Pacific Symposium on Biocomputing 13:453-464(2008)

USING DNA DUPLEX STABILITY INFORMATION FOR TRANSCRIPTION FACTOR BINDING SITE DISCOVERY

RALUCA GORDÂN, ALEXANDER J. HARTEMINK

Duke University, Dept. of Computer Science,
Box 90129, Durham, NC 27708, USA
E-mail: {raluca,amink}@cs.duke.edu

Transcription factor (TF) binding site discovery is an important step in understanding transcriptional regulation. Many computational tools have already been developed, but their success in detecting TF motifs is still limited. We believe one of the main reasons for the low accuracy of current methods is that they do not take into account the structural aspects of TF-DNA interaction. We have previously shown that knowledge about the structural class of the TF and information about nucleosome occupancy can be used to improve motif discovery. Here, we demonstrate the benefits of using information about the DNA double-helical stability for motif discovery. We notice that, in general, the energy needed to destabilize the DNA double helix is higher at TF binding sites than at random DNA sites. We use this information to derive informative positional priors that we incorporate into a motif finding algorithm. When applied to yeast ChIP-chip data, the new informative priors improve the performance of the motif finder significantly when compared to priors that do not use the energetic stability information.

1. Introduction

An important step in deciphering eukaryotic transcriptional regulatory control is the discovery of TF binding sites. Although the amount of TF binding data and the number of *de novo* motif discovery tools have been increasing over the last few years, the problem of finding and characterizing TF binding sites is far from being solved. Most DNA motif discovery tools focus on finding overrepresented motifs in sets of sequences believed to be bound by certain TFs. Recent tools also use cross-species conservation information, and thus look for overrepresented and conserved motifs. However, these tools do not take into account structural aspects of the physical interaction between DNA molecules and TFs.

We have shown previously that using structural information, such as the structural class of the TF[1] or nucleosome occupancy information[2] can significantly improve the accuracy of motif finders. In this paper, we explore another aspect of the TF-DNA interaction: the stability of the DNA double helix. During transcription, the two DNA strands must be separated so that the RNA polymerase can

Pacific Symposium on Biocomputing 13:453-464(2008)

slide along the DNA molecule and synthesize a nascent protein. Since proximal promoter regions, containing the TATA box and binding sites for *general* TFs, are located immediately upstream of the transcribed gene where transcription is initiated, one would expect these regions to have a low DNA duplex stability. It is not clear, however, whether a low or high DNA duplex stability at *specific* TF binding sites would be more beneficial for transcription initiation.

Some regulatory proteins bind DNA in a single-strand specific manner (*e.g.* the FBP protein in human[3]). However, the crystal structure of many TF-DNA complexes reveals interactions between TFs and both strands of DNA. This suggests that destabilization of the double helix could actually prevent the TFs from binding to their specific sites on the DNA.

Taking this into account, we hypothesis that TF binding sites occur preferentially in regions with high DNA duplex stability. To test this hypothesis, we consider a set of high-confidence TF binding sites in yeast and compare the duplex stability of these binding sites against the stability of randomly selected sites from the same genomic regions. As a measure of stability we use the *helix destabilization profiles* of Bi and Benham.[5] These profiles contain, for each position in a DNA molecule, the incremental free energy needed to separate the base-paired nucleotides at that position.

We will show that the distribution of the average energy needed to separate the base pairs in TF binding sites is significantly different than the distribution of the average energy needed to destabilize random sites, so we use these distributions to derive informative positional priors that we incorporate into our framework for DNA motif discovery, PRIORITY.[1] Intuitively, the first prior simply guides the search towards DNA sites that have a high energy of destabilization, while the second prior gives more weight to motifs with a higher energy of destabilization in the set of bound sequences than in the genome overall. We show that both energy-based priors significantly improve the performance of motif finding.

2. Data and methods

2.1. *TF binding data*

We use the *Saccharomyces cerevisiae* chromatin immunoprecipitation (ChIP-chip) data published by Harbison *et al.*,[7] who profiled 203 TFs in several environmental conditions. For each TF profiled under each condition, we define its bound sequence-set to be those intergenic sequences (probes) reported to be bound with p-value ≤ 0.001. Of the 307 resulting sequence-sets, we use only the 156 sets that contain at least 10 sequences each, and correspond to 80 TFs with known binding sites (as summarized by Harbison *et al.*,[7] or as reported earlier[11,12]). Each sequence-set is identified as *TF_condition* (*e.g.* Mbp1_YPD).

