Bailey *et al. BMC Bioinformatics* 2010, **11**:179
http://www.biomedcentral.com/1471-2105/11/179

BMC Bioinformatics

RESEARCH ARTICLE Open Access

The value of position-specific priors in motif discovery using MEME

Timothy L Bailey*, Mikael Bodén, Tom Whitington and Philip Machanick

Abstract

Background: Position-specific priors have been shown to be a flexible and elegant way to extend the power of Gibbs sampler-based motif discovery algorithms. Information of many types–including sequence conservation, nucleosome positioning, and negative examples–can be converted into a prior over the location of motif sites, which then guides the sequence motif discovery algorithm. This approach has been shown to confer many of the benefits of conservation-based and discriminative motif discovery approaches on Gibbs sampler-based motif discovery methods, but has not previously been studied with methods based on expectation maximization (EM).

Results: We extend the popular EM-based MEME algorithm to utilize position-specific priors and demonstrate their effectiveness for discovering transcription factor (TF) motifs in yeast and mouse DNA sequences. Utilizing a discriminative, conservation-based prior dramatically improves MEME's ability to discover motifs in 156 yeast TF ChIP-chip datasets, more than doubling the number of datasets where it finds the correct motif. On these datasets, MEME using the prior has a higher success rate than eight other conservation-based motif discovery approaches. We also show that the same type of prior improves the accuracy of motifs discovered by MEME in mouse TF ChIP-seq data, and that the motifs tend to be of slightly higher quality those found by a Gibbs sampling algorithm using the same prior.

Conclusions: We conclude that using position-specific priors can substantially increase the power of EM-based motif discovery algorithms such as MEME algorithm.

Background

Short, linear sequence motifs in protein or nucleic acid sequences are of considerable interest to biologists. This type of sequence pattern is often indicative of important biological sequence signals such as transcription factor binding sites (TFBSs) or splice junctions in nucleotide sequences, and of sumoylation sites and stabilization domains in proteins. Consequently, there has been long and continuing interest in developing software algorithms that can automatically discover functional sequence motifs in sets of biopolymer sequences suspected to harbor one or more common sequence signals.

Biological sequence motifs are often quite subtle, and discovering them in a set of sequences is often impossible since real motifs may be indistinguishable from random artifacts. This has encouraged the development of specialized motif discovery algorithms that can effectively utilize information in addition to the sequences themselves.

One successful approach for improving motif discovery using auxiliary data has been to incorporate evolutionary *conservation* information into the discovery process [1,2]. This approach typically augments the set of input sequences with one or more phylogenetic relatives of each of the original sequences. Such motif discovery algorithms are designed to emphasize motifs that are conserved across related species, on the assumption that such motifs are more likely to be functional.

Another fruitful approach has been to utilize biological information to select a "negative" set of sequences, and to modify the search process to seek motifs that are relatively over-represented in the "positive" sequences. This second approach can be also viewed as discovery of *discriminative* motifs [3,4]. Using negative sequences has the effect of steering the motif discovery process away from sequence patterns that are due to effects unrelated to the particular type of motif being sought. This is desirable when searching for binding site in genomic sequences

* Correspondence: t.bailey@imb.uq.edu.au
[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia
Full list of author information is available at the end of the article

due to the extremely non-random nature of genomic DNA.

A third approach for improving motif discovery has been to seek motifs whose presence in sequences is correlated with some biological signal such as mRNA level. These approaches typically use a *regression* model, and look for motifs that minimize the residual error between a biological signal associated with each input sequence and a motif-based mathematical model of the signal [5,6].

Recently, a general approach has been proposed that allows the incorporation of almost any type of auxiliary information into the class of motif discovery algorithms based on Gibbs sampling [7]. The additional information is converted into a measure of the likelihood that a motif starts at each position in each sequence in the input. This is measure is referred to as a "position-specific prior" (PSP). Gibbs sampling algorithms optimize a Bayesian sequence model, and the PSP approach allows the (summarized) auxiliary information to bias the optimization toward real motifs.

The PSP approach has several advantages. Firstly, it can directly and simultaneously incorporate multiple types of auxiliary data into motif discovery. Secondly, it cleanly separates the problem model optimization from any issues arising from trying to incorporate heterogenous data into the biological sequence model. Thirdly, the PSP methodology can sometimes avoid the severe increase in computational complexity suffered by many of previous approaches to incorporating auxiliary information into motif discovery.

The PSP approach has shown great promise in extending the power of Gibbs sampling-based motif discovery algorithms [7]. For example, a "discriminative conservation" ($\mathcal{DC}$) prior has been shown to be extremely effective for discovering TFBS motifs in yeast sequences when used with a Gibbs sampling algorithm [8]. Incorporation of nucleosome positioning and transcription factor structural class information into a PSP has also proved useful in the discovery of yeast TFBS motifs [9]. However, the benefits of PSPs to EM-based algorithms (such as MEME) has yet to be studied.

In this paper, we describe extending MEME to enable it to use position-specific priors. Like Gibbs sampling-based algorithms, the popular MEME motif discovery algorithm [10] uses a Bayesian probabilistic model in the search for motifs. To allow comparison with previous work, we study the affect of using the $\mathcal{DC}$ PSP with MEME. This PSP combines evidence of evolutionary conservation with the ability of a motif to discriminate between sequences binding the TF and those that do not.

To explore the benefits of using PSPs with MEME, we focus on discovering TFBS motifs in chromatin immunoprecipitation (ChIP) data for yeast and mouse transcription factors. We show that using the $\mathcal{DC}$ PSP greatly improves MEME's ability to discover motifs in an extremely well-studied example of 156 sequence sets derived from TF ChIP-chip (ChIP followed by microrarray) experiments in yeast. In fact, using this PSP, MEME discovers the correct TF motif in more of the yeast ChIP-chip datasets than six other algorithms that use conservation information, including the Gibbs sample using the same PSP. We further show that using the $\mathcal{DC}$ PSP, MEME discovers more *accurate* motifs in mouse ChIP-seq (ChIP followed by sequencing) data [11].

Methods

We describe the enhancements to MEME required for reading in and utilizing a file containing a position-specific prior corresponding to the input DNA or protein sequences. We cover how PSP information is utilized during each of MEME's three major phases. We also describe how MEME converts a prior on motifs of a width w_0 to a prior on motifs of width $w \neq w_0$ in order to allow MEME to discover motifs of a width different than that for which the prior was derived. (Further implementation details are given in Additional file 1.)

