

coMOTIF: a mixture framework for identifying transcription factor and a coregulator motif in ChIP-seq Data

Mengyuan Xu, Clarice R. Weinberg, David M. Umbach and Leping Li*

Biostatistics Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA

Associate editor: John Quackenbush

ABSTRACT

Motivation: ChIP-seq data are enriched in binding sites for the protein immunoprecipitated. Some sequences may also contain binding sites for a coregulator. Biologists are interested in knowing which coregulatory factor motifs may be present in the sequences bound by the protein ChIP'ed.

Results: We present a finite mixture framework with an expectation–maximization algorithm that considers two motifs *jointly* and simultaneously determines which sequences contain both motifs, either one or neither of them. Tested on 10 simulated ChIP-seq datasets, our method performed better than repeated application of MEME in predicting sequences containing both motifs. When applied to a mouse liver Foxa2 ChIP-seq dataset involving ~12 000 400-bp sequences, coMOTIF identified co-occurrence of Foxa2 with Hnf4a, Cebpa, E-box, Ap1/Maf or Sp1 motifs in ~6–33% of these sequences. These motifs are either known as liver-specific transcription factors or have an important role in liver function.

Availability: Freely available at <http://www.niehs.nih.gov/research/resources/software/comotif/>.

Contact: li3@niehs.nih.gov

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 3, 2011; revised on June 8, 2011; accepted on June 23, 2011

1 INTRODUCTION

A ChIP-seq experiment profiles the genome-wide binding of one transcription factor. Because transcription factors may work together to regulate gene expression, biologists are interested in knowing which putative coregulatory factor motifs may also be present in the sequences/peaks that were bound by the protein immunoprecipitated. With the primary factor and one coregulatory factor, identifying which sequences contain both motifs, either one or neither of them should provide useful information about gene regulation.

Over the years, many computational tools for motif discovery have been developed. Most of them employ local search techniques typically such as Gibbs sampling, expectation–maximization (EM) or related strategies. The Gibbs sample strategy, first described for sequence analysis by Lawrence *et al.* (1993) and refined by Liu *et al.* (1995), has been implemented in many motif discovery tools such

as AlignACE (Roth *et al.*, 1998), BioProspector (Liu *et al.*, 2001), Motif Sampler (Thijs *et al.*, 2001), GLAM (Frith *et al.*, 2004), NestedMICA (Down and Hubbard, 2005), A-GLAM (Kim *et al.*, 2008), BayesMD (Tang *et al.*, 2008), GIMSAN (Ng and Keich, 2008), info-gibbs (Defrance and van Helden, 2009) and HMS (Hu *et al.*, 2010). The EM algorithm, first applied to motif discovery by Lawrence and Reilly (1990), also has many implementations including the widely used MEME (Bailey and Elkan, 1995) as well as GreedyEM (Blekas *et al.*, 2003), the discriminative PSSM approach (Segal *et al.*, 2003), fdrMotif (Li *et al.*, 2008) and GADEM (Li, 2009). The performance of some of these tools has been assessed (Tompkins *et al.*, 2005). Stormo (2010) recently reviewed motif discovery using EM and Gibbs sampling techniques.

These motif discovery algorithms are univariate in that they identify one motif one at a time (although a tool may run its algorithm repeatedly and output several motifs from a single run after sequentially masking motifs already identified). Application of such univariate motif discovery tools that require sequential masking to the identification of both the primary and cofactors of immunoprecipitated proteins can be problematic especially if the two motifs are similar. The order in which the motifs are masked can have a major effect on the result. For example, a univariate approach would have trouble distinguishing a full site motif from its half site counterpart. Furthermore, a univariate approach would need to combine results for two individual motifs to obtain a composite picture of which motifs appear on which sequences.

Methods that consider the joint distribution of multiple motifs in the sequences include Gibbs Recursive Sampler (Thompson *et al.*, 2003, 2007), cisModule (Zhou and Wong, 2004) and EMCModule (Gupta and Liu, 2005). In contrast to the univariate approaches, those methods simultaneously search for the binding sites for K different motifs, represented by K position weight matrices (PWMs), where K can be either predefined or unknown. Those methods aim to identify clusters of binding sites that are close to each other, e.g. in a 100- or 200-bp window. Such a cluster is often referred to as a *cis*-module. Multiple copies of the same motif are allowed to coexist in a *cis*-module. Those methods are well suited for promoter sequences of coexpressed genes on which multiple transcription factor binding sites tend to colocalize. They may not be ideal for sequences, as ChIP-seq data, that are expected to lack the well-structured *cis*-modules of promoter sequences.

Existing tools are not specifically designed for simultaneously identifying the motifs for both the primary factor and a coregulatory factor in large-scale ChIP-seq data. Consequently, we developed a finite mixture framework not only to determine simultaneously

*To whom correspondence should be addressed.

which sequences contain both motifs, either single motif or neither of them, but also to model the coexistence of the two motifs in a sequence by the joint distribution of the two.

2 METHODS

2.1 Overview

Our model determines which sequences contain both motifs, either single motif or neither of them. It allows either motif to be absent or to be present in either the plus or reverse complement (RC) orientation in any sequence—nine categories of sequences in all. The coexistence of the two motifs in a sequence is modeled by their joint distribution. In essence, our proposed method extends the univariate approach to a bivariate approach within the mixture model framework to explicitly account for the nine categories. To avoid confusion, we refer to the categories as *states*. Under this mixture model framework, we obtain the nine corresponding posterior probabilities for each sequence and estimate the proportions of the entire set of sequences that fall into each of the nine states.