2.2. *DNA duplex stability data*

The B-form structure of the DNA double helix is not invariant. At specific sites, local DNA strand separation must occur for certain processes to take place (*e.g.* initiation of transcription or replication). The problems of characterizing the duplex stability of DNA molecules and finding the locations most susceptible to strand separation have been studied intensively by Benham and collaborators.[4,5,6]

Although eukaryotic chromosomes are linear, it is easier to understand the process of duplex destabilization in the context of circular DNA. These molecules have a constant *linking number*, defined as the number of times either strand links through the closed circle formed by the other strand.[5] All conformational rearrangements that do not break the strands must preserve this constant. The case of linear DNA molecules is similar because they are partitioned into topological domains consisting of closed loops within a chromosome, and these loops have fixed linking numbers in the relaxed state.[5]

Due to transient strand breakage and re-ligation, the actual linking number of a DNA molecule can deviate from the linking number in the relaxed state, a phenomenon known as *DNA superhelicity*. In general, DNA superhelicity is negative *in vivo* (*i.e.* the actual linking number is smaller than the linking number in the relaxed state) and therefore imposes untwisting torsional stresses on the DNA that can destabilize the double helix at specific sites, a phenomenon called *SIDD (stress-induced duplex destabilization)*.[5]

Bi and Benham[5] developed an approximate method for analyzing local destabilization in superhelically stressed DNA molecules. The method uses statistical mechanics and nearest neighbor energetics of local denaturation to find all states with free energy below a certain threshold, among the 2^N possible states for a DNA molecule of size N. Each state can be viewed as a binary array of size N, with each position indicating the state of the base pair at that position (denatured or not). Next, the authors use the ensemble of low energy states to derive a measure of destabilization called the *(helix) destabilization profile*. For each position j in a DNA molecule $\mathcal{X}$, the destabilization profile $G(\mathcal{X}, j)$ represents the incremental free energy needed to separate the base pair at that position.

We use Bi and Benham's online tool WebSIDD[6] to compute the destabilization profiles for all 6140 DNA probes in the yeast TF binding data. Accurate estimation of the energy profile requires that it be computed within a larger genomic context, because the stacking interactions of neighboring base pairs may have non-local influence on the energy profile. For this reason, when computing the profile for each probe, we include 1000 base pairs upstream and downstream.

2.3. *Average destabilization energy at TF binding sites vs. random sites*

To compute the average energy of destabilization at TF binding sites we use the 4312 high-confidence sites reported by MacIsaac *et al.*[8] The width of these binding sites varies from 5 to 13 nucleotides. Since in our study we primarily search for motifs of size 8—whose length can be refined later using criteria such as information content—we restrict our attention to the 2740 binding sites of size 7 to 9 nucleotides. For every resulting binding site $\mathcal{B}$ we compute the energy of destabilization $G(\mathcal{B})$ as the average of the destabilization profiles $G(\mathcal{B}, j)$ for all positions j in the site.

We build a histogram of the energies of the 2740 binding sites, normalize the values to get a valid probability distribution, and then use a moving average to obtain a smooth distribution of energy values, plotted as a CDF in Figure 1. For every energy value e, this distribution represents the probability of a DNA site $\mathcal{S}$ having that energy, given that $\mathcal{S}$ is a true TF binding site, *i.e.* $P(G(\mathcal{S}) = e \mid \mathcal{S} \in \text{TFBS})$, where TFBS is the set of all binding sites.

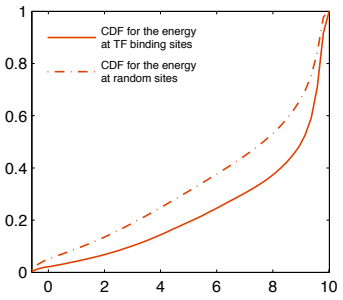

Figure 1. The cumulative distribution functions (CDFs) for the average energy of destabilization at TF binding sites (solid) versus random DNA sites (dashed). A two-sample Kolmogorov-Smirnov test indicates these two distributions to be different at a p-value of 2×10^{-68}.