Incorporating position-specific priors into MEME

The basic task of biological sequence motif discovery is, given a set of DNA or protein sequences, to determine which positions in the sequences are motif occurrences (sites). MEME does this using a statistical sequence model that it creates based on certain hints provided by the user about the number of sites expected in each sequence and the width of the motif sites. The parameters of the model are referred to here collectively as φ. MEME discovers motifs by optimizing the statistical parameters of the model using the Expectation Maximization (EM) algorithm. The statistical parameters of the model include a position-specific probability matrix (PSPM) representation of the motif, referred to here as θ. The PSPM specifies the probability of each possible letter (amino acid or nucleotide) at each of the w positions in the motif. A motif is a model of aligned words of a specific length k (k-mers), each from a different sequence, representing the likelihood of a given letter appearing at each position.

To discover a motif, MEME proceeds in three phases. In Phase I, MEME determines good *starting points* for the EM algorithm. Since MEME automatically determines the width of the motif and the number of sites, this first phase actually selects starting points for various combinations of motif width and number-of-sites. In Phase II, MEME runs the EM optimization algorithm from each starting point determined in the first phase in order to produce a candidate PSPM representation of the motif. In Phase III, MEME scores each of the candidate motif PSPMs found by EM. To do this, it uses the candidate motif PSPM to predict motif sites, and calculates the statistical significance of the relative entropy of the predicted sites. MEME outputs the candidate motif with the highest statistical significance. Once the motif has been selected and output, MEME "probabilistically erases" the motif's sites, and begins the process again at Phase I in order to discover further motifs.

We describe below how we have enhanced MEME to utilize PSP information in each of its three phases. In what follows, we assume that MEME has been provided with a set of n sequences $\mathbf{X} = \{X_1, ..., X_n\}$. For notational convenience we assume, without loss of generality, that all input sequences have the same length, L.

Overview of position-specific priors

By default, MEME assumes that every position in every sequence is equally likely *a priori* to be a motif site. Position-specific priors allow the user to change this assumption, thereby causing the sequence model to favor motifs that include high-probability sites over those that do not. A PSP defines, for each position in a given set of n sequences, our prior belief that a site (for any motif) starts at that position. To express this in notation, we introduce the binary "missing information" variables $\mathbf{Z} = \{Z_{i,j}\}$, where $Z_{i,j} = 1$ if a site starts at position j in sequence X_i, and $Z_{i,j} = 0$ otherwise. We can then specify a PSP completely by the set of values $\mathbf{P} = \{P_{i,j}\}$, where

$$P_{i,j} = Pr(Z_{i,j} = 1) \text{ for } i \in [1, \ldots, n], j \in [1, \ldots, L].$$

For convenience, we define the special value $P_{i,0}$ to be the prior probability there is *no* motif site anywhere in sequence X_i. To complete our definition of what a PSP is, we add the assumption that a PSP is tied to a particular motif width, w_0. Therefore, the meaning of $P_{i,j}$ is the prior probability of any motif *of width* w_0 having a site at position j in sequence X_i. (We discuss later how MEME derives PSPs of different widths from a fixed-width PSP given in its input.)

MEME only allows sites that fit completely within a sequence, so we require that the last w - 1 positions in a sequence have $P_{i,j} = 0$. MEME can require every sequence to have one site (OOPS sequence model) or it can allow sequence to have zero or one sites (ZOOPS sequence model). Clearly, this implies (based on our definition of $P_{i,0}$, above) that $P_{i,0} = 0$ in the OOPS sequence model. For the ZOOPS model, we allow $P_{i,0}$ to have any value in the range [0, ..., 1]. MEME has one more model–the ANR model–that allows any number of motif sites any sequence. We have not yet implemented PSPs for this model.

MEME searches for motifs in the protein or DNA sequences given in its input. However, MEME can also search for DNA motifs that may have sites on either strand. In that case, we index the sites on the opposite strand from $-L$ to -1 and we then define $Z_{i,j}$ and $P_{i,j}$ for j $[-L, ..., L]$. In order for the $P_{i,j}$ to define a probability distribution, they must all lie in the the range [0, ..., 1] and, for the OOPS and ZOOPS sequence models they must sum to 1 for $i = 1, ..., n$, where n is the number of sequences in the input to MEME. For all sequence models, the sum over site position, j, runs from 0 to L (rather than $-L$ to L) in the protein and single-stranded DNA cases. Note that we define $P_{i,j} = 0$ for all values of j where a motif would not fit entirely within the sequence.

Our implementation of PSPs in MEME has one additional constraint. When we are considering motifs that may occur on either DNA strand (the strand given in the input sequences or its reverse complement), we require that the PSP be *symmetrical*. That is, we require that $P_{i,j} = P_{i,-j}$ for all sequences X_i and sequence positions j. This restriction seems reasonable to us, since the prior probability of any DNA motif in a set sequences by definition is the same as that of its reverse complement motif.

Providing position-specific priors to MEME

MEME can now read PSPs in a format described in Additional file 1. When a PSP file is not provided, MEME assumes, as before, a uniform prior over motif site positions. PSPs can be generated using Hartemink software [8] followed by conversion to MEME's PSP format as described in Additional file 1. The MEME PSP format requires that the set of prior values, $\{P_{i,j}\}$ for $i = 1, ..., N$ and $j = 0, ..., L$, be specified, and that they obey all the constraints described above. For any sequences for which priors are not supplied in the PSP file, priors are calculated as uniform priors. The MEME PSP format includes the width, w_0, of the motifs for which the prior is designed. If MEME is run in double-stranded mode on DNA, the symmetry restriction allows us to generate the PSP for the second strand automatically.

Renormalizing position-specific priors for motifs of different widths

As mentioned above, a PSP is tied to a particular motif width for which it is derived. When MEME is considering motifs of width w, different from w_0, the one specified in the PSP input file, it renormalizes the PSP. The renormalization attempts, in a heuristic fashion, to extend the information captured by the PSP about motifs of width w_0 to a PSP suitable for motifs of width w. Renormalization also insures that the new PSP obeys all of the constraints described above. In particular, when $w > w_0$, there are fewer legal positions for motif sites in a given sequence, so the constraint that the $P_{i,j}$ sum to 1 would be violated without renormalization.

For motifs that are wider than the width for which the input PSP was designed, the renormalized PSP uses the geometric mean of $P_{i,j}$ for all width-w_0 sites that are completely contained by a width-w site. The intuition behind this definition is that the information in each of the completely contained sites should be included in our estimate of the prior probability of the longer site containing them. When $w > w_0$, a width-w site at position j completely contains width-w_0 sites starting at positions j through $j + w - w_0$. If we let $c = w - w_0 + 1$ be the number of shorter sites a longer site contains, our renormalized PSP, PSP$_{(w)}$, is computed as

$$P_{i,j,w} = \left(\prod_{k=0}^{c-1} P_{i,j+k} \right)^{1/c}, \tag{1}$$

for i [1, ..., n] and j [1, ..., $L - w + 1$]. To keep computation costs reasonable, and because the value of information contained in a prior on shorter motifs decreases as the width of the longer motif increases, we constrain $c \leq w_0$ in Eqn. 1.

