# Nucleosome Occupancy Information Improves *de novo* Motif Discovery

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

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

**Abstract.** A complete understanding of transcriptional regulatory processes in the cell requires identification of transcription factor binding sites on a genome-wide scale. Unfortunately, these binding sites are typically short and degenerate, posing a significant statistical challenge: many more matches to known transcription factor binding sites occur in the genome than are actually functional. Chromatin structure is known to play an important role in guiding transcription factors to those sites that are functional. In particular, it has been shown that active regulatory regions are usually depleted of nucleosomes, thereby enabling transcription factors to bind DNA in those regions [1]. In this paper, we describe a novel algorithm which employs an informative prior over DNA sequence positions based on a discriminative view of nucleosome occupancy; the nucleosome occupancy information comes from a recently published computational model [2]. When a Gibbs sampling algorithm with our informative prior is applied to yeast sequence-sets identified by ChIP-chip [3], the correct motif is found in 50% more cases than with an uninformative uniform prior. Moreover, if nucleosome occupancy information is not available, our informative prior reduces to a new kind of prior that can exploit discriminative information in a purely generative setting.

## 1 Introduction

Finding functional DNA binding sites of transcription factors (TFs) on a genome-wide scale is a crucial step in understanding transcriptional regulation. Despite an explosion of data about TF binding from high-throughput technologies like ChIP-chip [3, 4, and many more], DIP-chip [5], PBM [6], and gene-expression arrays [7, 8, and many more], *de novo* motif finding remains a difficult problem. The fundamental reason for this is that the binding sites of most TFs are short, degenerate sequences which occur frequently in the genome by chance. The 'signal' of functional sites (which are bound *in vivo*) is overwhelmed by the 'noise' due to the non-functional sites (which are not bound *in vivo*). Distinguishing functional sites from non-functional ones, and inferring the true motif recognized by the TF, is thus a challenge.

Many probabilistic motif discovery methods have been developed to tackle the problem of motif discovery [9, 10]. The standard approach is to look for a pattern common

* These authors contributed equally to this work.

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 107–121, 2007.

to the bound sequences that is statistically enriched with respect to the background distribution of all intergenic sequences. If, in addition to the set of bound sequences, a set of unbound sequences is available, a stronger criterion might insist that the pattern be able to discriminate between the two sets [11, 12, 13, 14, 15]. Unfortunately, due to the low signal-to-noise ratio of binding sites mentioned earlier, these methods generally suffer from low specificity and sensitivity [16].

Often, DNA sequences that match known TF motifs do not appear to be functional *in vivo* TF binding sites. One explanation is that not all parts of the genome are equally accessible to TFs *in vivo*. In particular, since DNA is wound over histone octamers called nucleosomes, the positioning of these nucleosomes provides a possible mechanism for differential access of TFs to potential binding sites [1, 2, 17, 18, 19, 20]. Our goal in this paper is to leverage knowledge about nucleosome positioning to improve *de novo* motif finding.

If we knew exactly what parts of the genome were occupied by nucleosomes in the exact environmental conditions for which we have *in vivo* TF binding data, we could bias our search for TF binding sites to the areas that are free of nucleosomes. Unfortunately, no high-resolution nucleosome occupancy data is available for any organism on a whole-genome scale. In the case of yeast, Yuan *et al.* [20] have reported high-resolution nucleosome occupancy data using tiling arrays, but only for chromosome III. On the other hand, Lee *et al.* [1] have published occupancy data for the whole genome, but it is of low resolution: they report only the average occupancy over each intergenic region.

Recently, Segal *et al.* [2] developed a computational model based on high-quality experimental nucleosome binding data to calculate the average nucleosome occupancy at each nucleotide position in the yeast and chicken genomes. This occupancy is purported to be intrinsic to the DNA sequence, and hence independent of *in vivo* conditions. The authors claim that their predictions explain about 50% of observed *in vivo* nucleosome positions. Here, we use predictions from their model to build informative priors over DNA sequence positions that can be used to improve the accuracy of motif finding. We formulate two different nucleosome occupancy priors: the first is based directly on the predictions of Segal *et al.*, while the second adopts a discriminative perspective, comparing nucleosome occupancy in bound versus unbound sequences. When nucleosome occupancy information is not available, the first prior simplifies to an uninformative uniform prior, but the second simplifies to a new kind of informative prior that can exploit discriminative information in a purely generative setting. This represents a novel approach to discriminative motif discovery that retains the computational benefits of a generative formulation. As we shall see, each of our three informative priors improves upon the uninformative uniform prior.

We choose Gibbs sampling as the search method in our algorithms, but in principle, the priors can be used with any search strategy. Our choice of a position specific scoring matrix (PSSM) [21] as a model for the motif is also arbitrary since our priors can be applied while learning any type of motif model. The purpose of this paper is not to demonstrate the benefits of one search strategy over another or one motif model over another, but to demonstrate the utility of nucleosome occupancy data in constructing informative priors for motif discovery.

## 2 Motif Discovery

In this section, we describe the popular generative formulation of the problem of motif discovery, derive the objective function we seek to optimize, and explain the search methodology that we use to optimize this objective function.

