A Fast, Alignment-Free, Conservation-Based Method for Transcription Factor Binding Site Discovery

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

Department of Computer Science, Duke University, Durham, NC 27708-0129
{raluca,lee,amink}@cs.duke.edu

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 incorporating conservation information into TF motif 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 does not require sequence alignments, nor the phylogenetic relationships between the orthologous sequences, and yet it is more effective on real biological data than methods that do.

1 Introduction

With recent advances in DNA sequencing technologies, the number of closely related genomes being sequenced [1, 2, 3] has increased tremendously. Consequently, this has led to an increased emphasis on comparative studies focused on detecting functional elements in intergenic DNA sequences. Functional elements, including TF binding sites, are known to evolve at a slower rate than non-functional elements, and therefore DNA sites that are well conserved in orthologous regulatory regions are considered good candidates for TF binding sites.

A plethora of 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, or by incorporating the conservation information into the motif finder itself. The former

* These authors contributed equally to this work.

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 98–111, 2008.

approach has a major limitation: motifs that are not well conserved are likely to be missed. Most conservation-based motif finders therefore take the latter approach. These methods can be further divided into two main categories: 1) 'single gene, multiple species', and 2) 'multiple genes, multiple species'. Methods in the first category (*e.g.*, FootPrinter [4], the phylogenetic Gibbs sampler of Newberg *et al.* [5]) 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.* [1], Converge [6, 7], PhyloCon [8], PhyME [9], PhyloGibbs [10], OrthoMEME [11], EMnEM [12], CompareProspector [13]) are designed to search for motifs that are both over-represented in a set of given sequences (from a reference species) and conserved across related organisms. Our method falls into this category, so for the rest of the paper we will focus only on 'multiple genes, multiple species' approaches.

Most conservation-based approaches to TF binding site discovery rely on multiple or pair-wise alignments of orthologous regulatory regions to assess whether a particular DNA site is conserved across related organisms [1, 6, 7, 9, 10, 12, 13]. However, since binding sites are usually short, sometimes degenerate, and often in reverse orientation or even relocated, alignment algorithms may not correctly align the binding sites within orthologous regulatory sequences. Especially when the sequences are very divergent, the background 'noise' of diverged non-functional regions may be stronger than the 'signal' of conserved motifs, preventing a correct alignment. In Fig. 1 we illustrate four scenarios where motifs in orthologous sequences are not correctly aligned, and thus would most likely be missed by alignment-based motif finders. When a motif changes position or orientation, as in Fig. 1(c,d), correct alignment of motifs may even be impossible.

In consequence, motif finding algorithms based on alignments of orthologous promoter regions will only work when the promoters in the reference species align well with the promoters in the related species (*e.g.*, this is not true for many promoters in *S. cerevisiae* and their orthologs in the non-*sensu stricto Saccharomyces* species used in our analysis). 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 to be conserved in related *Saccharomyces* species [14, 15], a study by Siggia [16] 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 his study, the sequences were aligned using a method by Morgenstern [17]).

Here, we describe a novel, alignment-free method for conservation-based motif discovery. We relax the definition of conserved DNA sites and 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 start with a set of sequences believed to be bound by a common TF in the reference organism. Using orthologous sequences from related organisms, we compute a conservation score for each word and use it to bias our search towards conserved DNA sites. Our method outperforms current conservation-based motif discovery methods

Fig. 1. Examples of conserved TF binding sites in aligned [14] orthologous yeast sequences that can be missed by alignment-based motif discovery programs. The sites matching the motifs of the respective TFs are 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. (The sequences in the figure correspond to *S. cerevisiae*, *S. paradoxus*, *S. kudriavzevii*, *S. mikatae*, *S. bayanus*, and *S. kluyveri*. Due to lack of experimental data, we can only assume the depicted binding sites are functional in organisms other than *S. cerevisiae*.)

in both speed and accuracy. We further show that if negative examples (*i.e.*, sequences believed *not* to be bound by the TF) are also available, we can further improve the performance of our algorithm by considering conservation across those regions as well.