Next, for each high-confidence binding site $\mathcal{B}$ of size 7 to 9 nucleotides we randomly select 20 DNA sites of the same size, from the same intergenic sequence as $\mathcal{B}$. We compute the energy of destabilization for each of the 54,800 random sites, and use these values to build the distribution of energies for random DNA sites, plotted as a CDF in Figure 1. For every energy value e, this distribution gives us the probability of a DNA site $\mathcal{S}$ having that energy, *i.e.* $P(G(\mathcal{S}) = e)$.

We can now use Bayes rule to compute the probability that a DNA site $\mathcal{S}$ is a TF binding site, given its energy:

$$P(\mathcal{S} \in \text{TFBS} \mid G(\mathcal{S})) = \frac{P(G(\mathcal{S}) \mid \mathcal{S} \in \text{TFBS}) \times P(\mathcal{S} \in \text{TFBS})}{P(G(\mathcal{S}))} \qquad (1)$$

The only unknown term on the right side of Eq. (1) is the prior probability of $\mathcal{S}$ being a TF binding site. We estimate this term using the frequency of random DNA sites that have a significant overlap with any of the known TF binding sites, as reported by MacIsaac *et al.*[8]

Given that the distributions of the average energy of destabilization are significantly different for true TF binding sites compared to random sites, we can

leverage this information to improve TF binding site discovery. More precisely, we use $P(\mathcal{S} \in \text{TFBS} \mid G(\mathcal{S}))$, as defined in Eq. (1), to derive informative positional priors that we incorporate into PRIORITY,[1] our generative framework for identifying motifs in sets of DNA sequences.

2.4. *The PRIORITY framework*

Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be a set of n DNA sequences reported to be bound by the same TF. For simplicity, we assume that each DNA sequence contains at most one binding site of the TF, and we use a vector $\boldsymbol{Z}$ to denote the starting location of the binding site in each sequence: $Z_i = j$ if there is a binding site starting at location j in $\boldsymbol{X}_i$. Since the TF binding data may have been affected by experimental errors, we also allow for the DNA sequences to contain no binding sites, and in this case we adopt the convention that $Z_i = 0$.

We model the TF binding sites as position-specific scoring matrices (PSSMs) of length W parameterized by $\boldsymbol{\phi}$, and we assume that the rest of the sequence follows some background model parameterized by $\boldsymbol{\phi}_0$. We fixed the length W of the binding sites to be 8, and the background model $\boldsymbol{\phi}_0$ to be a third order Markov model trained on all intergenic regions in yeast.

The goal of our motif finding algorithm is to find the $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ that maximize the joint posterior distribution of all the unknowns given the data. Assuming independent priors $P(\boldsymbol{\phi})$ and $P(\boldsymbol{Z})$ over $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ respectively, our objective is:

$$\arg\max_{\boldsymbol{\phi},\boldsymbol{Z}} P(\boldsymbol{\phi}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\phi}_0) = \arg\max_{\boldsymbol{\phi},\boldsymbol{Z}} P(\boldsymbol{X} \mid \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{\phi}_0) \times P(\boldsymbol{\phi}) \times P(\boldsymbol{Z}) \quad (2)$$

We use Gibbs sampling to sample repeatedly from the posterior over $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ with the hope that we are going to visit those values of $\boldsymbol{\phi}$ and $\boldsymbol{Z}$ that maximize the posterior probability. Gibbs sampling is a Markov chain Monte Carlo (MCMC) method that approximates sampling from a joint posterior distribution by sampling iteratively from individual conditional distributions.[9] For faster convergence, we apply collapsed Gibbs sampling[10] and integrate out $\boldsymbol{\phi}$ to sample only the Z_i:

$$\begin{aligned} P(Z_i \mid \boldsymbol{Z}_{[-i]}, \boldsymbol{X}, \boldsymbol{\phi}_0) &= P(\boldsymbol{X} \mid Z_i, \boldsymbol{Z}_{[-i]}, \boldsymbol{\phi}_0) \times P(Z_i) / P(\boldsymbol{X} \mid \boldsymbol{Z}_{[-i]}, \boldsymbol{\phi}_0) \\ &\propto P(\boldsymbol{X}_i \mid \boldsymbol{Z}, \boldsymbol{\phi}_0) \times P(Z_i) \end{aligned} \quad (3)$$

Most motif discovery algorithms based on Gibbs sampling strategies implicitly assume a uniform prior over the possible starting locations Z_i of a binding site in each sequence $\boldsymbol{X}_i$, and thus sample only according to the likelihood term. Our algorithm has a great advantage over other motif finders: it allows the incorporation of informative positional priors.

