Integrating quantitative information from ChIP-chip experiments into motif finding

HEEJUNG SHIM

Department of Statistics, 1300 University Avenue, University of Wisconsin, Madison, WI 53705, USA

SÜNDÜZ KELEŞ*

Department of Statistics and Department of Biostatistics and Medical Informatics, 1300 University Avenue, University of Wisconsin, Madison, WI 53705, USA

keles@stat.wisc.edu

SUMMARY

Identifying binding locations of transcription factors (TFs) within long segments of noncoding DNA is a challenging task. Recent chromatin immunoprecipitation on microarray (ChIP-chip) experiments utilizing tiling arrays are especially promising for this task since they provide high-resolution genome-wide maps of the interactions between the TFs and the DNA. Data from these experiments are invaluable for characterizing DNA recognition profiles (regulatory motifs) of TFs. A 2-step paradigm is commonly used for performing motif searches based on ChIP-chip data. First, candidate bound sequences that are in the order of 500–1000 bp are identified from ChIP-chip data. Then, motif searches are performed among these sequences. These 2 steps are typically carried out in a disconnected fashion in the sense that the quantitative nature of the ChIP-chip information is ignored in the second step. More specifically, all bound regions are assumed to equally likely have the motif(s), and the motifs are assumed to reside at any position of the bound regions with equal probability. We develop a conditional two-component mixture (CTCM) model that relaxes both these common assumptions by adaptively incorporating ChIP-chip information. The performances of the new and existing methods are compared using simulated data and ChIP-chip data from recently available ENCODE studies (Consortium, 2004). These studies indicate that CTCM efficiently utilizes the information available in the ChIP-chip experiments and has superior sensitivity and specificity especially when the motif of interest has low abundance among the ChIP-chip bound regions and/or low information content.

Keywords: ChIP-chip data; Conditional mixture models; Piecewise constant linear regression; Prior models; Regulatory motif finding; Tiling microarrays.

1. INTRODUCTION

Annotation of the transcription factor (TF) binding sites in a given genome is crucial for building genome-wide regulatory networks of coding DNA to elucidate complex interactions among genes. Each TF
recognizes a unique family of short sequence elements, usually between 5 and 20 bp in length. A recent technological innovation, termed ChIP-chip for chromatin immunoprecipitation coupled with microarray analysis, has enabled researchers to identify regions of a given genome that are bound by specific TFs by creating high-resolution genome-wide maps of the in vivo interactions between DNA-associated proteins and DNA. This technology has been successfully applied using both high-density oligonucleotide arrays (a.k.a. tiling arrays) (Affymetrix arrays, Cawley and others, 2004; NimbleGen arrays, Kim and others, 2005) and spotted 2 color microarrays (Ren and others, 2000; Weinmann and others, 2001). An immediate merit of ChIP-chip experiments is that they generate data sets which are valuable for identifying DNA sequence recognition profiles (motifs) of the TFs.

Motif finding based on ChIP-chip data traditionally consists of 2 consecutive steps. In the first step of the ChIP-chip data analysis, one is concerned with deducing tens or hundreds of bound regions using the ChIP-chip data on millions of probes on a microarray. The noisy microarray technology and the scarce number of replicate experiments make this task challenging. Several statistical and computational methods have been developed for the identification of bound regions from ChIP-chip data (Cawley and others, 2004; Kim and others, 2005; Ji and Wong, 2005; Li and others, 2005; Keles and others, 2006; Keles, 2007). Although substantial differences exist among these methods due to modeling assumptions and the targeted array platforms, for example, single-channel or 2-channel high-density oligonucleotide arrays, they all aim to partition the tiled regions of the genome into regions that are either bound or unbound. Bound regions refer to the set of probes which, as a whole, show evidence of enrichment in the treatment sample in comparison to the control sample, whereas the unbound regions are hypothesized to be equally enriched in both groups. Due to the spatial structure of the ChIP-chip data, enrichment occurs over the interval of a set of probes rather than for singletons. As typical outcome, each method reports a set of coordinates corresponding to start and end sites of the peaks and a summary measure for each peak. These coordinates are obtained controlling the false discovery rate (Benjamini and Hochberg, 1995) at a user-specified threshold using either an empirically driven cutoff or multiple testing procedures. The nature of the summary measures varies according to the analysis methods but all these, directly or indirectly at one point in the analysis process, assign a measure of enrichment evidence to each probe (e.g. posterior probability of binding; Ji and Wong, 2005; Keles, 2007).

