Simultaneously Learning DNA Motif along with Its Position and Sequence Rank Preferences through EM Algorithm

ZhiZhuo Zhang,[1] Cheng Wei Chang,[2] Willy Hugo,[1] Edwin Cheung,[2] Wing-Kin Sung,[1,2] ⋆

[1]National University of Singapore & [2]Genome Institute of Singapore
[1]{zhizhuo, hugowill, ksung}@comp.nus.edu.sg & [2]{changcw99, cheungcwe}@gis.a-star.edu.sg

Abstract. Although de novo motifs can be discovered through mining over-represented sequence patterns, this approach misses some real motifs and generates many false positives. To improve accuracy, one solution is to consider some additional binding features (i.e. position preference and sequence rank preference). This information is usually required from the user. This paper presents a *de novo* motif discovery algorithm called SEME which uses pure probabilistic mixture model to model the motif's binding features and uses expectation maximization (EM) algorithms to simultaneously learn the sequence motif, position and sequence rank preferences without asking for any prior knowledge from the user. SEME is both efficient and accurate thanks to two important techniques: the variable motif length extension and importance sampling. Using 75 large scale synthetic datasets, 32 metazoan compendium benchmark datasets and 164 ChIP-Seq libraries, we demonstrated the superior performance of SEME over existing programs in finding transcription factor (TF) binding sites. SEME is further applied to a more difficult problem of finding the co-regulated TF (co-TF) motifs in 15 ChIP-Seq libraries. It identified significantly more correct co-TF motifs and, at the same time, predicted co-TF motifs with better matching to the known motifs. Finally, we show that the learned position and sequence rank preferences of each co-TF reveals potential interaction mechanisms between the primary TF and the co-TF within these sites. Some of these findings were further validated by the ChIP-Seq experiments of the co-TFs.

Key words: Motif Finding, Expectation Maximization, Importance Sampling, Binding Preference

Program and Supplementary online: `http://biogpu.ddns.comp.nus.edu.sg/~chipseq/SEME/`

⋆ This work was supported in part by the MOEs AcRF Tier 2 funding R-252-000-444-112.

1 Introduction

Motif finding is an important classical bioinformatics problem. Given a set of biopolymer sequences (DNA or proteins), the motif finding problem aims to identify the recurring patterns (motifs) in them. Motif finders can generally be classified into two approaches: combinatorial searching and probabilistic modeling. The former approach enumerates the consensus patterns which are over-represented in the set of sequences. Using indexing data structures (e.g suffix tree [22], suffix array [16], and hash table [23]), it can efficiently identify short consensus motifs. Weeder [22], Trawler [7], YMF [29], DREME [2] are a few examples representing this line of approach. On the other hand, (most) probabilistic modeling approaches represent motifs using position weighted matrices (PWM) [28]. A PWM represents a length-w DNA motif as a $4 \times w$ matrix. It is more informative than a consensus pattern but it is also more difficult to compute. The high computational complexity of probabilistic modeling approach is a formidable bottleneck for its practical use. Expectation maximization [3] and Gibbs sampling [25] are the two most common approaches to find a PWM but they require long running time. Recently, some hybrid algorithms combined both approaches to get a good balance between accuracy and efficiency (e.g. [27], [17] and [15]).