Pacific Symposium on Biocomputing 13:453-464(2008)

2.5. *Building an energy-based positional prior*

Given a DNA sequence $\boldsymbol{X}_i$ and the energy profile $G(\boldsymbol{X}_i, j)$ we derive an informative positional prior in two steps. First, for each W-mer $X_{i,j}^W$ that starts at position j in sequence $\boldsymbol{X}_i$ we compute an energy-based score that reflects the prior probability of the W-mer being a TF binding site:

$$S_{\mathcal{E}}(\boldsymbol{X}_i, j) = P\left(X_{i,j}^W \in \text{TFBS} \mid \langle G(\boldsymbol{X}_i, j)\rangle_W\right) \tag{4}$$

where $\langle G(\boldsymbol{X}_i, j)\rangle_W$ is the average energy of destabilization for the W-mer that starts at position j in sequence $\boldsymbol{X}_i$:

$$\langle G(\boldsymbol{X}_i, j)\rangle_W = \frac{1}{W}\sum_{k=0}^{W-1} G(\boldsymbol{X}_i, j+k) \tag{5}$$

The score $S_{\mathcal{E}}$ can then be calculated from the distributions of the average energy of destabilization, as described in Eq. (1).

The second step in the derivation of the positional prior is to build a valid probability distribution $P(Z_i = j)$ using the energy-based score $S_{\mathcal{E}}$. Note that the values $S_{\mathcal{E}}(\boldsymbol{X}_i, j)$ themselves do not define a probability distribution over j, as they may not sum to 1. In addition, according to our model, we allow for the sequence $\boldsymbol{X}_i$ to contain no binding sites. In this case, none of the positions in $\boldsymbol{X}_i$ can be the starting locations of binding sites, so we must have:

$$P(Z_i = 0) \propto \prod_{u=1}^{l_i - W + 1} (1 - S_{\mathcal{E}}(\boldsymbol{X}_i, u)) \tag{6}$$

where l_i is the length of sequence $\boldsymbol{X}_i$. 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 of the other locations in $\boldsymbol{X}_i$. Formally, we write:

$$P(Z_i = j) \propto S_{\mathcal{E}}(\boldsymbol{X}_i, j) \prod_{\substack{u=1 \\ u \neq j}}^{l_i - W + 1} (1 - S_{\mathcal{E}}(\boldsymbol{X}_i, u)) \quad \text{for} \quad 1 \leq j \leq l_i - W + 1 \tag{7}$$

We then normalize $P(Z_i)$ using the same proportionality constant in Eqs. (6) and (7), so that under the assumptions of our model we have: $\sum_{j=0}^{l_i - W + 1} P(Z_i = j) = 1$, for $1 \leq i \leq n$. Finally, we incorporate this energy-based positional prior into our search algorithm PRIORITY, and we refer to the resulting algorithm as PRIORITY-$\mathcal{E}$.

To visualize how the positional prior $\mathcal{E}$ can improve TF binding site discovery, we show in Figure 2 the score $S_{\mathcal{E}}$ from which the prior $\mathcal{E}$ is computed, over four DNA probes from the sequence-set corresponding to TF Mbp1 profiled in YPD. We notice that most of the Mbp1 sites, depicted as black boxes on the DNA

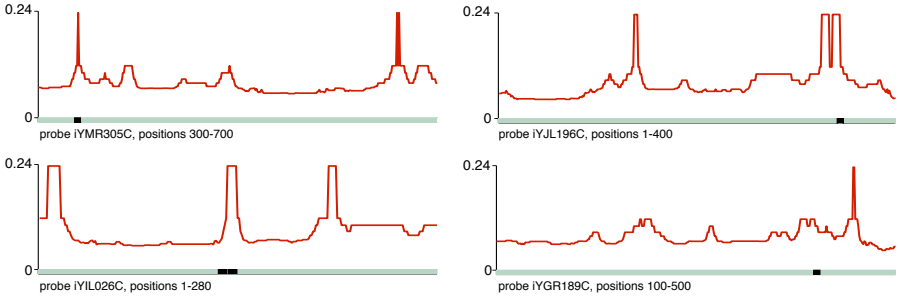

Figure 2. The energy-based score $S_{\mathcal{E}}$ used to compute the $\mathcal{E}$ prior. The x-axes represent DNA probes from the sequence-set Mbp1_YPD. The black boxes on the DNA sequences represent matches to the Mbp1 motif, ACGCGT.

