$)
CHA4	1pyi: 10 non-specific bases inserted between the TCGG,CGGA half-sites Nöel *et al.* (6) ($CGG - N_{10} - CCG$)
THI2	1pyi: 10 non-specific bases inserted between the TCGG,CGGA half-sites MacIsaac *et al.* (2) ($CGG - N_{12} - CCG$)
WAR1	2ere: 21 non-specific bases inserted between the CCGG,CCGG half-sites Kren *et al.* (7) ($CCG - N_{23} - CGG$)
RTG1, RTG3	R-box: GGTCAC consensus sequence with polarization 0.7 flanked by 2 non-specific bases. Jia *et al.* (GGTCAC)
CAD1, CIN5, YAP1, YAP3, YAP5, YAP6, YAP7, ARR1	1gd2: AP-1 prior is made by removing column 7 Fujii *et al.* (9), Fernandes *et al.* (10) (AP-1: GTTA-C/G-TAAC, ATF/CREB: GTTA-CG-TAAC)

The initial priors are labeled with their PDB codes. Note that the PDB consensus sequences were modified in some cases. Literature references provide evidence for the consensus sites in parentheses.

1. De Rijcke, M., Seneca, S., Punyammalee, B., Glansdorff, N. and Crabeel, M. (1992) *Mol.Cell.Biol.*, **12**, 68–81.
2. MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., Stormo, G.D. and Fraenkel, E. (2006) *BMC Bioinformatics*, **7**, 113.
3. Karpichev, I.V. and Small, G.M. (1998) *Mol.Cell.Biol.*, **18**, 6560–6570.
4. Akache, B., MacPherson, S., Sylvain, M. and Turcotte, B. (2004) *J.Biol.Chem.*, **279**, 27855–27860.
5. Le Crom, S., Devaux, F., Marc, P., Zhang, X., Moye-Rowley, W.S. and Jacq, C. (2002) *Mol.Cell.Biol.*, **22**, 2642–2649.
6. Noël, J. and Turcotte, B. (1998) *J.Biol.Chem.*, **273**, 17463–17468.

7. Kren, A., Mamnun, Y.M., Bauer, B.E., Schüller, C., Wolfger, H., Hatzixanthis, K., Mollapour, M., Gregori, C., Piper, P. and Kuchler, K. (2003) *Mol.Cell.Biol.*, **23**, 1775–1785.

8. Jia, Y., Rothermel, B., Thornton, J. and Butow, R.A. (1997) *Mol.Cell.Biol.*, **17**, 1110–1117.

9. Fujii, Y., Shimizu, T., Toda, T., Yanagida, M. and Hakoshima, T. (2000) *Nat.Struc.Biol.*, **7**, 889–893.

10. Fernandes, L., Rodrigues-Pousada, C. and Struhl, K. (1997) *Mol.Cell.Biol.*, **17**, 6982–6993.

SI Materials and Methods

Gibbs Sampling with Informative Priors. Full description of the PhyloGibbs algorithm, including comprehensive tests on synthetic and yeast data sets, can be found in Siddharthan *et al.* (1). Here we extend the relevant formulas to the case of informative priors. The PhyloGibbs algorithm assigns a posterior probability $P(C|S)$ to each configuration C which is a partition of the input sequence S into a set of nonoverlapping sequence windows corresponding to different PWMs and background. For phylogenetically unrelated sequences, a sequence window is simply a contiguous segment of bases with a fixed width m. For related species, sequence windows can extend to include aligned bases from the other sequences (1). In what follows we restrict ourselves to the single species case. The posterior probability of the configuration C can be computed using Bayes's theorem:

$$P(C|S) = \frac{P(S|C)P(C)}{\sum_{C'} P(S|C')P(C')}, \qquad [3]$$

where $P(C)$ is the prior probability of the configuration C, and $P(S|C)$ is the probability of the input sequence given C. We run PhyloGibbs for a fixed number of TFs (usually one) and a fixed total number of sequence windows, effectively setting $P(C) = 0$ for all configurations outside of this subspace. $P(S|C)$ is given by:

$$P(S|C) = P(S \notin C|B) \prod_{c \in C} P(S_c), \qquad [4]$$

where $P(S_c)$ is the probability that sequences assigned to a TF with index c in the current configuration are drawn from a common PWM, and $P(S \notin C|B)$ is the probability of the background sequence (not occupied by any sequence windows). The background sequence is assumed to be generated by a Markov model of order $k = 0, 1, ..$. $P(S_c)$ is given by the integral over all possible PWMs:

$$P(S_c) = \int_{w_i^\alpha > 0, \quad \sum_{\alpha=1}^4 w_i^\alpha = 1} dw \; P(S_c|w)P(w), \qquad [5]$$

where $P(S_c|w) = \prod_{i=1}^m \prod_{\alpha=1}^4 (w_i^\alpha)^{n_i^\alpha}$ is the probability that all sequences assigned to the same TF are sampled from a particular (but unknown) PWM, and $P(w)$ describes our prior knowledge of the PWM. w_i^α is the probability of base α at position i in the PWM w, n_i^α is the number of times base α is found at position i among all sequences in S_c, and m is the site width. The integral is taken over all PWM components, subject to the normalization constraint:

$$\int_{w_i^\alpha > 0, \quad \sum_{\alpha=1}^4 w_i^\alpha = 1} dw \; .. = \prod_{i=1}^m \left[\int dw_i^1 \int dw_i^2 \int dw_i^3 \int dw_i^4 \; \delta(w_i^1 + w_i^2 + w_i^3 + w_i^4 - 1) \; .. \right].$$

When the informative prior is available,

$$P(w) = \prod_{i=1}^m \frac{\Gamma(\bar{n}_i)}{\prod_{\alpha=1}^4 \Gamma(\bar{n}_i^\alpha)} \prod_{\alpha=1}^4 (w_i^\alpha)^{\bar{n}_i^\alpha - 1}, \qquad [6]$$

where $\Gamma(x)$ is the gamma function, $\bar{n}_i^\alpha$ is the number of prior counts of base α at position i, and $\bar{n}_i = \sum_{\alpha=1}^4 \bar{n}_i^\alpha$. Eq. **6** is a product of Dirichlet distributions which is a generalization of the prior with the constant pseudocount. We assume that the total number of prior counts is independent of the PWM column i ($\bar{n}_i = \bar{n}, \forall i$), and that the minimum possible value of any $\bar{n}_i^\alpha$ is one (*i.e.* any base is allowed at any PWM position). Taking the integral in Eq. **5** with the help of the standard formula:

$$\int_{w^\alpha>0, \quad \sum_{\alpha=1}^4 w^\alpha=1} dw \prod_{\alpha=1}^4 (w^\alpha)^{n^\alpha-1} = \frac{\prod_{\alpha=1}^4 \Gamma(n^\alpha)}{\Gamma(\sum_{\alpha=1}^4 n^\alpha)}, \qquad [7]$$

we obtain:

$$P(S_c) = \prod_{i=1}^m \left[\frac{\Gamma(\bar{n})}{\Gamma(n+\bar{n})} \prod_{\alpha=1}^4 \frac{\Gamma(n_i^\alpha + \bar{n}_i^\alpha)}{\Gamma(\bar{n}_i^\alpha)} \right], \qquad [8]$$

where n is the total number of sequences assigned to TF c, and $\bar{n}$ is the total number of prior counts as discussed above.