For motifs that are shorter than those for which the PSP was designed ($w < w_0$), MEME does not renormalize the input PSP. In this case, it simply uses the input PSP as though it were designed for the shorter width motifs, setting PSP$_{(w)}$ = PSP. This has the implication that some potential motif sites at the ends of sequences will be ignored when searching for shorter motifs, since their $P_{i,j}$ values will remain zero even though they are legal starting positions for the shorter motif. For example, if the PSP width is 8 and MEME is searching for motifs of width 7, the last possible position for a motif in each sequence will have $P_{i,j} = 0$. This seems more sensible than setting the value of $P_{i,j}$ based, say, on the value of $p_{i,j\text{-}1}$ since the width-8 PSP contains no explicit information on the prior probability of a site starting at position j. This is because removing the first letter of the word starting at position j - 1 might result in a word with a much lower prior probability. In any case, we expect that useful priors will tend to be relatively short (6 to 10) in relationship to the lengths of the sequences containing the motifs. In what follows, we always assume that the PSP has been normalized to the current motif width being considered by MEME, so we drop the width notation from PSP$_{(w)}$ and $P_{i,j,w}$.

MEME Phase I: Finding Starting Points

To find starting points for EM, MEME converts each subsequence of the data into a "starting" PSPM and calculates a score for it using an algorithm that approximates one step of EM followed by the scoring phase. Creation of the starting PSPM from a subsequence has been previously described [10]. Each such PSPM, θ_M is then used to calculate the probability under the motif model of every potential site in the input sequence, $Pr(site|\theta_M)$. Previously, for the OOPS and ZOOPS models, the single site with the highest likelihood from each sequence was determined. For the OOPS model, these sites were then assigned a score. For the ZOOPS model, these sites were sorted in decreasing order by their likelihoods, and the top t sites for successively larger values of t were scored.

To incorporate PSPs into this phase of MEME, sites are now sorted by a value proportional to their posterior probabilities, $Pr(site|\theta_M)Pr(\theta_M)$, where $Pr(\theta_M)$ is the prior probability of the potential site being a real site, as specified by the PSP. That is, if the site starts at position j in sequence X_i, then $Pr(\theta_M) = P_{i,j}$. We found this approach was not sufficient to insure that the best starting points for EM were found, but that incorporating the PSP into *scoring* the sets of sites with the highest posterior probabilities helped significantly (data not shown). Consequently, the prior probability of each site is now used by MEME when it scores the predicted sites, as described in the next paragraph.

The final score for a potential starting point is a weighted version of the log likelihood ratio (LLR) of its set of predicted sites. The LLR of a set of sites is normally computed by aligning the sites, counting the number of times each letter occurs in each column of the aligned sites, and normalizing the counts to frequencies. To calculate the weighted LLR, MEME scales the individual priors independently in each sequence so that the largest of $P_{i,j}$ in each sequence is 1. These scaled priors are then used as weights on the counts of the numbers of letters in each column of the motif.

In more detail, the weighted LLR is computed by MEME as follows. First, MEME computes weights

$$P'_{i,j} = \frac{P_{i,j}}{P_i^*},$$

where P_i^* is the maximum value of $P_{i,j}$ in sequence X_i. The weighted count, $c_{a,k}$, of letter a in position k of the motif, is computed by adding the weight $P'_{i,j}$, $0 \le P'_{i,j} \le 1$, to $c_{a,k}$ when the site at position j in sequence X_i has letter a in position k of the site. Thus, sites with higher prior values will contribute more to the weighted counts than sites with low prior values. (Note that, with the uniform prior, all the weights $P'_{i,j}$ are 1, so this results in the $c_{a,k}$ being unweighted counts.) The weighted counts are then turned into weighted frequencies by dividing by N_{wt}, $f_{a,k} = c_{a,k}/N_{wt}$, where N_{wt} is the sum of the weights, $P'_{i,j}$, of all sites included in the alignment. We now define a new motif model in terms of parameters $\theta'_M = \{f_{a,k}\}$. Let $p_a = Pr(a|\theta_B)$ be the probability of letter a under the zero-order Markov background model supplied as an input to MEME (the default if none is supplied is a zero-order model based on the letter frequencies of the sequence data). If the weights were all equal to 1, the LLR of the set of sites under this new model would be

$$\begin{aligned} llr &= \log \frac{Pr(sites|\theta'_M)}{Pr(sites|\theta_B)} \\ &= \log \prod_{k=1}^{w} \prod_{a} \left(\frac{f_{a,k}}{p_a} \right)^{c_{a,k}} \\ &= \sum_{k=1}^{w} \sum_{a} c_{a,k} \log \frac{f_{a,k}}{p_a}. \end{aligned} \qquad (2)$$

We refer to Eqn. 2 as the "weighted LLR" of the set of sites when the weights on the sites are not all equal to 1.

For each potential starting PSPM, MEME computes the LLR using Eqn. 2 on different numbers of predicted sites, t. MEME does this by considering only the t predicted sites with the largest posterior probabilities for successively larger values of t. For the OOPS model, the only value of t tried is the number of input sequences, $t = n$.

MEME repeats this entire process for successively larger values of w. For each combination of t and w, MEME runs EM using the potential PSPM that has the largest weighted log likelihood ratio. EM is described in the next section.

Table 1: Definition of terms used in describing the MEME algorithm

n	number of input sequences
L	length of input sequences
$\mathbf{X} = \{X_1, ..., X_n\}$	the set of n input sequences
w	width of a MEME motif
$m = L - w + 1$	number of positions for a site
γ	probability of a site in any sequence
θ	PSPM model of motif;
$\mathbf{P} = \{P_{i,j}\}$	position-specific prior (PSP)
w_0	width for which input PSP is defined
$\mathbf{Z} = \{Z_{i,j}\}$	missing information variables for i [1, n],j [-L, L]
$Z^{(t)}$	expectation of **Z** at EM iteration t
$P_{i,j}^{(t)} = Pr(Z_{i,j} = 1 \mid \varphi^{(t)})$	prior probability given PSP & model
$\varphi^{(t)}$	model parameters at EM iteration t
$\varphi = \{\theta, \gamma, \mathbf{P}\}$	all sequence model parameters

MEME Phase II: Expectation Maximization

MEME uses EM to maximize the expectation of the joint likelihood of the sequence model given the sequences **X** and the *missing information* variables **Z** (refer to Table 1). EM proceeds by iterating an E-step followed by an M-step. The only change required to MEME's existing EM implementation is the replacement of uniform assumption of site positions with the position-specific prior in the E-step.