### 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. Although in reality a sequence might have multiple binding sites, for simplicity we model only one binding site in each sequence. Because the experimental data might be erroneous, we also model the possibility of some sequences not having any binding site. 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 there is a binding site starting at location $j$ in $\boldsymbol{X}_i$ and we adopt the convention that $Z_i = 0$ if there is no binding site in $\boldsymbol{X}_i$. We assume that the TF motif can be modeled as a 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 $m_i$, 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:

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

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

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

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:

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

### 2.2 Optimization Strategy and Scoring Scheme

As others before us have done, 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. 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 [23]. Applying the collapsed Gibbs sampling strategy developed by Liu [24] for faster convergence, we can integrate out $\boldsymbol{\phi}$ and sample only the $Z_i$. This results in the following expression for sampling $Z_i$ from its conditional distribution assuming the prior on $\boldsymbol{Z}$ to be independent of the PSSM parameters $\boldsymbol{\phi}$:

$$P(Z_i \mid \boldsymbol{Z}_{[-i]}, \boldsymbol{X}, \boldsymbol{\phi}_0) = \frac{P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\phi}_0)}{P(\boldsymbol{Z}_{[-i]} \mid \boldsymbol{X}, \boldsymbol{\phi}_0)} = \frac{P(\boldsymbol{Z}) \int_{\boldsymbol{\phi}} P(\boldsymbol{X} \mid \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{\phi}_0) P(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi}}{P(\boldsymbol{Z}_{[-i]}) \int_{\boldsymbol{\phi}} P(\boldsymbol{X} \mid \boldsymbol{\phi}, \boldsymbol{Z}_{[-i]}, \boldsymbol{\phi}_0) P(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi}}$$

where $\boldsymbol{Z}_{[-i]}$ is the vector $\boldsymbol{Z}$ without $Z_i$. Proceeding analogously to the derivation of Liu [24], we compute the integrals using a Dirichlet prior on $\boldsymbol{\phi}$. We further simplify the sampling expression by dividing it by $P(Z_i = 0, \boldsymbol{X}_i \mid \boldsymbol{\phi}_0)$ which is a constant at a particular sampling step. This results in the following sampling distribution for a particular location $j$ within sequence $\boldsymbol{X}_i$, similar to the predictive update formula as described in [25]:

$$P(Z_i = j \mid \boldsymbol{Z}_{[-i]}, \boldsymbol{X}, \boldsymbol{\phi}_0) = \frac{P(Z_i = j) \times \left( \prod_{a=1}^{W} \phi_{a, X_{i,j+a-1}} \right)}{P(Z_i = 0) \times P(X_{i,j}, \ldots, X_{i,j+W-1} \mid \boldsymbol{\phi}_0)} \tag{2}$$

for $1 \leq j \leq m_i - W + 1$, and

$$P(Z_i = j \mid \boldsymbol{X}, \boldsymbol{\phi}_0) = 1 \tag{3}$$

for $j = 0$, where $\boldsymbol{\phi}$ is calculated from the counts of the sites contributing to the current alignment $\boldsymbol{Z}_{[-i]}$, plus the pseudocounts as determined by the Dirichlet prior. More details are provided in [26].

The joint posterior distribution after each iteration can be calculated as:

$$P(\boldsymbol{\phi}, \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\phi}_0) \propto P(\boldsymbol{X} \mid \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{\phi}_0) \times P(\boldsymbol{\phi}) \times P(\boldsymbol{Z}) \tag{4}$$

To simplify computation, we divide the above expression by the constant probability $P(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{0}, \boldsymbol{\phi}_0)$ and use the logarithm of the resulting value as a score for the motif.

To maximize the objective function and hence the score, we run the Gibbs sampler for a predetermined number of iterations after apparent convergence to the joint posterior and output the highest scoring PSSM at the end. We report only a single motif $\boldsymbol{\phi}$ to enable us to evaluate the algorithm and compare it with other popular methods. In principle, however, since we are using an MCMC sampling method, we could instead perform Bayesian model averaging over many samples from the posterior and report a mean motif (or multiple motifs if there are multiple modes in the distribution).

## 3 Informative Positional Priors for Motif Discovery

The basic Gibbs sampling approach mentioned above has been used in several motif finders, often with additional parameters and heuristics [27, 28, 29]. However, all these methods use an uninformative prior over the locations $\boldsymbol{Z}$ at which the TF is supposed to bind within the DNA sequences. In a recent paper, we showed how information about the TF's structural class could be leveraged to produce informative priors over $\boldsymbol{Z}$ that significantly help motif discovery [26]. Here, we describe other informative choices for

$P(\boldsymbol{Z})$ which we will henceforth refer to as 'positional priors'. We introduce a prior $\mathcal{N}$ based solely on nucleosome occupancy, a prior $\mathcal{DN}$ incorporating nucleosome occupancy information from both bound as well as unbound sequences, and a discriminative prior $\mathcal{D}$, which is a special case of $\mathcal{DN}$ when nucleosome occupancy information is unavailable. To assess the utility of these priors, we compare their performance to the performance of an uninformative uniform prior $\mathcal{U}$, keeping all other aspects of the algorithm identical.

### 3.1 Building a Positional Prior

The four positional priors mentioned above can be constructed in a similar fashion from different probabilistic scores. We use the term 'probabilistic score' in the remainder of the paper to denote the probability of a particular $W$-mer being a binding site of transcription factor $T$: $S_{i,j}$ = $P(X_{i,j}^{W}$ is a binding site of $T)$, where $X_{i,j}^{W}$ denotes the $W$-mer $X_{i,j}X_{i,j+1}\cdots X_{i,j+W-1}$.