It is interesting to note that $P(S_c)$ is related to the information score, defined as (2):

$$I(n_i^\alpha) = -\frac{1}{n} \log P(n_i^\alpha | b^\alpha), \qquad [9]$$

where $P(n_i^\alpha|b^\alpha)$ is the probability of observing n_i^α counts of base α at position i in the alignment of n sequences of length m (given the background model which assigns probabilities b^α to base α regardless of its position in the PWM):

$$P(n_i^\alpha | b^\alpha) = \prod_{i=1}^m n! \prod_{\alpha=1}^4 \frac{(b^\alpha)^{n_i^\alpha}}{n_i^\alpha!}. \qquad [10]$$

If all n_i^α are sufficiently large for the Stirling approximation to hold, the information score can be rewritten in a more familiar form (2):

$$I(n_i^\alpha) = \sum_{i=1}^m \sum_{\alpha=1}^4 w_i^\alpha \log\left(\frac{w_i^\alpha}{b^\alpha}\right), \qquad [11]$$

where $w_i^\alpha = n_i^\alpha/n$ is the frequency of base α at position i in the binding site. From Eq. **8**, the ratio of probabilities $P(S_c)$ for two independent sequence alignments (corresponding to two different PhyloGibbs configurations) is given by:

$$\frac{P(S_c)}{P(\widetilde{S}_c)} = \frac{\exp[(n+\bar{n}-4) I(n_i^\alpha + \bar{n}_i^\alpha - 1)]}{\exp[(\widetilde{n}+\bar{n}-4) I(\widetilde{n}_i^\alpha + \bar{n}_i^\alpha - 1)]} A(n_i^\alpha, \widetilde{n}_i^\alpha), \qquad [12]$$

where n_i^α ($\widetilde{n}_i^\alpha$) are the base counts in the first and second alignment respectively, n ($\widetilde{n}$) is the number of aligned sequences ($n = \sum_{\alpha=1}^4 n_i^\alpha,\ \widetilde{n} = \sum_{\alpha=1}^4 \widetilde{n}_i^\alpha,\ \forall i$), and

$$A(n_i^\alpha, \widetilde{n}_i^\alpha) = \prod_{i=1}^m \frac{(\widetilde{n}+\bar{n}-1)(\widetilde{n}+\bar{n}-2)(\widetilde{n}+\bar{n}-3)}{(n+\bar{n}-1)(n+\bar{n}-2)(n+\bar{n}-3)} \prod_{\alpha=1}^4 (b^\alpha)^{n_i^\alpha - \widetilde{n}_i^\alpha}. \qquad [13]$$

Note that $A(n_i^\alpha, \widetilde{n}_i^\alpha) = 1$ if $n = \widetilde{n}$ and the background is uniform: $b^\alpha = b = 0.25,\ \alpha = 1 \ldots 4$.

In this case Eq. **12** reduces to:

$$\frac{P(S_c)}{P(\widetilde{S}_c)} = \exp\{(n + \bar{n} - 4)[I(n_i^\alpha + \bar{n}_i^\alpha - 1) - I(\widetilde{n}_i^\alpha + \bar{n}_i^\alpha - 1)]\}. \quad [\mathbf{14}]$$

In the absence of any prior information it is convenient to simply set $\bar{n}_i^\alpha = 1$. With this choice of the pseudocount Eq. **14** becomes:

$$\frac{P(S_c)}{P(\widetilde{S}_c)} = \exp\{n[I(n_i^\alpha) - I(\widetilde{n}_i^\alpha)]\}. \quad [\mathbf{15}]$$

Therefore, the purpose of Gibbs sampling is to find an alignment of sequences with the highest information score. In other words, the process of maximizing the posterior probability $P(C|S)$ in Eq. **3** amounts to searching for clusters of sites whose alignments produce the highest information score with respect to the background model.

When the informative prior is available Eq. **14** becomes:

$$\frac{P(S_c)}{P(\widetilde{S}_c)} = \exp\{(n + \bar{n})[I(n_i^\alpha + \bar{n}_i^\alpha) - I(\widetilde{n}_i^\alpha + \bar{n}_i^\alpha)]\}. \quad [\mathbf{16}]$$