By only examining the over-representation of sequence patterns, the previous generation motif finders often miss some real motifs and generate many false positives. On the other hand, additional information for the input sequences are found to be helpful to improve motif finding. For example, some transcription factor (TF) binding motifs (e.g. TATA-box) are localized to certain intervals with respect to the transcription start site(s) (TSS) of the gene. In this case, the position information can help to filter spurious sites. In protein binding microarray (PBM) [4] data, the *de Bruijn* sequences are ranked by their binding affinities and we expect the correct motif occurs in the high ranking sequences; such data has a rank preference. In the ChIP-Seq data [30], the ChIPed TF's motif (ChIPed TF is the TF pulled down in the ChIP experiment) prefers to occur in sequences with high ChIP intensity and also near the ChIP peak summits (thus having both position and rank preference). Hence, if we know the position preference and the sequence rank preference of the TF motifs in the input sequences, we can improve motif finding. In fact, many existing motif finders already utilize such additional information. MDscan [18] only considers high ranking sequences to generate its initial candidate motifs. Other programs allow users to specify the prior distribution of position preference or sequence rank preference [3, 22, 2, 15, 12] or add such preferences as a prior knowledge component in their scoring functions [5, 21, 17, 13, 9]. However, the users may not know the correct prior(s) to begin with. Even worse, different motifs may have different preferences. For example, in ChIP-Seq experiments, some motifs prefer to occur in high ranking sequences and at the center of the ChIP peak summit while others do not.

To resolve such problem, we propose a novel motif finding algorithm called SEME (**S**ampling with **E**xpectation maximization for **M**otif **E**licitation). SEME assumes the set of input sequences is a mixture of two models: a motif model and a background model. It uses EM-based algorithm to learn the motif pattern (PWM), position preference and sequence rank preference at the same time; instead of asking users to provide them as inputs. SEME does not assume the presence of both preferences but automatically detect them during the motif refinement process through statistical significance testing. We also observe that EM algorithms are generally slow in analyzing large scale high throughput data. Speeding up EM using suffix tree was recently proposed [24] but the technique cannot be applied when one wants to also learn the position and sequence rank preferences. To improve the efficiency, SEME developed two EM procedures. The two EM procedures are based on the observations that the correct motifs usually have a short conserved pattern in it and majority of the sites in the input sequences are non-motif sites. For the first EM procedure, called extending EM (EEM), starts by finding all over-represented short l-mers and then attempts to include and refine the flanking positions around the l-mers within the EM iterations. This way, SEME recovers the proper motif length within a single run thus saving a substantial amount of time by avoiding multiple runs with different motif length (as done in many existing motif finders [3, 22, 17, 15, 12]). The second EM procedure, called the re-sampling EM (REM), tries to further refine the motif produced by EEM. It is based on a theorem similar to importance sampling [11], which stated that the motif parameters can be learned unbiasedly using a biased subsampling. By this principle, we can sample more sites which are similar to the EEM's motif and less sites from the background. This way, REM is able to learn the correct motifs using significantly less background sites. In our implementation, REM is capable to produce the correct TF motifs using approximately 1% of the sites normally considered in a normal EM procedure.

Using 75 large scale synthetic datasets, we show that SEME is better both in terms of accuracy and running time when compared to MEME, a popular EM-based motif finding program [3]. We found that MEME is unable to find motifs with gap regions while SEME's EEM procedure can successfully extend the motifs to include them. In the real experimental datasets, we perform comparison using 32 metazoan compendium datasets and 164 ChIP-Seq libraries. SEME consistently outperformed seven existing motif finding programs that we compare with. In general, we found that SEME not only finds more TF motifs but also gives more accurate results (as evaluated using either PWM divergence, AUC score or STAMP's p-value [20]). When we compare the programs to find co-regulated TF (co-TF) motifs[1] from 15 ChIP-Seq datasets, the superior performance of SEME is more pronounced. We propose that SEME's ability to learn the underlying motif binding preference is crucial in its performance. We further confirmed the correctness of the position and sequence rank preference of the coTF motifs learned by SEME on three ChIP-Seq datasets. The actual ChIP-Seq data of the predicted co-TFs clearly shows that SEME managed to infer the correct preferences. We also show that such preferences provide biological insights on the mechanism of the ChIPed TF–coTF interactions.