For each sequence $\boldsymbol{X}_i$, we wish to define a prior probability distribution over all possible starting locations $j$ of a binding site in that sequence, i.e. $P(Z_i = j)$. We notice that the values $S_{i,j}$ themselves do not define a probability distribution over $j$, because they may not sum to 1. As mentioned in Section 2.1, we model each sequence $\boldsymbol{X}_i$ as containing at most one binding site of $T$. If $\boldsymbol{X}_i$ has no binding site, none of the positions of $\boldsymbol{X}_i$ can be the starting location of a binding site of $T$ so it must be that:

$$P(Z_i = 0) \propto \prod_{u=1}^{m_i-W+1} (1 - S_{i,u}) \tag{5}$$

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 binding site should start at any of the other locations in $\boldsymbol{X}_i$. Formally, we write:

$$P(Z_i = j) \propto S_{i,j} \prod_{\substack{u=1\\u\neq j}}^{m_i-W+1} (1 - S_{i,u}) \qquad \text{for} \quad 1 \leq j \leq m_i - W + 1 \tag{6}$$

We then normalize $P(Z_i)$ assuming the same proportionality constant in (5) and (6), so that under the assumptions of our model we have:

$$\sum_{j=0}^{m_i-W+1} P(Z_i = j) = 1 \qquad \text{for} \quad 1 \leq i \leq n \tag{7}$$

### 3.2 Uniform Prior ($\mathcal{U}$)

This is the simplest form of positional prior. It is built using a uniform probabilistic score $U_{i,j}$ which assigns equal probabilities to a $W$-mer $X_{i,j}^{W}$ being a binding site of $T$ or not:

$$U_{i,j} = 1 - U_{i,j} = 0.5 \qquad \text{for} \quad 1 \leq j \leq m_i - W + 1 \tag{8}$$

If we substitute $S_{i,j}$ with $U_{i,j}$ in equations (5) and (6) and normalize $P(Z_i = j)$, we get a uniform prior $\mathcal{U}$:

$$P(Z_i = j) = \frac{1}{m_i - W + 2} \qquad \text{for} \quad 0 \leq j \leq m_i - W + 1 \tag{9}$$

### 3.3 Nucleosome Occupancy Prior ($\mathcal{N}$)

A uniform prior is a common choice for a positional prior and most motif finding algorithms implicitly use such a prior. In reality though, as mentioned earlier, certain DNA regions are inaccessible to TFs due to the presence of nucleosomes at those locations.

We would like to bias the search in a probabilistic manner towards nucleosome-free areas. For this purpose, we use the nucleosome occupancy predicted by the computational model developed by Segal *et al.* [2]. This model outputs the probability $N_{i,j}$ of each nucleotide $X_{i,j}$ in the input sequences being occupied by nucleosomes (for now the model is only designed for sequences in yeast or chicken). Assuming nucleosome occupancy indicates inaccessibility, we calculate the average probability of the $W$-mer $X_{i,j}^W$ being accessible to the TF as:

$$A_{i,j} = 1 - \frac{1}{W} \sum_{k=0}^{W-1} N_{i,j+k} \tag{10}$$

Alternatively one could use the maximum instead of the average occupancy over the $W$ nucleotides when computing $A_{i,j}$, but averaging reduces the effect of outliers. Having defined $A_{i,j}$, we can now build the positional prior $\mathcal{N}$ as described in Section 3.1, using $A_{i,j}$ as the probabilistic score $S_{i,j}$.

### 3.4 Discriminative Nucleosome Occupancy Prior ($\mathcal{DN}$)

The formulation of the above probabilistic score $A_{i,j}$ has a drawback: What if a particular $W$-mer is prone to be highly accessible throughout the genome? For instance, certain promoter elements which are required for the assembly of general TFs and are not related to the specific TF in question, might be depleted of nucleosomes. The prior $\mathcal{N}$, in that case, could indicate a high prior belief in that $W$-mer being a TF binding site regardless of the fact that the $W$-mer is equally accessible in the rest of the genome as in the bound set $\boldsymbol{X}$.

Most large-scale high-throughput experimental methods like ChIP-chip, DIP-chip, and PBM give rise to two sets of DNA sequences: those bound by the profiled transcription factor $T$ (positive sequences $\boldsymbol{X}$) and those not bound (negative sequences which we denote as $\boldsymbol{Y}$). The use of the negative set along with the positive set to enrich the motif signal has been shown previously to be beneficial in improving specificity [12, 14, 15]. In the referenced methods, if a $W$-mer is present in the negative set for transcription factor $T$, it is generally treated as an instance of a non-binding site and hence, penalized. However, in an *in vivo* situation, a $W$-mer matching the true motif of $T$ might occur in the negative set but be inaccessible for $T$ due to the presence of a nucleosome at that position. In that case, it should not be treated as a negative data point.

Here, we devise a new discriminative prior which takes into account both these issues. For each $W$-mer $X_{i,j}^W$, we ask the following question: “Of all the accessible occurrences of this word, how many occur in the positive set?” To answer this question, we subject each accessible $W$-mer to a Bernoulli trial. Unfortunately, we cannot tell for sure whether a particular location is accessible or not, because we only know the probability that each location is accessible. Thus, we count the number of accessible sequences in expectation, by weighing each occurrence of the $W$-mer according to how accessible it is. For this purpose, we introduce two functions $r_{k,l}$ and $r'_{k,l}$ defined on the set of all possible $W$-mers $\sigma$:

$$r_{k,l}(\sigma) = \begin{cases} 1 & : \quad X_{k,l}^{W} = \sigma \\ 0 & : \quad X_{k,l}^{W} \neq \sigma \end{cases} \quad \text{and} \quad r'_{k,l}(\sigma) = \begin{cases} 1 & : \quad Y_{k,l}^{W} = \sigma \\ 0 & : \quad Y_{k,l}^{W} \neq \sigma \end{cases} \tag{11}$$