Here we rewrote the pseudocounts as $\bar{n}_i^\alpha = 1 + \bar{n}_i^{'\alpha}$ ($\bar{n}_i^{'\alpha} = 0$ in the uninformed case), and dropped the primes. The ability of the Gibbs sampler biased in this way to find the "true" binding sites strongly depends on how closely the "true" counts n_i^α correspond to the informative prior. For example, if our guess for the informative priors is so poor that it is actually complementary to n_i^α, we will have $I(n_i^\alpha + \bar{n}_i^\alpha) = 0$, resulting in assigning the lowest probability to the correct answer. On the other hand, if both n_i^α and $\bar{n}_i^\alpha$ are sampled from the same PWM (*i.e.* $I(n_i^\alpha + \bar{n}_i^\alpha) = I(n_i^\alpha)$) the log-probability of the "true" alignment will be amplified by a factor of $(n + \bar{n})/n$ compared to the uninformed case (cf. Eqs. **15** and **16**).

Introducing accurate prior information also biases the algorithm towards the correct binding sites. In particular, as shown above, in the absence of the informative priors the Gibbs probability of the alignment of n sites drawn from a PWM divided by the probability of the alignment of n sites drawn from the background is given by $\exp\{nI(n_i^\alpha)\}$. By definition, $\exp\{-nI(n_i^\alpha)\}$ is the background model probability of observing n_i^α counts in the alignment of n sites of width m. Given an input sequence of length L and the information score $I(n_i^\alpha)$, we expect to find

$$\binom{N}{n} \int_{\widetilde{w}_i^\alpha > 0,\quad \sum_{\alpha=1}^4 \widetilde{w}_i^\alpha = 1,\quad I(\widetilde{n}_i^\alpha) \geq I(n_i^\alpha)} d\widetilde{w}\ \exp\{-nI(\widetilde{n}_i^\alpha)\} \simeq B \exp\{n[1 + \log(N/n) - I(n_i^\alpha)]\} \quad [\mathbf{17}]$$

alignments with the information score $I(n_i^\alpha)$ or higher by chance. Here, $N \simeq L$ is the number of allowed site positions in the input sequence, and the integral in

$$B = \exp\{nI(n_i^\alpha)\} \int_{\widetilde{w}_i^\alpha > 0,\quad \sum_{\alpha=1}^4 \widetilde{w}_i^\alpha = 1,\quad I(\widetilde{n}_i^\alpha) \geq I(n_i^\alpha)} d\widetilde{w}\ \exp\{-nI(\widetilde{n}_i^\alpha)\}$$

is taken over the hypervolume in the weight space on which all sets of weights are constrained to have the information score of at least $I(n_i^\alpha)$. The intergal in Eq. **17** reflects the fact that in the absence of informative priors the algorithm cannot distinguish between the alignment of true sites and *any other* alignment which happens to have a high enough information score by chance. Thus we need to compute the total probability that any alignment of the background sites have the information score of $I(n_i^\alpha)$ or higher. This probability is given by $B\exp\{-nI(n_i^\alpha)\}$, where B can be approximately interpeted as the total number of acceptable alignments (those that pass the information score test). It follows from Eq. **17** that with N greater than $N_{max} = n\exp[I(n_i^\alpha) - 1 - \log(B)/n]$ it is no longer theoretically possible to always converge to the true alignment during Gibbs sampling. In practive the threshold is even lower because deep local maxima make sampling convergence to the true global maximum progressively more difficult.

If we use the counts $\bar{n}_i^\alpha$ sampled from the same PWM as the true binding sites to construct the informative prior, the Gibbs probability relative to the alignment with zero information score becomes $\exp[(n+\bar{n})I(n_i^\alpha)]$, where $\bar{n} = \sum_{\alpha=1}^{4}\bar{n}_i^\alpha,\ \forall i$ (cf. Eq. **16**). Such a high score can only be attained by the alignment of background sites if it has the counts n_i^α which happen to match the informative prior by chance. The probability of such an alignment is simply $\exp\{-nI(n_i^\alpha)\}$, resulting in $N_{max} = n\exp[I(n_i^\alpha)-1]$. The advantage over the uninformed case is that here we have to observe an alignment of n background sites with specific counts n_i^α for the algorithm to miss all the true sites, whereas in the uninformed case any alignment with a high enough information score will do. We expect that $B \gg 1$ in most cases, making it possible to find true binding sites in much longer sequences. Besides a higher probability assigned to the cluster of true sites compared to the uninformed case, we observe faster rates of convergence towards the global maximum during sampling (data not shown). Intuitively, the probability landscape is biased towards the true sites by the informative prior such that the search space is significantly reduced.