2 Methods

In this section, we describe the generative formulation of motif discovery widely used to find significant motifs in sets of promoters of co-regulated genes. In earlier work [18, 19, 20, 21], we have introduced PRIORITY, a framework for incorporating additional information into motif discovery using informative positional priors. Here, we develop a method for incorporating conservation information across multiple species into our framework. It is important to note that the present paper 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 that is orders of magnitude more efficient than methods proposed to date. Consequently, the methods introduced here can also be adapted to other motif finders beyond PRIORITY.

2.1 Sequence Model and Objective Function

Assume we have n DNA sequences $\boldsymbol{X}_1$ to $\boldsymbol{X}_n$ believed to be commonly bound by some TF. For simplicity, we model at most one binding site in each sequence. This is analogous to the zero or one occurrence per sequence (ZOOPS) model in MEME [22]. Let $\boldsymbol{Z}$ be a vector of length n denoting the starting location of the binding site in each sequence: $Z_i = j$ if a binding site starts at location j in $\boldsymbol{X}_i$ and we adopt the convention that $Z_i = 0$ if $\boldsymbol{X}_i$ contains no binding site. We assume that the TF motif can be modeled as a position specific scoring matrix (PSSM) of length W while the rest of the sequence follows some background model parameterized by $\boldsymbol{\phi}_0$. The PSSM can be described by a matrix $\boldsymbol{\phi}$ where $\phi_{a,b}$ is the probability of finding base b at location a within the binding site for $1 \leq b \leq 4$ and $1 \leq a \leq W$.

Thus if the sequence $\boldsymbol{X}_i$ is of length l_i, and $\boldsymbol{X}_i$ contains a binding site at location Z_i, we can compute the probability of the sequence given the model parameters as:

$$P(\boldsymbol{X}_i \mid \boldsymbol{\phi}, Z_i > 0, \boldsymbol{\phi}_0) = P(X_{i,1}, \ldots X_{i,Z_i-1} \mid \boldsymbol{\phi}_0) \times \left(\prod_{a=1}^{W} \phi_{a,X_{i,Z_i+a-1}} \right) \times P(X_{i,Z_i+W}, \ldots X_{i,l_i} \mid \boldsymbol{\phi}_0)$$

and if it instead does not contain a binding site as:

$$P(\boldsymbol{X}_i \mid \boldsymbol{\phi}, Z_i = 0, \boldsymbol{\phi}_0) = P(X_{i,1}, X_{i,2} \ldots X_{i,l_i} \mid \boldsymbol{\phi}_0)$$

We wish to find $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ that maximize the joint posterior distribution of all the unknowns given the data. Assuming priors $P(\boldsymbol{\phi})$ and $P(\boldsymbol{Z})$ over $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ respectively, our objective function is:

$$\arg\max_{\boldsymbol{\phi},\boldsymbol{Z}} P(\boldsymbol{\phi}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\phi}_0) = \arg\max_{\boldsymbol{\phi},\boldsymbol{Z}} P(\boldsymbol{X} \mid \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{\phi}_0) P(\boldsymbol{\phi}) P(\boldsymbol{Z}) \tag{1}$$

2.2 Optimization Strategy and Scoring Scheme

We use Gibbs sampling to sample repeatedly from the posterior over $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ with the hope that we are likely to visit those values of $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ with the highest posterior probability. Proceeding analogously to the derivation of Liu [23], collapsing $\boldsymbol{\phi}$, we get the final distribution for sampling Z_i:

$$P(Z_i = j \mid \boldsymbol{Z}_{[-i]}, \boldsymbol{X}, \boldsymbol{\phi}_0) = \frac{P(Z_i = j) \times \left(\prod_{a=1}^{W} \phi_{a,X_{i,j+a-1}} \right)}{P(Z_i = 0) \times P(X_{i,j}, \ldots, X_{i,j+W-1} \mid \boldsymbol{\phi}_0)}$$

for $1 \leq j \leq l_i - W + 1$, and $P(Z_i = j \mid \boldsymbol{X}, \boldsymbol{\phi}_0) = 1$ for $j = 0$, where $\boldsymbol{\phi}$ is calculated from the counts of the sites contributing to the current alignment $\boldsymbol{Z}_{[-i]}$, which is the vector $\boldsymbol{Z}$ without Z_i. In practice, we run the Gibbs sampler, which we call PRIORITY [18], for a predetermined number of iterations after apparent convergence to the joint posterior and output the highest scoring PSSM at the end. We use the single best motif to evaluate the algorithm and compare it with other popular methods.