We now define a new probabilistic score $C_{i,j}$ as:

$$C_{i,j} = \frac{\sum\limits_{k,l} A_{k,l} r_{k,l}(X_{i,j}^{W})}{\sum\limits_{k,l} A_{k,l} r_{k,l}(X_{i,j}^{W}) + \sum\limits_{k,l} A'_{k,l} r'_{k,l}(X_{i,j}^{W})} \tag{12}$$

where $A'_{i,j}$ is the accessibility score calculated for the set $\boldsymbol{Y}$ analogous to the calculation of $A_{i,j}$ for $\boldsymbol{X}$ in (10). Using $C_{i,j}$ as our probabilistic score $S_{i,j}$, we can now build the positional prior $\mathcal{DN}$ as described in Section 3.1. In practice, we notice that $C_{i,j}$ can have some false peaks due to $W$-mers that occur very rarely in the genome. In such cases, when the $W$-mer occurs in $\boldsymbol{X}_i$ at some position $j$, $C_{i,j}$ becomes large due to a small denominator. This effect can be alleviated by adding pseudocounts to the expression in (12).

### 3.5 Simple Discriminative Prior ($\mathcal{D}$)

To assess the importance of incorporating nucleosome occupancy information in discriminative motif discovery, we now consider a special case of $\mathcal{DN}$. We assume we have no nucleosome occupancy information, i.e., each $A_{i,j} = c$ and $A'_{i,j} = c$, where $c$ is some arbitrary constant. Equation (12) then reduces to a new probabilistic score $D_{i,j}$:

$$D_{i,j} = \frac{\sum\limits_{k,l} r_{k,l}(X_{i,j}^{W})}{\sum\limits_{k,l} r_{k,l}(X_{i,j}^{W}) + \sum\limits_{k,l} r'_{k,l}(X_{i,j}^{W})} \tag{13}$$

In other words, we calculate the probability $D_{i,j}$ of $X_{i,j}^{W}$ being a binding site of $T$ as the number of occurrences of $X_{i,j}^{W}$ in $\boldsymbol{X}$ relative to the total number of occurrences of $X_{i,j}^{W}$ in both sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ without looking at accessibility. Again, we add pseudocounts while computing $D_{i,j}$ and then calculate a positional prior $P(Z_i = j)$ as described in Section 3.1 by substituting $D_{i,j}$ for $S_{i,j}$. We refer to this positional prior as $\mathcal{D}$.

Note that in computing $\mathcal{D}$ we use only the datasets $\boldsymbol{X}$ and $\boldsymbol{Y}$ and not any nucleosome occupancy information. Other motif discovery algorithms that make use of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ formulate the problem in a discriminative manner, and attempt to learn a motif that appears more often in the positive set than in the negative set. Since these models optimize a discriminative objective function over the sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, they have to deal with a large search space and typically are prone to many local optima. Such methods often require an ‘intelligent guess’ as a seed matrix to initialize the search so as to avoid poor local optima. In addition, at every step of the search algorithm, they have to evaluate the parameters of the model on each sequence in *both* sets. Hence, the time complexity of these algorithms is much worse compared to generative models which iterate only over the positive set. Here, however, our generative model framework remains generative and all the discriminative information is captured in our prior.

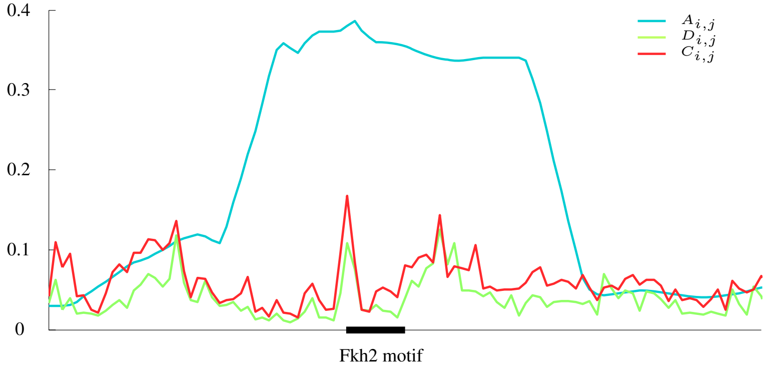

**Fig. 1.** Plot of $A_{i,j}$, $D_{i,j}$, and $C_{i,j}$ used to compute the priors $\mathcal{N}$, $\mathcal{D}$, and $\mathcal{DN}$, respectively. The $x$-axis represents part of an intergenic DNA region from a sequence-set for Fkh2 profiled under the YPD condition in a ChIP-chip experiment [3]. The intergenic region spans positions 770845 to 770945 in Chromosome XVI. The Fkh2 binding site shown in the figure starts at position 770887.

### 3.6 Informative Priors in Action

To visualize how informative priors might be helpful in identifying TF binding sites, we show in Figure 1 the values of $A_{i,j}$, $D_{i,j}$, and $C_{i,j}$ used to compute the priors $\mathcal{N}$, $\mathcal{D}$, and $\mathcal{DN}$ over a portion of a DNA sequence obtained from an Fkh2 ChIP-chip experiment. As can be seen from the figure, in this instance all three priors give a good indication of where a Fkh2 binding site is likely to exist, even before information from the likelihood is taken into account. Of course, this may not happen all the time so we use the remainder of the paper to assess more precisely the relative utility of these priors.