Eq. **3** allows us to calculate the probability of any configuration C given the input sequence S. Since the space of all allowed configurations is exponentially large, PhyloGibbs employs simulating annealing to search for the configuration $C^\star$ with the maximum posterior probability. During the simulated annealing phase the prior counts are permanently assigned to a TF with a fixed index. This assignment is not affected by the sampling moves. The simulated annealing phase is followed by the tracking phase which is designed to estimate the posterior probability $p(s,c)$ that a site s belongs to a TF with index c. The counts from the informative prior form a stable tracking cluster with which additional sequences sampled from S may be associated with probability $p(s,c)$. Thus adding the informative prior biases the simulated annealing search towards the sites whose specificity matches that of the prior (cf. Eq. **16**).

The PhyloGibbs code is available at www.biozentrum.unibas.ch/∼nimwegen/cgi-bin/phylogibbs.cgi.

Yeast Sequence Data. Intergenic sequences for the probes bound with $p < 0.001$ were downloaded from the supplementary web site for Harbison *et al.* (http://18.68.8.35/Harbison/) (3). The probe sequences correspond to the March 2003 release of the yeast genome. Many TFs have experiments for more than one environmental condition. Thus, if the PWM was predicted by Harbison *et al.* we used the same environmental condition, otherwise we chose the experiment with the highest number of bound intergenic regions. In a few cases information about regulated genes available from the literature was used to assemble a set of upstream promoter sequences.

The upstream sequences extended in the $5'$ direction from the gene ORF and terminated either at the next ORF (regardless of its orientation) or at 1,000 bp. In particular, for NDT80 we collected the upstream sequences for the genes expressed in mid-late and late sporulation phases (4). For PPR1 we used the genes likely to be involved in the pyrimidine pathway (*URA1* through *URA8*, *URA10*).

Structural Database of Protein-DNA Complexes. We downloaded all structures with at least one protein and one nucleic acid chain from the May 2006 release of the Protein Data Bank (PDB; www.rcsb.org). For each of these structures we checked if a corresponding "biological unit" file from the Nucleic Acids Database (NDB; http://ndbserver.rutgers.edu) was available. In order to make biological units crystallographic symmetry transformations are applied to half-structures, or if multiple copies of the same protein-DNA complex are present in the PDB file separate files are made for each copy (in such cases we simply used the first file). We required that the final structure have two DNA chains. We then attempted to pair these chains by using base-to-base distance cutoffs and sequence complementarity. If the attempt was successful, the structure of the protein-DNA complex was added to the structural database. In addition to the automatic processing, some structures were manually modified so that their DNA chains could be recognized as complementary in the subsequent analysis. In the end, the database contained 515 structures of proteins bound to double-stranded DNA, 252 of which were classified as transcription factors (5).

Pfam Classification. All protein sequences in the final database of structures were classified using Pfam. Pfam is a database of protein domain families (6). It contains manually curated multiple sequence alignments for each family (Pfam-A) and the corresponding profile hidden Markov models (profile HMMs) (7). Profile HMMs are built in Pfam using the HMMER package (hmmer-2.3.2, http://hmmer.janelia.org), which can also be used to search an HMM database for the matches to a query protein sequence. We searched for domain matches in all protein sequences from the structural database using a collection of Pfam-A HMMs
(command line: hmmpfam -E 0.1 Pfam_ls pdb.fasta >& pdb.hmmer), and restricted all subsequent Pfam searches only to those HMMs with at least one hit in the structural database. HMMER uses bit scores to evaluate the statistical significance of the match:

$$S = \log_2 \frac{P(seq|HMM)}{P(seq|null)}, \qquad [18]$$

where $P(seq|HMM)$ is the probability that a sequence is generated by the HMM, and $P(seq|null)$ is the probability that a sequence is generated by the null model based on aligned nonhomologous sequences. More often, an e-value is reported instead of the bit score: it is defined as the expected number of false positives with bit scores at least as high as the current bit score. By definition, the e-value is proportional to the total size of the sequences in the Pfam database. If multiple protein domains of the same type are detected, the domain bit scores sum up to the total bit score, and the e-values are reported both for the whole protein and for each separate domain.