For OOPS and ZOOPS models, the parameters of the sequence model are $\varphi = \{\theta, \gamma, \mathbf{P}\}$. EM re-estimates the PSPM, θ, but holds fixed the PSP, **P**. The additional parameter, γ, represents the probability that a randomly chosen sequence in the dataset contains a motif site. This is always equal to 1 for the OOPS model, and is estimated by EM for the ZOOPS model.

The E-step of EM computes new estimates of the conditional probabilities of the missing variables **Z**, conditioned on the current estimate of the model parameters,

$$Z^{(t)} = \underset{(\mathbf{Z}|\mathbf{X},\phi^{(t)})}{E}[\mathbf{Z}] \tag{3}$$

where $\varphi^{(t)}$ is the parameter estimate at the start of the current iteration, *t*, of EM. The current estimate of the probability of each possible site based only on the model is $Pr(Z_{i,j} = 1|\varphi^{(t)})$. For notational convenience, we define variables that represent this probability for *j* [1, ..., *L*],

$$P_{i,j}^{(t)} = \begin{cases} (1-\gamma)+\gamma P_{i,0,} & j=0 \\ \gamma P_{i,j,} & 0<|j|\le m \\ 0, & |j|>m \end{cases}$$

where $m = L - w + 1$ is the number of places a motif site will fit in a sequence.

With these definitions, the computation in the E-step of the new estimates of the conditional probabilities of missing variables **Z** for the OOPS and ZOOPS models can be written as

$$\begin{aligned} Z_{i,j}^{(t)} &= Pr(Z_{i,j}=1 \mid X_i, \phi^{(t)}) \\ &= \frac{Pr(X_i|Z_{i,j}=1,\phi^{(t)}P_{i,k}^{(t)}}{\sum\limits_{k=0}^{m} Pr(X_i|Z_{i,k}=1,\phi^{(t)})P_{i,k}^{(t)}}, \end{aligned} \tag{4}$$

for *i* [1, ..., *n*] and *j* [0, ..., *m*]. When searching both DNA strands, the sum in the denominator in Eqn. 4 goes from *-m* to *m*, and we define $Z_{i,j}^{(t)}$ for *j* [*-m*, ..., 0, ...,*m*].

The M step re-estimates φ by solving

$$\phi^{(t+1)} = \underset{\phi}{\operatorname{argmax}} \underset{(\mathbf{Z}|\mathbf{X},\phi^{(t)})}{E}[\log Pr(\mathbf{X},\mathbf{Z} \mid \phi)] \tag{5}$$

The M-step of the EM algorithm in MEME is unchanged. See Bailey and Elkan [10] for more details on how the terms in Eqn. 4 and Eqn. 5 are computed.

MEME Phase III: Scoring the Motifs

The scoring phase of the MEME algorithm assigns scores to the motifs discovered by EM. The criterion is based on the statistical significance of the log-likelihood ratio (Eqn. 2) of the most likely sites for the motif in the sequence dataset. Unlike the starting point phase (Phase I), the scoring phase of MEME computes the *unweighted* LLR, even when using non-uniform positional priors. This choice was motivated by tests which showed that the *weighted* LLR performed no better, so we chose to keep this part of the MEME algorithm unchanged (data not shown). Although the scoring phase of MEME was not changed as a result of incorporating PSPs, it has not been documented previously, so we describe it briefly here.

The significance measure used to rank motifs takes into account the LLR of the motif, its width and the number of sites it contains. The sites of a candidate motif are those with the largest final values of $\mathbf{Z}^{(t)}$. For the OOPS model, MEME scores the motif consisting of these sites. With the ZOOPS model, MEME sorts the sites by decreasing $Z_{i,j}$ value, and scores each prefix of the sorted list.

MEME scores a motif consisting of a set of sites as follows. The LLR of each column of the aligned sites is computed, and the *p*-value of the column-LLR is computed based on the background Markov model using the dynamic programming method of Hertz and Stormo [12]. These *p*-values are then multiplied together and the *p*-value of the resulting product is computed as described in Bailey and Gribskov [13]. (Computing this column-LLR based *p*-value requires far less time than directly computing the *p*-value of the total LLR of the motif.) To make the scores of various motif widths and numbers of sites compatible, MEME multiplies the *p*-value of the motif by the number of possible ways to select positions for the given number of sites in the set of sequences, **X**. This final score is referred to as the *E*-value of the motif.

Measuring the Benefits of using PSPs

To evaluate the benefit of using PSPs in motif discovery, we search for motifs in sets of sequences predicted to bind different TFs in yeast and in mouse. The yeast data is from 156 ChIP-chip experiments each measuring the binding of a single TF [14]. The mouse data is from 13 ChIP-seq experiments measuring binding of a TF [11]. The yeast TF data has been used extensively as a test case for evaluating motif discovery algorithms, so using it allows us to easily compare MEME with PSPs to a large number of other algorithms. Since ChIP-seq data is inherently of a higher quality than ChIP-chip data, the mouse TF data allows us to measure the benefit of using PSPs on a slightly easier motif discovery task. The mouse data covers 13 TFs–Nanog, Oct4, Sox2, Smad1, E2f1, Tcfcp2l1, Ctcf, Zfx, Stat3, Klf4, Esrrb, c-Myc and n-Myc.

We measure accuracy of MEME both with and without the use of PSPs. The PSP we use is the discriminative conservation prior ($\mathcal{DC}$), which has previously been shown to be very effective for discovering TF binding site motifs in the yeast dataset [8]. The $\mathcal{DC}$ prior is based on the degree to which the 8-mer starting at position $X_{i,j}$ is conserved in the input sequence set **X** and a set of orthologous sequences from other species, compared with a negative set of sequences and their orthologs. For comparison, we measure the accuracy of the PRIORITY motif discovery algorithm using the $\mathcal{DC}$ PSP (PRIORITY- $\mathcal{DC}$). We also compare with previously published results on the yeast dataset.

On the yeast data, we use the $\mathcal{DC}$ PSP as reported by Gordân *et al* [8]. This prior is based on intergenic regions from *S. cerevisiae* that have a ChIP-chip fluorescence *p*-value ≤ 0.001 and the orthologous regions from the six related organisms *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*. The negative sequences are all *S. cerevisiae* intergenic regions with a *p*-value ≥ 0.5 and their orthologs.