Our approach begins by using two different known PWMs to represent the starting models for the two motifs and a high order Markov background model to represent the background. We utilize the EM algorithm to iteratively update the parameters of the two PWMs and the nine state/mixing proportions. For each sequence, we compute nine likelihood scores (referred to as weight scores), one for each of nine states. For the noise/background state (both motifs absent), the likelihood is computed using a ninth-order Markov background model. For each of the four states with a single motif present (motif1+, motif1−, motif2+ and motif2−), we compute the likelihood of the motif starting at each possible location along the sequence using the respective PWM, with the region outside the motif location represented by the background model. For each of the remaining four states with both motifs present (motif1+, motif2+; motif1−, motif2+; motif1+, motif2−; and motif1−, motif2−), we compute the sequence likelihood of the two motifs starting at each possible pair of locations and model the regions outside the two motifs as background, giving us a matrix of weight scores. Together, those weight scores are then used to compute the probability of each sequence being in one of the nine states. Summing up each of the nine probabilities across all sequences followed by standardization provides the estimates for the nine proportions in the entire data. The weight scores are also used to update the two PWMs. The updated PWMs are subsequently used to compute new weight scores. This iterative process is continued until the changes in the parameters (PWMs and the nine proportions) are small (e.g. <0.0001). Finally, each sequence is classified into the state with the largest posterior probability. Below we give simplified version of our algorithm. The complete description of the finite mixture model and the EM algorithm is provided in the Appendix in Supplementary Material.

2.2 The finite mixture model

Suppose the observed data from a ChIP-seq experiment consist of N independent DNA sequences S_i , $i \in \{1, 2, \dots, N\}$, each with length L_i . Let $S = \{S_i, i = 1, 2, \dots, N\}$ denote the entire set of observed sequences. Let $Z_i = (Z_{i,1}, Z_{i,2})$ indicate the state of motif 1 and motif 2, respectively, in sequence S_i . Z_i is unknown ('missing'). The nine possible states are denoted as (0,0), (0,1), (0,2), (1,0), (2,0), (1,1), (1,2), (2,1) and (2,2), where 0, 1 and 2 indicate motif absent, motif present in plus strand and motif present in RC strand, respectively. Let $\pi_{jk} = P(Z_i = (j, k))$ be the probability that sequence S_i is in state (j, k) , $\sum_{j,k} \pi_{jk} = 1$, and $j \in \{0, 1, 2\}$ and $k \in \{0, 1, 2\}$. Thus, π is a vector of mixing parameters—the proportions of sequences in each of the nine states.

Also unobserved for each sequence are the locations of the motifs' start sites, if any. Let $Y_i = (Y_{i,1}, Y_{i,2})$ denote the locations of start sites for motif 1 and motif 2, respectively, in sequence S_i . $Y_{i,1} \in \{0, 1, 2, \dots, (L_i - w_1 + 1)\}$

and $Y_{i,2} \in \{0, 1, 2, \dots, (L_i - w_2 + 1)\}$ where w_1 and w_2 are lengths of the two motifs, respectively. We assign the start site location as 0 if the respective binding site is absent from S_i . With ChIP-seq data, we expect start sites for motif 1 (the primary motif) to be centrally enriched and model the prior probability of $Y_{i,1}$ by a Gaussian distribution with mean near $L_i/2$ and appropriate standard deviation. Similarly, one could model prior probabilities for the starting site locations when both are present in a sequence by having $Y_i = (Y_{i,1}, Y_{i,2})$ follow a bivariate Gaussian distribution. Alternatively, if there is no reason to think that one location is more likely than another, one could model the distribution of the start sites by a uniform distribution.

Let $\theta_1 = \{\theta_{1rb}, r = 1, 2, \dots, w_1\}$ and $\theta_2 = \{\theta_{2rb}, r = 1, 2, \dots, w_2\}$ denote the PWMs for motif 1 and motif 2, respectively, $b = \{a, c, g, t\}$. Let θ_0 denote the model for the background sequence. One could model the background in various ways. One could use an independent identically distributed multinomial distribution $\theta_0 = \{\theta_{0b}\}$ at each position, where θ_{0b} is the relative frequency of the four nucleotides. Alternatively, one could use a higher order Markov model estimated from a background set, e.g. the entire genome.

The overall parameters $\Theta = \{\theta_0, \theta_1, \theta_2, \pi\}$ include the background distribution, the PWMs of the two motifs and the mixing proportions. Recall that both Z and Y are unobserved ('missing'). In order to find the maximum likelihood estimators of the parameters Θ , that is the values of the parameters that maximize the log likelihood of the observed data $L(\Theta|S) = \sum_i \ln P(S_i|\Theta)$, we use the observed data S together with the missing data $\{Y_i\}$ and $\{Z_i\}$ to form the complete data likelihood and employ the EM algorithm of Aitkin and Rubin (1985).

