fdrMotif:
Identifying cis-elements by an EM Algorithm Coupled with False Discovery Rate Control

Leping Li¹,* , Robert L Bass², and Yu Liang¹,*
¹Biostatistics Branch, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA
²Computational Biology Facility, National Institute of Environmental Health Sciences, NIH, DHHS, Research Triangle Park, North Carolina 27709, USA

Abstract

Motivation—Most de novo motif identification methods optimize the motif model first and then separately test the statistical significance of the motif score. In the first stage, a motif abundance parameter needs to be specified or modeled. In the second stage, a z-score or p-value is used as the test statistic. Error rates under multiple comparisons are not fully considered.

Methodology—We propose a simple but novel approach, fdrMotif, that selects as many binding sites as possible while controlling a user-specified false discovery rate (FDR). Unlike existing iterative methods, fdrMotif combines model optimization (e.g., position weight matrix (PWM)) and significance testing at each step. By monitoring the proportion of binding sites selected in many sets of background sequences, fdrMotif controls the FDR in the original data. The model is then updated using an expectation (E) and maximization (M)-like procedure. We propose a new normalization procedure in the E-step for updating the model. This process is repeated until either the model converges or the number of iterations exceeds a maximum.

Results—Simulation studies suggest that our normalization procedure assigns larger weights to the binding sites than do two other commonly used normalization procedures. Furthermore, fdrMotif requires only a user-specified FDR and an initial PWM. When tested on 542 high confidence experimental p53 binding loci, fdrMotif identified 569 p53 binding sites in 505 (93.2%) sequences. In comparison, MEME identified more binding sites but in fewer ChIP sequences than fdrMotif. When tested on 500 sets of simulated “ChIP” sequences with embedded known p53 binding sites, fdrMotif, compared to MEME, has higher sensitivity with similar positive predictive value. Furthermore, fdrMotif is robust to noise: it selected nearly identical binding sites in data adulterated with 50% added background sequences and the unadulterated data. We suggest that fdrMotif represents an improvement over MEME.

1 INTRODUCTION

Different sites bound by the same transcription factor are often not exact matches. The position weight matrix (PWM) is a widely used motif model that represents the variability of the binding sites. A number of de novo motif discovery programs such as MEME (Bailey and Elkan, 1994), AlignACE (Roth et al., 1998), Consensus (Hertz and Stormo, 1999), BioProspector (Liu et al., 2001), motifSampler (Thijs et al., 2001), MDScan (Liu et al., 2002) and DWE (Smith et al., 2005) are based on a PWM formulation. These programs identify over-represented motifs in a set of sequences, e.g., loci from immunoprecipitation (ChIP) coupled

*To whom correspondence should be addressed.
with microarray (ChIP-on-chip) experiments or promoter sequences of genes that either share a common function or have a similar pattern of expression across different experimental conditions.

A motif may be present in only a subset of the sequences, each of which can have one or multiple copies of the motif. Since the early 1990s, many authors of the de novo methods (i.e., Bailey and Elkan, 1994; Liu et al., 1995; Thijs et al., 2001) have realized that the simple “one motif occurrence per sequence” model is inadequate. Various approaches have been developed to allow flexibility in their models to allow “zero or multiple occurrences per sequence”.

Most of these methods use a two-stage procedure: 1) optimizing the PWM (itself an iterative process), and 2) selecting binding sites based on a statistical test of significance for each subsequence’s motif score. These two stages are carried out separately.

In the first stage, a motif abundance parameter needs to be specified or modeled. For instance, MEME introduced a global parameter, the expected maximal number of appearances (MAXP) of the motif over the entire set of sequences. In the expectation (E) step, the probability that a binding site starts at each position is calculated based on Bayes formula. For the “zero or multiple occurrences per sequence” model, MEME normalizes the sum of probabilities over all positions in all sequences to MAXP, with a constraint that the sum of the probabilities within a motif can not exceed 1. In a Gibbs Motif Sampler, Liu et al. (2001) introduced two parameters in their threshold sampler in which high and low thresholds were used in the predictive update stage. Non-overlapping segments with scores above the high threshold are automatically selected whereas those with scores between the two thresholds are selected probabilistically for updating the PWM. Later, the group introduced a motif abundance parameter with a Beta prior distribution in the Bernoulli sampler model (Jensen et al., 2004). Thijs et al. (2002) modeled the number of copies of the motif in a sequence as a random variable. This variable is iteratively sampled under its posterior distribution. The choices for these parameters might be set by “trial-and-error”.

In the second stage, most de novo methods decide if a segment is a motif or not. A z-score or p-value is often used as the test statistic as in BioProspector and MEME, respectively. For example, MEME reports motifs with a p-value less than $10^{-5}$ by default. In BioProspector, a segment is declared to be a motif when its score is five standard deviations above the null motif score distribution mean.

Potential drawbacks of the two-stage approaches include 1) a motif abundance parameter needs to be specified or modeled; 2) the cut point for p-value or z-score can be arbitrary and errors for multiple comparisons may not be adequately controlled.