In the second step of the ChIP-chip data analysis, popular motif-finding algorithms such as Bailey and Elkan (1995) and Liu and others (2001) are used to search for enriched subsequences in these regions. Input to these algorithms are genomic sequences from the identified peak regions. There is a vast number of motif-finding methods but a large portion of these, at the core, relies on the two-component mixture (TCM) model proposed by Lawrence and Reilly (1990) and Lawrence and others (1993). Let $X_{i,k} \in \{A, C, G, T\}$ denote the nucleotide at position $k$ of sequence $i$ and $L_i$ denote the length of this sequence. Let $X_i = \{X_{i,k}\}_{k=1}^{L_i}$ represent all nucleotides in sequence $i$. The observed data are represented by $N$ i.i.d. random variables $\{X_1, \ldots, X_N\}$. The data-generating distribution is assumed to be composed of 2 components, namely the background and the motif component. The background component is typically a low-order Markov chain, for example, third order, over the DNA alphabet. The motif model is described by a product multinomial where each position of the binding site is assumed to have an independent multinomial distribution. This resulting representation for the binding site is also referred to as a position weight matrix (PWM) model. Here, a position weight matrix refers to 4 by length of the binding site columns where each column corresponds to a multinomial distribution (Stormo, 1982). Additionally, these models are described via a hidden random variable $Z$ that takes on values 0 or 1 depending on whether or not the sequence under consideration has 0 or more copies of the motif of interest. Thus, the distribution of the $i$th sequence is given by

$$\Pr(X_i) = \sum_{j=0}^{1} \Pr(X_i \mid Z_i = j) \pi_j,$$
where \( \Pr(Z_i = j) = \pi_j \) is assumed to be the same for all the sequences under consideration. For the case where the \( i \)th sequence has a single copy of the motif, we have

\[
\Pr(X_i \mid Z_i = 1) = \sum_{l=1}^{L_i - W + 1} \Pr(X_i \mid Y_{il} = 1, Y_{ik} = 0, k \in \{1, \ldots, L_i - W + 1\} \setminus l, Z_i = 1) \frac{1}{(L_i - W + 1)},
\]

where \( W \) denotes the length of the binding site. Here, each \( Y_{ij}, j \in \{1, \ldots, L_i\} \), takes on the value 0 or 1 depending on whether or not the motif starts at this particular \( j \)th position. Currently, all the available methods assume a discrete uniform distribution for the start site, that is \( \Pr(Y_{ij} = 1 \mid Z_i = 1) = 1/(L_i - W + 1) \).

The above 2-step framework has 2 limitations that are apparent from the underlying model: (1) each region is assumed to equally likely have the motif(s) of interest and (2) motifs are assumed to reside at any position of the bound regions with equal probability. In this paper, we address relaxation of these 2 assumptions by developing an adaptive framework that integrates quantitative information from ChIP-chip experiments into motif finding. The second assumption is especially limiting if the regions considered are long. Although the binding sites localize at a particular position of the 500-to 1000-bps-long regions (namely close to the center of the peak), currently available motif-finding algorithms ignore this information and treat each position of the sequences equally likely to be a motif start site.

In the remainder of the paper, we empirically investigate the amount of quantitative information available in ChIP-chip data using real case studies. We then present a conditional two-component mixture (CTCM) model that incorporates ChIP-chip data into motif finding beyond the use of bound sequences. The performance of this modeling framework is compared with the 2-step analysis framework using simulated data and data from ChIP-chip case studies.

### 2. Regions with High ChIP-chip Scores are More Likely to Have the Motif(s) of Interest

To motivate our approach further, we first investigate the amount of quantitative information available in ChIP-chip data using the Stat1 and c-Jun ChIP-chip experiments from the Encyclopedia of DNA Elements (ENCODE) regions. ENCODE is a public research consortium devoted to identifying all functional elements in the human genome sequence and currently focuses on 1% of the whole human genome sequence. Snyder lab from Yale identified 345 Stat1 and 200 c-Jun binding sites in the ENCODE regions using Nimblegen tiling arrays. We scored both the Stat1 and the c-Jun bound regions by their corresponding position weight matrices from TRANSFAC (Wingender, 1994) using the PATSER program (Hertz and Stormo, 1999). We then investigated the correlations between the PATSER scores of the top scoring subsequences in each region and the corresponding ChIP-chip scores. Top-left panel of Figure 1 displays these correlations for Stat1 as a function of the top scoring ChIP-chip peaks. The \( p \)-values, each corresponding to the statistical test that the correlation is 0, are also displayed along with the correlations. As depicted in this plot, there is a significant correlation between ChIP-chip scores and motif scores. For the top 20 regions, this correlation does not seem significant simply because all have almost identical ChIP-chip scores. As we consider more regions, ChIP-chip scores start to vary and the correlation becomes significant. It is also interesting to note that after about 80 regions, the correlation drops significantly. This is the point where additional regions start to sporadically have the motif of interest. To illustrate that what we see in these bound regions is different from what we would observe for the unbound regions, we chose a set of unbound regions from the ENCODE data, matching in number and length to the bound sequences, and repeated the same analysis. The top-right panel of Figure 1 displays the correlations and the corresponding \( p \)-values. Notably, correlations between the 2 scores are
Fig. 1. Correlations between ChIP-chip scores and motif scores as a function of top scoring ChIP-chip peaks. The \( p \)-value at each \( k \) corresponds to the null hypothesis that the correlation is 0. Both ChIP-chip data are from the ENCODE tracks of University of California at Santa Cruz genome browser. Stat1: Motif scores are based on top matching subsequences. c-Jun: Motif scores are based on top 3 matching subsequences.

significantly low and are never significant for these unbound regions. Similar results are observed when the analysis is repeated with c-Jun ChIP-chip data as shown in the 2 bottom panels of Figure 1.