## 4 Results

We compiled ChIP-chip data published by Harbison *et al.* [3], who profiled the intergenic binding locations of 203 yeast TFs under various environmental conditions: YPD, and one or more of Alpha, But14, But90, H202Hi, H202Lo, Pi-, RAPA, or SM over 6140 intergenic regions. These intergenic regions range from 48 to 1553 nucleotides and have an average length of 433 nucleotides. For each TF profiled under each condition, we define its bound sequence-set to be those intergenic sequences reported to be bound with $p$-value $< 0.001$. We restrict our attention to sequence-sets of size at least 10, which yields 242 sequence-sets, encompassing 148 TFs. Of these sequence-sets, 156 correspond to the 80 TFs with a consensus binding motif in the literature (as summarized by Harbison *et al.* at the time their paper was published, or as earlier reported by Dorrington and Cooper [30] or Jia *et al.* [31]), and these 156 are used throughout the remainder of the paper to compare the performance of various motif finding algorithms.

We incorporate the $\mathcal{U}$, $\mathcal{N}$, $\mathcal{D}$, and $\mathcal{DN}$ priors into our Gibbs sampling framework—implemented in PRIORITY [26]—and refer to the resulting algorithms as PRI-$\mathcal{U}$, PRI-$\mathcal{N}$, PRI-$\mathcal{D}$, and PRI-$\mathcal{DN}$, respectively. For evaluation purposes, we fix the motif-width $W$ to 8 in all our runs, although in practice one could certainly explore more values of $W$. As a background model, we use a third order Markov model trained on all intergenic regions in yeast. We run each algorithm 10 times from different random starting points for each sequence-set 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 matches the literature consensus for the corresponding TF. We use a variation of the inter-motif distance measure described by Harbison *et al.* and consider a motif learned by an algorithm to be correct if it is at a distance less than 0.25 from the literature consensus.[1] Different distance cut-offs give different results, but we notice the general trend across all programs remains the same.

Because we are primarily interested in quantifying the extent to which these new informative priors improve *de novo* motif discovery, the results presented in the main portion of the manuscript are limited to a comparison of PRI-$\mathcal{N}$, PRI-$\mathcal{D}$, and PRI-$\mathcal{DN}$ versus PRI-$\mathcal{U}$. However, to ensure that PRI-$\mathcal{U}$ is not simply a 'straw man', but represents a reasonable point of comparison, we have also compiled results from three other popular motif discovery programs as reported by Harbison *et al.* : AlignACE [27], MEME [22], and MDscan [32] (see Supplementary Material). Using the same criterion for success (the top-scoring motif should match the literature consensus), AlignACE is successful in 16 of the 156 sequence-sets, MEME in 35, MDscan in 54, and PRI-$\mathcal{U}$ in 46. AlignACE has one disadvantage over the others in that it uses a first-order Markov model of the background, but each of the three existing methods has advantages over PRI-$\mathcal{U}$: AlignACE considers many motif widths; MEME considers many motif widths, uses sophisticated heuristics to initialize its search, and uses a fifth-order Markov model of the background; and MDscan makes significant use of the $p$-values from the ChIP-chip experiments. Despite these disadvantages, PRI-$\mathcal{U}$ performs admirably, even without an informative prior, and therefore represents a reasonable point of comparison. Since everything about the algorithm is the same apart from the choice of prior, PRI-$\mathcal{U}$ permits the most accurate quantification of the utility of our new informative priors, and so we use it in the remainder of the paper as a baseline when comparing the performance of PRI-$\mathcal{N}$, PRI-$\mathcal{D}$, and PRI-$\mathcal{DN}$.

Figure 2 summarizes the results of the four algorithms on 156 sequence-sets. Overall, while PRI-$\mathcal{U}$ finds the correct motif in 46 sequence-sets, PRI-$\mathcal{DN}$ finds the correct motif in 69 sequence-sets, resulting in an improvement of 50% over baseline. To break down these results more carefully, we divide the sequence-sets into four groups based on the success/failure of PRI-$\mathcal{U}$ and PRI-$\mathcal{DN}$ (corresponding to the four quadrants in Figure 2). This grouping reveals that the $\mathcal{DN}$ prior never performs worse than the $\mathcal{U}$ prior, a claim that is also true of the $\mathcal{D}$ prior, but not of the $\mathcal{N}$ prior. To better understand the performance of these two priors in relation to the $\mathcal{DN}$ prior, we now consider each group in detail:

[1] The distance is normalized to lie between 0 and 1; see Supplementary Material for details about the distance calculation.