2.3 The EM algorithm

The EM algorithm is used to find estimates that locally maximize the observed data likelihood given the model. It consists of iterating two steps—E-step and M-step—based on the pseudo-complete data likelihood. Suppose at iteration (t) , the current estimated values of parameters are $\Theta^{(t)} = \{\theta_0^{(t)}, \theta_1^{(t)}, \theta_2^{(t)}, \pi^{(t)}\}$. During the E-step, one calculates the conditional distribution of the missing data (Z, Y) given the observed data S and the current parameter estimates $\Theta^{(t)}$,

$$P(Z_i, Y_i | S_i, \Theta^{(t)}) = \frac{P(S_i, Z_i, Y_i | \Theta^{(t)})}{P(S_i | \Theta^{(t)})}$$

Let $W_{jk}^{(i,t)}(l_1, l_2)$ denote the probabilities or *weight scores* that motif 1 and motif 2 are in state (j, k) and motif 1 starts at position l_1 and motif 2 starts at position l_2 on the i -th sequence at iteration (t) .

$$W_{jk}^{(i,t)}(l_1, l_2) = P(Z_i = (j, k), Y_i = (l_1, l_2) | S_i, \Theta^{(t)}),$$

where $j \in \{0, 1, 2\}$, $k \in \{0, 1, 2\}$ and $l_1 = 0, 1, 2, \dots, (L_i - w_1 + 1)$, $l_2 = 0, 1, 2, \dots, (L_i - w_2 + 1)$. When either index j or k is zero, the corresponding l_j or l_k can only be zero. In other words, location 0 indicates that the respective binding site is absent from S_i . For instance, $W_{00}^{(i,t)} = P(Z_i = (0, 0), Y_i = (0, 0) | S_i, \theta_0^{(t)})$ is the current estimate of the probability that sequence S_i contains neither of the two motifs. Note that for each sequence W_{00} is a single score, whereas W_{10} contains $(L_i - w_1 + 1)$ scores and similarly for W_{20} , W_{01} and W_{02} . On the other hand, W_{11} contains $(L_i - w_1 + 1) \times (L_i - w_2 + 1)$ scores excluding overlapping sites and similarly for W_{12} , W_{21} and W_{22} .

Summed over the relevant possible values of the starting sites (l_1, l_2) , these weight scores serve as the expected values of the missing indicator Z_i and Y_i , that is, the probabilities that sequence S_i is in one of nine states and the two motifs are in specific locations given the current parameter estimates. The prior distributions on the locations are used in calculating these weight scores. The detailed mathematical description of the weight scores is given in the Appendix in Supplementary Material.

These weight scores are then used in the M-step to update the estimates for the model parameters at step $(t+1)$, namely, $\Theta^{(t+1)}$. These estimates maximize the expected log likelihood of the pseudo-complete data with respect to the conditional distribution of the missing data (Z, Y) , or, in

notation, the estimates maximize $E_{(Z,Y|S,\Theta^{(t)})} \ln P(S,Z,Y|\Theta)$, which is described in the Appendix in Supplementary Material.

In the M-step, the EM algorithm maximizes the above expected log likelihood over choices of Θ to find the next estimates $\Theta^{(t+1)}$. Maximizing the expected log likelihood with respect to π , we obtain the estimates for π and similarly for the two PWMs (θ_1 and θ_2).

$$\pi_{jk}^{(t+1)} = \frac{1}{N} \sum_i \sum_{(l_1, l_2)} W_{jk}^{(i,t)}(l_1, l_2),$$

$$\theta_{1rb}^{(t+1)} = \frac{\sum_i \gamma_{1rb}^{(i,t)} + \delta}{\sum_i \sum_b \gamma_{1rb}^{(i,t)} + 4\delta}, \text{ where } r=1, 2, \dots, w_1,$$

$$\theta_{2rb}^{(t+1)} = \frac{\sum_i \gamma_{2rb}^{(i,t)} + \delta}{\sum_i \sum_b \gamma_{2rb}^{(i,t)} + 4\delta}, \text{ where } r=1, 2, \dots, w_2,$$

Here δ is a small pseudo-count (0.0001) to avoid setting a probability to the boundary value of zero and γ_{1rb} and γ_{2rb} represent the expected count of each base at each location within the designated motif location at the current iteration. See the Appendix in Supplementary Material for full definitions of γ_{1rb} and γ_{2rb} and other variables and the complete description of the EM algorithm.

We apply a convergence criterion and stop the EM iterations when $\max|\Theta^{(t+1)} - \Theta^{(t)}| < \varepsilon$, where ε is a selected small number, e.g. 2×10^{-5} , where the maximal difference is taken over all the elements in the matrix-valued parameters.

For notational simplicity, the background in the above algorithm was assumed to follow an independent multinomial distribution at each location, with probabilities that would correspond to frequencies observed in the dataset. Since θ_0 remains quite stable through the EM iterations, to save computation time we did not update it. As a more flexible alternative, in the analysis below of the actual data we used a high-order (e.g. ninth) Markov background model estimated from the entire mouse genome, and we did not update it during estimation.

2.4 Reducing sampling space for computational efficiency of the EM

For the version of the algorithm just described, calculating the joint distribution of the weight score $W_{jk}^{(i,t)}(l_1, l_2)$ for every possible pair of locations (l_1, l_2) requires a great deal of CPU time. For each 400-bp sequence at each E-step, the algorithm calculates approximately $(400+400) \times (400+400)$ such scores (the sum is due to plus and RC strands). For a typical ChIP-seq dataset containing tens of thousands of sequences and hundreds of iterations, the calculations become computationally burdensome. Here, we propose a technique to reduce the number of calculations by restricting attention to the most likely motif start sites.