3. CTCM MODEL

In order to directly incorporate quantitative ChIP-chip data into motif finding, we consider modeling the sequence data conditional on the ChIP-chip scores. Before presenting our model, we will give a brief
review of the TCM model (Bailey and Elkan, 1995) and set the notation. TCM model assumes that there is an arbitrary number of nonoverlapping occurrences of the motif in each sequence. A given sequence $X_i$ is generated by repeatedly deciding whether to insert a background nucleotide or a motif of width $W$. A rate parameter $\lambda$ denotes the probability that a given position is a motif start site rather than a background site. Due to high computational complexity of the TCM model, Bailey and Elkan (1995) uses a slightly modified mixture model. In this modified version, the original data set is converted to a new pseudo–data set by breaking up each sequence $X_i$ into overlapping subsequences of length $W$. Let $\tilde{X}_{i,j}$ denote the subsequence of length $W$ starting at position $j$ in sequence $X_i$. Let $Z_{i,j}$ be the indicator variable denoting whether a motif starts at position $j$ in sequence $X_i$, that is, whether sequence $\tilde{X}_{i,j}$ is a realization from the position weight matrix. The pseudo–data set is modeled as though it comes from a TCM model assuming that all the $\tilde{X}_{i,j}$ are independent of each other. $\lambda'$ represents the mixing proportion in the modified model. Although the independence assumption is violated due to the overlap of the sequences, Bembom and others (2007) illustrated that this approximation performs comparable to the exact likelihood inference and the benefit of using the exact TCM likelihood is only confined to high motif abundance scenarios. Therefore, we will also make use of this modified TCM model in our model development. For the sake of presentation, we use a zeroth-order Markov chain for the background distribution. The likelihood of the subsequence $\tilde{X}_{i,l}$ conditional on the hidden variable $Z_{i,l}$ that denotes whether or not the $l$th segment in sequence $i$ has a motif is given by

$$
\Pr(\tilde{X}_{i,l} \mid Z_{i,l} = 1, \Psi_W) = \prod_{k=1}^{W} \prod_{j=1}^{4} p_{kj} I_{(X_{i,l+k-1-j})}, \quad \Pr(\tilde{X}_{i,l} \mid Z_{i,l} = 0, \Psi_0) = \prod_{k=1}^{W} \prod_{j=1}^{4} p_{0j} I_{(X_{i,l+k-1-j})},$

where $\Psi_0$ refers to parameters of the background model and $\Psi_W$ refers to parameters of the motif product multinomial model. Thus, $\Psi_0$ consists of the background multinomial probabilities $p_{0j}, j = 1, \ldots, 4$, and $\Psi_W$ consists of the product multinomial parameters $p_{kj}, k = 1, \ldots, W$ and $j = 1, \ldots, 4$. We will let $\Psi$ to denote all the parameters, $\Psi_0, \Psi_W, \pi_0$, in the model.

### 3.1 CTCM formulation

Let $T = (T_1, \ldots, T_N)$, where $T_i = (T_{i,1}, \ldots, T_{i,L_i})$, denote the ChIP-chip scores for each base pair in each sequence. These scores are driven by the method used to analyze the ChIP-chip data. They could correspond to the posterior probabilities of binding (Ji and Wong, 2005; Keleş, 2007), smoothed averages of the probe-level intensities, or simply some choice of probe-level test statistics. As noted in Section 1, current tiling array designs do not usually provide single base pair–level resolution due to monetary limitations; therefore, the extent of the available ChIP-chip information is limited by the resolution of the array used. Here, all the bases within a probe are assigned the ChIP-chip value of the probe. ChIP-chip scores for the regions in between the probes are obtained by interpolation with simple averaging. We assume that conditional on the motif occurrence and location random variables, the sequence data are independent of the ChIP-chip data. This explicitly corresponds to the assumption $\tilde{X}_{i,l} \perp T_{il} \mid Z_{i,l}$. Next, consider the following conditional probability model:

$$
\Pr(\tilde{X}_{il} \mid T_{il}, \Psi) = \sum_{z=0}^{1} \Pr(\tilde{X}_{il} \mid Z_{il} = z, \Psi_0, \Psi_W) \Pr(Z_{il} = z \mid T_{il}, \Psi_f), \quad (3.1)
$$

where $\Psi_f$ denotes the parameters of the conditional distribution of $Z$ given $T$, and the conditional independence assumption makes the first term in the sum independent of the ChIP-chip scores. Next, we consider 3 alternatives for modeling $\Pr(Z_{il} = 1 \mid T_{il})$. 