We create our own $\mathcal{DC}$ PSP for each of the 13 mouse datasets. For each dataset, the positive sequences are 200 base pair (bp) regions centered on the ChIP-seq peaks reported by Chen *et al.* [11]. We use the `mafFrags` program to obtain orthologous sequences for sixteen additional species from the *multiz17way* alignment [15]. We obtain negative sequences and their orthologs for constructing the mouse $\mathcal{DC}$ PSP by extracting 100 bp regions on either side of each positive sequence. We use the $\mathcal{DC}$ -PSP creation scripts provided by the Hartemink group [8] to create the mouse PSPs from the positive and negative sequence and ortholog sets. (More detail including the list of other species is in Additional file 1.)

To measure the accuracy of motif discovery on the yeast datasets, we utilize the same metric as previous researchers [8,14]. This metric compares the single motif reported by a motif discovery algorithm to a known motif for the TF by computing the scaled Euclidean distance between the PSPMs for the motifs. The distance is scaled so that the maximum distance is 1, and the minimum distance is 0. The scaled Euclidean distance between PSPMs f and g is defined as

$$D(f,g)=\frac{1}{w}\sum_{i=1}^{w}\sqrt{\sum_{a\in\{A,C,G,T\}}\frac{(f_{a,i}-g_{a,i})^2}{2}} \qquad (6)$$

where $f_{a,\,i}$ and $g_{a,\,i}$ are the probabilities of base a at position i in the two motifs. We use the same known PSPMs as used by previous researchers [8], and the same criterion for successful motif discovery–scaled Euclidean distance <0.25. Since the reported motif may be of a different length or on the opposite DNA strand from the known motif, we actually compute the minimum value of D for all possible alignments of the reported motif (or its reverse complement) with the known motif, with the minimum overlap the length of the shorter motif.

Our evaluation of motifs discovered in the yeast ChIP-chip datasets utilizes a human-curated set of motifs that represents the consensus predictions of many motif discovery algorithms on those datasets. Such a "gold standard" set of motifs does not exist for the 13 mouse ChIP-seq datasets. Consequently, we take a different approach to measuring the accuracy of motifs discovered in those datasets.

With the mouse ChIP-seq datasets, our underlying measure of motif quality is the amount of correlation between a motif-based binding affinity score and a ChIP-based binding score. We believe that a high correlation between an *in vivo* measure of TF affinity and a motif-based *in silico* measure is indicative of an accurate TF binding motif. (We describe our two binding affinity scores and the correlation measure we use in the next paragraph.) For each ChIP-seq dataset, we measure this correlation in a cross-validation setting, discovering motifs on randomly chosen sets of positive and negative sequences, and computing the correlation measure on held-out sequences. To compare algorithms, we compare our correlation-based quality measure between motifs found on the same sample of sequences.

The details of our evaluation of motifs in the mouse ChIP-seq data are as follows. Our ChIP-based estimate of binding by the ChIP-ed TF at a genomic location is the "peak score" reported by Chen *et al.* [11], and is the normalized count of the number of sequence tags overlapping the peak's genomic location. This is our best direct evidence that the TF was bound in the neighborhood of the peak. Each positive sequence is assigned the peak score of the peak it contains. Our motif-based measure of binding by the ChIP-ed TF is for each positive sequence is its "Average Motif Affinity" (AMA) [16] score. The AMA score is justified as a measure of TF binding affinity on theoretical grounds [17], and it has been used for motif discovery [5] where it showed strong correlation with gene expression, and for motif enrichment analysis [18,19] where it showed strong correlation with TF binding. Because the AMA score estimates the average binding affinity of a region of DNA, it captures contributions from multiple sites in a given region. Our motif quality measure is the Spearman correlation coefficient (CC) between the ranks of the held-out positive sequences sorted by their AMA and peak scores, respectively. We use a rank-based statistic because it is less sensitive than a correlation between the original values to dissimilarities in the distributions being compared [20]. To compare pairs of algorithms, we use each algorithm to learn one motif in each of 50 random samples consisting of 100 positive sequences and 200 negative sequences. We then apply the sign test to the quality measures of pairs of motifs learned on the same input set to decide if one algorithm discovers significantly better motifs.

All yeast runs use a third-order Markov background model, for consistency with reported PRIORITY results. We let MEME search for a single motif of width from 7 to 12, with sites on either DNA strand. For all mouse runs, MEME and PRIORITY-$\mathcal{DC}$ use a fifth order background Markov model, computed from the negative set, and search for motifs of widths 8 to 20. To compute AMA, we use the AMA program, which is part of the MEME suite of programs [21], using the same background model as we use in motif discovery. In all cases, we use PRIORITY 2.0.0 with its default settings, except for changing to a fifth-order background model for the mouse runs. We test MEME with both the OOPS and ZOOPS models, with and without the $\mathcal{DC}$ PSP, with sites on either strand.

Results and Discussion

Improved motif discovery using MEME with PSPs in yeast TF ChIP-chip datasets

Our evaluation of the effect of adding PSPs to MEME starts with measuring improvement in finding TF motifs in yeast ChIP-chip datasets. We run MEME using the OOPS and ZOOPS models with and without the $\mathcal{DC}$ prior on each of 156 ChIP-chip datasets, and compare the single reported motif PSPM to the known PSPM for the TF pulled down in the ChIP-chip experiment. Success is defined as scaled Euclidean distance <0.25 between the reported PSPM and the known PSPM for the TF. Note that, to insure our results are directly comparable to the results reported by Gordân *et al.* [8], we use the script provided by them to compute the scaled Euclidean distance, which reports a distance of 1.0 (the maximum) if the found motif does not contain a region of width six with average information content at least 1 bit.)

The improvement in the number of motifs successfully discovered is quite dramatic. Using the $\mathcal{DC}$ PSP with MEME more than doubles the number of yeast TF motifs successfully discovered (Table 2). The most successful approach is using MEME with the ZOOPS model with the $\mathcal{DC}$ prior (ZOOPS-$\mathcal{DC}$), which discovers the correct motif in 81 of the 156 datasets. Without the $\mathcal{DC}$ prior, MEME with the ZOOPS model only discovers the correct motif in 39 of the datasets.

The accuracy of motif discovery by several other algorithms using these same yeast TF ChIP-chip datasets and success metric has been reported previously [8,22], allowing us to compare our current results more broadly. As seen in Table 2, the success rate of ZOOPS-$\mathcal{DC}$ (81 motifs found) is substantially higher than a number of conservation-based EM or Gibbs sampler motif discovery algorithms (PhyloCon [22], PhyME [23], PhyloGibbs [1], Converge [22], PRIORITY-C [8]).