For ChIP-seq sequences, not all locations are motif binding sites. Thus, one could consider restricting the search to only a few of the $2(L-w+1)$ sites, e.g. the highest scoring non-overlapping ones, and still hope to accurately estimate the model parameters Θ . In each EM step, for every sequence, we sequentially identified all possible *non-overlapping sites* starting from the highest scoring site for each motif separately, using the current estimate of the corresponding PWM. Those individually identified non-overlapping highest scoring sites also served as candidate sites for joint occurrence of the two motifs (for a detailed description of this approach, see Supplementary Material). All other sites are considered ‘background’ and not used to update the model parameters. This technique reduces the number of candidate sites in a sequence from ~ 800 to ~ 40 ($L=400$ and $w=10$). For a dataset containing thousands of sequences, this technique reduces the computational time from a few hours to a few minutes (see Section 3.3).

Real sequences often contain repetitive elements such as ‘AAAA...’, ‘TTTT...’ and ‘ATAT...’. These repetitive elements may lead the algorithm to converge to such patterns. In MEME, this difficulty was addressed by two

heuristic algorithms (SQUASH and SMOOTH) (Bailey and Elkan, 1995) to ensure the total probability within a subsequence from all sites was at most 1.0. By considering only the non-overlapping sites, our proposed technique addresses this problem implicitly.

Lastly, we would like to point out that this pragmatic approach to reducing the sampling space allows the highest scoring non-overlapping sites to be freely updated at each EM step (see Section 4 for example). Nevertheless, to increase confidence that one has achieved a global maximum of the likelihood, we advise repeating the algorithm several times using distinct starting values for the mixing parameters.

2.5 Classification and binding site declaration

Let Θ^* be the parameter estimates after convergence of the EM algorithm. We use a two-step Bayes optimal classification procedure (Duda *et al.*, 2000) to identify the embedded likely binding sites in sequences. First, we classify each sequence to one of the nine possible states by assigning the state with the largest posterior probability $P(Z_i = (j, k) | S_i, \Theta)$,

$$P(Z_i = (j, k) | S_i, \Theta) = \sum_{(l_1, l_2)} W_{jk}^{(i)}(l_1, l_2), \quad j=0, 1, 2 \text{ and } k=0, 1, 2.$$

In a second step, given the state assigned to a sequence at the first step, if the assigned state is not (0,0) we classify its subsequence(s) with the largest likelihood for the respective PWM(s), if any, as the likely binding site(s) for the corresponding motif(s).

Alternatively, one can identify all binding sites in a sequence using a classical testing procedure by first determining the likelihood score distribution of the PWMs using methods such as the probability generating function (Staden, 1989). Details can be found in the Supplementary Material.

3 RESULTS

3.1 Simulation study

To test the performance of our method, we created five simulated ChIP-seq datasets; in each of them $\sim 90\%$ of the sequences contained an Hnf4a (motif 1) binding site, and 10, 20, 30, 40 and 50% of the sequences (randomly and independently selected) contained an Hnf1a (motif 2) binding site. Each dataset initially consisted of the same 1000 genomic regions (400 bp each) randomly selected from the latest mouse genome. For each dataset, Hnf4a and Hnf1a binding sites were inserted at random non-overlapping locations into those sequences using motifs repeatedly generated from the corresponding PWM-determined multinomial distributions [TRANSFAC (Wingender, 2008); M00134 for Hnf4a and MM00132 for Hnf1a]. In a separate experiment, we simulated another five ChIP-seq datasets; in each of them $\sim 90\%$ of sequences contained an Hnf4a binding site, and 10–50% contained a Foxa2 binding site. Our expectation was that because the Hnf1a motif is long and has relatively high information it should be easy to find, whereas the Foxa2 motif is short and AT-rich (as is the mouse genome) and consequently should be challenging to find.

We ran our EM algorithm on all datasets. In each run, we used the PWMs from which the inserted motifs were simulated as the starting PWMs (similar results were obtained with other Hnf4a/Hnf1a and Hnf4a/Foxa2 starting PWMs). We used $\pi_{jk}^{(0)} = 1/9$; that is, each of the nine states was considered equally likely. Also, to better model the background, we used the ninth-order Markov background model estimated from the mouse genome and did not update the background model during the optimization.

For the five Hnf4a/Hnf1a datasets, the algorithm successfully estimated the two motif PWMs simultaneously, as suggested by