2.3 Incorporation of Conservation Information

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 will now show how conservation information across related organisms can be incorporated as an informative prior over $\boldsymbol{Z}$.

Assume that we have sequence information from k related organisms. Thus for each sequence $\boldsymbol{X}_i$ in the original species, we have an orthologous sequence $\boldsymbol{X}_i^{(s)}$ where $1 \leq s \leq k$. These sequences may be obtained via a genome alignment or by searching for regions near orthologous genes. A sequence may even be empty if no such region is found in the genome of the corresponding organism.

In this paper, we apply our method to ChIP-chip data [6] from *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 from Siepel *et al.* [14]. We describe two different ways in which this information can be used; the first uses the alignments, while the second does not.

Alignment-based conservation prior

Using multiple alignments of the seven yeast species mentioned earlier, Siepel *et al.* [14] have published a conservation track that is freely available at the UCSC genome browser. This track reports the probability of every position in the *S. cerevisiae* genome being conserved based on a program called PhastCons that fits a two-state phylogenetic HMM to aligned orthologous sequences by maximum likelihood. We use these conservation track probabilities to define a score $\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)$ for the W-mer at position j in the bound sequence $\boldsymbol{X}_i$ as:

$$\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j) = \frac{1}{W} \sum_{t=0}^{W-1} Ph(\boldsymbol{X}_i, j+t) \tag{2}$$

where $Ph(\boldsymbol{X}_i, j)$ is the probability of conservation reported by PhastCons at position j in sequence $\boldsymbol{X}_i$. In practice, while computing $\mathcal{S}_{\mathcal{T}}$, we scale the output of the PhastCons program linearly to lie between 0.1 and 0.9 to avoid singularities in the model. We assume that $\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)$ reflects the probability of the W-mer starting at position j in sequence $\boldsymbol{X}_i$ being a binding site. Note that the values $\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)$ themselves do not define a probability distribution over j. As mentioned earlier, we model each sequence $\boldsymbol{X}_i$ as containing at most one binding site. If $\boldsymbol{X}_i$ has no binding site, then none of the positions in $\boldsymbol{X}_i$ can be the starting location of a binding site. On the other hand, if $\boldsymbol{X}_i$ has one binding site at position j, not only must a binding site start at location j, but also no such binding site should start at any other location in $\boldsymbol{X}_i$. Using a little algebra, we can write:

$$P(Z_i = 0) \propto 1 \qquad \text{and} \qquad P(Z_i = j) \propto \frac{\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)}{1 - \mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)} \quad \text{for} \quad 1 \leq j \leq l_i - W + 1 \tag{3}$$

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 $1 \leq i \leq n$. We call this prior $\mathcal{T}$.

Alignment-free conservation prior

In Section 1, we outlined some of the shortcomings of using alignments to detect conserved binding sites. Due to the short length of most binding sites, multiple alignment algorithms are likely to misalign functional sites that are actually conserved across species (Fig. 1). We therefore describe an alignment-free prior that searches 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. We define a conservation score $\mathcal{S}_\mathcal{C}$ for the W-mer at position j in the bound sequence $\boldsymbol{X}_i$ as:

$$\mathcal{S}_\mathcal{C}(\boldsymbol{X}_i, j) = \frac{1}{k} \sum_{s=1}^{k} I[X_{ij}^W \in \boldsymbol{X}_i^{(s)}] \tag{4}$$

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 to 1. To avoid singularities, as before, we scale $\mathcal{S}_\mathcal{C}$ linearly so that the values lie between 0.1 and 0.9.

We have also explored refinements of this simple approach that weigh sequences based on evolutionary distance, or account for imperfect matches while searching for occurrences of W-mers in orthologous sequences. These extensions did not perform better so we stick here to the simplest version (but see Section 4 for further discussion).