sequences in Figure 2, correspond to peaks of the energy score $S_{\mathcal{E}}$, so they also correspond to peaks of the prior $P(Z_i = j)$. Thus, when prior $\mathcal{E}$ is used for sampling the starting locations of putative binding sites (see Eq. (3)), the locations of the true Mbp1 sites already have a high weight, even before the likelihood information is taken into account.

2.6. *Building a discriminative energy-based positional prior*

In Figure 2 we notice that matches to the Mbp1 motif correspond to peaks of the energy-based score. However, $S_{\mathcal{E}}$ has a number of other peaks that do not correspond to Mbp1 sites. This is not surprising since we cannot expect all the high-energy sites in these DNA sequences to be binding sites of the profiled TF, Mbp1. The other peaks may correspond to binding sites of other TFs, or to other DNA elements that have a high energy of destabilization. To address this issue we build a second informative prior, $\mathcal{DE}$, which uses the energy profiles in a discriminative manner. To do this we need, in addition to the set $\boldsymbol{X}$ of bound sequences, another set $\boldsymbol{Y}$ that contains sequences believed *not* to be bound by the TF in question. Both sets of sequences can be obtained from large-scale experimental methods like ChIP-chip.

The prior $\mathcal{DE}$ is derived similarly to the derivation of the simple energy prior $\mathcal{E}$, but using a new score that takes into account the energy of putative binding sites in both the positive (bound) and the negative (unbound) sequences. For a W-mer $X_{i,j}^W$ starting at position j in sequence $\boldsymbol{X}_i$, the discriminative energy score is defined as the ratio between the sum of the simple energy score for the occurrences of $X_{i,j}^W$ in the positive set, and the sum of the energy score for the occurrences of

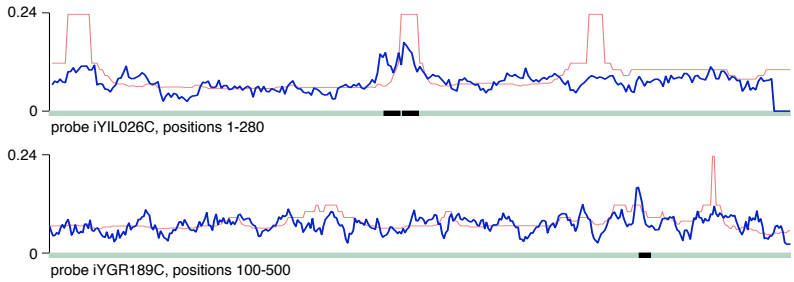

Figure 3. The discriminative energy score $S_{\mathcal{DE}}$ used to compute the $\mathcal{DE}$ prior. The x-axes represent DNA probes from the sequence-set Mbp1_YPD. The lighter curves represent the simple energy score $S_{\mathcal{E}}$ over the same DNA sequences. The black boxes on the DNA sequences represent matches to the Mbp1 motif, ACGCGT.

the same W-mer in both the positive and negative sets:

$$S_{\mathcal{DE}}(\boldsymbol{X}_i, j) = \frac{\sum\limits_{(k,l):X^W_{kl}=X^W_{ij}} S_{\mathcal{E}}(\boldsymbol{X}_k, l)}{\sum\limits_{(k,l):X^W_{kl}=X^W_{ij}} S_{\mathcal{E}}(\boldsymbol{X}_k, l) + \sum\limits_{(k,l):Y^W_{kl}=X^W_{ij}} S_{\mathcal{E}}(\boldsymbol{Y}_k, l)} \tag{8}$$

Using the discriminative score $S_{\mathcal{DE}}$ instead of the simple score $S_{\mathcal{E}}$ we build a valid probability distribution $P(Z_i = j)$, as described in Section 2.5. We call the new prior $\mathcal{DE}$, and we refer to our algorithm with this informative prior PRIORITY-$\mathcal{DE}$.