Table 1. Summary results for the simulated ChIP-seq data

Hnf4a/Hnf1a									
All overlapping sites					Highest scoring non-overlapping sites (short-cut method)				
Estimated proportion (actual proportion ^a)			ΔPWM ^b		Estimated proportion (actual proportion)			ΔPWM	
Background % (%)	Hnf4a % (%)	Hnf1a % (%)	Hnf4a	Hnf1a	Background % (%)	Hnf4a % (%)	Hnf1a % (%)	Hnf4a	Hnf1a
14.2 (10.2)	84.8 (89.3)	21.3 (10.1)	0.0486	0.2786	18.5 (10.2)	80.9 (89.3)	13.8 (10.1)	0.0559	0.2339
12.5 (10.2)	86.5 (89.5)	25.5 (19.4)	0.0415	0.1036	15.0 (10.2)	83.4 (89.5)	21.0 (19.4)	0.0426	0.0921
10.5 (10.1)	88.9 (89.6)	31.1 (28.5)	0.0304	0.1041	13.8 (10.1)	85.5 (89.6)	27.6 (28.5)	0.0300	0.1185
11.8 (9.9)	85.7 (89.8)	44.6 (40.8)	0.0527	0.0583	14.4 (9.9)	81.7 (89.8)	41.2 (40.8)	0.0577	0.0618
8.8 (9.4)	89.3 (90.4)	52.1 (50.2)	0.0373	0.0526	10.7 (9.4)	85.8 (90.4)	50.0 (50.2)	0.0438	0.0485
Hnf4a/Foxa2									
All overlapping sites					Highest scoring non-overlapping sites (short-cut method)				
Estimated proportion (actual proportion)			ΔPWM		Estimated proportion (actual proportion)			ΔPWM	
Background % (%)	Hnf4a % (%)	Foxa2 % (%)	Hnf4a	Foxa2	Background % (%)	Hnf4a % (%)	Foxa2 % (%)	Hnf4a	Foxa2
11.7 (10.1)	87.3 (89.7)	14.8 (10.0)	0.0403	0.2940	16.2 (10.1)	83.5 (89.7)	12.3 (10.0)	0.0399	0.3037
10.3 (11.2)	89.7 (88.5)	27.1 (20.5)	0.0345	0.1253	14.8 (11.2)	85.0 (88.5)	19.8 (20.5)	0.0316	0.1016
10.9 (9.7)	88.3 (90.2)	33.2 (29.0)	0.0379	0.1085	14.8 (9.7)	84.4 (90.2)	25.9 (29.0)	0.0453	0.0951
11.3 (10.6)	87.3 (89.4)	43.1 (40.8)	0.0339	0.0897	14.2 (10.6)	83.9 (89.4)	36.0 (40.8)	0.0371	0.1060
14.2 (9.9)	85.7 (89.7)	47.6 (47.6)	0.0432	0.0860	16.9 (9.9)	82.6 (89.7)	39.6 (47.6)	0.0451	0.1058

^aThe actual proportion appearing in each simulated dataset differed slightly from the targeted 10%, 20%, etc. due to sampling variability.
^bThe maximal absolute element by element difference over all the elements in the matrix-valued parameters between the true PWM and the estimated PWM.

the acceptably small maximal differences. An exception was for Hnf1a when the proportion actually present was 10% (Table 1). The estimated proportions of sequences containing pure ‘noise’ ranged from 9% to 14%, well in line with the simulated true background proportions (~10%). The estimated proportions (85–89%) of sequences containing an Hnf4a binding site, though slightly low, also agreed well with the simulated true proportions (~90%). The estimated proportions of sequences containing an Hnf1a (motif 2) binding site also agreed well with the simulated true sample proportions, except for the first dataset where the true proportion was low (10.1% versus estimated 21.3%). Similar results were obtained for the five Hnf4a/Foxa2 datasets.

To assess the performance of our method, we checked the proportion of the sequences in each dataset that were correctly classified. For example, a sequence in the true state (1,2) had to be classified as (1,2) to be correct. The median proportions of the sequences that were correctly classified for the five Hnf4a/Hnf1a and Hnf4a/Foxa2 datasets ranged from ~60% to 80% and ~40 to 80%, respectively (Supplementary Tables S1 and S2). For Foxa2a, those values are disappointing but expected, as it is AT-rich and an AT-rich motif should be difficult to identify against an AT-rich background, especially at low abundance. Nonetheless, these results demonstrate that our method is capable of identifying both the primary and the coregulator motifs for the simulated data.

The above analyses were carried out with the version of our algorithm that used all overlapping sites in each sequence. We also performed the same analyses using only the highest scoring non-overlapping sites (see Section 2). As expected, the algorithm

slightly underestimated the motif proportions and overestimated the background proportion (Table 1), since only a small fraction of sites were used to update the motif parameters. Nonetheless, all estimated values for both the PWMs and motif proportions (Supplementary Tables S1–S4) are comparable to those using all sites.

3.2 Comparison with other tools

Our approach was not designed to detect *cis*-modules, and tools designed to detect them such as cisModule are not well suited to the problem that we address. To illustrate the latter point, we applied cisModule (Zhou and Wong, 2004) in two different ways to the five simulated Hnf4a/Foxa2 ChIP-seq datasets. Using cisModule in its default module-mode, CisModule did not find either the Hnf4a or the Foxa2 motif. This observation is not surprising since the two motifs were randomly inserted into the sequences and not necessarily near each other as required for *cis*-modules. Using cisModule to search for multiple motifs simultaneously (‘-s’ option), cisModule did find the primary Hnf4a motif but failed to find Foxa2, the less abundant and more AT-rich secondary motif (Supplementary Table S11).

Our method considers the two motifs jointly, while MEME is a univariate approach. Nonetheless, we ran MEME on the same five simulated Hnf4a/Foxa2 ChIP-seq datasets. For each run, we provided MEME with the same parameters as ours whenever possible—the consensus motif sequence (-cons), the correct motif length (-w), the correct model (zoops) and a seventh-order Markov

background model (a higher order requires a prohibitive amount of memory).

For each dataset, we carried out two separate MEME runs with the respective parameters, one for Hnf4a and one for Foxa2. In each run, MEME identified the abundant Hnf4a motif. However, MEME failed to identify the Foxa2 motif when its proportion in the data was <40%. When it succeeded, MEME did well in predicting which sequences contain the Hnf4a motif or the Foxa2 motif, but not both (Supplementary Table S2c). This is perhaps not surprising as MEME models each motif separately, whereas our method models the two motifs jointly.