Web site for Structure-Based PWM Prediction and Protein-DNA Homology Modeling. We have set up an interactive web site which enables users to employ structure-based PWM predictions in a range of sequence analysis projects (Protein-DNA Explorer: http://protein-dna.rockefeller.edu). Given a query protein sequence, the user can: (i) identify its DNA-binding domains by running HMMER (http://hmmer.janelia.org); (ii) for each domain, find the structural templates sorted by the interface score S_{hm}. The user can then download the template PDB files, or examine the number and type of interface mutations by inspecting the structures in Jmol (http://jmol.sourceforge.net). For each structural template the user can download a structure-based PWM prediction, or display it as a Weblogo image (http://weblogo.berkeley.edu). Structure-based PWM predictions can be used as the informative priors, constraints, or starting points in many sequence analysis algorithms. Furthermore, once the structural template has been chosen the user can identify the orthologous proteins in a number of pre-computed species by carrying out protein-DNA interface alignments. In this approach, proteins with the maximum conservation at the DNA binding interface are reported as putative orthologs even though they may not exhibit the highest overall sequence similarity. Finally, given a PWM of the unknown origin (*e.g.* independently discovered using bioinformatics methods) the user can check for statistically significant alignments to the database of structure-based PWMs, and determine which TF the input PWM is most likely to have come from.

1. Siddharthan, R., Siggia, E.D. and van Nimwegen, E. (2005) *PLoS Comput.Biol.*, **1**, e67.

2. Nimwegen, E., Zavolan, M., Rajewsky, N. and Siggia, E.D. (2002) *Proc.Nat.Acad.Sci.*, **99**, 7323–7328.

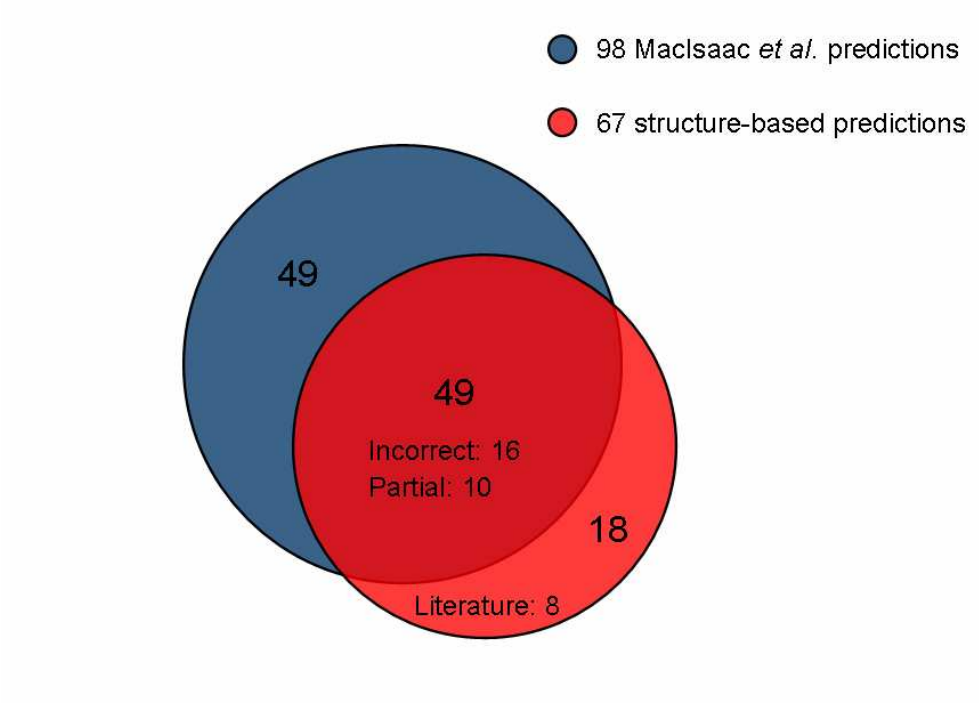
3. Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., MacIsaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004) *Nature*, **431**, 99–104.