Herein we propose a hybrid method, fdrMotif, which does not require a global abundance parameter and post-analysis significance testing. Like MEME, our method utilizes an EM-like algorithm with two iterative steps, the expectation (E) step and the maximization (M) step in model optimization. We introduce a new normalization procedure in the E-step. Our approach couples model optimization and significance testing during each iteration. fdrMotif selects as many binding sites as possible while controlling the false discovery rate (FDR), defined as the expected proportion of non-motif subsequences that are falsely declared as binding sites. FDR for error control was first proposed by Benjamini and Hochberg (1995) to prevent an overabundance of false positive results in multiple hypothesis testing. They developed a sequential procedure to control FDR for independent multiple testing and discussed its advantages over the family-wise error rate (FWER). Subsequently, a number of methods have been proposed for estimating the FDR (Benjamini and Yekutieli, 2001; Storey and Tibshirani, 2001; Storey 2002). These methods have started to gain wide attention in genome research (Zaykin et al., 2000; Tsai et al., 2003).
2 METHODS

2.1 Overview of the methodology

Our method is iterative and alternates between updating the PWM and significance testing. It starts with an initial PWM and a set of sequences (e.g., from ChIP experiments). We generate many sets of background (null) sequences under the input sequence probability model. At each model estimation step, we determine the number of binding sites in each sequence by performing statistical tests. The FDR in the original dataset is controlled by monitoring the proportion of background subsequences that are declared as binding sites. The PWM is updated using an EM-like algorithm with two iterative steps (the E and M steps) until convergence. In the E-step, our method normalizes the sum of the probabilities over all positions in a sequence to the number of binding sites found in the sequence. Details are given below and a schematic diagram is in supplementary material (Scheme 1).

2.2 False discovery rate (FDR)

Following the strategy of MEME, for a dataset of DNA sequences, $M$, we break it up conceptually into all $m$ overlapping subsequences of length $w$, where $w$ is the length of motif and is known. This new dataset is referred to as $x = (x_1, x_2, \ldots, x_m)$. For each subsequence, the null hypothesis is that the subsequence is not a binding site and the alternative hypothesis is that the subsequence is a binding site. Given a PWM, $\Theta = \{\theta_{ij}, l=1,\ldots, w \text{ and } j = A,C,G,T\}$; we compute the log likelihood score for each subsequence based on a multinomial probability model. We declare a subsequence significant (i.e., to be a binding site) if its score equals or exceeds a threshold, $\tau$. Following conventional notation (Benjamini and Hochberg, 1995), the possible outcomes of the $m$ tests are defined in Table 1.

The FDR is defined as

$$f_\theta(\tau) = \frac{V_{M,\theta}(\tau)}{R_{M,\theta}(\tau)} \quad (2.1)$$

It is easy to see that when $\tau$ is equal to or less than the smallest score among all subsequences, $f_\theta(\tau) = m_0 / m$; whereas $f_\theta(\tau) = 0$ when $\tau$ is larger than the largest score among all subsequences as no subsequences are selected, meaning that $R_{M,\theta}(\tau) = 0$, therefore, $V_{M,\theta}(\tau)$ must be 0 and by definition $f_\theta(\tau) = 0$. For a given $\Theta$, we would like to find $\tau^*$, such that $f_\theta(\tau^*) = \gamma_0$ for a prespecified $\gamma_0$, say $\gamma_0 = 0.05$. The solution to $f_\theta(\tau^*) = \gamma_0$ exists only when $\gamma_0 \leq m_0 / m$, assuming that $m_0 / m$ is monotonic. This assumption might not be strictly correct, especially when $m$ is small. A small change in $m_0$, $m$ or both could slightly alter the monotonic trend. To account for this possibility, fdrMotif considers a specified FDR bound is met when the computed FDR bound gives the largest $R_{M,\theta}(\tau)$ and is within 5% of the specified bound. Assuming that $f_\theta(\tau)$ is a continuous function oft, we formally define $\tau^*$, as

$$\tau^* = \inf_{\tau} \{\tau; f_\theta(\tau) \leq \gamma_0\} \quad (2.2)$$

This definition makes $\tau^*$ unique and selects as many motif subsequences as possible that satisfy the FDR constraint for a given $\Theta$ (see below).

If both $R_{M,\theta}(\tau)$ and $V_{M,\theta}(\tau)$ were observable, one can simply determine $\tau^*$ that satisfies:
However, in real applications, $R_M, \theta(\tau)$ is observable, but $V_M, \theta(\tau)$ is not. We replace the unknown random variable, $V_M, \theta(\tau)$, with a suitable estimator $\hat{V}_M, \theta(\tau)$. A similar approach was used in estimating the false selection rate (FSR) in variable selection (Wu et al., 2007). The number of uninformative variables among all selected variables is not observable; therefore, it was replaced by a suitable estimator. Let $B$ denote a set of background subsequences and $m_B$ be the number of subsequences. We assume that the subsequences in $B$ and the non-motif subsequences in $M$ have the same probability of being declared (falsely) as binding sites. Similar to set $M$, given $\Theta$, one can count the number of subsequences in set $B$ that score at or above, denoted as $R_B, \theta(\tau)$. $R_B, \theta(\tau) / m_B$ is the proportion of binding sites found in set $B$ and

\[
E \{V_M, \theta(\tau)\} = E \left\{ \frac{R_B, \theta(\tau)}{m_B} \cdot m_0 \right\} \text{ under our assumptions. Given } N \text{ sets of } B, E\{R_B, \theta(\tau)\} \text{ may be estimated by } \bar{R}_B, \theta(\tau) = N \sum_{b=1}^{N} R_{b, \theta(\tau)}. \text{ A schematic diagram is in supplementary material (Scheme 2). We propose } \hat{V}_M, \theta(\tau) = \frac{m_0}{m_B} \bar{R}_B, \theta(\tau). \text{ Consequently, one can determine } \tilde{\tau}^* \]