As in the case of $\mathcal{S}_\mathcal{T}(\boldsymbol{X}_i, j)$, $\mathcal{S}_\mathcal{C}(\boldsymbol{X}_i, j)$ is only the probability of the W-mer at position j in sequence $\boldsymbol{X}_i$ being a binding site. To convert these values into a positional prior, we substitute $\mathcal{S}_\mathcal{C}$ for $\mathcal{S}_\mathcal{T}$ in (3). After normalizing the resulting $P(Z_i)$ as shown earlier, we get a valid prior over $\boldsymbol{Z}$, which we call $\mathcal{C}$.

Priors with a discriminative perspective

The scores $\mathcal{S}_\mathcal{T}$ and $\mathcal{S}_\mathcal{C}$ used to compute the priors $\mathcal{T}$ and $\mathcal{C}$, respectively, reflect the probability that a W-mer at a certain position 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. We now describe a method for computing a prior that addresses the issue of conserved regions not specific to the profiled TF.

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 $\boldsymbol{Y}_1^{(s)}$ to $\boldsymbol{Y}_m^{(s)}$ where $1 \leq s \leq k$. We compute a discriminative score $\mathcal{S}_{\mathcal{DT}}(\boldsymbol{X}_i, j)$ by taking into account the conservation score $\mathcal{S}_\mathcal{T}$ over both sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows. For each W-mer in $\boldsymbol{X}$, we ask the following question: "Of all the conserved occurrences of this W-mer, what fraction occur in the bound set?". The motivation behind this is to ensure 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. Since we only know the probability that a certain location is conserved, we count the number of conserved W-mers in expectation, weighing each occurrence of the W-mer according to how conserved it is. Using the score $\mathcal{S}_{\mathcal{T}}$ derived over both sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, we calculate $\mathcal{S}_{\mathcal{DT}}$ as:

$$\mathcal{S}_{\mathcal{DT}}(\boldsymbol{X}_i, j) = \frac{\sum_{(q,r):X^W_{qr}=X^W_{ij}} \mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_q, r)}{\sum_{(q,r):X^W_{qr}=X^W_{ij}} \mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_q, r) + \sum_{(q,r):Y^W_{qr}=X^W_{ij}} \mathcal{S}_{\mathcal{T}}(\boldsymbol{Y}_q, r)} \tag{5}$$

As in the case of $\mathcal{S}_{\mathcal{T}}(\boldsymbol{X}_i, j)$ and $\mathcal{S}_{\mathcal{C}}(\boldsymbol{X}_i, j)$, we convert $\mathcal{S}_{\mathcal{DT}}$ into a positional prior which we call $\mathcal{DT}$. Similarly, we compute the discriminative score $\mathcal{S}_{\mathcal{DC}}$ using the conservation-based score $\mathcal{S}_{\mathcal{C}}$ across $\boldsymbol{X}$ and $\boldsymbol{Y}$, by substituting $\mathcal{S}_{\mathcal{C}}$ for $\mathcal{S}_{\mathcal{T}}$ in equation (5). We convert $\mathcal{S}_{\mathcal{DC}}$ into a positional prior which we call $\mathcal{DC}$.

Fig. 2 shows the scores $\mathcal{S}_{\mathcal{C}}$ and $\mathcal{S}_{\mathcal{DC}}$ over an intergenic sequence belonging to the sequence-set of Ste12. As can be seen, the prior computed with a discriminative perspective is effective in filtering out false peaks. Note that if we assume a constant level of conservation across all W-mers, then priors $\mathcal{C}$ and $\mathcal{T}$ simplify to the widely used uniform prior over $\boldsymbol{Z}$, which we call $\mathcal{U}$. Priors $\mathcal{DC}$ and $\mathcal{DT}$, 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 $\boldsymbol{Y}$; we have shown previously [19] the benefits of using such a discriminative prior. We incorporate these six priors $\mathcal{U}$, $\mathcal{T}$, $\mathcal{C}$, $\mathcal{D}$, $\mathcal{DT}$, and $\mathcal{DC}$ in PRIORITY and call the resulting programs, respectively,