To further demonstrate the advantage of our bivariate approach over a univariate approach, we analyzed several datasets using coMOTIF implementations of both approaches. We showed that coMOTIF's bivariate approach performed better than a one-motif-at-a-time approach (section VI in Supplementary Material and Table S7–S10).

Finally, we showed the superiority of coMotif over a simple scanning procedure using the five simulated Hnf4a/Hnf1a datasets. We used the likelihood ratio test implemented in FIMO (Grant *et al.*, 2011) and scanned each dataset twice, once with each of the two PWMs that we used to generate the datasets. We found that, regardless of the *P*-value cut-point used, this simple scanning procedure estimated the pair of marginal proportions of sequences containing Hnf4a or Hnf1a less accurately than coMotif did (Supplementary Table S12).

3.3 Computation time

We ran our algorithm on the five simulated Hnf4a/Foxa2 ChIP-seq datasets. Compared to using all sites, using only the highest scoring non-overlapping sites in the EM algorithm reduced the computation time by a factor of 200–600, depending on the speed of the convergence (Supplementary Table S4). The savings imply, as shown below, that it can be applied to large-scale genomic ChIP-seq data for discovery of a primary motif together with a coregulatory factor motif.

3.4 Applied to real ChIP-seq data

We applied our method to mouse liver Foxa2 ChIP-seq data (Hoffman *et al.*, 2010). The dataset consists of ~11 000 sequences, each 400 bp long. In these data, we attempted to identify Foxa2 motifs along with motifs of its putative coregulatory factors, one at a time. For each candidate coregulatory factor, the unknown parameter Θ included motifs for Foxa2 and for an unknown coregulator. To search for the unknown coregulator motifs, we carried out ~800 independent runs, each of which used a published Foxa2 PWM (Wederell *et al.*, 2008) as the starting value for the Foxa2 motif together with one of the 800 TRANSFAC PWMs as the starting value for a possible co-occurring coregulator motif. We used a ninth-order Markov model for the background and did not update it during the optimization. We set the maximal number of EM steps to 250, stopping sooner if the maximal element-by-element absolute difference between the PWMs from two consecutive EM steps was ≤ 0.0002 .

In all runs, the algorithm estimated that ~75–80% of the sequences contained the Foxa2 motif (the primary motif). The algorithm also estimated that ~32% of sequences contained an Hnf4a binding site and 23–29% contained a Cebpa binding site,

depending on the starting PWM used. These results indicate that both Hnf4a and Cebpa colocalize with Foxa2, suggesting that they may be Foxa2 coregulatory factors. Consistent with these results, Hoffman *et al.* (2010) identified ~12 000 Hnf4a sequences in mouse liver. Approximately, 21–29% of the Foxa2 ChIP-seq peak sequences overlap with the Hnf4a peak sequences by at least one-fourth (100 bp) of the nucleotides. Hnf4a, a liver-specific transcription factor, is important for liver function. Cebpa is also a liver-specific transcription factor. Schmidt *et al.* (2010) recently profiled the genome-wide distribution of Hnf4a and Cebpa binding sites in five vertebrates. Between 30% and 50% of Cebpa binding sites colocalized with Hnf4a binding sites in the ChIP-seq identified loci. Here we found that both Cebpa and Hnf4a binding sites also colocalize with Foxa2 binding sites.

Unlike Hnf4a and Cebpa, the majority of the runs with the other starting PWMs for the coregulator motifs converged to motifs distinct from their starting motifs; most often these were Hnf4a, Hnf4a-like or unknown motifs with low information content. While the EM algorithm performed as expected and found the secondary PWM that maximized the data likelihood, optimizing the PWM for a secondary motif evidently will be difficult for low-abundant motifs.



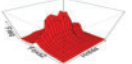


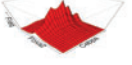
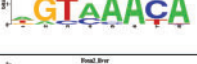
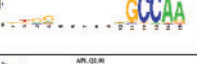

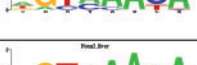







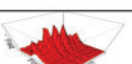
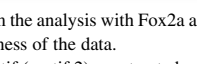
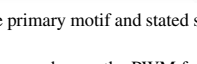
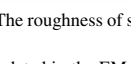
In the absence of an apparent analytical solution to this problem, we repeated the above analysis by simply fixing the PWM for the coregulator motif at its starting value with no iterative updating while updating the PWM for the primary motif and other parameters as before. Essentially, this approach takes the starting PWM for the coregulator motif and treats it as known. Since the PWMs for the coregulator motifs were fixed, many coregulator motifs were identified with a minimal proportion of 5%. To find out which of the coregulator motifs may be enriched in the ChIP-seq data, we compared their abundance in the ChIP-seq data with those from analyses using a set (~16 000) of randomly selected sequences of the same length (400 bp) taken from the mouse genome. Specifically, we scanned both datasets with the same coregulator PWMs. For each PWM and each dataset, we counted the number of sequences containing at least one binding site and the number of sequences without a binding site. For each PWM, we cross-classified dataset (ChIP-seq/random) by binding site (present/absent) and then calculated the enrichment *P*-value using Fisher exact test (section IX in Supplementary Material). As expected, both Hnf4a and Cebpa motifs were highly enriched in the Foxa2 ChIP-seq data compared with the negative control data. In addition, Ap1-like/Maf and several E-box containing motifs such as Ap4 and Neurod1, and Sp1 motif were also enriched (Table 2).