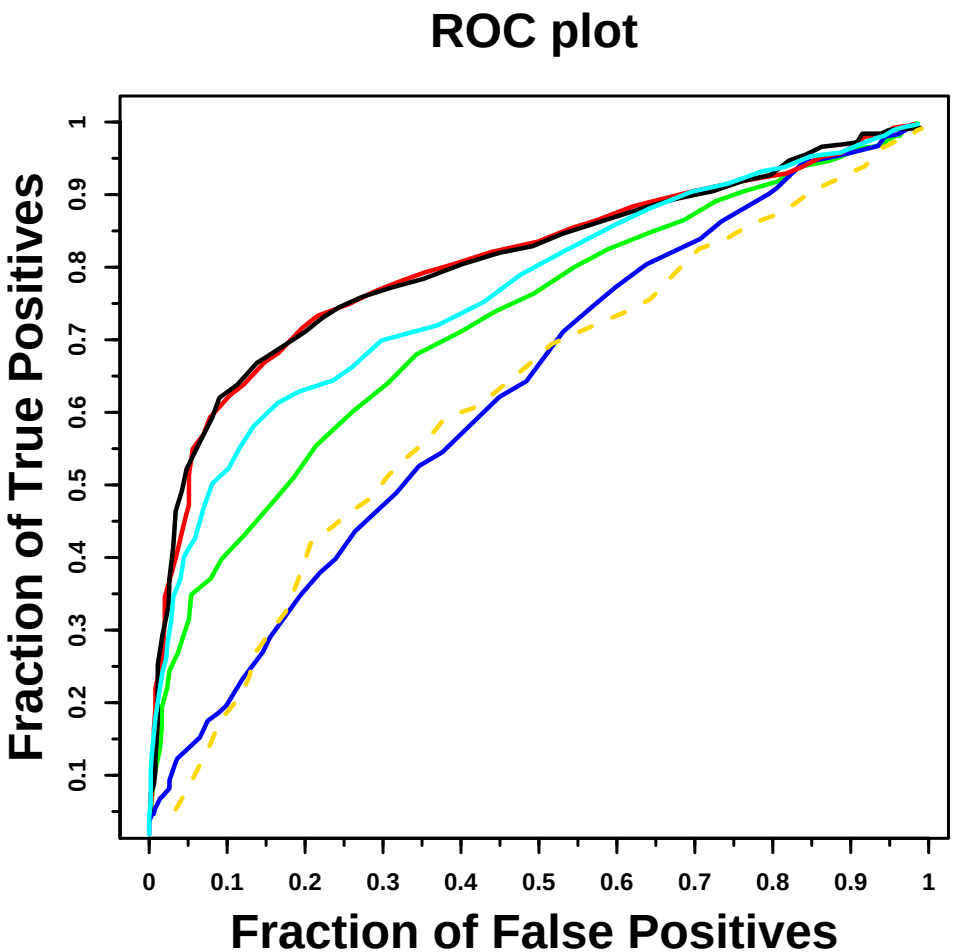
4. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) *Science*, **282**, 699–705.

5. Kummerfeld, S.K. and Teichmann, S.A. (2006) *Nucl.Acids Res.* **34**, D74–D81.

6. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) *Nucl.Acids.Res.*, **32**, D138–D141.

7. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.


98 MacIsaac *et al.* predictions
67 structure-based predictions
49
49
Incorrect: 16
Partial: 10
18
Literature: 8


ROC plot
Fraction of True Positives
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Fraction of False Positives