---

*Note: The text is a snippet from a scientific document discussing a method for integrating ChIP-chip data, focusing on the TCM model and its modifications. The text is quite technical and involves probability and statistics theory.*
**Beta prior on \( \lambda \) (M1).** The first alternative considers a beta prior formulation for \( \Pr(Z_{il}) \) as follows:

\[
Z_{il} \mid \lambda_{il} \sim \text{Bernoulli}(\lambda_{il}),
\]

\[
\lambda_{il} \mid T_{il} \sim \text{Beta} \left( 1, \frac{v}{v + T_{il}} \right),
\]

where \( v \) is the unknown scaling factor. Then, by integrating out \( \lambda_{il} \), we get

\[
\Pr(Z_{il} = 1 \mid T_{il}) = \frac{T_{il}}{v + T_{il}}.
\]

**Logistic regression model (M2).** The following logistic regression model

\[
\log \left( \frac{\Pr(Z_{il} = 1 \mid T_{il})}{1 - \Pr(Z_{il} = 1 \mid T_{il})} \right) = \beta_0 + \beta_1 T_{il}
\]

models the log-odds of having a motif starting at position \( l \) as a linear function of the ChIP-chip score at that position.

**Piecewise constant model (M3).** A semiparametric alternative to the above formulations is a piecewise constant model. In this model, the ChIP-chip scores are partitioned into \( k \) intervals with boundaries \( t_1, t_2, \ldots, t_k, t_{k+1} \). Within each interval, \( \Pr(Z_{il} = 1 \mid T_{il}) \) takes on an unknown constant value given by

\[
\Pr(Z_{il} = 1 \mid T_{il}) = \begin{cases} 
\alpha_1, & \text{if } T_{il} \in [t_1, t_2), \\
\alpha_2, & \text{if } T_{il} \in [t_2, t_3), \\
\vdots & \\
\alpha_k, & \text{if } T_{il} \in [t_k, t_{k+1}).
\end{cases}
\]

This model is capable of approximating both the linear and the nonlinear dependencies of \( Z \) on \( T \) and is therefore a semiparametric alternative to the above beta prior and logistic regression formulation that both assume restricted functional forms.

### 3.2 Inference for the CTCM model

Parameter estimation in all 3 formulations can be carried out using the standard expectation–maximization (EM) (Dempster and others, 1977) formulation. The models differ in terms of their computational complexities in estimating the parameters corresponding to the ChIP-chip conditional likelihood. Neither the beta prior nor the logistic regression formulation has a closed-form M-step. Thus, they require numerical optimization for estimating \( \Psi_f \). On the other hand, we can easily derive closed-form M-step updating formulas for the vector parameter \( \alpha \) in the piecewise constant model framework. The expected full-data likelihood of the conditional TCM model at the \( r \) step of the EM algorithm is given by

\[
Q(\Psi \mid \Psi^r) = \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E \left[ I_{\{Z_{il}=1\}} \mid \tilde{X}_{i,l}, T_{i,l}, \Psi^r \right] \log \Pr(Z_{i,l} = 1 \mid T_{i,l}, \Psi_f) \\
+ \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} E \left[ I_{\{Z_{il}=1\}} \mid \tilde{X}_{i,l}, T_{i,l}, \Psi^r \right] \log \Pr(\tilde{X}_{i,l} \mid Z_{i,l} = 1, \Psi)
\]
\[ + \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} \left( 1 - E[I[Z_{i,l} = 1] | \tilde{X}_{i,l}, T_{i,l}, \Psi^r] \right) \log (1 - \Pr(Z_{i,l} = 1 | T_{i,l}, \Psi_f)) \]

\[ + \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} \left( 1 - E[I[Z_{i,l} = 1] | \tilde{X}_{i,l}, T_{i,l}, \Psi^r] \right) \log \Pr(\tilde{X}_{i,l} | Z_{i,l} = 0, \Psi). \]

Here, \( \Psi^r \) denotes the parameter estimates after the \( r \)th step of the EM algorithm. Then, the EM algorithm steps are as follows:

E-step:

\[ \zeta_{i,l} = E[I[Z_{i,l} = 1] | \tilde{X}_{i,l}, T_{i,l}, \Psi^r] = \Pr(Z_{i,l} = 1 | \tilde{X}_{i,l}, T_{i,l}, \Psi^r) \]

\[ = \frac{\Pr(\tilde{X}_{i,l} | Z_{i,l} = 1, \Psi^r_m)\Pr(Z_{i,l} = 1 | T_{i,l}, \Psi^r_f)}{\Pr(\tilde{X}_{i,l} | Z_{i,l} = 0, \Psi^r_m)(1 - P(Z_{i,l} = 1 | T_{i,l}, \Psi^r_f)) + P(\tilde{X}_{i,l} | Z_{i,l} = 1, P^r)P(Z_{i,l} = 1 | T_{i,l}, \Psi^r_f)}, \]

where \( \Psi^r_m = (\Psi_0^r, \Psi_w^r) \) represents the back ground and position weight matrix parameter estimates after the \( r \)th iteration.