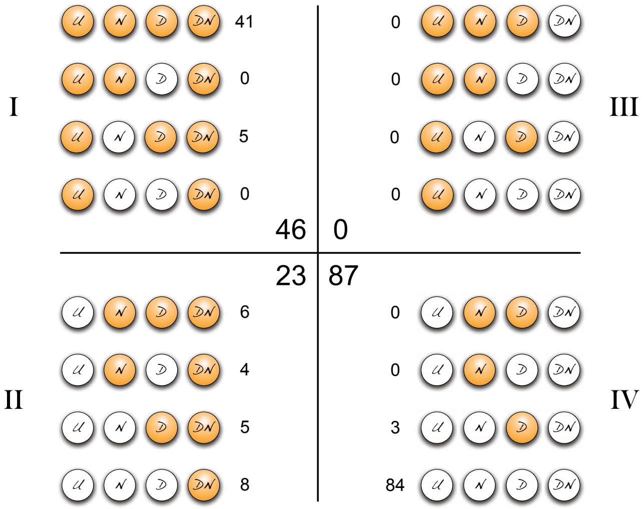

**Fig. 2.** Results of the four algorithms on 156 yeast sequence-sets produced by ChIP-chip experiments [3]. Each row of four balls corresponds to the four positional priors $\mathcal{U}$, $\mathcal{N}$, $\mathcal{D}$, and $\mathcal{DN}$. A filled ball indicates the situation where the respective prior succeeds in finding the true motif. There are $2^4 = 16$ possible combinations of successes/failures of the four priors shown by 16 rows of filled/empty balls. The number of cases resulting in each combination is indicated next to the respective row. The 16 combinations are divided into four quadrants, conditioned on the success/failure of $\mathcal{U}$ and $\mathcal{DN}$. The central numbers indicate the cardinality of each quadrant. As can be seen, some combinations like those in quadrant III do not occur.

**Group I:** PRI-$\mathcal{U}$ succeeds and PRI-$\mathcal{DN}$ succeeds.

This group corresponds to the upper-left quadrant of Figure 2 and it contains 46 sequence-sets corresponding to 31 TFs. For most sequence-sets in this group (41 of 46) all four algorithms find motifs matching the literature consensus. For the other 5 sequence-sets (Cin5_H202Lo, Ste12_Alpha, Ste12_YPD, Hsf1_H202Lo, and Skn7_YPD) PRI-$\mathcal{N}$ is the only algorithm that fails.

Let us look at the case of the TF Ste12 in more detail. In theory, the way the priors are formulated, PRI-$\mathcal{N}$ should work on TFs for which the nucleosome occupancy over the functional binding sites is lower, in general, than the nucleosome occupancy over the rest of the sequences in the set. For PRI-$\mathcal{DN}$ to succeed though, the nucleosome occupancy over the functional sites must be lower than the occupancy over the non-functional sites (that is, sites in the negative set). In both the Alpha and YPD conditions, the average nucleosome occupancy in the sequence-sets is lower than the nucleosome occupancy at the functional binding sites of Ste12. This explains why PRI-$\mathcal{N}$ fails. But according to the analysis of Segal *et al.* [2, Supplemental Figure 36], the average nucleosome occupancy at the functional sites of Ste12 is lower than the average occupancy at the non-functional sites. This clarifies why in spite of using the same nucleosome occupancy data, PRI-$\mathcal{DN}$ succeeds in finding the true motif of Ste12 in both conditions, although PRI-$\mathcal{N}$ does not. This suggests the importance of using nucleosome occupancy information in a discriminative setting.

**Group II:** PRI-$\mathcal{U}$ fails and PRI-$\mathcal{DN}$ succeeds.

This group corresponds to the lower-left quadrant of Figure 2 and it contains 23 sequence-sets corresponding to 19 TFs. In eight cases, PRI-$\mathcal{DN}$ is the only algorithm that succeeds in finding the true motif. This implies that neither $\mathcal{D}$ nor $\mathcal{N}$ alone is strong enough to identify the true motif, but the combination $\mathcal{DN}$ succeeds. In 9 other cases in this group, in addition to $\mathcal{DN}$, exactly one of $\mathcal{D}$ and $\mathcal{N}$ is successful. This suggests that in those cases, the improvement in $\mathcal{DN}$ comes mainly from the respective prior.

**Group III:** PRI-$\mathcal{U}$ succeeds and PRI-$\mathcal{DN}$ fails.

This group, corresponding to the upper-right quadrant of Figure 2, is empty. This implies that whenever the uniform prior succeeds, the $\mathcal{DN}$ prior also succeeds. Thus using this informative prior does not worsen the performance of the algorithm for any sequence-set.

**Group IV:** PRI-$\mathcal{U}$ fails and PRI-$\mathcal{DN}$ fails.

This group corresponds to the lower-right quadrant of Figure 2 and contains 87 sequence-sets corresponding to 50 TFs. For 84 of these 87 sequence-sets, none of the four algorithms finds motifs matching the literature consensus. For the remaining three cases (Msn2_H202Hi, Skn7_H202Lo, and Tec1_YPD) although PRI-$\mathcal{D}$ succeeds, PRI-$\mathcal{DN}$ seems to fail to find the true motif. However, the failure of PRI-$\mathcal{DN}$ seems to be the result of the program getting stuck in a local optimum in each case. When we score the three motifs found by PRI-$\mathcal{D}$ according to the posterior score obtained using the $\mathcal{DN}$ prior, we get a significantly higher score than the score reported by PRI-$\mathcal{DN}$ for the respective top motifs it learns (which do not score well according to the distance metric). The same reasoning applies for the failure of PRI-$\mathcal{N}$ for the sequence-sets of Msn2_H202Hi and Skn7_H202Lo. In the Tec1_YPD sequence-set, however, Tec1 binding sites have an average nucleosome occupancy of $\sim$89% which is higher than the average occupancy over all intergenic regions ($\sim$85%) causing the $\mathcal{N}$ prior to fail.

## 5 Discussion

Although it has been known for a while that nucleosomes control the binding activity of TFs by providing differential access to DNA binding sites [1, 2, 17, 18, 19, 20], we believe we are the first to use nucleosome occupancy information to more accurately predict *de novo* binding sites of TFs.

