Published online 4 January 2010

Nucleic Acids Research, 2010, Vol. 38, No. 6 ***e90***
doi:10.1093/nar/gkp1166

Finding regulatory DNA motifs using alignment-free evolutionary conservation information

Raluca Gordân*, Leelavati Narlikar and Alexander J. Hartemink*

Department of Computer Science, Duke University, Box 90129, Durham, NC 27708, USA

Received August 16, 2009; Revised October 30, 2009; Accepted November 23, 2009

ABSTRACT

As an increasing number of eukaryotic genomes are being sequenced, comparative studies aimed at detecting regulatory elements in intergenic sequences are becoming more prevalent. Most comparative methods for transcription factor (TF) binding site discovery make use of global or local alignments of orthologous regulatory regions to assess whether a particular DNA site is conserved across related organisms, and thus more likely to be functional. Since binding sites are usually short, sometimes degenerate, and often independent of orientation, alignment algorithms may not align them correctly. Here, we present a novel, alignment-free approach for using conservation information for TF binding site discovery. We relax the definition of conserved sites: we consider a DNA site within a regulatory region to be conserved in an orthologous sequence if it occurs anywhere in that sequence, irrespective of orientation. We use this definition to derive informative priors over DNA sequence positions, and incorporate these priors into a Gibbs sampling algorithm for motif discovery. Our approach is simple and fast. It requires neither sequence alignments nor the phylogenetic relationships between the orthologous sequences, yet it is more effective on real biological data than methods that do.

INTRODUCTION

Due to advances in DNA sequencing technologies, the number of closely related genomes being sequenced has increased tremendously (1–4). This has consequently led to the emergence of comparative studies focused on detecting functional elements in intergenic DNA sequences. Functional elements, including transcription factor (TF) binding sites, are known to evolve at a slower rate than non-functional elements, and therefore well-conserved DNA sites should be good candidates for TF binding sites.

Many algorithms use evolutionary conservation information for *de novo* TF motif discovery, either by filtering the putative regions according to their conservation levels and then applying conventional motif finders (5) or by incorporating the conservation information into the motif finder itself (5–7, and many more). The former approach, although straightforward, has the main drawback that any region with a conservation level below the chosen cutoff is completely ignored, and thus motifs that are not well-conserved are not found by such methods. Thus, most conservation-based motif finders take the latter approach and incorporate the conservation information into the algorithm itself.

These methods can be further divided into two main categories: 'single gene' and 'multiple genes'. Methods in the first category [e.g. FootPrinter (8), the phylogenetic Gibbs sampler of Newberg *et al.* (9)] take as input the regulatory region of a single gene, together with its orthologs from related organisms. Methods in the second category [e.g. the method of Kellis *et al.* (2), Converge (5), PhyloCon (6), PhyME (7), PhyloGibbs (10), OrthoMEME (11), EMnEM (12), Compare Prospector (13)] are designed to search for motifs that are both overrepresented in a set of sequences from a reference species and conserved across related organisms. Our method falls into this category, so for the rest of the article we will focus only on 'multiple genes' approaches.

*To whom correspondence should be addressed. Tel: +1 (919) 660-6514; Fax: +1 (919) 660-6519; Email: amink@cs.duke.edu
Correspondence may also be addressed to Raluca Gordân. Tel: +1 (617) 525-4753; Fax: +1 (617) 525-4705; Email: raluca@cs.duke.edu
Present addresses:
Raluca Gordân, Division of Genetics, Dept. of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.
Leelavati Narlikar, Centre for Modeling and Simulation, University of Pune, Pune 411007, India.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Conserved TF binding sites are sometimes misaligned

Most conservation-based approaches to TF binding site discovery rely on multiple or pairwise alignments of orthologous regulatory regions to assess whether a particular DNA site is conserved across related organisms (2,5,7,10–14). However, since binding sites are usually short, sometimes degenerate, and often in reverse orientation or even relocated (15), alignment algorithms may not correctly align the binding sites within orthologous regulatory sequences. Especially when the sequences are very divergent, the background 'noise' of non-functional regions may be stronger than the 'signal' of conserved motifs, preventing a correct alignment. In Figure 1, we illustrate four scenarios where known motifs in orthologous yeast sequences are not correctly aligned, and thus would almost surely be missed by alignment-based motif finders. When a motif changes position or orientation, as in Figure 1C and D, correct alignment of motifs may even be impossible. Alignment algorithms may also incorrectly align orthologous regulatory regions from higher organisms: Kheradpour *et al.* (16) have shown examples of *Drosophila melanogaster* Mef-2 binding sites that are not correctly aligned to the orthologous sites in 11 related *Drosophila* species.

In consequence, motif finding algorithms based on alignments of orthologous promoter regions are only applicable when the promoters in the species of reference align well with the promoters in the related species; this is not true, for example, for many promoters in *Saccharomyces cerevisiae* and their orthologs in non-*sensu stricto* *Saccharomyces* species. Even when the orthologous promoters align well, depending on the exact algorithm used to construct the alignments, different sites may appear to be conserved. For example, while some studies report a significant number of *S. cerevisiae* TF binding sites that are conserved in related *Saccharomyces* species (17,18), Siggia *et al.* (19) found that among 407 experimentally verified binding sites in *S. cerevisiae*, only about half appear to be conserved in an alignment of *sensu stricto* promoter sequences [in this study, the sequences were aligned using a method by Morgenstern (20)]. Similarly, Stark *et al.* (21) investigated the effect of alignment choice on finding conserved TF binding sites in *D. melanogaster* and found large discrepancies between the alignments (only 59% agreement).

Relaxed definition of conserved binding sites

To overcome the limitations of using aligned orthologous sequences to distinguish between conserved and non-conserved sites, we have designed a novel, alignment-free method for conservation-based motif discovery that relies on a relaxed definition of conserved DNA sites: we consider a site within a reference regulatory region to be conserved in an orthologous sequence if it occurs anywhere in that sequence, irrespective of orientation. We show that this simple definition of conserved sites can be used to detect regulatory motifs more successfully than using existing alignment-based approaches.

MATERIALS AND METHODS

In this section, we first provide a brief description of PRIORITY, a general framework for motif discovery (22–24). Based on a Gibbs sampling strategy, PRIORITY can easily incorporate additional biological data that may be relevant for finding DNA motifs. Second, we show how evolutionary conservation data (both alignment-based and alignment-free) can be incorporated into PRIORITY in the form of positional priors.

A Sequence iYLR213C, bound by Mac1

```
Scer: ...CGCCGATATTTTTGCTCACCTTTTTTTTTTGCTCATCG-AAAATTGTTATAGCG...
Spar: ...CACCGATATTTTTGCTCACCTTTTTTTT--GCTCATCG-AAAATTGTTA--GCG...
Skud: ...AGTCGATATTTTTGCTCATCTTTTTTTTTTGCTCATTGAAAAATTGCAATGGCG...
Sbay: ...CAGTGAAATTTTTGCTCATCGAATTTTTT--GCTCATCG---AAGTGTAAT-GCG...
```

B Sequence iYAR014C, bound by Tec1

```
Scer: ...ATATATATATATATATACATTCTATATATTCTTACCCAGATTCTTT-GAGGTAAGA...
Spar: ...ATATATATATATATA-----TGTACATTCTCACCTGGATTCTTTGGGGGTAAAA...
```

C Sequence iYKL054C, bound by Rpn4

```
Scer: ...TGGGGTAATTGGTAAGAGTTT-TT...GCCACTACTTTTTGCCACCATTT-CCC...
Spar: ...TGGGGTAATTGGTAAGAGTTTCTT...GCCACTATTTTTTGCCACCATTT-CCC...
Smik: ...-GGGGTAATTGGTAAGAGTTTCTT...GCCACTGTTTTTTGCCACCATTTTCCC...
Skud: ...TGGGGTAATTGGTAAGAGTTCCTT...GCCACT-TTTTTTGCCACCATTT--CC...
Sbay: ...TGTGGTAATTGGTAAGTTTTTCTT...GCCACT-TTTTTTGCCACCATTTTTCC...
Sklu: ...GTGGGAGGGTGGCAAATTTTTCTC...GACACAGT------CCATAAGCT-GCC...
```

D Sequence iYMR107W, bound by Leu3 and Ume6

```
Scer: ...CGCCTAGCCGCCGGAGCCTGCCGGTACCGGCTTGGCTTCAGTTGCTGATCTCGG...
Smik: ...TACCTAACAGCCGG----------TACCGGCTTGAATGCCGCCGTTGGCTTCCG...
```

Figure 1. Examples of conserved TF binding sites in aligned orthologous yeast sequences (18) that are likely to be missed by alignment-based motif discovery programs. The sites matching the motifs of the respective TFs are underlined and marked in color. (**A**) Alignment algorithms may incorrectly insert gaps in orthologous motif occurrences. (**B**) Non-functional regions that are conserved in closely related organisms may prevent a correct alignment of the binding sites. (**C**) Binding sites are sometimes free to change orientation, which is probably the case for the Rpn4 binding site in *S. kluyveri*. (**D**) Motifs may change their position relative to each other, as shown by the Leu3 and Ume6 sites.