[1] Other TFs which bind nearby and function together with the ChIPed TF

2 SEME Algorithm

SEME uses a probabilistic framework known as the two component mixture model (TCM) which is first proposed by MEME [3]. It assumes that the observed data is generated by two independent components: a motif model and a background model. Given an ordered list X of equal length DNA sequences, each site X_i in X is associated with a DNA sequence $X_i^{(seq)}$ and two integers: the rank of the sequence containing X_i ($X_i^{(rank)}$) and the position of the site X_i in the sequence ($X_i^{(pos)}$). We use an indicator variable Z_i to indicate if X_i is from the motif model or the background model, i.e., denote $Z_i = 1$ if X_i is from the motif model and 0 otherwise. The likelihood of an observed site X_i is written as:

$$Pr(X_i) = Pr(X_i|Z_i = 1)Pr(Z_i = 1) + Pr(X_i|Z_i = 0)Pr(Z_i = 0) \tag{1}$$

We use a naïve bayesian approach to combine three types of preferences (sequence, position, rank):

$$Pr(X_i|Z_i) = Pr(X_i^{(seq)}|Z_i)Pr(X_i^{(pos)}|Z_i)Pr(X_i^{(rank)}|Z_i) \tag{2}$$

For sequence preference, we model the motif site sequence with a position weight matrix (PWM) $\mathit{\Theta}$, and the background sequence with a 0-order markov model θ_0. $\mathit{\Theta}$ is a $4 \times w$ matrix where $\mathit{\Theta}_{j,a}$ is the probability that the nucleotide a occurs at position j. For any length-w sequence X_i, the probability that X_i is generated from the motif model and the background model are as follows.

$$Pr(X_i^{(seq)}|Z_i = 1) = Pr(X_i^{(seq)}|\mathit{\Theta}) = \prod_{j=1}^{w} \mathit{\Theta}_{j,X_{i,j}^{(seq)}} \tag{3}$$

$$Pr(X_i^{(seq)}|Z_i = 0) = Pr(X_i^{(seq)}|\overrightarrow{\theta_0}) = \prod_{j=1}^{w} \theta_{0,X_{i,j}^{(seq)}} \tag{4}$$

where $X_{i,j}^{(seq)}$ is the nucleotide in the j-th position of the site X_i.

The position and sequence rank preferences are modeled using multinomial distributions. The position preference models the preference of the motif site to certain positions. Similarly, the sequence rank preference tries to model if the motif site prefers the sequences with certain range of ranks assuming input sequences are ordered by some criteria. To this end, we discretize both the positions and sequence ranks into K bins. The probability a binding site occurs at the k-th position bin is denoted as α_k, for $k = 1, \ldots, K$, while the background distribution is assumed to be uniform. Precisely, for every X_i, we have

$$Pr(X_i^{(pos)} = k|Z_i = 1) = \alpha_k; Pr(X_i^{(pos)} = k|Z_i = 0) = \frac{1}{K}$$

Similarly for sequence rank preferences, the probability a motif site occurs at the k-th sequence rank bin is denoted as β_k and,

$$Pr(X_i^{(rank)} = k|Z_i = 1) = \beta_k; Pr(X_i^{(rank)} = k|Z_i = 0) = \frac{1}{K}$$

Let $Pr(Z_i = 1)$ be λ. The parameters of the mixture model in SEME are $\Phi = (\lambda,\Theta,\theta_0,\{\alpha_1, ..., \alpha_K\}, \{\beta_1, ..., \beta_K\})$. We estimated these parameters by maximizing the log likelihood $\sum_{i=1}^{n} \log Pr(X_i|\Phi)$ using expectation maximization (EM) procedure. Given a set of sequences X, the classical EM algorithm is as follows. It first gives an initial guess of the parameter $\Phi^{(0)}$. Then, it iteratively performs two steps: E-step and M-step. Given $\Phi^{(t-1)}$, the t-th iteration of the E-step estimates $Z_i^{(t)} = Pr(Z_i|\Phi^{(t-1)}, X)$. Then, given $Z_i^{(t)}$, the t-th iteration of the M-step computes $\Phi^{(t)} = \arg\max_{\Phi} \sum_{i=1}^{n} \log Pr(X_i, Z_i^{(t)}|\Phi)$. The E-step and M-step are iterated until $\Phi^{(t)}$ is converged.