Our results show that direct use of the nucleosome occupancy predictions of Segal *et al.* [2] as a positional prior does not help motif discovery much: PRI-$\mathcal{N}$ finds 51 correct motifs compared to the 46 found by PRI-$\mathcal{U}$. Motifs of some TFs are more prone to be occupied by nucleosomes than others. The example of Ste12 in Group I illustrates how the prior $\mathcal{N}$ can fail because of the high nucleosome occupancy at Ste12 functional sites. However, when we adopt a discriminative perspective on nucleosome occupancy, the prior $\mathcal{DN}$ succeeds in finding the true Ste12 motif. In fact, there is no sequence-set on which PRI-$\mathcal{N}$ succeeds, but PRI-$\mathcal{DN}$ fails. Overall, our results show that discriminative use of nucleosome occupancy information is extremely useful: PRI-$\mathcal{DN}$ finds 69 true motifs, 50% more than PRI-$\mathcal{U}$. Although in this paper we focus on the usefulness of nucleosome occupancy information, the $\mathcal{D}$ prior also improves motif discovery noticeably without this information: PRI-$\mathcal{D}$ finds 60 true motifs, 30% more than PRI-$\mathcal{U}$. In addition to the three programs AlignACE, MEME, and MDscan discussed earlier, Harbison *et al.* use three conservation-based algorithms to discover motifs: MEME_c, CONVERGE [3], and a method by Kellis *et al.* [33] which find 49, 56, and 50 correct motifs respectively (see Supplementary Material). Not only does PRI-$\mathcal{DN}$ perform much better than these programs, even PRI-$\mathcal{D}$ finds more correct motifs than the best of these programs. This suggests that our prior $\mathcal{D}$ will be quite useful in motif discovery problems even when nucleosome occupancy information is unavailable.

Our discriminative priors (both $\mathcal{D}$ and $\mathcal{DN}$) are novel in the way they incorporate discriminative information in a generative setting. Note that in a specific genome, for a particular $W$-mer $\sigma$ starting at position $j$ in $\boldsymbol{X}_i$, the denominator of (13) remains the same regardless of the sequences in $\boldsymbol{X}$, since it is nothing but the total number of occurrences of $\sigma$ in the whole genome. Similarly, for a particular nucleosome occupancy dataset (experimental or computational), the weighted sum of all accessible sites in the denominator of (12) remains the same for all possible sequence sets $\boldsymbol{X}$. Hence these numbers can be precomputed and stored in a table of size $4^W$. Then, for a particular sequence-set $\boldsymbol{X}$, computing the prior involves one pass (linear-time) over just the sequences in $\boldsymbol{X}$. No information needs to be explicitly computed from the negative set $\boldsymbol{Y}$, which is good because it changes as the positive set changes. In addition, since the actual algorithm only needs to sample over the positive set, the overall time and space complexities of the search are much less than the complexities of other discriminative approaches. In fact, it is practically impossible to compare the performance of PRI-$\mathcal{D}$ with these approaches since the size of the intergenic regions in yeast is about 3 megabases (and larger for metazoan genomes).

In this study, we have fixed $W$ to be 8. In the case of longer motifs, we could post-process the short motif learned by the algorithm and expand it appropriately on either side. Alternatively, we could build priors for multiple values of $W$ and, like most motif finders, run the algorithm with different motif lengths. A larger value for $W$ has certain consequences, however. First, the space required to store priors over $W$-mers is exponential in $W$. Second, as $W$ grows, the average probability of seeing a $W$-mer in the genome decreases, implying that pseudocounts used to smooth the prior become increasingly important (of course, this effect will be mitigated somewhat in larger genomes).

Throughout the paper, we have used PSSMs to model motifs. Although the PSSM is currently a popular choice for a motif model, recent biological [34] and computational [35, 36] findings indicate that more expressive (and hence, more complex) models might be more appropriate. Since our method assigns a prior on the locations within each sequence and not on any specific form of the motif model, it can be used to learn any motif model.

The nucleosome occupancy predictions from the model of Segal *et al.* attempt to capture the static, intrinsic nucleosome binding properties of the DNA. In reality, however, the positioning of nucleosomes changes dynamically as the environmental conditions change or even as the cell progresses through its cell-cycle. Nucleosomes covering certain functional sites might be displaced under specific conditions by other mechanisms to permit access to TFs. It is thus not surprising that Segal *et al.* note that according to their computational model, certain TFs have higher nucleosome occupancy at their functional sites than non-functional sites. If nucleosome occupancy data collected under the same environmental conditions in which the TFs are profiled were available, we would expect to get better results. Unfortunately, at this time high-resolution nucleosome occupancy data is limited. But as more data becomes available, we can incorporate it usefully into our approach.

In closing, we stress that incorporating informative priors over sequence positions is of great benefit to motif discovery algorithms. Low signal-to-noise ratio, especially in higher organisms, makes it difficult to successfully use algorithms based only on statistical overrepresentation. Narlikar *et al.* [26] have shown that using informative priors based on structural classes of TFs improves motif discovery and this paper shows that other kinds of informative priors improve motif discovery as well. Algorithms using conservation information across species [3, 33, 37, 38] are another example of successful incorporation of additional information for motif discovery. We note that although PRI-$\mathcal{DN}$ does better overall than the conservation based methods described earlier, there are certain motifs that one or more of these methods find but PRI-$\mathcal{DN}$ does not. This suggests that combining conservation and nucleosome occupancy might further improve the performance of motif finders. We are currently working toward a unified framework of informative priors based on nucleosome occupancy, TF structural class, and conservation.