It is important to note that the present article is not about the PRIORITY framework *per se*, but rather about a simple, but clever, method for exploiting conservation information for more accurate motif discovery. Consequently, the approach introduced here can easily be adapted to other motif finders. For example, Bailey *et al.* (personal communication) have now incorporated our alignment-free conservation-based priors into MEME (25).

The PRIORITY framework

Given n DNA sequences $\boldsymbol{X}_1$ to $\boldsymbol{X}_n$ believed to be bound by a common TF, PRIORITY searches for a DNA motif that occurs in most sequences and is overrepresented with respect to background. We model the motif as a position-specific scoring matrix [(PSSM) (26)] ϕ of length W, while the rest of the sequence follows some background model parameterized by ϕ_0. Let $\boldsymbol{Z}$ be a vector of length n denoting the starting location of the binding site in each sequence with the convention that $Z_i = 0$ if $\boldsymbol{X}_i$ contains no binding site. For simplicity, we model at most one binding site in each sequence.

We use collapsed Gibbs sampling (27) to find ϕ and $\boldsymbol{Z}$ that maximize the joint posterior distribution of all the unknowns given the data:

$$\underset{\phi, \boldsymbol{Z}}{\operatorname{argmax}}\ P(\phi, \boldsymbol{Z} \mid \boldsymbol{X}, \phi_0) = \underset{\phi, \boldsymbol{Z}}{\operatorname{argmax}}\ P(\boldsymbol{X}|\phi, \boldsymbol{Z}, \phi_0)P(\phi)\ P(\boldsymbol{Z})$$

We randomly initialize $\boldsymbol{X}$ and then sample repeatedly from the joint posterior over ϕ and $\boldsymbol{Z}$ for a predetermined number of iterations, while keeping track of the highest scoring PSSM. In each run of PRIORITY, we start from several random starting points (by default, the number of trials is 50) and output the highest scoring PSSM across all the trials. Further details are available in our earlier work (22–24,28) and in the Supplementary Data. All the results reported here were obtained using version 2.1.0 of our PRIORITY software implementation, which is available online.

Incorporating positional priors

The Gibbs sampling technique described above has been used in several motif finders, often with additional parameters and heuristics. Usually, these motif finders assume a uniform prior over the locations $\boldsymbol{Z}$. We believe that some positions are *a priori* more likely to be starting locations of TF binding sites, and therefore use informative positional priors $P(\boldsymbol{Z})$.

A positional prior can be built from any score $\mathcal{S}$ that defines, for each site of size W in the input sequences, the *a priori* probability of that site being a TF binding site: $\mathcal{S}(\boldsymbol{X}_i, j) = P$(the W-mer starting at position j in sequence $\boldsymbol{X}_i$ is a binding site). Given $\mathcal{S}(\boldsymbol{X}_i, j)$ for all positions j in the sequence $\boldsymbol{X}_i$, we define the positional prior $P(Z_i)$ as:

$$P(Z_i = 0) \propto \prod_u (1 - \mathcal{S}(\mathbf{X}_i, u)) \quad \mathbf{1}$$

$$P(Z_i = j) \propto \mathcal{S}(\mathbf{X}_i, j) \prod_{u \neq j} (1 - \mathcal{S}(\mathbf{X}_i, u)) \quad \mathbf{2}$$

for $1 \leq j \leq l_i - W + 1$, where l_i is the length of sequence $\boldsymbol{X}_i$. We then normalize $P(Z_i)$ so that under the assumptions of our model we have $\sum_{j=0}^{l_i - W + 1} P(Z_i = j) = 1$ for $1 \leq i \leq n$.

As described in the next sections, we substitute $\mathcal{S}$ with different scores based on evolutionary conservation to obtain different positional priors and thus different versions of the PRIORITY algorithm.

Derivation of alignment-free conservation score $\mathcal{S}_C$

Based on the relaxed definition of conserved binding sites, we develop the alignment-free conservation score $\mathcal{S}_C$ as described in Figure 2.

Let $\boldsymbol{X}_i$ be a DNA sequence from the reference species (i.e. the species for which we have TF binding data). Let $\boldsymbol{X}_i^{(s)}$ for $1 \leq s \leq k$ be orthologous sequences from related species, obtained via a genome alignment or by searching for regions near orthologous genes. A sequence is permitted to be empty if no such region is found in the genome of the corresponding organism.

We compute the alignment-free conservation score $\mathcal{S}_C$ by searching orthologous sequences $\boldsymbol{X}_i^{(s)}$ for occurrences of all W-mers present in $\boldsymbol{X}_i$. We assume that a W-mer has a high probability of being conserved if it occurs in most of the orthologous sequences regardless of its orientation or specific position in the sequences. For the W-mer at position j in the bound sequence $\boldsymbol{X}_i$, the score is defined as:

$$\mathcal{S}_C(\mathbf{X}_i, j) = \frac{1}{k} \sum_{s=1}^{k} I\left[X_{ij}^W \in \boldsymbol{X}_i^{(s)}\right] \quad \mathbf{3}$$

Figure 2. Computation of the alignment-free conservation score. For the W-mer starting at position j in sequence $\boldsymbol{X}_i$, the conservation score $\mathcal{S}_C(\boldsymbol{X}_i, j)$ is the fraction of sequences orthologous to $\boldsymbol{X}_i$ that contain the word X_{ij}^w, in either forward or reverse orientation.

where $I[\cdot]$ is an indicator function and X_{ij}^W denotes the W-mer at position j in sequence $\boldsymbol{X}_i$. In other words, the score $\mathcal{S}_\mathcal{C}(\boldsymbol{X}_i, j)$ is directly proportional to the number of orthologous sequences in which the W-mer X_{ij}^W appears. The values of $\mathcal{S}_\mathcal{C}$ range from 0 to1. To avoid singularities, we scale $\mathcal{S}_\mathcal{C}$ linearly so that the values lie between 0.1 and 0.9. We convert this score into a positional prior by substituting $\mathcal{S}_\mathcal{C}$ for $\mathcal{S}$ in (1) and (2) and then normalizing $P(Z_i)$. We call this prior $\mathcal{C}$, and the Gibbs sampling algorithm that uses this prior, PRIORITY-$\mathcal{C}$.

Derivation of alignment-based conservation scores $\mathcal{S}_\mathcal{A}$ and $\mathcal{S}_\mathcal{T}$

The alignment-based score $\mathcal{S}_\mathcal{A}$ is built directly from the aligned orthologous sequences $X_i, X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, \ldots, X_i^{(k)}$. First, for every position j in sequence $\boldsymbol{X}_i$, we compute the fraction of orthologous sequences in which the nucleotide at position j is conserved according to the alignment. Next, we average this fraction over the W-mer starting at position j in sequence X_i and call this score $\mathcal{S}_\mathcal{A}(\boldsymbol{X}_i, j)$.

The score $\mathcal{S}_\mathcal{T}$ is based on the conservation track computed by Siepel *et al.* (18) from multiple alignments of seven yeast species, and available at the UCSC genome browser (29). For every position j in sequence $\boldsymbol{X}_i$, the track specifies the probability of that position being conserved (as computed by PhastCons from the multiple alignments). We define $\mathcal{S}_\mathcal{T}(\boldsymbol{X}_i, j)$ as the average of the conservation track over the W-mer starting at position j in sequence $\boldsymbol{X}_i$.

As with $\mathcal{S}_\mathcal{C}$, we scale $\mathcal{S}_\mathcal{A}$ and $\mathcal{S}_\mathcal{T}$ linearly to lie between 0.1 and 0.9 instead of 0 and 1 to avoid singularities in the model. We then convert the scores into positional priors by substituting $\mathcal{S}_\mathcal{A}$ and $\mathcal{S}_\mathcal{T}$ for $\mathcal{S}$ in (1) and (2) and normalizing $P(Z_i)$. We call the new positional priors $\mathcal{A}$ and $\mathcal{T}$.