In this work, we developed four phases in the SEME pipeline (see Figure 1). To search for a good starting point, SEME first enumerates a set of over-represented short l-mers (phase 1) and extends each short l-mer to a proper length PWM motif by the extending EM (EEM) procedure (phase 2). The PWM reported by the extending EM procedure will approximate the true motif when its starting l-mer captures the conserved region of the motif. To further refine EEM's PWM motif, SEME applies the re-sampling EM (REM) procedure (phase 3). It is an importance sampling version of the classical EM algorithm which greatly speed up the EM iterations. Finally, the refined PWM motifs are scored and filtered for redundancies (phase 4). Below, we briefly describe these four phases (see the Supplementary section 3 for details).

```
SEME Pipeline
Require: A set of input DNA sequences (fasta format)
Ensure: Return a set of non-redundant motifs $M$
Identify a set of over-represented short $l$-mers $Q$ in $X$ ;
for every $q \in Q$ do
    Extend $q$ to full length PWM motif $\Theta^{EEM}$ using extending EM procedure;
    Refine $\Theta^{EEM}$ to a more accurate PWM $\Theta^{REM}$ using re-sampling EM procedure;
    Add $\Theta^{REM}$ to candidate motif set $M$;
end for
Compute empirical scores (AUC or Z-score) for all the PWMs in $M$;
Sort all the PWMs in $M$ and filter lower scoring redundant PWMs in $M$;
```

Fig. 1. Algorithm description for SEME Pipeline.

Identifying Over-represented *l*-mers. In the first phase, SEME computes the frequencies of all short l-mers ($l = 5$ by default) in the input sequences and the background. If no background sequences are provided, a 1^{st}-order markov model will be learned from the input sequences as the background model. We output all l-mers whose frequencies in the input sequences are higher than in the background to the next phase.

Extending EM. For each l-mer q obtained from the first phase, the aim of the extending EM (EEM) procedure is to extend the l-mer to a longer motif which maximizes the likelihood of observing the sites with q. In this phase, the EEM procedure only needs to study the sites containing the l-mer q, i.e., $Y = \{X_i \in X \mid X_i^{(seq)}$ matches $(N)^{w-|q|}q(N)^{w-|q|}\}$ ("N" is a wild char), and w is the maximum length of a motif. For example, if l-mer is "GGTCA" and the predefined longest possible motif length is 10, then EEM considers only those sites in X matching the string pattern "NNNNNGGTCANNNNN". The EEM procedure first initializes the parameters $\Phi^{(0)} = (\lambda^{(0)}, \Theta^{(0)}, \theta_0^{(0)}, \{\alpha_1^{(0)}, \ldots, \alpha_K^{(0)}\}, \{\beta_1^{(0)}, \ldots, \beta_K^{(0)}\})$ where $\lambda^{(0)}$ is the estimated percentage of Y not from background, $\Theta^{(0)}$ is PWM representing q, $\theta_0^{(0)}$ is the frequency of A,C,G,T in Y excluding the conserved l-mer q, $\alpha_i^{(0)} = \beta_i^{(0)} = 1/K$ for $i = 1, \ldots, K$ (uniform distribution). Then, it performs E-step (expectation) and M-step (maximization) iteratively.

```
Extending EM
Require: $l$-mer $q$, maximum allowed motif length $w$ , input sequences $X$
Ensure: final extended PWM $\Theta^{(t)}$
$Y := \{X_i \in X \mid X_i^{(seq)}$ matches $(N)^{w-|q|}q(N)^{w-|q|}\}$;
Initialize the parameter set $\Phi^{(0)}$ for the mixture model;
t:=1;
repeat
    E-step: $\forall X_i \in Y$, compute the expectation of $Z_i^{(t)}$ using the parameter set $\Phi^{(t-1)}$ in the last iteration;
    M-step: update the parameter set $\Phi^{(t)}$ by maximizing log likelihood $Pr(Y, Z^{(t)}|\Phi)$;
    if length of $\Theta^{(t)} < w$; then
        Find a position $j$ which maximizes the log likelihood increment in Equation5 and denote $J$ to be the corresponding nucleotide distribution of position $j$;
        if $J$ is significantly different from the background distribution $\overrightarrow{\theta_0}^{(t)}$ using Chi-square test; then
            Use $J$ as the distribution in position $j$ of PWM $\Theta^{(t)}$;
        end if
    end if
    t:=t+1;
until PWM $\Theta^{(t)}$ converges;
The columns representing the $l$-mer $q$ in $\Theta^{(t)}$ are diluted;
```