**Acknowledgments.** The authors thank Eran Segal for sending them the nucleosome occupancy model, Jason Lieb for sharing unpublished data, and Uwe Ohler for useful discussions and suggestions. The research presented here was supported by a National Science Foundation CAREER award and an Alfred P. Sloan Fellowship to A.J.H.

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

## References

[1] Lee,C., Shibata,Y., Rao,B., Strahl,B., Lieb,J. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nature Genetics*, 36(8): 900–905.

[2] Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thastrom,A., Field,Y., Moore,I., Wang,J., and Widom,J. (2006) A genomic code for nucleosome positioning, *Nature*, 442(7104):772–778.

[3] Harbison,C., *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, 431:99–104.

[4] Lee,T., *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804.

[5] Liu,X., Noll,D., Lieb,J., and Clarke,N. (2005) DIP-chip: Rapid and accurate determination of DNA binding specificity, *Genome Research*, 15(3):421–427.

[6] Mukherjee S., Berger M., Jona G., Wang X., Muzzey D., Snyder M., Young R., and Bulyk M. (2004) Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays, *Nature Genetics*, 36(12):1331–1339.

[7] Spellman,P., Sherlock,G., Zhang,M., Iyer,V., Anders,K., Eisen,M., Brown,P., Botstein,D., and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, 9:3273–3297.

[8] Kim,S., Lund,J., Kiraly,M., Duke,K., Jiang,M., Stuart,J., Eizinger,A., Wylie,B., and Davidson,G. (2001) A gene expression map for *Caenorhabditis elegans*, *Science*, 293:2087–2092.

[9] Wasserman,W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements, *Nat Rev Genet*, 5(4):276–287.

[10] Siggia,E. (2005) Computational methods for transcriptional regulation, *Current Opinion in Genetics and Development*, 15:214–221.

[11] Workman,C. and Stormo,G. (2000) ANN-Spec: A method for discovering transcription factor binding sites with improved specificity, *Pac. Symp. Biocomput.*, 467–478.

[12] Segal,E., Barash,Y., Simon,I., Friedman,N., and Koller,D. (2002) From sequence to expression: A probabilistic framework, *RECOMB '02*.

[13] Sinha,S. (2002) Discriminative motifs, *RECOMB '02*.

[14] Hong,P., Liu,X., Zhou,Q., Lu,X., Liu,J., and Wong,W. (2005) A boosting approach for motif modeling using ChIP-chip data, *Bioinformatics*, 21(11):2636–2643.

[15] Sinha,S. (2006) On counting position weight matrix matches in a sequence, with application to discriminative motif finding, *Bioinformatics*, 22(14):e454–463.

[16] Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol.*, 23(1):137–144.

[17] Almer,A., Rudolph,H., Hinnen,A., and Horz,W. (1986) Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements, *Embo. J.*, 5:2689–2696.

[18] Mai,X., Chou,S., and Struhl,K. (2000) Preferential accessibility of the yeast *his3* promoter is determined by a general property of the DNA sequence, not by specific elements, *Cell Biol.*, 20:6668:6676.

[19] Sekinger,E., Moqtaderi,Z., and Struhl,K. (2005) Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast, *Mol. Cell*, 18:735–748.

[20] Yuan,G., Liu,Y., Dion,M., Slack,M., Wu,L., Altschuler,S., and Rando,O. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*, *Science*, 309:626–630.

[21] Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences, *Nucleic Acids Research*, 12:505–519.

[22] Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *ISMB '94*, AAAI Press, Menlo Park, California, pp. 28–36.

[23] Gelfand,A. and Smith,A. (1990) Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85:398–409.

[24] Liu,J. (1994) The collapsed Gibbs sampler with applications to a gene regulation problem, *Journal of the American Statistical Association*, 89:958–966.

[25] Liu,J., Neuwald,A., and Lawrence,C. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *Journal of the American Statistical Association*, 90:1156–1170.

[26] Narlikar,L., Gordân,R., Ohler,U., and Hartemink,A. (2006) Informative priors based on transcription factor structural class improve *de novo* motif discovery, *Bioinformatics*, 22(14):e384–e392.

[27] Roth,F., Hughes,J., Estep,P., and Church,G. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation, *Nature Biotech.*, 16:939–945.

[28] Liu,X., Brutlag,D., and Liu,J. (2001) BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, *Pac Symp Biocomput.*, 127–138.

[29] Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P., and Moreau,Y. (2002) A Gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes, *Journal of Computational Biology*, 9:447–464.

[30] Dorrington,R.A. and Cooper,T.G. (1993) The DAL82 protein of *Saccharomyces cerevisiae* binds to the DAL upstream induction sequence (UIS), *Nucleic Acids Research*, 21(16):3777-3784.

[31] Jia,Y., Rothermel,B., Thornton,J. and Butow,R.A. (1993) 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.

[32] Liu,X., Brutlag,D., and Liu,J. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments, *Nature Biotech.*, 20:835–839.

[33] 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.

[34] Bulyk,M., Johnson,P., and Church,G. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Research*, 30:1255–1261.

[35] Agarwal,P. and Bafna,V. (1998) Detecting non-adjacent correlations within signals in DNA, *RECOMB '98*

[36] Barash,Y., Elidan,G., Friedman,N., and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites, *RECOMB '03.*

[37] Miller,W., Makova,K., Nekrutenko,A., and Hardison,R. (2004) Comparative Genomics, *Annu. Rev. Genom. Human. Genet.*, 5:15–56.

[38] Siddharthan,R., Siggia,E., and Nimwegen,E. (2005) PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny, *PLoS Comput. Biol.*, 1(7):e67.