M-step: For \( M3 \), parameter estimates of the piecewise constant regression model are given by

\[ \hat{\alpha}_k' = \frac{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} \zeta_{i,l} I[T_{i,l} \in I_k]}{\sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} I[T_{i,l} \in I_k]}, \]

where \( I_k = [t_k, t_{k+1}) \) and

\[ \hat{\alpha}_k = \frac{1}{\sigma_k^2 - W + 1}, \]

where \( \hat{\alpha}_k \) and \( \hat{\alpha}_k' \) correspond to the estimates of the mixing proportions in the original and modified CTCM models. The parameter \( \alpha_k' \) is a generalization of the rate parameter \( \lambda' \) of the TCM model. Moreover, we have

\[ \hat{p}_{0j} = \frac{N_{0j}}{\sum_{j=1}^{4} N_{0j}}, \quad j \in \{1, \ldots, 4\}, \]

\[ \hat{p}_{wj} = \frac{N_{wj}}{\sum_{j=1}^{4} N_{wj}}, \quad j \in \{1, \ldots, 4\}, \quad w \in \{1, \ldots, W\}, \]

where

\[ N_{wj} = \begin{cases} \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} \zeta_{il} I(X_{i,l+w-1} = j), & w \neq 0, \\ \sum_{i=1}^{N} \sum_{l=1}^{L_i-W+1} [(1 - \zeta_{il}) \sum_{k=1}^{W} I(X_{i,l+k-1} = j)], & w = 0. \end{cases} \]

### 3.3 A technical note

The use of \( M2 \) formulation for our CTCM model warrants some caution in especially small simulated data sets. Under the condition of perfect linear separability (e.g. sequences with and without the motif are
perfectly linearly separable based on their ChIP-chip scores), the proposed EM algorithm can converge to nonfinite estimates for the vector parameter \( \beta \). This is due to the fact that M-step of the EM algorithm can reduce to exactly a logistic regression log-likelihood and the problem of nonfinite estimates is a known issue (Albert and Anderson, 1984) in the fitting of logistic regression models. Due to the factorization of the likelihood, consistent estimates for the other parameters of the model are still attainable in this pathological case. We prove this property of the EM algorithm for the logistic regression CTCM model algebraically in Section 6 of the supplementary material available at Biostatistics online. From a practical point of view, this is not a concern as such perfect separability is hardly ever encountered in the ChIP-chip data from tiling arrays.

3.4 Model selection for the piecewise constant model

We note that M3 conditional formulation requires the partitioning of the ChIP-chip score into bins and setting of the bin boundaries. The number of parameters in this model increases as the number of bins increase. Therefore, model selection schemes are needed to prevent overfitting. We utilize the Bayesian information criteria (BIC) to select the number of bins. Given the bin numbers, the bins are set to equal widths to cover the range of the ChIP-chip scores. As will be illustrated in our case study, the visual inspection of the histogram of the ChIP-chip scores often suggests natural bin boundaries.

3.5 Previous algorithms utilizing ChIP-chip data in motif finding

Currently, there are only a small number of motif-finding approaches (Liu and others, 2002; Hong and others, 2005) tailored toward ChIP-chip identified regions. These approaches were developed before the tiling array technology; therefore, they have shortcomings in utilizing ChIP-chip data from tiling arrays. MDSCAN (Liu and others, 2002), originally developed for ChIP-chip data from cDNA arrays, is not based on the TCM model. Instead, it first generates seeds that are in the form of nonredundant \( W \)-mers from the top \( k \) \( \in (3, 20) \) sequences (\( k \) regions with the highest ChIP-chip score). For each seed, all the \( m \) matches (subsequences with at least \( m \) matches to the seed) in the top \( k \) regions are found and then are used to generate a position weight matrix. Then, each matrix is evaluated based on an approximate maximum a posteriori scoring function, and the resulting top 10–50 matrices are further calibrated using the rest of the sequences that were not among the top \( k \). Although this approach respects the ranking of the ChIP-chip scores, it does not take into account the second issue, that is motifs are more likely to reside near the center of the peak. More recently, Hong and others (2005) developed a discrimination-based approach to enhance motif models using a boosting technique. This approach utilizes the correlation between ChIP-chip scores and motif occurrence scores. Similar to MDSCAN, the information regarding the most likely positions of the motif in a given sequence is not incorporated.

