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:

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

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:

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

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:

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

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:

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

When the informative prior is available,

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

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:

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

we obtain:

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

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):

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

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):

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

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):

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

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:

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

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

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

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:

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

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:

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

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:

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

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

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

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

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

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:

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

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.