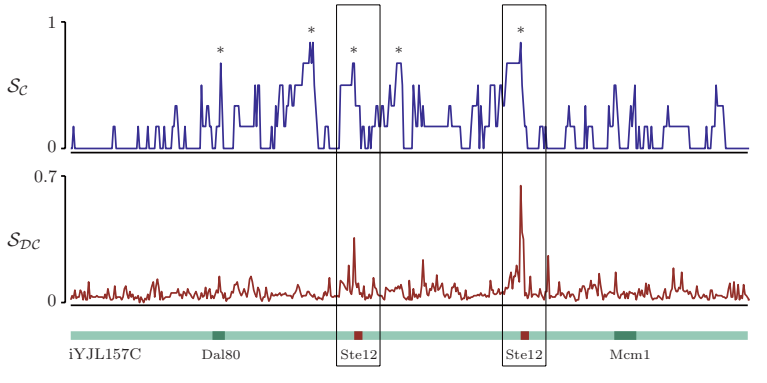

Fig. 2. Scores $\mathcal{S}_{\mathcal{C}}$ and $\mathcal{S}_{\mathcal{DC}}$ computed over intergenic region iYJL157C. Binding sites of Dal80, Ste12, and Mcm1 are shown as annotated by MacIsaac *et al.* [7]. iYJL157C belongs to the sequence-set bound by Ste12 during a ChIP-chip experiment [6]. The score $\mathcal{S}_{\mathcal{DC}}$ is therefore computed from this sequence-set and a sequence-set that is not bound (see text). $\mathcal{S}_{\mathcal{C}}$ has five big peaks, marked with asterisks. Two of them correspond to the start of Ste12 binding sites, one to the start of the Dal80 binding site. The two remaining peaks correspond to conserved `A-T` rich regions. However, the score $\mathcal{S}_{\mathcal{DC}}$ has only two large peaks and both correspond to the start of Ste12 binding sites. This shows that prior $\mathcal{DC}$ is more specific to the profiled TF and effectively filters non-specific peaks corresponding to `A-T` rich regions or other conserved sites.

PRIORITY-$\mathcal{U}$, PRIORITY-$\mathcal{T}$, PRIORITY-$\mathcal{C}$, PRIORITY-$\mathcal{D}$, PRIORITY-$\mathcal{DT}$, and PRIORITY-$\mathcal{DC}$.

3 Results

We compiled ChIP-chip data published by Harbison *et al.* [6], who profiled the intergenic binding locations of 203 yeast TFs under various environmental conditions over 6140 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 ≤ 0.001 in that condition. Similarly, for each TF we define $\boldsymbol{Y}$ to be all intergenic sequences bound with p-value ≥ 0.5. We consider all sequence-sets $\boldsymbol{X}$ of size at least 10 that are bound by TFs with a consensus binding motif in the literature (as used by Harbison *et al.* [6], or as reported in [24, 25]). This leaves us with 156 sequence-sets corresponding to 80 TFs profiled under various conditions. The analysis that follows is performed on those 156 sequence-sets.

It is common practice for methods to be evaluated on synthetically generated promoter data. However, in our framework, the informative priors capture information of biological relevance from true genomic sequences. Therefore, evaluating our method on simulated data is not appropriate.

3.1 Comparison of Priors

Table 1 shows the performance of the six priors when incorporated into PRIORITY[1], on the 156 ChIP-chip sequence-sets with known motifs. Three main conclusions can be drawn from the results in Table 1:

1. Overall, it appears that alignment-based conservation information (at least when used in the form of $\mathcal{T}$) is only slightly more useful than using no information. However, PRIORITY-$\mathcal{T}$ finds 10 motifs that PRIORITY-$\mathcal{U}$ does not, and PRIORITY-$\mathcal{U}$ finds 8 motifs that PRIORITY-$\mathcal{T}$ does not (data available in Supplementary Material). In examining the former 10 cases, it seems the information in the alignment helps. In most of the latter 8 cases, however, we notice that PRIORITY-$\mathcal{T}$ reports motifs with low information content. A closer examination reveals that some of them are weak matches to the literature consensus but do not satisfy our stringent success criterion. It is possible that the alignments produce misleading peaks in the prior at regions other than (or in addition to) the binding sites of the TF, thereby diluting the true motif signal. In the rest of the cases, we believe the alignment is faulty, *i.e.*, the binding sites do not get aligned correctly. Interestingly, one of these 8 sequence-sets corresponds to TF Mac1 and contains the sequence iYLR213C (see Fig. 1).