4. Results

4.1 Incorporating quantitative ChIP-chip information compensates for increased noise

It is now well established that signal-to-noise ratio in motif searches decreases dramatically as the lengths of the sequences become longer or the proportion of sequences with the motif(s) decreases rapidly (Keich and Pevzner, 2002). Therefore, it is vital to take into account as much prior information as possible when performing motif finding with large number of long sequences. We compared CTCM to TCM of MEME and MDSCAN on 2 case studies. Using publicly available c-Jun and Stat1 ChIP-chip data from the ENCODE project, we generated sequence data sets containing \( \pm 100, 300, 500 \) flanking bases of the mid points of the peaks (i.e. probe with the highest ChIP-chip score within the peak) and varied the
number of bound regions used in the analysis as \( N = 20, 80, 200 \). Details of these data sets are provided in Section 1 of the online supplementary material available at *Biosatistics* online. With this analysis, we aim to illustrate that one can decrease the signal-to-noise ratio in the sequence data set but can preserve roughly the same performance by utilizing quantitative ChIP-chip information in an adaptive way. In the comparisons below, we feed each method the correct motif width as our focus here is not the identification of the right width. Selection of the motif width \( W \) is an integral part of motif finding and has been extensively studied in Bailey and Elkan (1995), Keleș and others (2003) and Bembom and others (2007).

A closer look into MEME software code revealed that MEME allows sequence-specific weights that are used to smooth the E-step of the algorithm. These weights are restricted between 0 and 1 and are kept constant during the model fitting process. Although the weights are available in the code, we did not come across their explanation or use in the MEME methodology papers. Nonetheless, we decided to take advantage of these for comparison purposes since smoothing of the E-step aims to down-weight or up-weight contributions of the individual sequences via weights. Note that MEME only uses one weight per sequence, thereby does not incorporate information on likely motif start sites within a sequence. There are no guidelines for setting the weights in MEME, therefore we experimented with \( w_1 = T_i / \sum_{j=1}^{N} T_j \) and \( w_2 = T_i / \max_{j \in \{1, \ldots, N\}} T_j \). CTCM (M1, M2, M3), MEME, and weighted versions of MEME (MEME(w1) and MEME(w2)) were employed with a starting value that is similar to the motifs of interest. MDSCAN does not accept as input a starting position weight matrix. So, in order to make the MDSCAN comparisons fair, we considered top 5 position weight matrices estimated by MDSCAN and used the best one, for example, the one closest to the true PWM in the Euclidean distance. Mean of the componentwise squared distances between the known and the estimated position weight matrices of c-Jun is displayed in Figure 2. As depicted in this figure, CTCM has the smallest distance between the true and the estimated position weight matrices in all cases regardless of the specific CTCM model (M1, M2, or M3) used. Investigating this analysis more closely, we note that CTCM’s performance is fairly constant in the c-Jun analysis,

![Fig. 2. c-Jun: illustration of the sequence length and sample size effect in the motif analysis of ChIP-chip identified regions. \( [N, L] \), \( N = 20, 80, 100 \) and \( L = 100, 300, 500 \), refer to \( N \) highest scoring peak regions and \( L \) flanking bases to the right and of the mid point of the peak. Mean squared error (MSE) is calculated by averaging the squared distance between the components of the true and the estimated position weight matrices. MEME(wk), \( k = 1, 2 \), are 2 weighted versions of MEME. MDSCAN* refers to the use of the best position weight matrix estimate in the MSE sense among the top 5 reported by MDSCAN. M1, M2, and M3 refer to beta prior, logistic regression, and piecewise constant linear model formulations of CTCM.](image-url)
especially for the logistic regression (M2) and piecewise constant linear model (M3), as the sample size and length of the sequences are increased. On the other hand, MEME and its weighted versions and MD-SCAN have increased mean squared distance especially when the length of the sequences are increased from 100 to 300 and 500 bp. This eludes to the fact that including quantitative information on the potential locations of the motifs provides compensation for the increased noise level. More specifically, utilizing ChIP-chip scores directly, CTCM reduces the search space of motif occurrences. The histogram of ChIP-chip scores and the estimates of the parameter $\alpha$ within each bin in the CTCM (M3) model are provided in Figure 3 for the cases $N = 20, L = 100$ (top panel) and $N = 80, L = 300$ (bottom panel) of the c-Jun analysis. In these analysis, the number of bins were set based on the visual inspection of the histograms of ChIP-chip scores. The estimates of $\alpha$ indicate that some of the bins could potentially be collapsed. As illustrated in these plots, the probability that a given base is a motif start site highly depends on the ChIP-chip scores. In the bottom panel of Figure 3, we observe that probe sequences with a ChIP-chip score of 1.52 or higher are twice as more likely to have a motif instance than the probe sequences with a score of 1.52 or smaller. We also note significant differences between weighted and unweighted versions of MEME where the weighted versions often have worse performance than the unweighted version. We elaborate on the potential reasons of these differences within the context of our simulations in Section 4.2.