To illustrate the advantages of the new discriminative prior over the simple energy prior, we show in Figure 3 the score $S_{\mathcal{DE}}$ over the last two DNA sequences in Figure 2 (see the Supplementary Material for plots of $S_{\mathcal{DE}}$ over all four DNA sequences). We notice that in both sequences the highest $S_{\mathcal{DE}}$ peaks correspond to Mbp1 sites. In the first sequence, the simple score $S_{\mathcal{E}}$ has two peaks that do not correspond to Mbp1 sites: the peak on the left corresponds to a Mot3 motif, and the peak on the right to a Swi5 motif. The score $S_{\mathcal{DE}}$ does not contain these two peaks because of its specificity for the profiled TF, which in this case is Mbp1. In the second sequence, the highest peak of $S_{\mathcal{E}}$ is misleading: it corresponds to an imperfect match to the Swi4/Swi6 motif. $S_{\mathcal{DE}}$, however, does not have a peak at this position. Instead, it indicates the correct location of the Mbp1 binding site.

The energy-based priors $\mathcal{E}$ and $\mathcal{DE}$ are derived from distributions of the average energy of destabilization for both known TF binding sites and random DNA sites. When using these priors to find the binding motif of a certain TF, one might worry that occurrences of this motif may have been included in the training data (*i.e.* the set of known binding sites) and therefore the algorithms may be successful simply because they are being tested on some of the data that was used for training. One way to overcome this issue is to remove all the binding sites of the TF

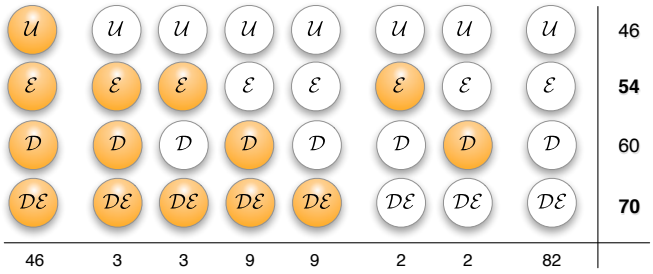

Figure 4. Summary of the results obtained by PRIORITY with priors $\mathcal{U}$, $\mathcal{E}$, $\mathcal{D}$, and $\mathcal{DE}$. Each column represents a possible combination of successes (filled balls) and failures (empty balls) for the four priors. Out of the 16 possible combinations, we only depict those that occur in at least one of the 156 sequence-sets. The number of sequence-sets falling into each category is indicated below the respective column. The last column contains the total number of successes for each algorithm.

in question from the set of known binding sites, compute the two energy distributions, derive the priors, and then apply the algorithms for that TF. We did exactly this and noticed that the two energy distributions were virtually unchanged. This makes sense since the set of binding sites is very large (2740 sites), so leaving out the sites of a particular TF does not influence the distribution of average energy significantly.

3. Results

To assess the performance of PRIORITY-$\mathcal{E}$ and PRIORITY-$\mathcal{DE}$ we use the 156 sequence-sets compiled from the ChIP-chip data of Harbison *et al.*[7] (see Section 2.1). For each sequence-set we run the algorithms 10 times from different random starting points for 10,000 sampling iterations and report the top-scoring motif among the 10 runs. We consider an algorithm to be successful for a sequence-set only if the top-scoring motif is at a distance less than 0.25 from the literature consensus. For details about the distance function, see Narlikar *et al.*[2]

We first compare the performance of the energy-based positional priors with that of a uniform prior $\mathcal{U}$ and a simple discriminative prior $\mathcal{D}$. These two priors are similar to $\mathcal{E}$ and $\mathcal{DE}$, respectively, except that they do not use information about the destabilization energy. We build the uniform prior using a flat score $S_{\mathcal{U}} = 0.5$. The simple discriminative prior $\mathcal{D}$ is calculated similarly to $\mathcal{DE}$, but using the uniform score $S_{\mathcal{U}}$ instead of the energy score $S_{\mathcal{E}}$ in Eq. (8). We incorporate the priors into our framework PRIORITY and refer to the new algorithms as PRIORITY-$\mathcal{U}$ and PRIORITY-$\mathcal{D}$. The results of the four algorithms on the 156 sequence-sets are summarized in Figure 4 and presented in detail in the Supplementary Material.