Because $m_0$ cannot be observed, we seek a bound that is free of $m_0$. Let $\pi_0$ denote the proportion of the non-motif subsequences in $M$, $\pi_0 = m_0 / m$ and $m_0 < m$. Since only a small fraction of the subsequences in $M$ are binding sites, $\pi_0 \approx 1$ and the following relationship holds,

\[
\frac{\bar{R}_B, \theta(\tau)}{m_B} \cdot \frac{m_0}{m} = \pi_0 \cdot \frac{m}{m_B} \cdot \frac{\hat{R}_B, \theta(\tau)}{\hat{R}_M, \theta(\tau)} \leq \frac{m}{m_B} \cdot \frac{\bar{R}_B, \theta(\tau)}{\bar{R}_M, \theta(\tau)}
\]

Now, all variables in the right side of the equation are estimable. One may control the upper bound of the FDR to $\gamma_0$, instead of the exact FDR, although the upper bound is as close to the desired FDR as $\pi_0$ is close to 1. Finally, $\tilde{\tau}^*$ can be defined as

\[
\tilde{\tau}^* = \inf \left\{ \tau; \frac{m}{m_0} \cdot \frac{\bar{R}_B, \theta(\tau)}{\bar{R}_M, \theta(\tau)} \leq \gamma_0 \right\}
\]

### 2.3 Estimation of $\tau^{**}$

In Section 2.2, we assumed that the subsequences in $B$ and the non-motif subsequences in $M$ have similar structure. To honor this assumption, we generated the set $B$ from a $4^{th}$-order Markov model estimated from the sequences in set $M$. When the Markov background model is estimated from the input data, a $4^{th}$-order should capture the local dependence in the data, but not high enough to capture the structure of the embedded motifs. The subsequences in $B$ are referred to as background sequences. Base composition and local dependency are preserved.
by the Markov model. To obtain \( \hat{R}_{\Theta}(\tau) \) we generated \( N \) (e.g., \( N = 10 \)) replicates of \( B \), each of which contains \( m_B \) (here \( m_B = m \)) subsequences of length \( w \).

In most applications, the number of occurrences of a motif in a sequence is small. Instead of examining all \( m + m_B \) subsequences in sets \( M \) and \( B \) together, we examine only some of them to determine \( \hat{\tau}^* \). We choose \( C_{\text{max}} \) subsequences from each sequence, \( (K_M + K_B)C_{\text{max}} \) in total, to examine, where \( K_i \) is the number of sequences in set \( i, i = B, M \), and \( C_{\text{max}} \) is a pre-set number of highest scoring non-overlapping subsequences from all \( 2(L - w + 1) \) subsequences (both plus and the complementary strands) in each sequence of length \( L \) (for notational convenience, we assume all sequences have the same length). Note that we still need to compute the scores for all \( m \) and \( m_B \) subsequences. We set \( C_{\text{max}} = 10 \) solely for computational efficiency and this parameter can be changed by a user. By sorting the scores of the \( (K_M + K_B)C_{\text{max}} \) chosen subsequences in \( M \cup B \) in descending order, we examine the scores of subsequences upward and stop when the equation 2.5 is satisfied. The subsequences with scores \( \hat{\tau}^* \) or larger are declared binding sites. This rule maximizes the number of subsequences that are declared binding sites in \( M \) while controlling the upper bound of the FDR.

### 2.4 Update PWM

The above procedure describes how to determine the PWM score threshold, \( \hat{\tau}^* \), to satisfy the FDR constraint for a given \( \Theta \). In this section, we discuss how \( \Theta \) is updated iteratively using an EM-like algorithm. Our updating procedure is similar to that of MEME; however, important differences exist. In MEME, subsequence scores are normalized so that their sum is equal to a global parameter, MAXP. The global parameter is fixed in advance and does not change during the iterations. Unlike MEME, our method identifies \( R_{\Theta}(\hat{\tau}^*) \) significant sites at each iteration, while satisfying the FDR constraint. That is, we estimate the number of binding sites in each sequence at each iteration and use that information when updating the PWM.

Let \( r_{ij}(\hat{\tau}^*) \) be the number of binding sites in sequence \( i, i = 1, 2, \cdots, K_M \), where \( K_M \) is the number of sequences in set \( M \) and \( \rho_{ij} \) be the probability that a binding site starts at position \( j \) in sequence \( i \). We compute \( \rho_{ij} \) using Bayes formula as in MEME (supplementary S1). Our procedure has the following characteristics:

\[
\hat{\tau}^* = \arg \max \sum_{i=1}^{2(L-w+1)} \sum_{j=1}^{\min \{L,w\}} \rho_{ij} \quad \text{and} \quad \sum_{i=1}^{K_M} r_{ij}(\hat{\tau}^*) = R_{\Theta}(\hat{\tau}^*),
\]