Derivation of discriminative scores

A ChIP-chip experiment gives rise to sequences $\boldsymbol{X}$ that are bound by the profiled TF as well as sequences $\boldsymbol{Y}$ that are not bound. Assume we are given m such unbound sequences. As in the case of $\boldsymbol{X}$, we have orthologous sequences $Y_1^{(s)}$ to $Y_m^{(s)}$ where $1 \le s \le k$. We compute a discriminative score $\mathcal{S}_{\mathcal{DC}}(\boldsymbol{X}_i, j)$ by taking into account the conservation score $\mathcal{S}_\mathcal{C}$ over both sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows:

$$\mathcal{S}_{\mathcal{DC}}(\mathbf{X}_i, j) = \frac{\sum_{(q,r):X_{qr}^W = X_{ij}^W} \mathcal{S}_\mathcal{C}(\mathbf{X}_q, r)}{\sum_{(q,r):X_{qr}^W = X_{ij}^W} \mathcal{S}_\mathcal{C}(\mathbf{X}_q, r) + \sum_{(q,r):Y_{qr}^W = X_{ij}^W} \mathcal{S}_\mathcal{C}(\mathbf{Y}_q, r)}$$

Intuitively, this prior ensures a high score for W-mers that are conserved only in the bound set but not W-mers that are conserved in general throughout the genome. Please refer to the Supplementary Data for details.

As in the case of previous conservation scores, we convert $\mathcal{S}_{\mathcal{DC}}$ into a positional prior which we call $\mathcal{DC}$. Similarly, we compute the discriminative scores $\mathcal{S}_{\mathcal{DT}}$ and $\mathcal{S}_{\mathcal{DA}}$ using the scores $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{A}$ across both bound and unbound sequences, and convert the scores into positional priors $\mathcal{DT}$ and $\mathcal{DA}$, respectively.

RESULTS

Our motif finding approach can be applied to any set of DNA sequences believed to be bound by a common TF. It is common practice for motif finding to be evaluated on synthetically generated promoter data. However, our framework uses informative priors that capture information of biological relevance from true genomic sequences so evaluation with simulated data is not appropriate.

Instead, we gather TF binding data from chromatin immunoprecipitation (ChIP-chip) experiments performed by Harbison *et al.* (5) in *S. cerevisiae*. We choose this data because it contains experiments for a large number of TFs, both with known and unknown consensus binding motifs. More precisely, Harbison *et al.* profiled the intergenic binding locations of 203 TFs under various environmental conditions over 6140 yeast intergenic regions. For each TF, we define its sequence-set $\boldsymbol{X}$ for a particular condition to be those intergenic sequences reported to be bound with P-value $\le$0.001 in that condition. Similarly, we define set Y to be all intergenic sequences reported to be bound with P-value $\ge$0.5. We restrict our attention to sequence-sets $\boldsymbol{X}$ of size at least 10. This results in 238 sets that can be subdivided into 156 sets corresponding to TFs with known binding motifs [as summarized by Harbison *et al.* (5) at the time their article was published or as reported by Dorrington and Cooper (30), Jia *et al.* (31), Zhao *et al.* (32), Liu *et al.* (33) or Tan *et al.* (34)], which we can use for assessment of accuracy, and 82 sequence-sets without known motifs, which we can use for prediction of novel motifs.

Conservation information is most useful when used in an alignment-free manner

The PRIORITY framework easily incorporates evolutionary conservation information in the form of positional priors. As described in detail in the 'Materials and Methods' section, a positional prior over a DNA sequence $\boldsymbol{X}_i$ in a sequence-set $\boldsymbol{X}$ is computed from a probabilistic score $\mathcal{S}(\boldsymbol{X}_i, j)$, which is simply an indication of the *a priori* probability that the W-mer at position j in sequence $\boldsymbol{X}_i$ represents a TF binding site.

Here, we compute probabilistic scores from evolutionary conservation information derived from the UCSC genome browser. For each intergenic region in *S. cerevisiae*, we obtain orthologous sequences from six related organisms (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli* and *S. kluyveri*) based on the MULTIZ and BLASTZ alignments (18). We compute the conservation-based scores $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{A}$, which are alignment-based, and $\mathcal{S}_\mathcal{C}$, which is alignment-free.

Comparison of scores. We first evaluate the conservation scores on 62 sequence-sets that correspond to TFs with minimally degenerate consensus motifs reported in the literature. More precisely, we chose the sequence-sets for which the known TF motif (or its core) satisfied the

following criteria: (i) it was six to eight positions long, (ii) it was not too degenerate (i.e. it had at most four non-degenerate variants) and (iii) it occurred, in the forward or reverse orientation, at least 10 times in the sequence-set. We chose minimally degenerate consensus binding motifs with the hope that matches to these motifs are functional binding sites and not spurious matches. We imposed the other restrictions to insure that there were enough occurrences of the motifs to conduct our analysis. The exact sequence-sets and W-mers used in this analysis are available in Supplementary Table S1.

For each of the 62 sequence-sets, we compute the average conservation score for all possible words of length W, where W is the size of the reported TF motif. Next, we look at the W-mers that match the motif and use their ranks to evaluate how promising each conservation score is: the higher the rank, the better the conservation score. The alignment-free score $\mathcal{S}_\mathcal{C}$ is better than the alignment-based scores $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{A}$: when sorted in decreasing order by conservation score, the W-mers that match the TF motifs are ranked highest according to $\mathcal{S}_\mathcal{C}$ in the majority of sequence-sets (Figure 3). Therefore, the alignment-free score $\mathcal{S}_\mathcal{C}$ provides a better indication of which DNA sites are truly conserved, and thus likely to be functional, than the scores based on alignments.

Comparison of priors. Next, we use the conservation scores $\mathcal{S}_\mathcal{C}$, $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{A}$ to derive the informative positional priors $\mathcal{C}$, $\mathcal{T}$ and $\mathcal{A}$, respectively. We incorporate each prior into PRIORITY and run the resulting algorithms (PRIORITY-$\mathcal{C}$, PRIORITY-$\mathcal{T}$ and PRIORITY-$\mathcal{A}$) on the 156 sequence-sets with known TF motifs. To assess the contribution of the conservation priors, we also run PRIORITY with a uniform prior that we call $\mathcal{U}$. We count the number of successes for each algorithm, considering an algorithm successful on a particular sequence-set if the reported motif matches the known literature motif [according to a standard inter-motif distance criterion (5,28)]. PRIORITY is a stochastic algorithm, so different runs may give slightly different results. For this reason, we run each version of PRIORITY 50 times and report the median number of successes, along with the first and the third quartiles (Figure 4).