[1] All the results reported here were obtained with PRIORITY 2.0.0, which implements an improved sampling strategy compared to PRIORITY 1.0.0. This improves the results of baseline priors $\mathcal{U}$ and $\mathcal{D}$ over the results reported earlier [19, 20, 21].

Table 1. Number of motifs correctly identified by PRIORITY when using the six priors described in Section 2. Each version of PRIORITY is run with the default settings (motif width set to 8, and using a third order Markov model to describe the background). Then, for each of the 156 sequence-sets, the top scoring motif is compared with the literature consensus. We call an algorithm 'successful' on a particular sequence-set if this motif is less than a distance of 0.25 from the literature consensus according to the widely used inter-motif distance [6].

Priors	$\mathcal{U}$	$\mathcal{T}$	$\boldsymbol{\mathcal{C}}$	$\mathcal{D}$	$\mathcal{DT}$	$\boldsymbol{\mathcal{DC}}$
Number of successes	58	60	**69**	68	71	**76**

2. Our alignment-free approach, PRIORITY-$\mathcal{C}$, does significantly better than PRIORITY-$\mathcal{T}$ and PRIORITY-$\mathcal{U}$. Since the computation of $\mathcal{S}_{\mathcal{C}}$ depends only on the presence of W-mers across orthologous sequences, this approach is impervious to the alignment artifacts described in Fig. 1, and hence seems to better pick up the true motif signal.
3. In each of the three priors $\mathcal{U}$, $\mathcal{T}$, and $\mathcal{C}$, adopting a discriminative perspective helps find the true motif in many more instances. PRIORITY-$\mathcal{DC}$ does the best: it finds the true motif in 76 sequence-sets across 50 TFs. In fact, there is no sequence-set on which PRIORITY-$\mathcal{DC}$ fails to find the true motif but PRIORITY-$\mathcal{D}$ or PRIORITY-$\mathcal{DT}$ is successful. This shows that, at least on these sequence-sets, conservation information used in this manner does not harm motif discovery.

Since PRIORITY-$\mathcal{T}$ is not much better than PRIORITY-$\mathcal{U}$ (nor is PRIORITY-$\mathcal{DT}$ much better than PRIORITY-$\mathcal{D}$), we will henceforth focus on the performance of our alignment-free motif finders PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$.

3.2 PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ Are More Accurate than Current Conservation-Based Methods

In this section we compare the results of PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$ with the results of six conservation-based motif finders: MEME_c [6], a method of Kellis *et al.* [1], Converge [7], PhyloCon [8], PhyME [9], and PhyloGibbs [10]. All methods fall into the 'multiple genes, multiple species' category, and thus search for motifs that are both over-represented in a set of bound sequences from a species of reference, and conserved across related species. We did not compare with other methods from this category [11, 12, 13, 26] 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; some have been reported to perform worse than methods we include in our analysis; and some were simply not available. We provide more detailed descriptions of all algorithms in the Supplementary Material, along with specific reasons why an algorithm was not selected for comparison in cases where that applies.

Table 2. Number of successfully identified motifs for different conservation-based methods. For each of the 156 sequence-sets, we use the same criterion of success as in Section 3.1.

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	56
Converge	alignment-based; uses EM; incorporates conservation and evolutionary distances into the model	66
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
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	**76**

Table 2 shows the results of PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$ compared to the six conservation-based methods described above. For MEME_c and the method of Kellis *et al.* we use the results reported by Harbison *et al.* [6]; for Converge we use the results reported by MacIsaac *et al.* [7]. We ran PhyloCon version 3b with the default parameter setting and the parameter `s` set to 0.5, as in [7]. However, unlike [7], we did not preprocess the data or postprocess the results reported by PhyloCon. Both PhyME (version 1.2) and PhyloGibbs (version 1.0) were run with their respective default settings, a motif width of 8, and a third order Markov model to describe the background. As recommended by the authors of these programs, we used LAGAN [27] and Sigma [28] to compute alignments for PhyME and PhyloGibbs, respectively.

These results show that our algorithm PRIORITY-$\mathcal{DC}$ is more effective at finding the true motif than the other methods. Even when negative examples (*i.e.*, sequences believed *not* to be bound by the TF) are not available, PRIORITY with the simple conservation prior $\mathcal{C}$ still performs better than all six methods; when negative examples are available, the performance is higher yet.