where \( r_{ij}(\hat{\tau}^*) \leq C_{\text{max}} \) for each \( i \). The above equations are subject to two constraints: 1) \( 0 \leq \tau_i \leq 1 \), for \( 1 \leq i \leq K_M \), and \( 1 \leq j \leq 2(L - w + 1) \); 2) the total sum of \( \rho_{ij} \) in any window of length \( w \) for the plus strand and the corresponding window of length \( w \) for the reverse complementary strand is less than or equal to 1. Equations 2.6 state that for each sequence \( i \), the \( 2(L - w + 1) \) probabilities are normalized so that the sum over all positions in a sequence is equal to \( r_{ij}(\hat{\tau}^*) \) and the sum over all positions in all sequences is equal to the total number of binding sites found in the data, \( R_{\Theta}(\hat{\tau}^*) \). Note that in this step, all subsequences in \( M \), not just the \( K_M C_{\text{max}} \) ones, are used in normalization and in updating the PWM.

Finally, \( \Theta \) is updated according to the following formula,

\[
\hat{\theta}_{1x} = \frac{c_{l,x} + \delta}{R_{\Theta}(\hat{\tau}^*) + 4\delta},
\]
where \( c_{l,x} \) is the weighted count of nucleotide \( x \) at position \( l,x = A,C,G,T; l = 1,2,\ldots; w \); \( \delta \) is a small pseudo-count (0.0001) to avoid a zero denominator. Here, \( c_{l,x} \) is computed as follows,

\[
\begin{align*}
  c_{l,x} &= \sum_{i=1}^{K_B} \sum_{j=1}^{L-2w+1} \rho_{l,j} I \{ s_{i,j+l-1}=x \} \\
  &\text{where } I \{ s_{i,j+l-1}=x \} = 1 \text{ when the } (j+l-1)\text{th nucleotide in sequence } i \text{ matches nucleotide } x, \text{ otherwise } 0.
\end{align*}
\]

It can be seen that although this sums over all the subsequences, the ones that are not likely to be binding sites are appropriately down-weighted by \( \rho_{l,j} \).

To briefly summarize, our method consists of iterating the following two steps until \( \Theta \) converges or the number of iterations exceeds a maximum:

1. Given a PWM, \( \Theta \), we compute scores for the subsequences in the original set of sequences \( M \), and the subsequences in the \( N \) background sets, \( B \). We select the \( C_{\max} \) highest scoring non-overlapping subsequences from each sequence to determine the smallest \( \tau^* \) that satisfies equation 2.5.

2. Given the number of binding sites found in each sequence by step 1, normalize the sum of the probabilities over all positions in the sequence to this number to obtain \( \rho_{l,j} \). Then update \( \Theta \) using \( \rho_{l,j} \) as in equation 2.7.

Our convergence criterion is \( \max_l \left( \max_x \left| \hat{d}_{l,x}^{(t+1)} - \hat{d}_{l,x}^{(t)} \right| \right) < d \), e.g., \( d = 0.001 \).

3. DATA

3.1 ChIP data (set \( M \))

Lacking benchmark sequence data, we used experimental data for a well-studied transcription factor binding site. The data are the 542 high-confidence p53 binding loci identified by ChIP-PET experiments (Wei et al., 2006) with average sequence length 1187 bp. Most of the identified loci are likely true p53 binding sites (Wei et al., 2006). To examine our method’s sensitivity to adulteration, we created 6 additional datasets by systematically adding 5%, 10%, 20%, 30%, 40%, and 50% randomly generated sequences to the original ChIP data. We refer to the original and these six adulterated datasets as datasets 1 to 7, respectively. We used a 4th-order Markov model estimated from the ChIP data to generate random sequences with the same average length as the ChIP sequences. Importantly, the same random sequences were also included in all datasets having more random sequences. This feature eliminates the complete independence of the adulteration in datasets 2-7 because, for instance, the same 5% random sequences in dataset 2 are included in datasets 3-7. We assume that no p53 binding sites are present in the simulated sequences and regard them as “noise”.

We also tested fdrMotif on several additional ChIP datasets including the estrogen receptor \( \alpha \) binding sites (Lin et al., 2007) and the CTCF-binding sites (Kim et al., 2007).

3.2 Background sequences (set \( B \))

Our method requires many sets of background sequences. Ideally, the background sequences should contain few or no binding sites. The background sequences may be simulated or real sequences. If simulated, we generated \( N = 10 \) sets of random background sequences using the same 4th-order Markov model that we generated the random sequences described above. Each set contains the same number of sequences of the same length as the original (ChIP) data (e.g., \( K_B = 542 \) for the p53 set).
3.3 Initial PWMs

Our method requires an initial PWM, which can be constructed from known binding sites or a consensus sequence or taken from databases such as TRANSFAC (Knuppel et al., 1994) and JASPER (Sandelin et al., 2004). The initial PWM for p53 was generated according to Li et al. (2007) with motif length (number of columns) 20. The initial PWMs for CTCF and EREα were adapted from Kim et al. (2007) and Lin et al. (2007), respectively.