Fig. 2. Pseudocode for Extending EM procedure.

In each iteration of the M-step, the EEM procedure will also try to include one additional column into $\Theta^{(t)}$ if such extension improves the likelihood. Precisely, for each position $j = 1, \ldots, 2w - |q|$ not in $\Theta^{(t)}$, we show that the maximum increment of the log likelihood before and after including the position j is $G(j)$ where

$$G(j) = \sup_J \sum_{X_i \in Y} Z_i^{(t)} \log(\frac{Pr(X_{i,j}^{(seq)}|J)}{Pr(X_{i,j}^{(seq)}|\theta_0^{(t)})}) \tag{5}$$

where J is any probability distribution over the nucleotides {A,C,G,T}.

```
Re-sampling EM
Require: the extended PWM $\Theta^{(EEM)}$, sampling rate $\mu$, input sequences $X$
Ensure: Final refined PWM $\Theta^{(t)}$
Initialize the parameter set $\Phi^{(0)}$ for the mixture model;
$X_Q$ :=$\{X_i \in X \, | Q(X_i) = 1\}$ according to the probability $Pr(Q(X_i) = 1) = min\{4^w \mu Pr(X_i|\Theta^{(EEM)}), 1\}$;
t:=1;
repeat
    E-step: $\forall X_i \in X_Q$, compute $Z_i^{(t)}$ using the parameter set $\Phi^{(t-1)}$ in the last iteration;
    M-step: update $\Phi^{(t)}$ by maximizing the weighted log likelihood $\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i|\Phi)}{Pr(Q(X_i)=1)}$;
    if the position distribution of $\{Z_i^{(t)}\}$ is significantly different from uniform distribution then
        include position preference in the model;
    end if
    if sequence rank distribution of $\{Z_i^{(t)}\}$ is significantly different from uniform distribution then
        include sequence rank preference in the model;
    end if
    t:=t+1;
until $\Theta^{(t)}$ converge;
```

Fig. 3. Pseudocode for Re-sampling EM procedure.

While the length of $\Theta^{(t)}$ is less than w, we extend the PWM $\Theta^{(t)}$ to include position j which brings the largest $G(j)$. To avoid over-fitting, the selected column also has to be tested (Chi-square) significantly different from the background frequency θ_0. The EEM procedure ends when PWM Θ converges. Finally, the columns in Θ representing the l-mer q will be further diluted (by setting all $[1.0, 0.0, 0.0, 0.0]$ columns representing "A" to $[0.5, \frac{0.5}{3}, \frac{0.5}{3}, \frac{0.5}{3}]$—other nucleotides are handled similarly) before Θ is returned as the output of the EEM procedure. In Supplementary section 1.1, we confirmed that EEM consistently recovers the correct motif length.

Re-sampling EM. The EEM procedure identifies an approximate motif model $\Theta^{(EEM)}$ with a proper motif length. This motif can be further refined using the classical EM algorithm to improve accuracy. However, when the input data X is big, this step will be slow. Using the idea of importance sampling, we proposed the re-sampling EM (REM) procedure which reduces the running time by running EM algorithm on a subsample of the original data.