4 DISCUSSION AND CONCLUSION

A ChIP-seq dataset should be enriched in binding sites (motifs) for the protein immunoprecipitated. Some of the sequences may also contain binding sites for a transcriptional coregulator, yet to be identified. To learn about possible coregulatory motifs, we proposed a finite mixture model fitted by applying an EM algorithm not only to identify a coregulatory motif, but also to simultaneously determine which sequences contain both motifs, either one or neither of them.

coMOTIF uses two known PWMs as the starting points for the EM algorithm to elucidate the two motifs. Since the identity of the coregulators may not be known, coMOTIF allows a user to use a set

Table 2. Foxa2 and its putative coregulatory motifs identified in mouse liver Foxa2 ChIP-seq data

Motif1/Motif2	Motif1 logo	Motif2 logo	Motif1 proportion (%)	Motif2 proportion (%)	Site joint distribution	Motif2 enrichment P-value
Foxa2/Hnf4a			73.8	32.8		6.0×10^{-185}
Foxa2/Cebpa			75.0	25.3		5.0×10^{-60}
Foxa2/Nf1 ^a			78.8	14.8		5.1×10^{-63}
Foxa2/Ap1 ^a			76.3	7.0		6.2×10^{-36}
Foxa2/Neurod1 ^a			75.8	6.5		8.8×10^{-33}
Foxa2/Ap4 ^a			78.6	6.9		9.8×10^{-32}
Foxa2/Sp1 ^a			78.5	6.2		1.4×10^{-16}

The Foxa2 motif was obtained from the analysis with Fox2a as the primary motif and stated secondary motif as the coregulator motif. The roughness of some of bivariate distribution plots was due in part to the sparseness of the data.
^aThe PWM for the coregulator motif (motif 2) was treated as known, whereas the PWM for motif 1 (Foxa2 in this case) was fully updated in the EM.

of PWMs as the candidate PWMs and runs it one pair at a time. The set of PWMs could be all the PWMs in a database such as Transfac (Wingender, 2008)

To our knowledge, coMOTIF is unique in considering the *joint* distribution of the two motifs within a sequence and estimating the proportion of sequences in each of nine states defined by cross-classifying whether each motif is absent, present on the plus strand or present on the reverse complementary strand. Since coMOTIF models the coexistence of two motifs in a sequence *jointly* and does not allow motif overlaps, intuitively, it should perform better than the *one-motif-at-a-time* approaches, especially when the two motifs share some resemblance. Our test supports this intuition (section VI in Supplementary Material and Table S7–S10).

A simpler mixture model coupled with an EM algorithm was previously proposed for motif discovery by Bailey and Elkan (1994) and implemented in MEME+. However, our method differs fundamentally from MEME+ in that our framework simultaneously considers the *joint* distribution of two motifs, the presence of either single motif and none (background), whereas MEME+ considers only a single motif and background. Consequently, our method allows nine states, whereas MEME+ allows only three. Furthermore, our algorithm works in the sequence space rather than the (overlapping) subsequence space as in MEME+. Nevertheless, our method can be viewed as an extension to the mixture model of MEME+.

Our proposed method is also fundamentally different from the *cis*-module-based approaches (Gupta and Liu, 2005; Thompson *et al.*, 2003; Zhou and Wong, 2004). The *cis*-module-based approaches are

well suited for sequences that are enriched in multiple transcription factor binding sites such as promoter sequences of coexpressed genes. We aim to identify simultaneously the motif for the protein that was immunoprecipitated and a coregulatory factor motif in ChIP-seq data. The two motifs must be different and can co-exist anywhere in a sequence with or without a modular structure. Moreover, our framework also allows sequences with just one of the two motifs or neither. We use a single mixture framework to simultaneously estimate all nine proportions. We believe that this framework is well suited for identifying transcription factor and its coregulator motifs in ChIP-seq data.

Our approach that finds motifs using a mixture model with a single ChIP-seq dataset is also distinct from discrimination-based approaches that find motifs by contrasting two different datasets. For example, Mason *et al.* (2010) developed a contrast motif finder that could be adapted to finding cofactor motifs using pairs of datasets, though they focused on discerning context-dependent motifs for the same transcription factor.

Modeling the joint distribution of two binding sites within a sequence can be computationally challenging, especially for a large dataset. To greatly reduce the computational time, we proposed to use only the highest scoring non-overlapping sites for updating the motif parameters (both PWM and proportions). This option makes our method practical for genome-wide ChIP-seq data analysis.

With simulated datasets, we demonstrated that the results from using only the non-overlapping sites were comparable to those from using all sites. Intuitively, this technique makes sense since only

a small fraction of all overlapping sites are likely binding sites in a typical ChIP-seq sequence. Importantly, this procedure does not restrict the EM algorithm from reaching its local maximum as the identification of the highest scoring non-overlapping sites is updated at each EM step, as exemplified in Supplementary Table S6. We also showed that coMOTIF is relatively robust to the starting PWMs (section V in Supplementary Material and Table S5).

We investigated the performance of our method on several simulated datasets. To make the ChIP-seq simulations realistic, we used background sequences randomly taken from the mouse genome. In all simulations, the primary motif was present in ~90% of the sequences, whereas the ‘co-regulator’ motif was present in 10–50% of the sequences. Both long and conserved coregulators such as Hnf1a and short and AT-rich coregulators such as Foxa2 were considered. In both cases, the primary and cofactor motifs were successfully identified (Supplementary Table S3). We showed that our method was superior to MEME in identifying the coexistence of two motifs in a sequence while it performed comparably in identifying a single motif in a sequence. We also showed with simulated data that cisModule is not well suited to the problem that coMOTIF addresses. In addition, coMOTIF performed better than a simple scanning procedure for estimating the proportions of sequences containing a common primary motif and containing a less abundant coregulatory motif (section VIII in Supplementary Material).