3.4 Simulated “ChIP” sequence data

The number of binding sites and the locations of the binding sites in ChIP data are not known. For performance comparison, sequences with known binding site information are needed. Over the years, 138 p53 binding sites have been experimentally identified (Horvath et al., 2007). These known p53 binding sites constitute two half sites separated by 0-13 nucleotides. Among them, 66 contain no nucleotides between the two half sites. Since fdrMotif and the competing methods such as MEME do not allow gap between the two half sites, we only consider the 66 sites in this simulation study. We simulated 66 sequences, each (length of 250bp) which was generated from a 4th-order Markov model that was estimated from the 543 p53 ChIP-PET sequences, as previously described. For each sequence, a location within the sequence was randomly chosen and the 20-bp segment to the 3′ end of the site was replaced by a randomly selected binding site from the 66 known sites without replacement (no binding sites were used twice). Each sequence contains exactly one of the 66 known p53 binding sites. Each 66 sequences form a set. We independently generated 500 such sets.

To test the sensitivity to noise in the data, we also generated another 500 independent sets of sequences, each of which contains 66 sequences with one binding site each and 16 sequences (20% added noise) without any embedded known binding sites.

4. RESULTS

4.1 ChIP Data

4.1.1 P53—For the original p53 ChIP dataset and each of the six adulterated datasets, we ran fdrMotif analysis by controlling the FDR upper bound at 10%. For each run, fdrMotif converged within approximately 20 iterations. The results are summarized in supplementary Table s1. The total numbers of binding sites found in the seven datasets are comparable. As “noise” in the data increased, fdrMotif selected fewer binding sites in the ChIP sequences and more in the background sequences. Overall, fdrMotif identified at least one binding site in 91.1-93.2% of ChIP sequences regardless of the noise level in the data. A majority (83.2-84.5%) of these sequences contained one predicted p53 binding site, 6.8-10.0% contained multiple predicted p53 binding sites, and 6.8-8.9% contained no binding sites.

Although fdrMotif found slightly fewer binding sites as the level of adulteration increased, generally speaking, nearly the same sites were identified in all seven datasets. For example, the 536 binding sites selected in the ChIP sequences in dataset 7 (50% adulteration) are all included among the 569 binding sites selected in dataset 1 (no adulteration). These results suggest that our method is robust to “noise” in the data. The logo plots (Crooks et al., 2004) of the binding sites selected in ChIP sequences for all seven datasets are in supplementary Figures s1-2 and are consistent to the canonical p53 motif.

As a comparison, we carried out MEME analyses on the same seven datasets using their anr model that, like ours, allows zero or multiple occurrences of a motif per sequence. We also provided a consensus sequence of “gggcatgcccgggcatgccc” as the starting point for MEME. The same 4th-order Markov model that we used to generate the background sequences was provided to MEME as the background model. Without knowing the expected number of
binding sites (MAXP) in each dataset, we carried out four MEME analyses by setting MAXP to 570, 600, 750, and 1000, respectively. The results are in supplementary Table s1. In general, MEME consistently identified more binding sites than fdrMotif. In particular, MEME identified multiple p53 binding sites in more sequences than fdrMotif. However, MEME identified fewer sequences (ranging from 63.8% to 79.7% depending on MAXP setting) containing exactly one binding site than fdrMotif (83.2-84.5%). For comparable number of total binding sites identified, fdrMotif identified p53 binding sites in 91.1-93.2% sequences compared to 89.3-91.7% sequences by MEME.

Taken together, the results for the p53 ChIP data suggest that, compared with MEME, our method identified fewer binding sites in total, but more sequences containing at least one binding sites and fewer containing zero, two or more binding sites, suggesting that method may have slightly higher sensitivity.

4.1.2 EREα—When FDR bound was controlled at 10%, fdrMotif identified 505 full EREα sites on 1234 ChIP loci (Lin et al., 2007). The average number of binding sites selected in the 10 sets of 1234 background sequences was 50 (4%). This result suggests that although the upper bound of the FDR was controlled at 10%, the estimated false positive rate at the sequence level was much lower. When the FDR upper bound was controlled at 15%, fdrMotif identified 610 full EREs with an estimated false positive rate of 7.2%. The logo plots for the binding sites are in supplementary Figure s3.

4.1.3 CTCF—We also applied fdrMotif to a subset of human loci from ChIP-chip experiment (Kim et al., 2007). Among all 13,804 ChIP identified loci, only those that are less than 400bp are used in this analysis. This resulted in a total of 1156 sequences with an average length of 351bp. fdrMotif identified 687 CTCF binding sites when FDR upper bound was controlled at 10%. The average number of binding sites selected from the 10 sets of background sequences was 68 (5.9%). The logo plots (supplementary Figure s4) from those binding sites closely resemble that in Kim et al. (2007).

4.2 Simulated “ChIP” data

We generated 500 sets of “ChIP” sequences, each of which contains 66 simulated sequences with 66 embedded known p53 binding sites, one per sequence. For each dataset, we carried out two fdrMotif and two MEME analyses, respectively. For fdrMotif, the upper bounds of FDR were controlled at 5% and 6%, respectively, whereas for MEME, MAXP was set to 74 and 100, both of which are larger than the number of known binding sites in the data. The same 4th-order Markov model that was used to generate the “ChIP” sequences was provided to MEME as the background model. These settings should be optimal for MEME for these datasets. The two FDR upper bounds were chosen so that the number of false positives from fdrMotif and MEME was comparable.