The ZOOPS-$\mathcal{DC}$ approach also performs at least as well on the yeast datasets as the Gibbs sampler-based algorithm PRIORITY, when PRIORITY is given the same $\mathcal{DC}$ PSP as MEME. The developers of the PRIORITY algorithm (and of the PSP concept) reported a success rate of 76 out of 156 on the yeast datasets (result shown in Table 2 above the horizontal line). However, since Gibbs sampling algorithms are stochastic–their outputs vary from run to run–we wished to place error bars on PRIORITY-$\mathcal{DC}$'s success rate. We therefore ran PRIORITY-$\mathcal{DC}$ ten times on each yeast dataset. The success rate varied from 65 to 74 correct motifs, with an average success rate of 69 (sd = 3), as shown in the last line of Table 2. The fact that we did not observe any PRIORITY-$\mathcal{DC}$ run with as high a success rate as previously reported [8] may be a result of the stochastic nature of the algorithm, or may be due to us using a more recent version of PRIORITY (Version 2.0.0).

Improved motif discovery using MEME with PSPs in mouse TF ChIP-seq datasets

As an additional check on the value of using PSPs with MEME, we measure the improvement in TF motif discovery on 13 mouse TF ChIP-seq datasets. Our evalua-

Table 2: Performance of motif discovery algorithms on yeast TF ChIP-chip datasets.

Algorithm	Description	Successes	Proportion of Successes
PhyloCon	local alignment of conserved regions	19	12%
PhyME	alignment-based; uses EM	21	13%
MEME_c	MEME run with non-conserved bases masked	49	31%
PhyloGibbs	similar to PhyME but uses Gibbs sampling	54	35%
Kellis *et al*.	alignment-based	56	36%
Converge	alignment-based; uses EM	66	42%
PRIORITY-$\mathcal{C}$	Gibbs sampler with conservation-based priors	69	44%
PRIORITY-$\mathcal{DC}$	Gibbs sampler with discriminative conservation-based priors	76	49%
MEME: OOPS	MEME with OOPS model	36	23%
MEME: ZOOPS	MEME with ZOOPS model	39	25%
MEME: OOPS-$\mathcal{DC}$	MEME with OOPS model and $\mathcal{DC}$ priors	73	47%
MEME: ZOOPS-$\mathcal{DC}$	MEME with ZOOPS model and $\mathcal{DC}$ priors	81	52%
PRIORITY-$\mathcal{DC}$	Gibbs sampler with discriminative conservation-based priors	69 (3)	44%

The table shows the number motifs (out of 156) successfully discovered by the named algorithms. The results in the top half of the table are taken from Gordân *et al*. [8]. Results in the bottom half are for new experiments performed by us. Each algorithm is allowed to report one motif, and success is declared if the scaled Euclidean distance to the known PSPM is <0.25. Proportions (out of 156) successes are rounded to the nearest integral percent.

tion of mouse data is intended to demonstrate that the results generalize to a data set of different properties–a higher eukaryote, with sequence data derived from a different technology. We measure the correlation between the ChIP-seq peak score ranks of the sequences, and the AMA score ranks assigned using the discovered motif. To insure that this measurement is unbiased, we measure the correlation using held-out sequences, which are not used in discovering the motif. We compare *pairs* of motif discovery algorithms by sampling from all the sequence data (positive and negative), and applying a paired significance test (sign test) to the pairs of correlation scores.

On the mouse datasets, using the $\mathcal{DC}$ prior improves the accuracy of the motifs discovered by MEME (see Table 3), although the improvement is slight compared to that seen in the yeast ChIP-chip datasets. The OOPS model with the $\mathcal{DC}$ prior has significantly better accuracy than without the prior on 3 of the 13 datasets, and shows no significant difference in accuracy on the other 10, according to the sign test. Similarly, the $\mathcal{DC}$ prior improves the ZOOPS on 4 of the 13 datasets, but degrades the performance on 3 datasets. MEME using the $\mathcal{DC}$ prior finds better motifs for TFs c-Myc and n-Myc with both the OOPS and ZOOPS models. The motifs for three other TFs (Stat3, Zfx and E2f1) are improved using the $\mathcal{DC}$ prior with one or the other of the two sequence models. These results indicate that using

Table 3: Performance of motif discovery algorithms on mouse TF ChIP-seq datasets.

	OOPS- $\mathcal{DC}$ vs. OOPS				**ZOOPS- $\mathcal{DC}$ vs. ZOOPS**				**OOPS- $\mathcal{DC}$ vs. ZOOPS- $\mathcal{DC}$**				**OOPS- $\mathcal{DC}$ vs. PRIORITY- $\mathcal{DC}$**			
TF	**W**	**L**	**T**	**p-value**	**W**	**L**	**T**	**p-value**	**W**	**L**	**T**	**p-value**	**W**	**L**	**T**	**p-value**
Nanog			×	3.4e-01		×		1.3e-03	×			7.7e-03	×			2.1e-09
Oct4			×	1.0e-01			×	3.4e-01	×			5.8e-07	×			4.5e-14
Sox2			×	1.6e-01		×		1.6e-02			×	4.4e-01	×			1.3e-03
Smad1			×	1.0e-01			×	1.6e-01			×	2.4e-01			×	1.6e-01
E2f1			×	2.4e-01	×			4.5e-05		×		4.5e-05			×	4.4e-01
Tcfcp2l1			×	1.0e-01		×		7.7e-03		×		1.6e-02	×			1.9e-11
Ctcf			×	4.4e-01			×	2.4e-01			×	4.4e-01	×			8.9e-16
Zfx			×	1.0e-01	×			1.3e-03			×	1.6e-01	×			2.2e-10
Stat3	×			3.3e-03			×	4.4e-01			×	6.0e-02			×	1.6e-01
Klf4			×	1.6e-01			×	6.0e-02			×	1.0e-01			×	1.6e-01
Esrrb			×	6.0e-02			×	6.0e-02	×			3.3e-03		×		4.5e-14
c-Myc	×			3.3e-02	×			3.3e-03			×	2.4e-01		×		4.5e-05
n-Myc	×			1.5e-04	×			4.5e-05	×			1.6e-02		×		1.6e-08
Total	3	0	10		4	3	6		4	2	7		6	3	4	

The table compares the relative accuracy of pairs of motif discovery algorithms. Relative accuracy is measured by the correlation on held out sets of sequences of the sequence ranks based on ChIP-seq peak scores versus the ranks based on the motif-based AMA score. A check in the "win" or "W" ("loss" or "L") column indicates that the motifs found by the first (second) algorithm had significantly better Spearman rank correlation, as judged by the sign test on the 50 random repeats (p-value < 0.05). A check in the "tie" or "T" column indicates that there was no significant difference. The "Total" line shows the totals using the sign test to judge significance. OOPS, ZOOPS, OOPS- $\mathcal{DC}$ and ZOOPS- $\mathcal{DC}$ refer to MEME with those models and with or without the $\mathcal{DC}$ prior.