3.1. *Energy-based priors perform better than uniform prior*

An accurate quantification of the extent to which the energy-based priors improve motif discovery can be obtained by comparing PRIORITY-$\mathcal{E}$ and PRIORITY-$\mathcal{DE}$ with PRIORITY-$\mathcal{U}$.

We notice that PRIORITY-$\mathcal{E}$ is able to find 54 correct motifs, an improvement of 17% over the uniform prior. PRIORITY-$\mathcal{DE}$ performs even better: it finds the correct motif in 70 sequence-sets, 52% more than the uniform prior. Furthermore, we notice that in all the sequence-sets where PRIORITY-$\mathcal{U}$ succeeds, the energy-based priors also succeed, so they are never detrimental to motif discovery. We also mention that in the sequence-set Mbp1_YPD, from which the DNA sequences depicted in Figures 2 and 3 were extracted, PRIORITY-$\mathcal{U}$ is unable to find the correct Mbp1 motif, while both PRIORITY-$\mathcal{E}$ and PRIORITY-$\mathcal{DE}$ succeed.

The improvement of PRIORITY-$\mathcal{DE}$ over PRIORITY-$\mathcal{U}$ is remarkable: 70 correctly found motifs versus 46. We note, however, that this improvement is not due solely to the energy information, but also to the discriminative information. Out of the 24 motifs found by PRIORITY-$\mathcal{DE}$ and not found by PRIORITY-$\mathcal{U}$, 9 motifs are only detected when using the discriminative priors, so it is probably the discriminative information that causes the improvement in these cases. In 9 other sequence-sets, though, the $\mathcal{DE}$ prior is the only one to find the correct motif. This suggests that neither $\mathcal{E}$ nor $\mathcal{D}$ alone contains enough information to identify the true motif, though the combination $\mathcal{DE}$ is successful.

Figure 4 also reveals that there are four cases in which either $\mathcal{E}$ or $\mathcal{D}$ succeeds in finding the correct motif, but $\mathcal{DE}$ fails. We next discuss these cases in more detail. The two sequence-sets where PRIORITY-$\mathcal{E}$ is the only one that finds the correct motif are Met32_SM and Sip4_YPD. In both cases we notice that the occurrences of the true motif in the bound set have a high energy of destabilization, which explains the success of PRIORITY-$\mathcal{E}$, but the two motifs also have a high energy of destabilization overall in the genome, which explains why PRIORITY-$\mathcal{DE}$ fails. We also notice that the sequence-sets Met32_SM and Sip4_YPD contain very few occurrences of the Met32 and Sip4 motifs, respectively. We believe it is possible that some high-energy occurrences of these motifs in the unbound sets are in fact binding sites of the profiled TFs, but were not bound in the particular environmental conditions of the ChIP-chip experiments.

In two sequence-sets, the $\mathcal{D}$ prior succeeds while both energy-based priors fail: Skn7_H2O2Lo and Msn2_H2O2Hi. In the case of Skn7_H2O2Lo, both $\mathcal{E}$ and $\mathcal{DE}$ fail because they get stuck in local optima. If we score the motif found by $\mathcal{D}$ according to the posteriors obtained using $\mathcal{E}$ and $\mathcal{DE}$, we get significantly higher scores than the ones reported by PRIORITY-$\mathcal{E}$ and PRIORITY-$\mathcal{DE}$, respectively, for

their top motifs (which do not match the literature consensus). In the case of Msn2_H2O2Hi, the fact that PRIORITY-$\mathcal{DE}$ does not find the correct motif is due to the motif size, which by default is 8. If we set it to 6—the true size of the Msn2 motif—PRIORITY-$\mathcal{DE}$ succeeds. For the same sequence-set Msn2_H2O2Hi, the failure of PRIORITY-$\mathcal{E}$ seems to be the result of the algorithm getting stuck in a local optimum.

3.2. *Comparison with popular motif finders*

Finally, we present a comparison between the results of our algorithm with energy-based positional priors and the results of six popular motif finders, as reported by Harbison *et al.*[7]: AlignACE,[13] MEME,[14] MDscan,[15] and three methods that use evolutionary conservation information (MEME_c,[7] a method of Kellis *et al.*,[16] and Converge[7]). We emphasize, however, that the goal of this paper is not to introduce a new motif discovery tool, but to show that structural information typically disregarded by motif finders can significantly improve their performance.