When tested on the mouse liver Foxa2 ChIP-seq dataset, two known liver-specific transcription factor motifs, Hnf4a and Cebpa, were identified. Both motifs are relatively abundant in the Foxa2 ChIP-seq data. However, the majority of the other starting PWMs for the coregulators motifs converged to different motifs. Although this behavior was expected, it demonstrates that motifs with low abundance are difficult to identify. One solution to this problem, which is implemented as an option in the software, is to fix the PWM for the coregulator motif in the EM procedure at its starting value while updating the PWM for the primary motif and all other parameters. When the secondary PWM is not updated and its starting value is poorly specified, the estimated proportion of sequences containing the coregulator motif will be biased. Fortunately, multiple PWMs for the same transcription factor are often available in databases such as TRANSFAC. The results from different PWMs for the same TF may provide some insight. Knowing which coregulator motifs might be present in which sequences can be useful for generating hypotheses.

Thanks to new technologies such as protein-DNA microarray, SELEX and bacteria-1-hybrid (B1H) [see Stormo and Zhao (2010) for a recent review], a large amount of protein–DNA interaction data has been generated. Computational methods that consider not only the binding sequences, but also experimental binding affinity have led to PWMs with higher specificity than those based on sequences alone (Zhao *et al.*, 2009). Thanks to the advances in both new technologies and computational methods, high-quality PWMs will increasingly be available in databases such as UniProbe (Newburger and Bulyk, 2008).

In conclusion, we propose a finite mixture framework coupled with an EM algorithm to simultaneously model the *joint* distribution of two motifs and classify ChIP-seq sequences containing both motifs, either one or neither of them. We propose a procedure to reduce the sampling space in EM so that the method is applicable to large-scale genomic ChIP-seq data for a transcription

factor and a coregulator motif discovery. Finally, coMOTIF can also take a single PWM and automatically carry out the one-motif analysis as in MEME for a single motif finding. Both functionalities are described in the user manual that is included in the software.

ACKNOWLEDGEMENTS

We thank Grace Kissling and Xuting Wang for their comments on the manuscript and Frank Dai for generating the software configure files.

Funding: Intramural Research Program of the National Institutes of Health; National Institute of Environmental Health Sciences (ES101765-05).

Conflict of Interest: none declared.

REFERENCES

- Aitkin, M. and Rubin, D.B. (1985) Estimation and hypothesis testing in finite mixture models. *J. R. Statist. Soc. B*, **47**, 67–75.
- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
- Bailey, T.L. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn. J.*, **21**, 51–83.
- Blekas, K. *et al.* (2003) Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, **19**, 607–617.
- Defrance, M. and van Helden, J. (2009) info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics*, **25**, 2715–2722.
- Down, T.A. and Hubbard, T.J. (2005) NestedMICA, sensitive inference of over-represented. *Nucleic Acids Res.*, **33**, 1445–1453.
- Duda, R.O. *et al.* (2000) *Pattern Classification*. 2nd edn. John Wiley & Sons, Inc., NY.
- Frith, M.C. *et al.* (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, **32**, 189–200.
- Grant, C.E., *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
- Gupta, M. and Liu, J.S. (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **102**, 7079–7084.
- Hoffman, B.G. *et al.* (2010) Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Res.*, **20**, 1037–1051.
- Hu, M. *et al.* (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
- Kim, N.K. *et al.* (2008) Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics*, **9**, 262.
- Lawrence, C.E. and Reilly, A.A. (1990) An expectation maximization EM algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
- Lawrence, C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
- Li, L. (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J. Comput. Biol.*, **16**, 317–329.
- Li, L. *et al.* (2008) fdrMotif: identifying cis-elements by an EM algorithm coupled with false discovery rate control. *Bioinformatics*, **24**, 629–636.
- Liu, J.S. *et al.* (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
- Liu, X. *et al.* (2001) BioProspector, discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Mason, M.J. *et al.* (2010) Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, **26**, 2826–2832.
- Newburger, D.E. and Bulyk, M.L. (2008) UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acid Res.*, **37**, D77–D82.
- Ng, P. and Keich, U. (2008) GIMSAN: a Gibbs motif finder with significance analysis. *Bioinformatics*, **24**, 2256–2257.

- Roth,F.P. et al. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
- Schmidt,D. et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
- Segal,E. et al. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**, i273–i282.
- Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.*, **5**, 89–96.
- Stormo,G.D. (2010) Motif discovery using expectation maximization and Gibbs' sampling. *Methods Mol. Biol.*, **674**, 85–95.
- Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
- Tang,M.H. et al. (2008) BayesMD: flexible biological modeling for motif discovery. *J. Comput. Biol.*, **15**, 1347–1363.
- Thijs,G. et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
- Thompson,W. et al. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.
- Thompson,W.A. et al. (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res.*, **35**, W232–W237.
- Tompa,M. et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
- Wederell,E.D. et al. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
- Wingender,E. (2008) The transfac project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinformatics*, **9**, 326–332.
- Zhao,Y. et al. (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**, e1000590.
- Zhou,Q. and Wong,W.H. (2004) CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.