For each dataset, we monitored the numbers of true positives (TP), false negatives (FN), and false positives (FP) as defined in Tompa et al. (2005). The sensitivity (Sn), positive predictive value (PPV), the average site performance (ASP), and the performance coefficient (PC) all at the site level were computed as described in Tompa et al. (2005). We did not consider statistics at the nucleotide level as the majority of the subsequences are non-binding sites. For all 500 datasets, the summary results are listed in Table 2. Both the averages and standard errors of Sn, PPV, ASP, and PC are comparable for the two settings within each method. However, fdrMotif, compared to MEME, is more sensitive but has a similar positive predictive value. Similar results (Supplementary Table s2) were obtained for the other 500 datasets that contain 20% sequences without any embedded known sites.
4.3 Effect of starting PWMs

To study the effect of the starting PWMs on results, we tested eleven different starting p53 PWMs with various degree of degeneracy. The eleven starting PWMs converged to three maxima, among which two were similar and corresponded to p53 motif whereas the other was distinct and corresponded to a poly (A) consensus (supplementary Table s3). Between the two similar maxima, a majority of the binding sites overlapped, that is, 561 of the 570 (98.4%) binding sites found in one maximum were included in the 620 binding sites found in the other maximum. Interestingly, PWMs 1-7 all converged to a nearly identical maximum in terms of the number of binding sites and their locations despite of the large variations of degeneracy among them. On the other hand, PWM1 and PWMs 8-10 differed only in one position (position 3) in which the corresponding base is a ‘g’ in PWM1 and a non-‘g’ in PWMs 8-10 converged to different maxima. A logo plot of the additional 59 binding sites between the two maxima showed departure departure from the canonical p53 motif (not shown), suggesting that some of the additional sites might be noise. Among all eleven starting PWMs, PWM11 is the most degenerate one. Not surprisingly, it converged to a non-p53 motif (poly(A)).

4.4 Simulation study to compare normalization procedures

In EM-like algorithms, the probability that a binding site starts at each position is calculated based on Bayes formula using all motif scores. For the “one occurrence per sequence” model, one normalizes the sum of the probabilities over all positions in a sequence to 1 (referred to as seq-by-seq-1). For the “zero or multiple occurrences per sequence” model, one normalizes the sum over all positions in all sequences to MAXP (referred to as global). Alternatively, one may normalize the sum over all positions in a sequence to the number of binding sites in the sequence (referred to as seq-by-seq-anr, the procedure in fdrMotif). The formulas for the three normalizations are given in the supplementary text.

To compare three normalization procedures, we carried out two simulations; each employed simulated subsequence “scores” in 20 sequences of length 100. In simulation 1, sequences 1-10 each contained a single 5-bp binding site \( w = 5 \), sequences 11-15 each contained two 5-bp binding sites, and sequences 16-20 contained no binding sites. In simulation 2, each sequence contained exactly one binding site. The locations of all binding sites were randomly selected. The scores for the binding sites (motifs) and non-motif sites were sampled from uniform distributions \( U(0.5,1) \) and \( U(0,0.1) \), respectively. In both simulations, there were 20 binding sites in 20 sequences.

Since the number and locations of the binding sites in each sequence are known, the best normalization procedure should yield the largest overall posterior probabilities of the binding sites. We tested three different normalization procedures: global, seq-by-seq-1, seq-by-seq-anr. For all procedures, we enforced the constraints that the sum of probabilities within a sequence does not exceed 1 and that the sum of probabilities over all positions in all sequences is 20 after normalization. We summed the probabilities of the 20 known binding sites in the data. To calculate the mean and standard deviation of the sum using each procedure, we repeated each simulation 1000 times. With exactly one binding site per sequence (simulation 2), the three procedures gave almost identical results (supplementary Table s4). However, when sequences contain zero or multiple binding sites (simulation 1), seq-by-anr gave the largest sum. As expected, the seq-by-seq-1 normalization did not do well in simulation 1.

4.5 Effect of \( R_{ML} (\hat{\tau}^{**}) \) estimation on motif selection

To study the effect of \( R_{ML} (\hat{\tau}^{**}) \) estimation on results in real data, we repeated fdrMotif analyses on the p53 dataset with FDR=10% by setting \( R_{ML} (\hat{\tau}^{**}) \) to 100 in the first iteration using the same random seed. With \( R_{ML} (\hat{\tau}^{**}) \) arbitrarily fixed at 100 for the first iteration, fdrMotif

Bioinformatics. Author manuscript; available in PMC 2008 May 12.
selected 455 binding sites at the first iteration and 569 binding sites after convergence. Unrestricted, fdrMotif selected 565 binding sites at the first iteration and 571 binding sites after convergence. Only one of the 569 sites was not in the 571 sites selected without any restriction. Additional results are in supplementary Tables s5 and s6. These results suggest that misspecification of $R_{\mu,\psi}(T^{\psi})$ at the beginning may have little effect on the number and locations of binding sites identified by our iterative EM-like method. Since fdrMotif typically converges in a few iterations, persistent constraints will affect the results.

5 DISCUSSION

A sequence can have zero or multiple occurrences of a motif. MEME breaks the sequences into all (overlapping) subsequences of length $w$, where $w$ is the motif length, and searches these subsequences. This approach avoids the need to specify the number of binding sites in each sequence. Following this strategy, our method starts with a set of sequences to be searched (e.g., ChIP data) and an initial PWM. We also generate many sets (e.g., 10) of independently and identically distributed background sequences from a 4th-order Markov model that is estimated from the original sequences. The subsequences are then scored using the PWM based on a multinomial probability model. A score threshold is automatically chosen so that fdrMotif selects as binding sites as many subsequences in the original dataset as possible under the FDR constraint. The PWM is subsequently updated using an EM-like algorithm. This iterative procedure is repeated until the PWM converges or the maximal number of iterations has been reached. We would like to point out that the length of the PWM is fixed during these iterative procedures.