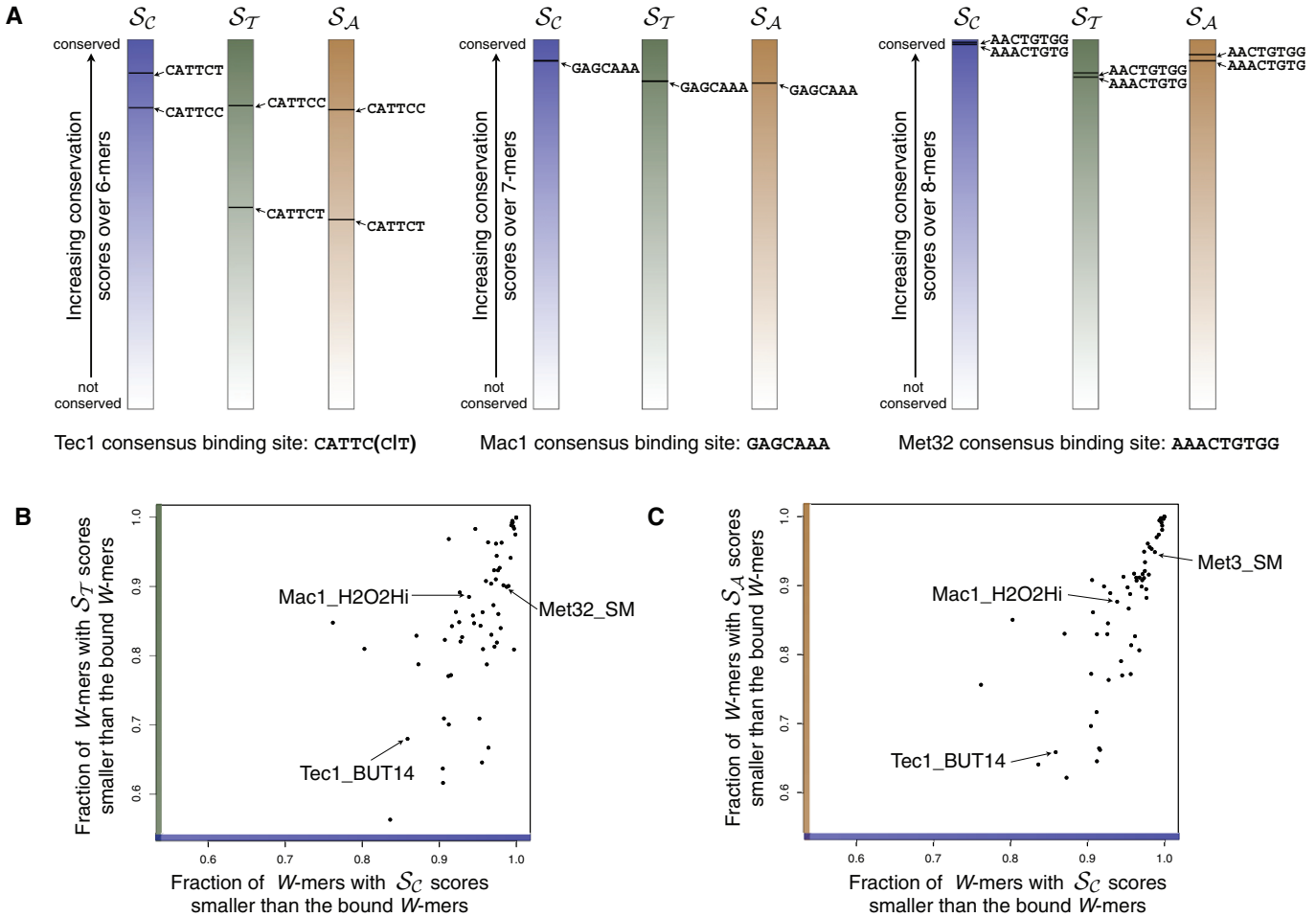

Figure 3. Comparison of alignment-free and alignment-based conservation scores. (**A**) $\mathcal{S}_\mathcal{C}$, $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{A}$ for all W-mers in three sequence-sets: Tec1_BUT14, Mac1_H2O2Hi and Met32_SM. Higher scores for the W-mers that match the TF motifs indicate better conservation. In all these sequence-sets, the alignment-free score $\mathcal{S}_\mathcal{C}$ is the highest for the W-mers that match the TF motifs. (**B**) The fraction of W-mers with conservation scores smaller than the bound W-mers (i.e. the ones that match the TF motifs). Higher fractions indicate better conservation scores, with a fraction of 1 being ideal. Of the 62 tested sequence-sets, both $\mathcal{S}_\mathcal{T}$ (**B**) and $\mathcal{S}_\mathcal{A}$ (**C**) outperform $\mathcal{S}_\mathcal{C}$ in only six sequence-sets.

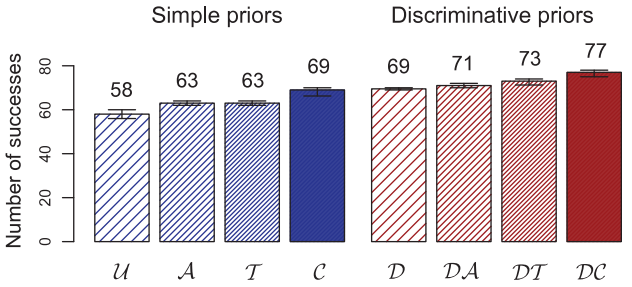

Figure 4. Number of successes obtained by PRIORITY with different positional priors on the 156 sequence-sets. Each algorithm was run 50 times with the default settings (Supplementary Data). The height of each bar represents the median number of successes among the 50 runs. The plot also shows, as confidence intervals, the first and third quartiles for each algorithm.

While all three conservation priors perform better than the uniform prior, the improvement is greater when the conservation information is used in an alignment-free manner: the median number of successes obtained by PRIORITY-$\mathcal{C}$ is 69, compared with 63 for PRIORITY-$\mathcal{T}$ and PRIORITY-$\mathcal{A}$ and 58 for PRIORITY-$\mathcal{U}$. Since the computation of the C prior uses our relaxed definition of conserved sites, we believe this approach is able to sidestep the alignment artifacts described in Figure 1, and hence pick up the true motif signal more often than the alignment-based priors $\mathcal{T}$ and $\mathcal{A}$. A detailed analysis for each sequence-set is available in the Supplementary Data.

Adopting a discriminative perspective improves results further

The scores $\mathcal{S}_{\mathcal{C}}$, $\mathcal{S}_{\mathcal{T}}$ and $\mathcal{S}_{\mathcal{A}}$ used to compute the priors $\mathcal{C}$, $\mathcal{T}$ and $\mathcal{A}$, respectively, reflect the probability that a W-mer is conserved. While it is true that regions bound by the TF are more likely to be conserved, it does not follow that every conserved region is more likely to be bound by the profiled TF. Some conserved regions could be binding sites of other TFs or other functional DNA elements. To address this, we developed the discriminative conservation scores $\mathcal{S}_{\mathcal{DC}}$, $\mathcal{S}_{\mathcal{DT}}$ and $\mathcal{S}_{\mathcal{DA}}$, which are specific to each profiled TF (see 'Materials and Methods' section). Intuitively, the discriminative conservation scores give higher weights to W-mers that are more conserved in the bound sequences than overall in the genome.

Comparison of scores. Figure 5 shows the scores $\mathcal{S}_{\mathcal{C}}$ and $\mathcal{S}_{\mathcal{DC}}$ over an intergenic sequence belonging to a sequence-set of Ste12. As can be seen, the conservation score computed with a discriminative perspective is effective in filtering out false conservation peaks, i.e. conserved DNA regions that are not specific to the profiled factor.

Comparison of priors. We use the scores $\mathcal{S}_{\mathcal{DC}}$, $\mathcal{S}_{\mathcal{DT}}$ and $\mathcal{S}_{\mathcal{DA}}$ to build the positional priors $\mathcal{DC}$, $\mathcal{DT}$ and $\mathcal{DA}$, respectively. Note that if we assume a constant level of conservation across all W-mers, then priors $\mathcal{C}$, $\mathcal{T}$ and $\mathcal{A}$ simplify to the widely used uniform prior U. Priors DC, $\mathcal{DT}$ and $\mathcal{DA}$, however, simplify to a special prior $\mathcal{D}$ that reflects the relative frequency of each W-mer in $\boldsymbol{X}$ versus both $\boldsymbol{X}$ and $\boldsymbol{Y}$ [see (23) for benefits of using the discriminative prior $\mathcal{D}$]. Here, we analyze the benefits of using conservation information in a discriminative manner.

Indeed, for each of the priors $\mathcal{C}$, $\mathcal{T}$, $\mathcal{A}$ and $\mathcal{U}$, adopting a discriminative perspective helps find the true motif in many more instances than without doing so (Figure 4). PRIORITY-$\mathcal{DC}$ is the most accurate, with a median number of successes of 77. A detailed analysis is available in the Supplementary Data.

For both simple and discriminative priors, the conservation information is most useful when used in an alignment-free manner: PRIORITY-$\mathcal{C}$ is superior to PRIORITY-$\mathcal{T}$ and PRIORITY-A, and PRIORITY-$\mathcal{DC}$ is superior to PRIORITY-$\mathcal{DT}$ and PRIORITY-$\mathcal{DA}$ (Figure 4). We will henceforth focus on the performance of our alignment-free motif finders: PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$.

PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ perform better than current conservation-based methods

In this section, we compare the results of PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$ with the results of seven conservation-based motif finders: MEME_c (5), a method of Kellis *et al*. (2), Converge (14), PhyloCon (6), PhyME (7), PhyloGibbs (10) and CompareProspector (13). All methods fall into the 'multiple genes' category, and thus search for motifs that are both overrepresented in a set of bound sequences from a species of reference and conserved across related species. Our alignment-free algorithms PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$ are more effective at finding the true motif than all these methods, with PRIORITY-$\mathcal{DC}$ performing best (Table 1).

We did not compare with a few other methods in the 'multiple genes' category (8,11,12,35) due to one or more of the following reasons: some are so computationally expensive that running them on all 156 sequence-sets was practically impossible; some are designed for only two related organisms; and some have been reported to perform worse than methods we have included in our analysis (see Supplementary Data for details).

PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ are much faster than current conservation-based methods

PRIORITY with the alignment-free conservation priors outperforms other methods not only in terms of accuracy, but also speed (Supplementary Figure S3). Since the running times of PRIORITY-C and PRIORITY-$\mathcal{DC}$ are comparable (with minor differences in the prior computation making PRIORITY-$\mathcal{C}$ slightly faster), we only discuss the times for PRIORITY-$\mathcal{DC}$ here.

The running time of PRIORITY-$\mathcal{DC}$ varies only slightly with increasing numbers of sequences, and PRIORITY-$\mathcal{DC}$ is faster than PhyloCon, PhyME and PhyloGibbs on all sequence-sets. On sets of 50 or more sequences, our algorithm is 2–3 orders of magnitude faster than the three methods. These results are not surprising: PRIORITY-$\mathcal{DC}$ uses the conservation information only during prior computation, while PhyloCon, PhyME and PhyloGibbs analyze the orthologous regions at every step of the algorithm. The running times of

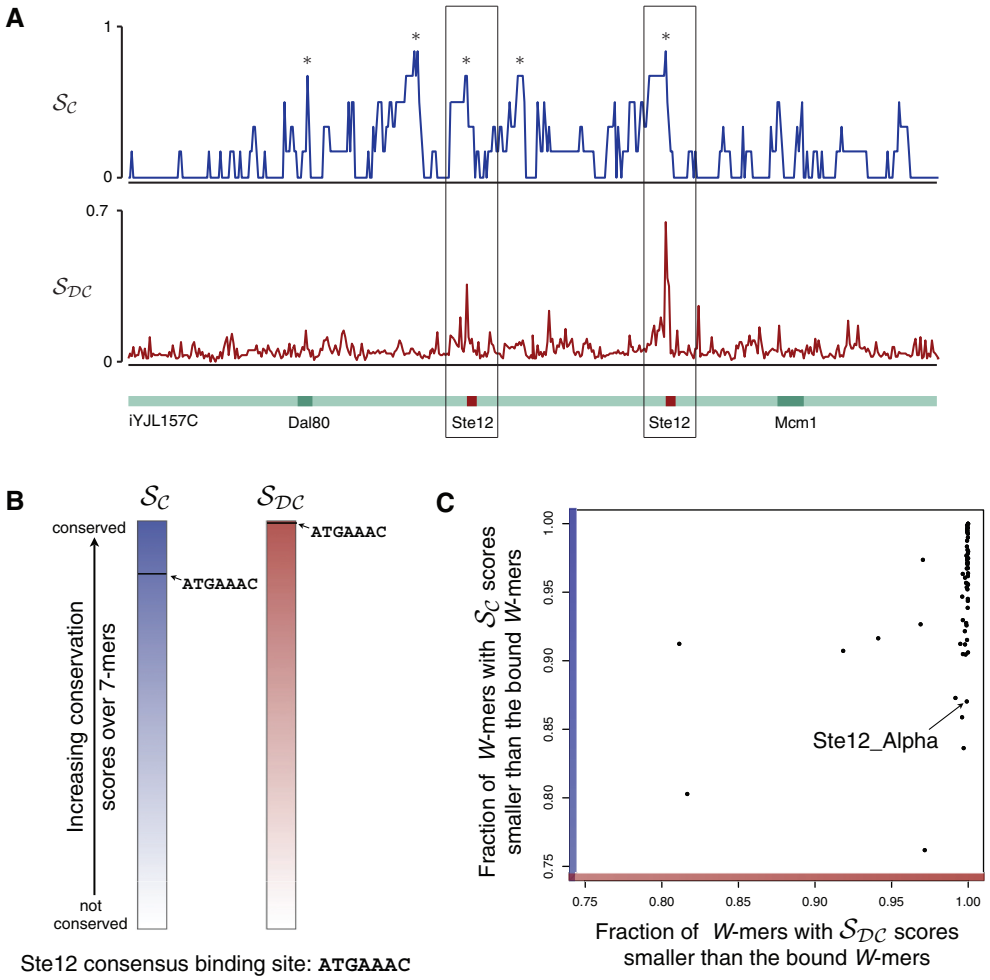

Figure 5. (**A**) Scores $\mathcal{S}_{\mathcal{C}}$ and $\mathcal{S}_{\mathcal{DC}}$ computed over intergenic region iYJL157C, bound by Ste12 after treatment with alpha factor (5). Binding sites of Dal80, Ste12 and Mcm1 are shown as annotated by MacIsaac *et al.* (14). $\mathcal{S}_{\mathcal{C}}$ has five tall peaks, marked with asterisks: two of them correspond to Ste12 binding sites, one to the Dal80 binding site and the two remaining peaks correspond to conserved A-T rich regions (not annotated here). The $\mathcal{S}_{\mathcal{DC}}$ score is specific to the profiled factor: it retains only the peaks corresponding to the Ste12 sites, and filters out non-specific peaks corresponding to A-T rich regions or other conserved sites. (**B**) $\mathcal{S}_{\mathcal{C}}$ and $\mathcal{S}_{\mathcal{DC}}$ for all *W*-mers in the Ste12_Alpha sequence-set. (**C**) The fraction of *W*-mers with conservation scores smaller than the bound *W*-mers. $\mathcal{S}_{\mathcal{DC}}$ outperforms $\mathcal{S}_{\mathcal{C}}$ in 57 of the 62 tested sequence-sets, and is essentially equally good in four others. Furthermore, the fraction corresponding to $\mathcal{S}_{\mathcal{DC}}$ is highly enriched towards the ideal fraction of 1.

Table 1. Number of successes for different conservation-based methods

Program	Description	Number of successes
MEME_c	Alignment-based; masks non-conserved bases and then applies MEME	49
Kellis *et al.*	Alignment-based; searches for significantly conserved 3-gap-3 motifs, then extends them	47
Converge	Alignment-based; uses EM; incorporates conservation and evolutionary distances into the model	68
PhyloCon	Locally aligns conserved regions into profiles, compares profiles and merges them using a greedy approach	19
PhyME	Alignment-based; uses EM; evolutionary model accounts for binding site specificities	21
PhyloGibbs	Alignment-based; similar to PhyME, but uses Gibbs sampling; searches for multiple motifs simultaneously	54
CompareProspector	Alignment-based; uses Gibbs sampling; biases the search towards conserved windows	64
PRIORITY-$\mathcal{C}$	Alignment-free; incorporates a prior based on conserved *W*-mers into a Gibbs sampler	**69**
PRIORITY-$\mathcal{DC}$	Alignment-free; incorporates a prior based on conserved *W*-mers in both bound *and* unbound sequences	**77**

For MEME_c and the method of Kellis *et al.* we use the results reported by Harbison *et al.* [5]; for Converge we use the results reported by MacIsaac *et al.* [14]. We ran PhyloCon, PhyME, PhyloGibbs and CompareProspector, as described in the Supplementary Data. For PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$, the numbers (shown in bold) represent medians across 50 runs of the algorithms (see main text for details).

CompareProspector and PRIORITY-$\mathcal{DC}$ are more comparable, but PRIORITY-$\mathcal{DC}$ scales better as the number of sequences grows, overtaking CompareProspector on sequence-sets of size 40 or more.

Predicting novel TF binding motifs using PRIORITY-$\mathcal{DC}$

Since PRIORITY-$\mathcal{DC}$ performed best among the tested algorithms, we next use it to predict novel yeast TF

binding motifs. The yeast ChIP-chip data set contains 82 sequence-sets with at least 10 bound probes, which correspond to TFs without a consensus binding motif at the time the ChIP-chip experiments were performed. Subsequently, a few groups (14,36) have reported putative motifs in these sequence-sets using computational approaches, but often with little discussion of their biological significance. Here, we discuss in detail novel motifs obtained using PRIORITY-$\mathcal{DC}$ on these sequences-sets, along with their implications in yeast biology.

For each sequence-set, PRIORITY-$\mathcal{DC}$ returns the top-scoring motif along with its log-posterior score (28). We assess the significance of a motif score by running PRIORITY on randomized sequence-sets and using the resulting scores to compute empirical distributions and consequently empirical *P*-values. For each novel motif reported by PRIORITY-$\mathcal{DC}$ we compute the associated *P*-value, which reflects our confidence in the accuracy of the motif. We choose a *P*-value cutoff of 2×10^{-7}, which corresponds to an estimated false discovery rate of 15% (see Supplementary Data for details). Out of the 82 motifs predicted for sequence-sets without a known TF binding motif, 16 motifs have a *P*-value smaller than the chosen cutoff (Figure 6). Since PRIORITY-$\mathcal{DC}$ is a stochastic algorithm, for each of the 16 predicted motifs we verified that even if we run the algorithm several times on a particular sequence-set, the reported motifs are highly similar (or exactly the same), and the *P*-values associated with the motif scores are always below the chosen cutoff.