3.3 PRIORITY-$\mathcal{C}$ and -$\mathcal{DC}$ Are Orders of Magnitude Faster than Current Conservation-Based Methods

PRIORITY with the conservation priors outperforms other methods not only in terms of accuracy, but also speed. In Fig. 3 we show a log-scale plot of the running

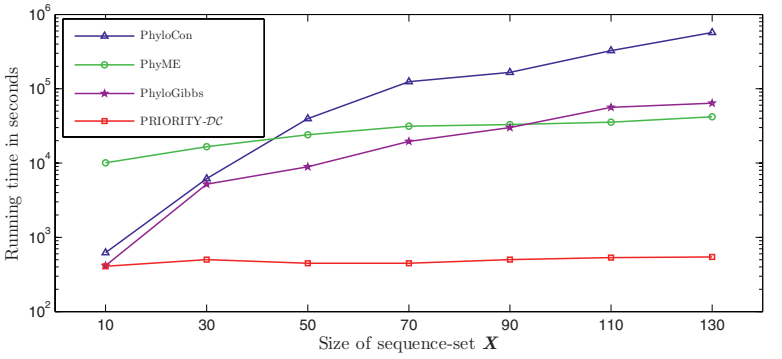

Fig. 3. Log-scale plot of running times of conservation-based algorithms on sequence-sets of increasing size. Running times for each algorithm include preprocessing steps (*i.e.*, alignment computation for PhyME and PhyloGibbs, and prior computation for PRIORITY-$\mathcal{DC}$). All programs were run on a 3.06GHz Intel Pentium 4 processor.

time of PhyloCon, PhyME, PhyloGibbs, and PRIORITY-$\mathcal{DC}$ for sequence-sets of varying size. Since the running times of PRIORITY-$\mathcal{C}$ and PRIORITY-$\mathcal{DC}$ are comparable (with minor differences in the prior computation making PRIORITY-$\mathcal{C}$ slightly faster), we only show the times for PRIORITY-$\mathcal{DC}$.

The running time of PRIORITY-$\mathcal{DC}$ varies only slightly with increasing number 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 becomes 2-3 orders of magnitude faster than the other three methods.

4 Discussion

We have presented a fast motif discovery algorithm that uses sequence conservation across related organisms without relying on alignments. Our method outperforms currently used conservation-based programs in both speed and accuracy.

We are not the first to use alignment-free conservation across species to find motifs. Elemento and Tavazoie [29] look for conserved regulatory elements by scanning a pair of related genomes for highly enriched W-mers, on the order of 400. Then they use a hypergeometric distribution to evaluate the significance of each of these W-mers in bound ChIP-chip sets. 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 PSSM. In addition, the approach of Elemento and Tavazoie is limited to pairs of related organisms, and thus the choice of organisms becomes crucial for the success of the algorithm.

In this paper, 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 more organisms improve performance.

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. 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.* [30], 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.

Our conservation-based approach is much faster than current methods. It only needed 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 Fig. 3 we showed 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 conservation-based methods increases substantially (data available in the Supplementary Material).

Currently, the derivation of our conservation-based priors 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, by weighting 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.*

[14]. However, conservation priors computed using the weighted sequences did not show any improvement over the initial conservation priors, $\mathcal{C}$ and $\mathcal{DC}$.

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 one mismatch when searching for conserved words. Since we do not know *a priori* the position in which a mismatch may occur, we allowed it to be anywhere in the W-mer. For example, 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 was that the signal of truly conserved sites was mixed with random noise due to the 24 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, when we know we should be searching for a gapped motif) may be used to restrict the mismatches to certain positions.

Here, we have successfully applied our algorithm on seven *Saccharomyces* species. We believe our approach is even more useful on higher organisms, where motif finding has proven difficult due to longer promoters and smaller fraction of functional elements. We are planning to apply our method on data from higher organisms, including worm, fly, and human.

Supplementary Material can be found at http://www.cs.duke.edu/∼amink/.