In MEME, the global normalization scales each site by the sum of all scores then multiplies by MAXP. The advantage of this approach is that knowledge of how the binding sites are distributed among all sequences is not needed. If the number of binding sites in each sequence were known at each PWM optimization step, it would be desirable to normalize the probabilities, sequence by sequence, so that the sum of probabilities over all positions in the sequence is equal to the number of sites in the sequence. However, this information is generally not known. In contrast, our method identifies the number of binding sites in each sequence at each estimation step, based on the current PWM at that step. Therefore, our normalization is achieved sequence by sequence under the same constraints as in MEME. Our normalization scheme sets the probability of every position in sequences without binding sites to zero, minimizing the contribution of non-motif sites to the PWM.

Our simulation results show that when the number of binding sites in each sequence is known, our seq-by-seq-anr normalization procedure results in the largest posterior probabilities for those binding sites. In real applications, however, the number of binding sites in each sequence is not known. At each iteration, our method selects as many binding sites as possible while satisfying the FDR constraint. In the early stage of model optimization, some true binding sites may be missed while non-motif sites may be selected as binding sites. One might expect that our normalization procedure may be sensitive to these “errors”. Surprisingly, for the p53 data, we found that misspecification of the number of binding sites in the first step had little on the final result. One likely explanation is that our method uses an EM-like algorithm that iteratively optimizes the PWM. Each optimization step is coupled with a FDR constraint on binding-site selection. However, large and persistent misspecifications will likely affect the results. This could happen when the starting PWM represents consensus (e.g., one of the four cells has 1 for all columns) but with misspecifications. We found that arbitrarily setting the number of binding sites selected to 1.5 times the number of sequences only in the E-step in the first iteration works well for all cases tested.
Our simulated “ChIP” data should favor methods that use log likelihood ratio (LLR) as the objective function such as MEME as the data were generated from the same model that was also used as the background model. For those data, fdrMotif, compared to MEME, had higher sensitivity with comparable positive predictive value.

Our method uses log likelihood as the scoring function. Significance testing is carried out by comparing the scores of subsequences in the ChIP dataset to those in the background dataset. For methods using LLR as the scoring function, a background model is needed. Modeling the background distribution may be challenging. The ChIP loci are distributed across the genome. This heterogeneity makes it impossible to accurately estimate the background distribution. Background models of different orders estimated from the same dataset or background models of the same order estimated from different regions of the genome can give different results. Alternatively, one might use background sequences either simulated or real as the null. Although this approach would not eliminate the arbitrariness of background, the effect of the background on the results can be minimized. As long as the background sequences are null, it is not necessary to assume that non-motif subsequences in ChIP data and the background sequences have the same distribution. Background sequences may be obtained from randomly selected coding sequences or simulated from input data. Advantages of using the background sequences generated by Markov models are: 1) the simulated background sequences are less likely to contain true motifs; 2) nucleotide composition and local dependence structure in the original data are retained. However, the disadvantages include: 1) it is computationally more costly; 2) the results may vary slightly when a different random seed is used. One may want to repeat the analysis several times.

Our method uses an EM algorithm to estimate the PWM. For problems with multiple optimal solutions, multiple starting points are recommended (Redner and Walker, 1984). Among the eleven different starting p53 PWMs tested, ten converged to p53 motif while the most degenerate one converged to a distinct maximum. For those that converged to p53 motif, seven belonged to one maximum whereas the other three belong to a slightly different maximum.

To estimate the FDR, one would need to know the proportion of non-motif subsequences that are falsely declared as significant. In practice, this proportion is not observable. Instead, we use many sets of simulated background sequences from the input sequence model to estimate the proportion. This idea is similar to the concept of adding pseudo variables to tune variable selection in the classical regression settings (Luo et al., 2004; Wu et al. 2006) and to compare different variable selection methods (Miller 2002). However, we believe that this idea has not yet been used in estimating the false discovery rate in sequence analysis. Further discussion of FDR is in supplementary s4.

7. CONCLUSION

We have proposed a novel method for identifying transcription factor binding sites in a set of sequences. Our method considers zero or multiple occurrences of the motif in a sequence. It selects as many binding sites as possible while controlling a user-specified FDR. The PWM is optimized using an EM-like procedure similar to that in MEME. No motif abundance parameter or post-analysis statistical significance test is needed. The choice for FDR is intuitive and errors in multiple comparisons are controlled. Furthermore, we propose an improved normalization procedure in the E-step. Results on real ChIP data and simulated data suggest that our method performs better than MEME. Our method is fast and robust to inclusion of sequences without the motif of interest. In summary, we believe that combination of the following features—combining model optimization and significance testing, normalizing subsequence probability sequence by sequence, and using background sequences as null sequences make fdrMotif a uniquely useful tool for motif identification.
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. We thank Grace Kissling and David Umbach for critical reading the manuscript and their insightful comments. We thank the anonymous reviewers for improving our manuscript. We also thank Xuting Wang and Douglas Bell for providing the known p53 binding data used in the simulations.