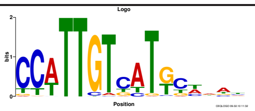
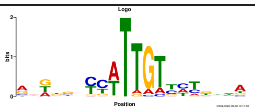
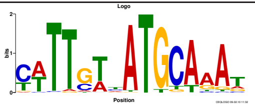
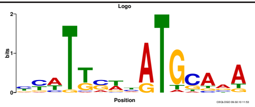
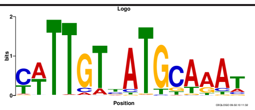
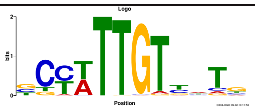
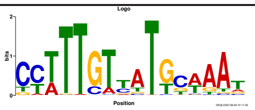
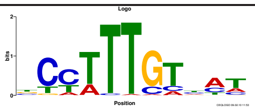
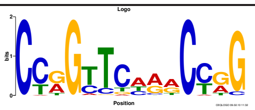
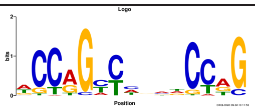
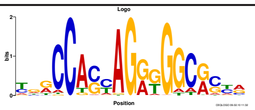
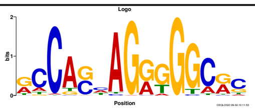
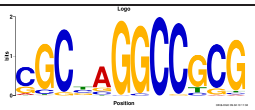
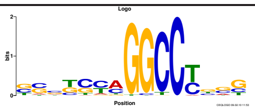
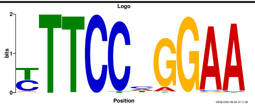
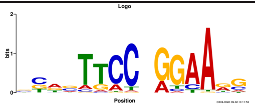
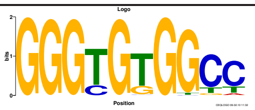
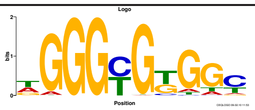
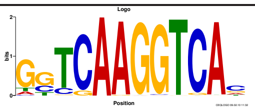
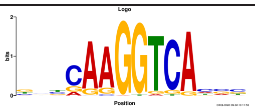
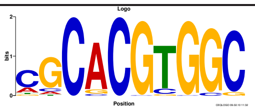
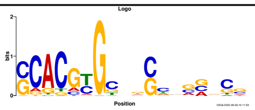
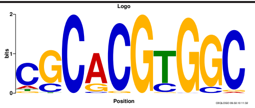
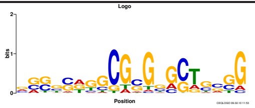
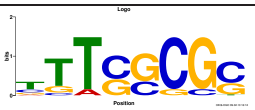
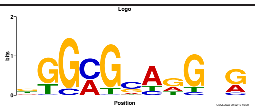
TF	Chen *et al.* motif	MEME motif	inter-motif distance
Nanog			0.11
Oct4			0.12
Sox2			0.054
Smad1			0.029
Tcfcp2l1			0.17
CTCF			0.043
Zfx			0.078
STAT3			0.12
Klf43			0.097
Esrrb			0.052
c-Myc			0.15
n-Myc			0.27
E2f1			0.30

Figure 1 Comparison of motifs found in mouse ChIP-seq datasets. The figure shows the motifs reported by Chen *et al.* [11] and those found by MEME in sequences identified as bound to the given transcription factor in 13 ChIP-seq experiments. The MEME motifs were found using 100 randomly chosen bound sequences and the OOPS-$\mathcal{DC}$ prior. The inter-motif distance (scaled Euclidean distance) is computed as described in Additional file 1.

the $\mathcal{DC}$ prior with MEME will likely improve the accuracy of TF motifs found in ChIP-seq data from higher eukaryotes.

As a further evaluation of our method with ChIP-seq data, we also directly compare the motifs found by MEME using the OOPS-$\mathcal{DC}$ prior with those reported by Chen *et al.* [11]. In Figure 1, we show the Chen *et al.* motifs along side the motif found by MEME using a random sample of 100 ChIP-seq peak sequences that achieved the highest value of our unbiased correlation-based quality measure. For all 12 of the 13 Chen *et al.* ChIP-seq experiments where they reported a motif, MEME using the OOPS-$\mathcal{DC}$ prior discovers a strongly similar motif. Those authors reported no motif for the E2f1 experiment, but the motif found by MEME resembles the TRANSFAC [24] E2f1 motif. We also show the scaled Euclidean distance (Eqn. 6) between each Chen *et al.* motif and the MEME motif in Figure 1. (Note that we do not require the aligned motif regions to have average information content of at least 1 bit in the inter-motif distance computation in Figure 1. Without this change, the inter-motif distances for Oct4, n-Myc and E2f1 would be reported as "1.0".) All 13 motifs discovered by MEME have distances less than or equal to 0.30 to the corresponding Chen *et al.* or TRANSFAC motif, and 11 out of 13 have distances below 0.17. We emphasize however, that this result only indicates that MEME is finding motifs similar to those found by those authors, and we believe that our correlation-based quality measure is more appropriate with this data.

To answer the question of whether using the $\mathcal{DC}$ prior with the OOPS or ZOOPS model is more appropriate with the mouse ChIP-seq data, Table 3 shows the relative accuracy of OOPS-$\mathcal{DC}$ versus ZOOPS-$\mathcal{DC}$. According to the sign test, OOPS-$\mathcal{DC}$ finds significantly more accurate motifs in more datasets (4 vs. 2) than ZOOPS-$\mathcal{DC}$. Although the sample size is small (13 datasets), it seems reasonable to conclude that the OOPS-$\mathcal{DC}$ approach will produce better motifs on average with ChIP-seq data.

A direct comparison of the accuracy of motifs found in the mouse datasets by OOPS-$\mathcal{DC}$ and PRIORITY-$\mathcal{DC}$ indicates that MEME with using the $\mathcal{DC}$ prior and the OOPS model has a slight edge. According to the sign test, OOPS-$\mathcal{DC}$ produces significantly better motifs for 6 of the 13 mouse ChIP-seq datasets, compared with PRIORITY-$\mathcal{DC}$. On 3 of the datasets, PRIORITY-$\mathcal{DC}$ produces more accurate motifs. This result is in agreement with our results using the yeast ChIP-chip datasets, where MEME using the $\mathcal{DC}$ prior and the ZOOPS model was (slightly) more successful than PRIORITY-$\mathcal{DC}$. As we expect, the ZOOPS model works better for ChIP-chip data, while the OOPS model works better for ChIP-seq for the examples we present here.