References

[1] Kellis, M., et al.: Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 432, 241–254 (2003)
[2] Cliften, P., et al.: Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 301, 71–76 (2003)
[3] Clark, A., et al.: Proposal for *Drosophila* as a model system for comparative genomics (2003), `http://flybase.net/.data/docs/CommunityWhitePapers/GenomesWP2003.html`
[4] Blanchette, M., Tompa, M.: FootPrinter: a program designed for phylogenetic footprinting. Nucleic Acids Research 31, 3840–3842 (2003)
[5] Newberg, L.A., et al.: A phylogenetic Gibbs sampler that yields centroid solutions for *cis*-regulatory site prediction. Bioinformatics 23, 1718–1727 (2007)
[6] Harbison, C., et al.: Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99–104 (2004)
[7] MacIsaac, K.D., et al.: An improved map of conserved regulatory sites for *Saccharomyces* cerevisiae. BMC Bioinformatics 7, 113 (2006)
[8] Wang, T., Stormo, G.D.: Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics 19, 2369–2380 (2003)
[9] Sinha, S., Blanchette, M., Tompa, M.: PhyME: A probabilistic algorithm for Finding Motifs in Sets of Orthologous Sequences. BMC Bioinformatics 5, 170 (2004)

[10] Siddharthan, R., Siggia, E.D., van Nimwegen, E.: PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. PLoS Comp. Biol. 1, e67 (2005)
[11] Prakash, A., Blanchette, M., Sinha, S., Tompa, M.: Motif discovery in heterogeneous sequence data. In: PSB 2004, pp. 348–359 (2004)
[12] Moses, A., Chiang, D., Eisen, M.: Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. In: PSB 2004, pp. 324–335 (2004)
[13] Liu, Y., et al.: Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Research 14, 451–458 (2004)
[14] Siepel, A., et al.: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005)
[15] Chin, C., Chuang, J.H., Li, H.: Genome-wide regulatory complexity in yeast promoters: Separation of functionally conserved and neutral sequence. Genome Res. 15, 205–213 (2005)
[16] Siggia, E.: Computational methods for transcriptional regulation. Current Opinion in Genetics & Development 15, 214–221 (2005)
[17] Morgenstern, B.: A space-efficient algorithm for aligning large genomic sequences. Bioinformatics 16, 1531–1539 (2000)
[18] Narlikar, L., Gordân, R., Ohler, U., Hartemink, A.: Informative priors based on transcription factor structural class improve *de novo* motif discovery. Bioinformatics 392, e384–e392 (2006)
[19] Narlikar, L., Gordân, R., Hartemink, A.: Nucleosome Occupancy Information Improves *de novo* Motif Discovery. In: Speed, T., Huang, H. (eds.) RECOMB 2007. LNCS (LNBI), vol. 4453, pp. 107–121. Springer, Heidelberg (2007)
[20] Narlikar, L., Gordân, R., Hartemink, A.: A Nucleosome-Guided Map of Transcription Factor Binding Sites in Yeast. PLoS Computational Biology 3, e215 (2007)
[21] Gordân, R., Hartemink, A.: Using DNA duplex stability information to discover transcription factor binding sites. In: PSB 2008, vol. 13, pp. 453–464 (2008)
[22] Bailey, T., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: ISMB 1994, pp. 28–36 (1994)
[23] Liu, J.: The collapsed Gibbs sampler with applications to a gene regulation problem. Journal of the American Statistical Association 89, 958–966 (1994)
[24] Dorrington, R.A., Cooper, T.G.: The DAL82 protein of *Saccharomyces cerevisiae* binds to the DAL upstream induction sequence (UIS). Nucleic Acids Research 21, 3777–3784 (1993)
[25] Jia, Y., Rothermel, B., Thornton, J., Butow, R.A.: A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. Molecular and Cellular Biology 17, 1110–1117 (1993)
[26] Li, X., Wong, W.H.: Sampling motifs on phylogenetic trees. PNAS 102, 9481–9486 (2005)
[27] Brudno, M., et al.: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 13, 721–731 (2003)
[28] Siddharthan, R.: Sigma: multiple alignment of weakly-conserved non-coding DNA sequence. BMC Bioinformatics 7, 143 (2006)
[29] Elemento, O., Tavazoie, S.: Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biology 6, R18 (2005)
[30] Cliften, P.F., et al.: Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. Genome Res. 11, 1175–1186 (2001)