The first three motifs shown in Figure 6 correspond to Dig1 profiled under different environmental conditions: Alpha (treatment with the alpha pheromone, which induces mating), BUT90 (butanol treatment for 90 min) and BUT14 (butanol treatment for 14 h, which induces filamentation). Dig1 is not currently known to bind DNA directly, but only through Ste12 or Tec1 during mating and filamentation, respectively (37). The predictions made by PRIORITY-$\mathcal{DC}$ are consistent with the literature: the motif found in the Dig1_Alpha sequence-set is a very good match to the Ste12 motif, while the motif found in Dig1_BUT14 matches the Tec1 motif. It is not clear what complexes bind DNA when cells are treated with butanol for a short duration (90 min in this case), so currently we cannot evaluate the prediction made by PRIORITY-$\mathcal{DC}$ in the Dig1_BUT90 sequence-set.

The next three motifs correspond to Fhl1 under different cellular conditions, and they all match the Rap1 consensus motif. Both Fhl1 and Rap1 associate with promoters of ribosomal protein genes (38), and a recent study (39) has shown that among the three main factors that control transcription of ribosomal protein genes (Rap1, Fhl1 and Ifh1), only Rap1 binds DNA directly. This evidence, together with evidence of direct interaction between Rap1 and Fhl1 (39), supports the hypothesis that Fhl1 binds DNA indirectly through Rap1, and thus our predictions for the sequence-sets of Fhl1 are very likely to be correct. Similarly, Sfp1 may bind DNA indirectly through Rap1 (40), and the Rap1 motif is indeed predicted by PRIORITY-$\mathcal{DC}$ in the Sfp1_SM sequence-set.

For Ime4_YPD and Mal33_H2O2Hi, we find the repetitive motif TG_n to be highly significant. Although these motifs may play a role in disrupting the chromatin structure (41), we do not believe they represent motifs of Ime4 or Mal33. To our knowledge, DNA binding motifs of these factors have not yet been reported in the literature. If we mask the TG repeats and run PRIORITY-$\mathcal{DC}$ again, in both sequence-sets we obtain motifs with lower scores, which do not pass our significance criterion.

As for predictions 10–16, in at least four of these cases we believe the motifs found by PRIORITY-$\mathcal{DC}$ are correct: Phd1_BUT90, Rfx1_YPD, Ydr026c_YPD and Stb1_YPD. The motif predicted for the Phd1_BUT90 sequence-set is consistent with the Phd1 motif reported in two recent *in vitro* studies (42,43). The DNA binding

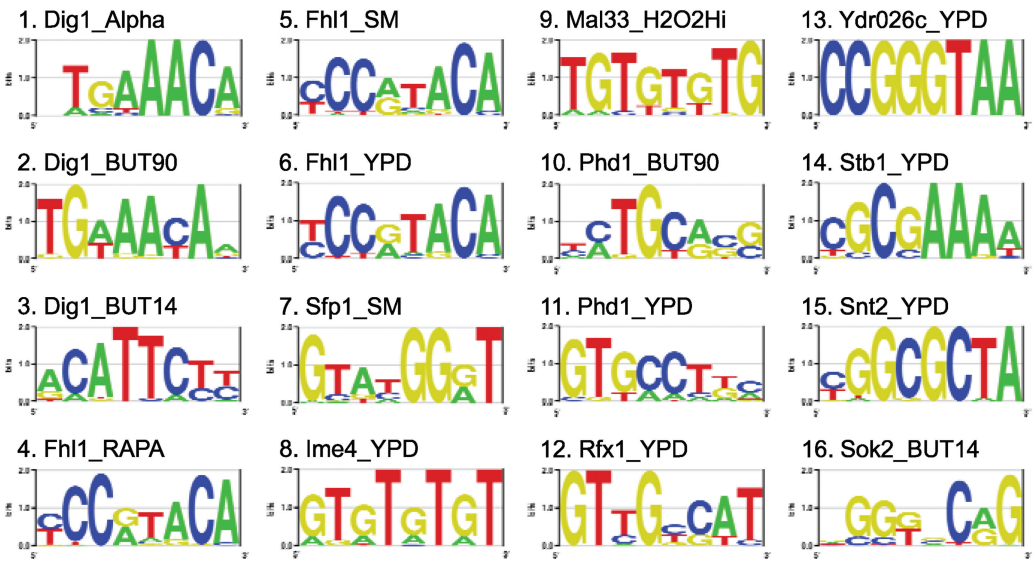

Figure 6. Novel TF binding motifs predicted by PRIORITY-$\mathcal{DC}$.

motif of Rfx1 has been recently listed in TRANSFAC 11.1, and also reported by Badis *et al.* (42) and Zhu *et al.* (43), matching the motif predicted by PRIORITY-$\mathcal{DC}$. The predicted motif for the putative TF Ydr026c is a very good match to the known Reb1 binding motif (44). Although an experimentally verified binding site is not currently available for Ydr026c, it is known that it has a strong similarity to the DNA-binding protein Reb1 (45). In the Stb1_YPD sequence-set, we predict a motif that matches the known Swi6 binding specificity (46), which is consistent with what is known in the literature: Stb1 plays a role in the regulation of MBF-specific transcription, and it is known to bind *in vitro* to Swi6 (47), a member of the MBF complex.

The motifs we predict for sequence-sets Phd1_YPD, Snt2_YPD and Sok2_BUT14 remain to be verified. We note, however, that our predictions for these sequence-sets are in good agreement with motifs obtained using other computational methods (5,14,36).

DISCUSSION

We are not the first to use alignment-free conservation across species to find motifs. Elemento and Tavazoie (48) look for conserved regulatory elements by scanning a pair of related genomes for on the order of 400 highly enriched W-mers. They then use a hypergeometric distribution to evaluate the significance of each of these W-mers in bound ChIP-chip sets. However, using this method, they are able to assign a W-mer that matches to the true motif to only 15 TFs. Since they limit their analysis to reporting W-mers, it is possible that they are not able to find TF motifs that have greater sequence variation. In contrast, though our scores $\mathcal{S}_{\mathcal{DC}}$ are also computed over W-mers, we use them only to construct positional priors; our Gibbs sampler returns a full PSSM, which handles sequence variations in the binding sites. In addition, the approach of Elemento and Tavazoie (48) is limited to pairs of related organisms, and thus the choice of organisms becomes crucial for the success of the algorithm.

One potential limitation of our approach is that the conservation priors are computed by counting only exact matches between the W-mers in the reference genome and W-mers in the related genomes. We have also tried computing priors similar to $\mathcal{C}$ and $\mathcal{DC}$ that allow for a mismatch when searching for conserved words. Specifically, an 8-mer was defined as 'conserved' in an orthologous sequence if the sequence contained either an exact match to that 8-mer or any of the 24 8-mers that differed at exactly one position. The effect of allowing one mismatch anywhere in the W-mer was that the signal of truly conserved sites was mixed with noise due to the 24 possible 8-mers, and overall these priors were not as effective as $\mathcal{C}$ and $\mathcal{DC}$. Allowing for more than one mismatch may further dilute the signal of conserved sites. However, prior knowledge about the structure of the binding site (for example, if we know we should be searching for a gapped motif) might be useful in restricting the mismatches to certain positions.

Our alignment-free approach can be used with both closely and distantly related organisms

In this article, we show how multiple unaligned genomes can be successfully used for motif discovery. Our method can be applied to any number of genomes. For instance, we independently computed six variant $\mathcal{DC}$ priors using: only the single closest species (*S. paradoxus*); the two closest species (*S. paradoxus* and *S. mikatae*); the three closest species (*S. paradoxus*, *S. mikatae* and *S. kudriavzevii*); and so on. PRIORITY-$\mathcal{DC}$ consistently found 69 or more motifs with each of these variant priors. The general trend indicated that adding more organisms improves performance. For other algorithms, however, the choice of related organisms is crucial for the success of the algorithm. PhyloGibbs, for example, works well when using the *sensu stricto Saccharomyces* species, but the performance drops dramatically if we include more distantly related species.