Out of the 156 sequence-sets, AlignACE is successful in 16, MEME in 35, MDscan in 54, MEME_c in 49, the method of Kellis *et al.* in 50, and Converge in 56, so our algorithm PRIORITY-$\mathcal{DE}$ outperforms all six methods, with a total of 70 correctly identified motifs. Furthermore, even the simpler PRIORITY-$\mathcal{E}$ outperforms five of the six methods.

4. Discussion

In this paper we demonstrate the benefits of using information about the DNA double-helical stability to detect TF binding sites. Using the energy profiles of Bi and Benham[5] as a measure of stability, we notice that in general more incremental free energy is needed to separate the DNA strands at TF binding sites compared to random sites across the genome. This is not surprising since TF binding sites are usually GC-rich. We stress, however, that the energy profiles we used in our analysis were computed using a complex method that takes into account not only individual base pairs, but also the neighboring effects of other base pairs in the same DNA region. Although there is some correlation between the energy profiles and the GC content of the DNA sequences, using an informative positional prior similar to $\mathcal{E}$ but derived from GC content instead of destabilization profiles did not show any improvement over the uniform prior.

One limitation of using helix destabilization energy is that the only eukaryotic organism whose profile has been made available is yeast. The online tool WebSIDD[5] could in principle be used to compute energy profiles for other eukaryotic genomes, but it is limited to sequences a few kilobases long and a downloadable version of the software is not currently available.

The improvement obtained using the energy-based priors demonstrates, once again, the importance of incorporating structural information into motif discovery algorithms; whenever structural information can be translated into a prior over sequence positions, it can be straightforwardly incorporated into our PRIORITY framework for DNA motif discovery. We have shown that useful positional priors can be derived from knowledge of TF structural class,[1] from nucleosome occupancy information,[2] and now from profiles of helix destabilization energy. The usefulness of each of these sources of information leads naturally to the question of the degree of redundancy among them; for instance, the positioning of nucleosomes may be correlated with DNA duplex stability. However, we observe that only some priors are successful on certain sequence-sets. As one example, although both the discriminative nucleosome prior $\mathcal{DN}^2$ and the discriminative energy prior $\mathcal{DE}$ succeed on 70 sequence-sets, in 10 of these sets, only one of the two succeeds, suggesting that combining the informative priors in a principled way—which is not a trivial task—has the potential to further improve motif discovery using informative positional priors.

Supplementary material is available at *www.cs.duke.edu/∼raluca/psb08/*.

Acknowledgments

This project began during a course taught by Bruce Donald, whom R.G. wishes to thank for his early advice. A.J.H. gratefully acknowledges funding for this work from an NSF CAREER award, an Alfred P. Sloan Fellowship, and awards from NIEHS and NIGMS.

References

1. L.Narlikar, R.Gordân, U.Ohler, A.Hartemink, *Bioinformatics* **22**, e384 (2006).
2. L.Narlikar, R.Gordân, A.Hartemink, *PLoS Comp. Bio.*, in press (2007).
3. R.Duncan *et al.*, *Genes Dev.* **8**, 465 (1994).
4. C.J.Benham, PSB 2001, 103 (2001).
5. C.Bi, C.J.Benham, CSB2003 (2003).
6. C.Bi, C.J.Benham, *Bioinformatics* **20**, 1477 (2004).
7. C.T.Harbison *et al.*, *Nature* **431**, 99–104 (2004).
8. K.D.MacIsaac *et al.*, *BMC Bioinformatics* **7**, 113 (2006).
9. A.Gelfand, A.Smith, *J. Amer. Statistical Assoc.* **85**, 398–409 (1990).
10. J.Liu, *J. Amer. Statistical Assoc.* **89**, 958–966 (1994).
11. R.A. Dorrington, T.G. Cooper, *Nucleic Acids Res.* **21**, 3777–3784 (1993).
12. Y. Jia *et al.*, *Mol. Cell. Biol.* **17**, 1110–1117 (1993).
13. F.Roth *et al.*, *Nature Biotech.* **16**, 939–945 (1998).
14. T.Bailey, C.Elkan, *ISMB '94*, AAAI Press, Menlo Park, pp. 28–36 (1994).
15. X.Liu, D.Brutlag, J.Liu, *Nature Biotech.* **20**, 835–839 (2002).
16. M.Kellis *et al.*, *Nature* **432**, 241–254 (2003).