Conclusions

Position specific priors are an elegant and flexible way to utilize prior information from heterogeneous sources to improve the discovery of sequence motifs. In addition to allowing information from multiple sources to be incorporated into a Bayesian motif discovery framework, positional priors can even incorporate information from negative examples (so-called "discriminative" priors). Furthermore, using PSPs does not increase the running time of the underlying motif discovery algorithm. This flexibility has the potential to extend the range of applications and sensitivity of motif discovery algorithms that can utilize PSPs. Although we only study DNA datasets, our modifications to MEME are not DNA-specific. MEME is freely available for academic use and downloading at http://meme.nbcr.net.

PSPs had previously been shown to be of benefit when used with a Gibbs sampling motif discovery algorithm. We have shown here that they can also be of great benefit to MEME, which is based on EM and a heuristic search for starting points. We focused on using a prior that combines evolutionary information gleaned from orthologous sequences with positively and negatively labeled sequences in a discriminative prior (the "discriminative conservation", $\mathcal{DC}$ prior). Using this PSP on well-studied sequence datasets from 156 yeast TF ChIP-chip experiments improves the performance of MEME dramatically–more than doubling the number of datasets where MEME identifies the correct TF binding motif as its first prediction. Furthermore, using the $\mathcal{DC}$ prior allows MEME to achieve prediction accuracies that are superior to a large number of motif discovery algorithms, without increasing its running time.

We also confirm the benefit of PSPs to MEME when applied to TF motif discovery in ChIP-seq data from a higher eukaryotic species (mouse). To increase the independence of this second test, we used a novel way to measure the accuracy of the discovered motifs that obviates the need to rely on a set of known motifs (a "gold standard"). Although the observed benefits were somewhat small, they were substantial enough to indicate that constructing a $\mathcal{DC}$ prior and utilizing it with MEME is worthwhile even for higher eukaryotic ChIP-seq derived data.

In follow-up work, we plan to investigate PSPs designed specifically for ChIP-seq data. One approach might be to create a PSP that encodes the increased prior probability of the primary motif being located near the center of the ChIP-seq peak. We also plan to investigate PSPs designed for motif discovery in protein sequences. For protein motifs, PSPs based on spaced triples rather than the *k*-mers used here for DNA PSPs might be more appropriate, given the larger protein alphabet. We also intend to implement PSPs for use with MEME's ANR model, which allows multiple repeats of a motif within a single sequence. We don't foresee any major difficulties in incorporating PSPs into the ANR model but have focused on the OOPS and ZOOPS models in this work in order to facilitate direct comparison with previous work by others on PSPs.

Additional material

Additional file 1 Additional details on algorithm implementation and methods of evaluation.

Authors' contributions

TLB designed the experiments, adapted the MEME algorithm to use PSPs and wrote the final draft. MB worked on the mouse evaluation method. TW assisted with processing the ChIP-seq data. PM did most of the programming, performed the experiments and wrote the initial draft.

Acknowledgements

TLB and PM are funded by NIH/NCRR award R01 RR021692. TLB and MB are funded by the ARC Centre of Excellence in Bioinformatics. The authors thank Alexander Hartemink and Raluca Gôrdan for providing the scripts for computing the $\mathcal{DC}$ prior.

Author Details

Institute for Molecular Bioscience, The University of Queensland, Brisbane 4072, Queensland, Australia

Received: 10 February 2010 Accepted: 9 April 2010
Published: 9 April 2010

References

1. Siddharthan R, Siggia ED, van Nimwegen E: **PhyloGibbs: a gibbs sampling motif finder that incorporates phylogeny.** *PLoS Comput Biol* 2005, **1(7)**:e67.
2. Fang F, Blanchette M: **FootPrinter3: phylogenetic footprinting in partially alignable sequences.** *Nucleic Acids Res* 2006:W617-W620.
3. Redhead E, Bailey T: **Discriminative motif discovery in DNA and protein sequences using the DEME Algorithm.** *BMC Bioinformatics* 2007, **8**:385.
4. Barash Y, Bejerano G, Friedman N: **A simple hyper-geometric approach for discovering putative transcription factor binding sites.** *Algorithms in Bioinformatics: Proc. First International Workshop, no. 2149 in LNCS* 2001:278-293.
5. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ: **Profiling condition-specific, genome-wide regulation of mRNA stability in yeast.** *Proc Natl Acad Sci USA* 2005, **102(49)**:17675-17680.
6. Bauer DC, Bailey TL: **Studying the functional conservation of cis-regulatory modules and their transcriptional output.** *BMC Bioinformatics* 2008, **9**:220.
7. Narlikar L, Gôrdan R, Hartemink AJ: **Nucleosome Occupancy Information Improves *de novo* Motif Discovery.** *11th Annual International Conference on Computational Biology, RECOMB 2007, San Francisco* 2007:107-121.
8. Gordân R, Narlikar L, Hartemink AJ: **A Fast, Alignment-Free, Conservation-Based Method for Transcription Factor Binding Site Discovery.** In *12th Annual International Conference on Computational Biology, RECOMB 2008* Edited by: Vingron M, Wong L. Springer-Verlag; 2008:98-111.
9. Narlikar L, Gôrdan R, Hartemink AJ: **A nucleosome-guided map of transcription factor binding sites in yeast.** *Plos Computational Biology* 2007, **3(11)**:e215.
10. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, United Kingdom, July 16-19, 1995* 1995, **3**:21-29.
11. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133(6)**:1106-1117.
12. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15(7-8)**:563-577.
13. Bailey TL, Gribskov M: **Combining evidence using *p*-values: application to sequence homology searches.** *Bioinformatics* 1998, **14**:48-54.
14. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431(7004)**:99-104.
15. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12(6)**:996-1006.
16. Buske FA, Bodén M, Bauer DC, Bailey TL: **Assigning roles to DNA regulatory motifs using comparative genomics.** *Bioinformatics* 2010, **26**:860-866.
17. Stormo GD: **Information content and free energy in DNA-protein interactions.** *J Theor Biol* 1998, **195**:135-137.
18. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation.** *Nucleic Acids Res* 2004, **32(4)**:1372-1381.

19. McLeay RC, Bailey TL: **Motif Enrichment Analysis: A unified framework and method evaluation.** *BMC Bioinformatics* 2010 in press.
20. Kruskal WH, Wallis WA: **Use of Ranks in One-Criterion Variance Analysis.** *J American Statistical Association* 1952, **47(260):**583-621.
21. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME Suite: tools for motif discovery and searching.** *Nucleic Acids Res* 2009:W202-W208.
22. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for Saccharomyces cerevisiae.** *BMC Bioinformatics* 2006, **7:**113.
23. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, **5:**170.
24. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006:D108-D110.

doi: 10.1186/1471-2105-11-179
Cite this article as: Bailey *et al.*, The value of position-specific priors in motif discovery using MEME *BMC Bioinformatics* 2010, **11**:179