The *sensu stricto* species (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii* and *S. bayanus*) provide most of the conservation information in the priors. However, since these species are closely related to *S. cerevisiae*, their intergenic regions may contain many non-functional conserved sites, simply because not enough evolutionary time has passed since the species diverged from their common ancestor. Although in this case many DNA sites will appear to be conserved—and thus functional—this does not pose a problem for our conservation-based algorithm because the information in the cobound sequences helps reduce the space of putative TF binding sites to those conserved DNA sites that also appear in most of the cobound sequences. Furthermore, the more distantly related species *S. castelli* and *S. kluyveri* provide some of the sequence divergence necessary for filtering out the conserved non-functional sites. According to a study by Cliften *et al.* (49), only a small number of the intergenic regions in the *S. castelli* and *S. kluyveri* genomes can be aligned to *S. cerevisiae* regions, and only after the corresponding orthologous genes have been identified. Even then, the conserved regulatory sites may be hard, if not impossible, to align correctly. Hence, alignment-based motif finders may not be able to fully exploit the information provided by the two distantly related species, while our alignment-free algorithm can.

PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ scale well with the size of the sequence-set and the number of organisms

Our conservation-based approach is much faster than current motif finding algorithms. It only needs a few minutes to compute a motif, even on the largest sequence-set, while other methods required days or in some cases months. Interestingly, other methods become slower precisely because they use conservation information, but our method actually speeds up: the informative prior computed from conservation information facilitates rapid convergence to the posterior, as evidenced by the fact that PRIORITY-$\mathcal{DC}$ reaches convergence faster than PRIORITY-$\mathcal{U}$ (data not shown).

In Supplementary Figure S3, we show that PRIORITY-$\mathcal{DC}$ scales well with the size of the sequence-set. A similar

analysis can be done by keeping the size of the sequence-set fixed but varying the number of orthologs for each sequence. The running time for PRIORITY-$\mathcal{DC}$ varies only slightly when we increase the number of orthologous sequences, while the running time of other methods increases substantially.

PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ do not require a phylogenetic tree

Currently, the derivation of our alignment-free priors $\mathcal{C}$ and $\mathcal{DC}$ does not take phylogenetic information into account, mainly because high-quality phylogenetic trees are usually hard to compute. However, when such a tree is available, our algorithm can easily incorporate the phylogenetic information into the priors. A simple approach is to weight the sequences in each organism (and thus the occurrences of W-mers in these sequences) according to the evolutionary distance between that organism and the reference organism. We have derived such a weighting scheme for the *Saccharomyces* species using the phylogenetic tree reported by Siepel *et al.* (18). However, on these data sets, the conservation priors computed using the weighted sequences did not show any improvement over the initial conservation priors, $\mathcal{C}$ and $\mathcal{DC}$. Other approaches for incorporating phylogenetic information may also be used. For example, one could use the branch length score developed by Stark *et al.* (21)—which takes into account phylogenetic information—as a probabilistic score (see 'Materials and Methods' section). This score would then be converted into a prior and incorporated into PRIORITY.

It is important to make the distinction between discovering the TF binding motifs (which is the focus of our analysis) and finding the exact locations of the binding sites of a TF across the genome based on a known model (this latter problem is sometimes called motif scanning). In our work, we show that using conservation information in the form of aligned sequences or phylogenetic models is not the best solution for improving the prediction accuracy of computational motif finders. In a recent analysis, Hawkins and Bailey (50) have shown that motif scanning also does not benefit significantly from alignment-based conservation information and complex phylogenetic models. In another recent paper, Ward and Bussemaker (51) have shown the advantages of using orthologous promoter sequences in an alignment-free manner to solve yet another problem: the discovery of functional regulatory targets of TFs.

Using TF binding data from higher organisms

PRIORITY is not restricted to *S. cerevisiae* data, but can also be used on TF binding data from higher organisms. Unfortunately, an experimental study similar to the one performed by Harbison *et al.* (5) in yeast is not currently available in more complex organisms, so a thorough analysis of the performance of our alignment-free approach on data from more complex organisms is not currently possible. We did, however, as a proof of principle, apply PRIORITY on fly, mouse and human TF binding data. On all the sequence-sets described below, PRIORITY was applied with the same parameters as on the yeast sequence-sets, except for the conservation-based priors and the background model, which are specific to the organism and the sequence-set. For each sequence-set, each version of PRIORITY was run 50 times, and for each run we computed the distance between the literature motif and and the PRIORITY motif. We consider that an algorithm finds the correct motif in a sequence-set if the median distance between the literature motif and the reported motif is <0.25 (the distance criterion used for the yeast data). Details on the literature motifs and the exact organisms used in each analysis are available in the Supplementary Data.

We used fly TF binding data from Zhou and Wong (52), who collected over 60 enhancer sequences controlling 20 different genes expressed during the early development of *D. melanogaster*. Based on known regulatory interactions, they built three sequences-sets, each of which contained all enhancers believed to be bound by one of the three TFs Bicoid (Bcd), Hunchback (Hb) and Krüppel (Kr). As described earlier for the yeast ChIP-chip sequence-sets, we used orthologous regulatory regions to compute conservation priors $\mathcal{C}$, $\mathcal{T}$ and $\mathcal{A}$, and we applied PRIORITY-$\mathcal{C}$, PRIORITY-$\mathcal{T}$, PRIORITY-$\mathcal{A}$ and PRIORITY-$\mathcal{U}$ on the three fly sequence-sets. (Note that we cannot compute discriminative priors since a set of unbound sequences is not available.) PRIORITY-$\mathcal{U}$ and PRIORITY-$\mathcal{A}$ find only one correct motif (Hb), PRIORITY-$\mathcal{T}$ finds two correct motifs (Hb and Kr), while PRIORITY-C finds all three motifs correctly (Supplementary Figures S4 and S5).

We also applied PRIORITY-$\mathcal{U}$, -$\mathcal{A}$, -$\mathcal{T}$, -$\mathcal{C}$, -$\mathcal{D}$, -$\mathcal{DA}$, -$\mathcal{DT}$ and -$\mathcal{DC}$ on 12 mouse ChIP-seq data sets from Chen *et al.* (53), as compiled by Bailey *et al.* (personal communication). All eight algorithms were successful in 9 of the 12 mouse sequence-sets (Supplementary Figures S6–S10). However, the alignment-free priors seem to perform better than the other priors: an analysis of the median distance between the PRIORITY motifs and the literature motifs showed that PRIORITY-$\mathcal{C}$ is at least as good as PRIORITY-$\mathcal{U}$ in 7 of the 9 successful sequence-sets, while PRIORITY-$\mathcal{T}$ and PRIORITY-$\mathcal{A}$ are at least as good as PRIORITY-$\mathcal{U}$ in only 4 of the 9 successful sets.

Discovery of human TF binding motifs can also benefit from using alignment-free conservation information. We applied PRIORITY-$\mathcal{U}$, PRIORITY-$\mathcal{A}$, PRIORITY-$\mathcal{T}$ and PRIORITY-$\mathcal{C}$ on three sequence-sets containing promoters bound by the human TFs HNF1, HNF4 and HNF6 in human hepatocytes, derived from ChIP-chip experiments by Odom *et al.* (54). Supplementary Figures S11 and S12 show the results on the three sequence-sets: PRIORITY-$\mathcal{U}$, PRIORITY-$\mathcal{A}$ and PRIORITY-$\mathcal{T}$ find the corrrect motif in only one sequence-set (HNF6), while PRIORITY-$\mathcal{C}$ is successful in the HNF4 and HNF6 sequence-sets.

Although the fly, mouse and human TF binding data sets discussed in this section are not as comprehensive as the yeast ChIP-chip data of Harbison *et al.* (5), the results show that incorporating conservation information in an alignment-free manner improves motif discovery not only in the case of yeast TFs but also TFs from more complex

organisms. Furthermore, our approach is not restricted to ChIP-chip data, but can be used on ChIP-seq data or any set of co-regulated sequences.

CONCLUSION

Sequence alignments are undoubtedly very useful for the analysis of genomic data. For example, many genes are detected in newly sequenced organisms based on their homology to genes in related, well-studied species. Once homologous genes are detected, one can also align their regulatory regions with the hope of finding conserved TF binding motifs. We show, however, that for this purpose, alignments of orthologous regions can be misleading. Due to the short length of most binding sites, alignment algorithms are very likely to misalign true functional sites that are actually conserved across species (Figure 1). Furthermore, different algorithms may build very different alignments, in which different DNA sites appear to be conserved, so choosing the alignment algorithm becomes crucial for finding the conserved TF binding sites. Our method overcomes these issues because it uses cross-species conservation information without relying on alignments. In doing so, it outperforms currently used conservation-based programs in both speed and accuracy.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health (P50-GM081883-01, R01-ES015165-01); DARPA (HR0011-08-1-0023, HR0011-09-1-0040); a National Science Foundation CAREER award (NSF 0347801); Alfred P. Sloan Research Fellowship (to A.J.H.). Funding for open access charge: DARPA HR0011-09-1-0040.