Fig. 3. c-Jun: histogram of the ChIP-chip scores of probes. Top panel: $N = 20$ and $L = 100$: probes within 100 bp to the right and left of the mid points of the top 20 c-Jun peaks are considered. Bottom panel: $N = 80$ and $L = 300$: probes within 300 bp to the right and left of the mid-points of the top 80 c-Jun peaks are considered. In both panels, vertical lines depict bin boundaries. Estimates of the bin-specific $\alpha$ parameters are reported for each bin.
An additional observation is that the performances of the logistic regression model (M2) and piecewise constant linear model (M3) are remarkably similar in this analysis. We do not expect such a result to hold for all the analysis. However, since the logistic regression model does not require model selection over the bin sizes, it might serve as a computationally easier alternative. The best performance for the Stat1 data is also achieved with the CTCM logistic regression model M2. We refer to Section 2 of the supplementary material available at *Biostatistics* online for a detailed investigation of the model fit to Stat1 data.

These case studies provide the proof of the principle that including quantitative information from ChIP-chip experiments into motif finding leads to better position weight matrix estimates that are robust against increased noise level in the sequence data. Notably, all the 3 formulations we proposed, namely beta prior, logistic regression, and piecewise constant model provide better means of utilizing ChIP-chip information than *ad hoc* weighting schemes that are possible via MEME. Overall, even though the logistic regression model is a restrictive parametric model, it performed as better or even better than the flexible semiparametric alternative piecewise constant linear model.

### 4.2 Comparative analysis of CTCM with MEME and MDSCAN based on simulated data

In order to investigate the operating characteristics of our CTCM model framework systematically, we performed simulation studies. The data-generation processes for these simulations are described in detail in Section 4 of the online supplementary material available at *Biostatistics* online. We varied 2 main parameters, namely, the degeneracy of the underlying position weight matrix and the abundance of the motif within the sequences and considered 3 different position weight matrix classes based on the information content to accommodate varying degrees of degeneracy. These 3 classes, listed in Table 1, are chosen to represent the characteristics of the the known position weight matrices from the JASPAR database (Sandelin and others, 2004). Our simulations included position weight matrices with the information content 0.965, 1.172, and 1.389 corresponding to roughly 25, 50, and 75 percentiles of the information content distribution of the matrices in the JASPAR database. All these position weight matrices have larger information contents than those of c-Jun and Stat1 position weight matrices. Average number of motifs within a sequence, \( a \), is varied in the set \( \{10\%, 30\%, 50\%, 70\%, 130\%\} \). A small abundance of 10\% is considered since recent case studies indicate that canonical motif might be present in as small as 10\% of the identified bound regions (Bieda and others, 2006).

For each scenario, we generated 50 data sets, analyzed them with different methods, and computed a receiving operating characteristic (ROC) curve for each method. In all the data sets, the motif occurrence probability based on the ChIP-chip data is generated from a logistic regression model (see supplementary Table 1 of the supplementary material available at *Biostatistics* online for details on the choice of the parameters). The logistic regression model was chosen as the simulation model since it fitted the case study data of Section 4.1. For each simulated data set, area under the receiving operating characteristic (AUROC) curve is computed using the trapezoid method implemented in the “R” package “caTools”. Box plots of AUROC curves for the factors GA repeat binding protein (GABPA), heterodimeric bHLH protein (TAL1-TCF3), and Broad-complex_2 are provided in the top-panel plots of Figure 4 for subsets

<table>
<thead>
<tr>
<th>Matrix</th>
<th>IC</th>
<th>Width</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA0011</td>
<td>0.965</td>
<td>8</td>
<td>Broad-complex_2</td>
</tr>
<tr>
<td>MA0091</td>
<td>1.172</td>
<td>12</td>
<td>TAL1-TCF3</td>
</tr>
<tr>
<td>MA0062</td>
<td>1.389</td>
<td>10</td>
<td>GABPA</td>
</tr>
</tbody>
</table>