Let $Q(\cdot)$ be the sampler function,where $Q(X_i) = 1$ if the sequence X_i is sampled; and 0 otherwise. In the Supplementary Theorem 1 (Supplementary section 3.4), we show that the log likelihood function $\log Pr(X, Z|\Phi)$ can be unbiasedly approximated by

$$\sum_{X_i \in X} \log Pr(X_i, Z_i|\Phi) = E_{X_Q}[\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i|\Phi)}{Pr(Q(X_i) = 1)}] \qquad (6)$$

where each sampled site is weighted by a factor $\frac{1}{Pr(Q(X_i)=1)}$. The theorem implies that we need only run the EM algorithm on X_Q. Moreover, in the M-step of the original EM, instead of maximizing $\log Pr(X, Z|\Phi)$, we maximize $\sum_{X_i \in X_Q} \frac{\log Pr(X_i, Z_i|\Phi)}{Pr(Q(X_i)=1)}$.

Although Equation 6 is true for any arbitrary sampler function $Q(\cdot)$, running EM using different $Q(.)$ yields different sampling efficiencies. For example, we can use a uniform random sampler, i.e., $Pr(Q(X_i) = 1) = \mu$ for every $X_i \in X$, where $\mu \in [0, 1]$ is the sampling ratio. This function is expected to only cover $100\mu\%$ of the correct motif sites from X which prohibits the use of small μ. In our work, we employ the idea of importance sampling. Our sampling function $Q(.)$ satisfies $Pr(Q(X_i) = 1) = min\{4^w \mu Pr(X_i|\Theta^{(EEM)}), 1\}$, where w is motif length. This sampling function gives higher probabilities for the sites that are more consistent to $\Theta^{(EEM)}$ thus it is expected to sample more from the correct motif sites and less from the background. (assuming $\Theta^{(EEM)}$ models more of the correct motif site signal than the background signal). This strategy is useful since we avoid most of the background sites in X. In fact, our simulation reveals that the REM procedure can achieve nearly 60% recall rate (of the correct motif sites) at the sampling ratio as small as $2^{-10}(\approx 0.001)$ and 90% recall rate at the sampling ratio of $2^{-5}(\approx 0.031)$ (see Supplementary section 1.2). We choose a default sampling ratio of 0.01 in all experiments of this paper.

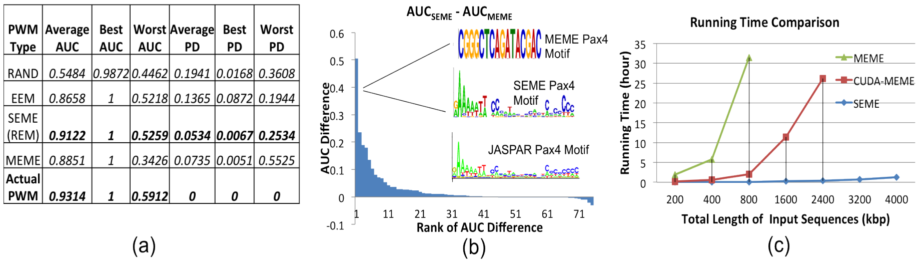
The position and sequence rank preferences are assumed to be non-existent at the beginning of the REM iterations (i.e., $Pr(X_i|Z_i) = Pr(X_i^{(seq)}|Z_i)$). The position and/or sequence rank preferences are considered only when the position and/or sequence rank distributions of $\{Z_i^{(t)}\}$ are significantly different from the uniform distribution (by Chi-square test). This strategy allows SEME to tell users which preference is really important for the predicted motif. Figure 3 is the pseudocode for this procedure.

Sorting and Redundancy Filtering. The PWMs output by REM are evaluated and sorted by empirical ROC-AUC (the area under the receiver-operator characteristic curve) or over-representation Z-score (representing the motif abundance) with the input data (details on each scoring are in the Supplementary section 4). The first score is preferred for the case when input sequences are short and most of sequences contain at least one motif site (e.g., ChIPed TF motif finding); for the other cases, we suggest to use the Z-score. We eliminate redundant PWMs from the sorted list as follows. When the sites of a PWM motif overlap with those of another PWM motif by more than 10%, we will treat the PWM motif with the lower score as redundant and remove it.