REFERENCES

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol 1994;2:28–36.
Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B 1995;57:289–300.
Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Stat 2001;29:1165–1188.
Crooks GE, et al. WebLogo: A sequence logo generator. Genome Res 2004;14:1188–1190. [PubMed: 15173120]
Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999;15:663–577.
Horvath MM, et al. Divergent evolution of human p53 binding sites: cell cycle versus apoptosis. PLoS Genet 2007;3:1284–1295.
Jensen ST, et al. Computational discovery of gene regulatory binding motifs: a Bayesian perspective. Statist. Sci 2004;18:188–204.
Kim TH, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 2007;128:1231–145. [PubMed: 17382889]
Knuppel R, et al. TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol 1994;1:191–198. [PubMed: 8790464]
Li L, et al. GAPWM: GAPWM: a genetic algorithm method for optimizing a position weight matrix. Bioinformatics 2007;23:1188–1194. [PubMed: 17341493]
Lin CY, et al. Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet 2007;3:867–885.
Liu X, et al. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput 2001;6:127–138. [PubMed: 11262934]
Liu XS, et al. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat. Biotechnol 2002;20:835–839. [PubMed: 12101404]
Luo X, et al. Tuning variable selection procedures by adding noise. Technometrics 2004;48:165–175.
Miller, AJ. Subset selection in regression. Chapman & Hall; London: 2002.
Redner RA, Walker HF. Mixture densities maximum likelihood and EM algorithm. SIAM Rev 1984;26:195–239.
Roth FP, et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol 1998;16:939–945. [PubMed: 9788350]
Sandelin A, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 2004;32:D91–D94. [PubMed: 14681366]
Smith AD, et al. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc. Natl. Acad. Sci. USA 2005;102:1560–1565. [PubMed: 15668401]
Storey JD. A direct approach to false discovery rate. J. R. Statist. Soc. Ser. B 2002;64:479–498.
Storey, JD.; Tibshirani, R. Technical Report. Vol. 28. Department of Statistics; Stanford University, CA: 2001. Estimating the positive false discovery rates under dependence, with applications to DNA microarrays.

Bioinformatics. Author manuscript; available in PMC 2008 May 12.
Tompa M, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol 2005;23:137–144. [PubMed: 15637633]

Thijs G, et al. A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 2001;17:1113–1122. [PubMed: 11751219]

Tsai CA, et al. Estimation of false discovery rates in multiple testing application to gene microarray data. Biometrics 2003;59:1071–1081. [PubMed: 14969487]

Wei CL, et al. A global map of p53 transcription-factor binding sites in the human genome. Cell 2006;124:207–219. [PubMed: 16413492]

Wu Y, et al. Controlling variable selection by the addition of pseudo variables. J. Am. Stat. Assoc 2007;102:269–279.

Zaykin DV, et al. Truncated product method for combining P-values. Genet. Epidemiol 2002;22:170–185. [PubMed: 11788962]
### Table 1

Possible outcomes of $m$ tests of significance

|                  | Declared non-binding sites | Declared binding sites | total |
|------------------|-----------------------------|------------------------|-------|
| Non-binding sites| $U_{M,θ}(τ)$                | $V_{M,θ}(τ)$           | $m_0$ |
| Binding sites    | $T_{M,θ}(τ)$                | $S_{M,θ}(τ)$           | $m_1$ |
| Total            | $m - R_{M,θ}(τ)$            | $R_{M,θ}(τ)$           | $m$   |

$m_0 / m_1$ is the number of non-binding site/binding site subsequences.

$U_{M,θ}(τ) / T_{M,θ}(τ)$ is the number of non-binding site/binding site subsequences declared not binding site.

$V_{M,θ}(τ) / S_{M,θ}(τ)$ is the number of non-binding site/binding site subsequences declared binding site.

$R_{M,θ}(τ)$ is the number of subsequences declared binding sites.

$S_{M,θ}(τ), T_{M,θ}(τ), U_{M,θ}(τ)$ and $V_{M,θ}(τ)$ are all unobservable random variables and are dependent on $Θ$ and $τ$. 

*Bioinformatics*. Author manuscript; available in PMC 2008 May 12.
Table 2
Mean and standard error of sSn, sPPV, sASP, and sPC from analyses of 500 datasets of simulated “ChIP” sequences

| statistics | fdrMotif (6%) | fdrMotif (7%) | MEME MAXP=74 | MEME MAXP=100 |
|------------|--------------|--------------|--------------|---------------|
| sSn        | 0.96 (0.01)  | 0.97 (0.02)  | 0.92 (0.03)  | 0.92 (0.03)   |
| sPPV       | 0.94 (0.03)  | 0.93 (0.03)  | 0.93 (0.04)  | 0.93 (0.04)   |
| sASP       | 0.95 (0.02)  | 0.95 (0.02)  | 0.92 (0.03)  | 0.92 (0.03)   |
| sPC        | 0.91 (0.03)  | 0.90 (0.03)  | 0.86 (0.04)  | 0.86 (0.04)   |

sSn: sensitivity at the site level; sPPV: positive predictive value at the site level; sASP: site level average performance; sPC: the site level performance coefficient. All statistics were computed according to Tompa et al. (2006).