Table 1. JASPAR matrices used in the simulations. Information content is defined as \( \frac{1}{W} \sum_{j=1}^{W} \sum_{i=1}^{4} [2 + p_{ij} \log 2(p_{ij})] \), where \( p_{ij} \) denotes the probability of \( i \)th nucleotide at \( j \)th position and \( W \) is the width of the motif. Information content takes on values between 0 and 2.
of the abundance parameter. These subsets are selected such that a wide range of information content and motif abundance combinations is spanned. All the simulation results are provided in Section 5 of the supplementary material (available at Biostatistics online, supplementary Figures 4, 5, and 6). Although in practice the final bin size will be determined by the BIC, we display performances of different bin sizes for illustrative purposes in Figure 4. Overall, we note that modeling the motif occurrence probability within each bin improves the performance compared to MEME, which can be thought as having a single bin. In Figure 4, only the results for M3 model of CTCM with 3 different bin sizes are reported. Application of M1 and M2 leads to comparable results with M2 with a better performance since it corresponds exactly to the simulation model.
In Figure 4, we first observe that even though model M3 does not correspond to the simulation model generating the ChIP-chip data, it is able to provide a good approximation to the data-generating mechanism and has good operating characteristics overall. When the motif abundance ($\alpha$) is low (0.1–0.5), incorporating quantitative information from ChIP-chip data gives strikingly better results with higher sensitivity and specificity than the default MEME that treats each sequence uniformly. Notably, our adaptive scheme that models the “weights” for each sequence is performing substantially better than MEME’s ad hoc weighting schemes. For the conserved motif GABPA, when we increase the motif abundance, MEME, MEME with second weighting scheme, and our adaptive models are performing about the same, with ROC curves rising sharply (area under the ROC curve is close to 1). However, to our surprise, MEME with the first weighting scheme is performing significantly worse than the unweighted scheme. Since the first set of weights have much smaller numerical values compared to the second set of weights, they reduce the motif occurrence probability of each sequence. In our adaptive framework, relative values of the ChIP-chip scores rather than their absolute values gain importance. If the ChIP-chip information does not add any power to motif discovery, $\alpha$ values are expected to have similar estimates for different bins. This flexibility is important for robustness against the measurement and modeling error in ChIP-chip scores.

We also note that for degenerate motifs such as Broad-complex_2, CTCM models are substantially better than MEME and the weighted versions of MEME even at high motif abundance cases.

We perform another comparative analysis using MDSCAN which is the only motif-finding method that is tailored toward improving enrichment-based motif finding with ChIP-chip data. Since publicly available MDSCAN only reports the identified motifs but does not allow thresholding, we base our comparisons directly on the motif estimates rather than generating ROC curves. The MSEs between the true and the estimated position weight matrices (we let MDSCAN to identify 5 motifs and choose the one with the best MSE) are computed for each method. The results displayed in the bottom-panel plot of Figure 4 clearly illustrate that MDSCAN provides worse position weight matrix estimates in terms of MSE and has larger variability especially when the motif is degenerate and moderately abundant. In these calculations, the top $k$ sequences’ parameter of MDSCAN is set to its true value from the data-generating mechanism. However, using top $k$ without accounting for the variability in the ChIP-chip data does not facilitate MDSCAN to identify the motif of interest accurately. MSE levels of CTCM are significantly smaller than those of MEME and its weighted versions at lower abundance and/or low information content scenarios. As the abundance and/or the information content increase, the MSE levels of CTCM, MEME, and MEME(w2) become small and comparable whereas MEME(w1) has substantially higher MSE as expected from the ROC curves.

5. DISCUSSION

We have presented a systematic way of incorporating quantitative ChIP-chip information into motif finding. Both the actual case studies and the extensive simulations suggest that principled use of ChIP-chip quantitative information compensates for the decreased signal-to-noise ratio when considering long sequences around the peak regions. Additionally, in the case of degenerate motifs with low information content, direct use of ChIP-chip data can boost the signal for the motif leading to a better characterization of the DNA recognition profile. We considered 3 models including a beta prior, logistic regression, and piecewise constant linear regression formulation. Among the 3, piecewise constant regression formulation seems favorable due to less model assumptions and computationally easier M-step. However, the logistic regression model performed as better or even better in the actual case studies. Therefore, logistic regression formulation provides an alternative without additional model selection issues, for example, bin sizes in the piecewise constant linear model. Our current implementation of the piecewise constant linear model does not impose any constraints on the ChIP-chip score-specific motif occurrence probabilities. However, a set of order constraints that enforces the bin with the highest ChIP-chip score to have the highest motif
occurrence probability might be desirable. Such a constraint might prevent ChIP-chip information from being overriden by overwhelmingly abundant spurious motif occurrences. In this paper, we focused on identification of individual motifs. However, the probability of sequence data given the start site can be easily replaced with a probabilistic model of motif module as in Zhou and Wong (2004) to allow for the identification of clusters of motifs.

Software availability: The CTCM model implementation will soon become a part of the “R” sequence analysis package “cosmo” (Bembom and others, 2007) and will be downloadable from Sündüz Keleş’s Web site www.stat.wisc.edu/~keles. The designated name for the R package is “SUCcESS” which stands for “Systematic Utilization of ChIP-chip Data for Estimating Sequence Signals”.

ACKNOWLEDGMENTS

This research has been supported by a Wisconsin Alumni Research Foundation grant to Sündüz Keleş from the University of Wisconsin, Madison. Conflict of Interest: None declared.

REFERENCES


Integrating ChIP-chip data


[Received December 11, 2006; accepted for publication March 8, 2007]