Conflict of interest statement. None declared.

REFERENCES

1. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
2. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **432**, 241–254.
3. Clark,A., Gibson,G., Kaufman,T., Myers,E. and O'Grady,P. (2003) Proposal for *Drosophila* as a model system for comparative genomics. http://insects.eugenes.org/species/about/genome-proposals/GenomesWhitePaper2003/.
4. *Drosophila* 12 Genomes Consortium. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.
5. Harbison,C., Gordon,D., Lee,T., Rinaldi,N., Macisaac,K., Danford,T., Hannett,N., Tagne,J., Reynolds,D., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
6. Wang,T. and Stormo,G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
7. Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170.
8. Blanchette,M. and Tompa,M. (2003) Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.
9. Newberg,L., Thompson,W., Conlan,S., Smith,T., McCue,L. and Lawrence,C. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis*-regulatory site prediction. *Bioinformatics*, **23**, 1718–1727.
10. Siddharthan,R., Siggia,E. and van Nimwegen,E. (2005) PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
11. Prakash,A., Blanchette,M., Sinha,S. and Tompa,M. (2004) Motif discovery in heterogeneous sequence data. In *Pacific Symposium on Biocomputing*, Vol. 9. World Scientific, New Jersey, pp. 348–359.
12. Moses,A., Chiang,D. and Eisen,M. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In *Pacific Symposium on Biocomputing*, Vol. 9. World Scientific, New Jersey, pp. 324–335.
13. Liu,Y., Liu,X., Wei,L., Altman,R. and Batzoglou,S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, **14**, 451–458.
14. MacIsaac,K., Wang,T., Gordon,D., Gifford,D., Stormo,G. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.
15. Ludwig,M.Z. (2002) Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.*, **12**, 634–639.
16. Kheradpour,P., Stark,A., Sushmita,R. and Kellis,M. (2007) Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.*, **17**, 1919–1931.
17. Chin,C., Chuang,J. and Li,H. (2005) Genome-wide regulatory complexity in yeast promoters: Separation of functionally conserved and neutral sequence. *Genome Res.*, **15**, 205–213.
18. Siepel,A., Bejerano,G., Pedersen,J., Hinrichs,A., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
19. Siggia,E. (2005) Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.*, **15**, 214–221.
20. Morgenstern,B. (2000) A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics*, **16**, 948–949.
21. Stark,A., Lin,M., Kheradpour,P., Pedersen,J., Parts,L., Carlson,J., Crosby,M., Rasmussen,M., Roy,S., Deoras,A. *et al.* (2000) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, **450**, 184–185.
22. Narlikar,L. and Hartemink,A. (2006) Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, **22**, 157–163.
23. Narlikar,L., Gordân,R. and Hartemink,A. (2007) Nucleosome occupancy information improves *de novo* motif discovery. In *Research in Computational Molecular Biology*. Springer-Verlag, New York, pp. 107–121.
24. Gordân,R. and Hartemink,A. (2008) Using DNA duplex stability information to discover transcription factor binding sites. In *Pacific Symposium on Biocomputing*, Vol. 13. World Scientific, New Jersey, pp. 453–464.
25. Bailey,T. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. In *Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, USA, pp. 21–29.
26. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
27. Liu,J. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, **89**, 958–966.
28. Narlikar,L., Gordân,R. and Hartemink,A. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.
29. Kent,W., Sugnet,C., Furey,T., Roskin,K., Pringle,T., Zahler,A. and Haussler,D. (2002) The human genome browser at UCSC. *Science*, **12**, 996–1006.

30. Dorrington,R. and Cooper,T. (1993) The DAL82 protein of *Saccharomyces cerevisiae* binds to the DAL upstream induction sequence (UIS). *Nucleic Acids Res.*, **21**, 3777–3784.
31. Jia,Y., Rothermel,B., Thornton,J. and Butow,R. (1997) A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. *Mol. Cell. Biol.*, **17**, 1110–1117.
32. Zhao,H., Butler,E., Rodgers,J., Spizzo,T., Duesterhoeft,S. and Eide,D. (1998) Regulation of zinc homeostasis in yeast by binding of the ZAP1 transcriptional activator to zinc-responsive promoter elements. *J. Biol. Chem.*, **273**, 28713–287120.
33. Liu,X., Lee,C., Granek,J., Clarke,N. and Lieb,J. (2006) Whole-genome comparison of Leu3 binding *in vitro* and *in vivo* reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, **16**, 1517–1528.
34. Tan,K., Feizi,H., Luo,C., Fan,S.H., Ravasi,T. and Ideker,T.G. (2008) A systems approach to delineate functions of paralogous transcription factors: role of the Yap family in the DNA damage response. *Proc. Natl Acad. Sci. USA*, **105**, 2934–2939.
35. Liu,X., Noll,D., Lieb,J. and Clarke,N. (2005) DIP-chip: Rapid and accurate determination of DNA binding specificity. *Genome Res.*, **15**, 421–427.
36. Habib,N., Kaplan,T., Margalit,H. and Friedman,N. (2008) A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS Comput. Biol.*, **4**, e1000010.
37. Chou,S., Lane,S. and Liu,H. (2006) Regulation of mating and filamentation genes by two distinct Ste12 complexes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **26**, 4794–4805.
38. Zhao,Y., McIntosh,K., Rudra,D., Schawalder,S., Shore,D. and Warner,J. (2006) Fine-structure analysis of ribosomal protein gene transcription. *Mol. Cell. Biol.*, **26**, 4853–4862.
39. Rudra,D., Mallick,J., Zhao,Y. and Warner,J.R. (2007) Potential interface between ribosomal protein production and pre-rRNA processing. *Mol. Cell. Biol.*, **27**, 4815–4824.
40. Marion,R.M., Regev,A., Segal,E., Barash,Y., Koller,D., Friedman,N. and O'Shea,E.K. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl Acad. Sci. USA*, **101**, 14315–14322.
41. Liu,H., Mulholland,N., Fu,H. and Zhao,K. (2006) Cooperative activity of BRG1 and Z-DNA formation in chromatin remodeling. *Mol. Cell. Biol.*, **26**, 2550–2559.
42. Badis,G., Chan,E.T., van Bakel,H., Pena-Castillo,L., Tillo,D., Tsui,K., Carlson,C.D., Gossett,A.J., Hasinoff,M.J., Warren,C.L. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–87.
43. Zhu,C., Byers,K., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D., Saulrieta,K., Smith,Z., Shah,M., Radhakrishnan,M. *et al.* (2009) High-resolution DNA binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
44. Liaw,P.C. and Brandl,C.J. (1994) Defining the sequence specificity of the *Saccharomyces cerevisiae* DNA binding protein REB1p by selecting binding sites from random-sequence oligonucleotides. *Yeast*, **10**, 771–87.
45. Guldener,U., Munsterkutter,M., Kastenmuller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al.* (2005) CYGD: The comprehensive yeast genome database. *Nucleic Acids Res.*, **33**, D364–D368.
46. Taba,M.R., Muroff,I., Lydall,D., Tebb,G. and Nasmyth,K. (1991) Changes in a SWI4,6-DNA-binding complex occur at the time of HO gene activation in yeast. *Genes Dev.*, **5**, 2000–2013.
47. Ho,Y., Costanzo,M., Moore,L., Kobayashi,R. and Andrews,B.J. (1999) Regulation of transcription at the *Saccharomyces cerevisiae* Start transition by Stb1, a Swi6-binding protein. *Mol. Cell. Biol.*, **19**, 5267–5278.
48. Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, **6**, R18.
49. Cliften,P., Hillier,L., Fulton,L., Graves,T., Miner,T., Gish,W., Waterston,R. and Johnston,M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11**, 1175–1186.
50. Hawkins,J. and Bailey,T.L. (2008) The statistical power of phylogenetic motif models. *Lect. Notes Bioinf.*, **4955**, 112–126.
51. Ward,L.D. and Bussemaker,H.J. (2008) Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**, i165–i171.
52. Zhou,Q. and Wong,W.H. (2004) CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–9.
53. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
54. Odom,D.T., Zizlsperger,N., Gordon,D.B., Bell,G.W., Rinaldi,N.J., Murray,H.L., Volkert,T.L., Schreiber,J., Rolfe,P.A., Gifford,D.K. *et al.* (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science*, **303**, 1378–1381.