3 Result

3.1 Profiling two novel EM procedures

SEME significantly outperforms MEME in recovering the planted PWM. To analyze SEME's performance, we extract all seventy-five motifs of



PWM Type	Average AUC	Best AUC	Worst AUC	Average PD	Best PD	Worst PD
RAND	*0.5484*	*0.9872*	*0.4462*	*0.1941*	*0.0168*	*0.3608*
EEM	*0.8658*	*1*	*0.5218*	*0.1365*	*0.0872*	*0.1944*
SEME (REM)	***0.9122***	***1***	***0.5259***	***0.0534***	***0.0067***	***0.2534***
MEME	*0.8851*	*1*	*0.3426*	*0.0735*	*0.0051*	*0.5525*
Actual PWM	***0.9314***	***1***	***0.5912***	***0***	***0***	***0***

(a)

(b)

(c)

Fig. 4. The empirical performance of SEME on synthetic datasets. (a) The accuracy of SEME's PWM (both EEM step's (unrefined) PWM and the REM's (final) PWM are listed). We quantify accuracy using the commonly used Area-Under-ROC Curve (AUC) score and PWM divergence (PD) . We show that EEM's predicted PWM is already significantly stronger than random; indicating the goodness of EEM's PWM as starting point for the subsequent REM step. The scores also show that SEME's PWMs are significantly better when compared to MEME's. (b) Based on the performances of SEME and MEME on the Pax4 motif dataset, we observe that MEME has serious difficulties in mining PWMs with long gap region within them. (c) The running time of SEME is shown against increasing input size. We observe that CUDA-MEME, the GPU enabled version of MEME, still runs slower than SEME running on normal CPU (it takes 1 day to handle ≈ 6000 sequences while SEME takes around 1 hour for 10000 sequences).

lengths > 9 in JASPAR[31] vertebrate core database. For each such motif, we generated a training dataset of 1000 random sequences of length 400bp where 500 of them have a motif instance. The instances are planted uniformly across all positions and sequences.

For each dataset, we run SEME (EEM only), SEME (EEM + REM), and MEME (the classical EM-based motif finder) and obtain the top 5 predicted PWMs from each program. To test the goodness of the predicted PWMs, we compare the PWM divergence between the predicted PWMs and the actual planted PWMs. We also generate independent testing sequences with length 400bp (1000 positive sequences with one implanted motif site, 1000 negative sequences without motif site), and compute the ROC-AUC value for each predicted PWM. Figure 4(a) shows the comparison result. As expected, the random PWMs have the worst AUC values while the actual planted PWMs have the best AUC values. EEM's predicted PWMs have significantly better discriminative capability (AUC) and similarity (less PWM divergence) to the actual planted PWM as compared to random PWMs. This indicates that EEM's PWMs are good starting points for the subsequent REM procedure. REM's predicted PWMs further improve the AUC score and are similar to the actual planted PWM (as indicated by the small PWM divergence).

Figure 4(a) also shows that SEME outperforms MEME. In fact, SEME is better than MEME in 42 out of 75 experiments (the cases with positive AUC differences in Figure 4(b)). The cases where SEME performed worse have relatively small AUC score differences (less than 0.04). We examined the Pax4 dataset in which SEME gains the highest improvement against MEME. The implanted JASPAR Pax4 motif is a diverged PWM of length 30. SEME successfully extended and recovered the full Pax4 motif; thanks to the ability of its EEM procedure to handle long gaps in its extension step. In contrast, MEME failed to model the long gaps due to their starting point finding procedure which assumes that all of the PWM positions are equally important.

SEME is more suitable in handling large scale data. We further generated 7 large datasets to observe the capability of SEME in handling large scale data. Each dataset consists of different number of sequences (from 500 to 10000, each of length 400bp). Figure 4(c) showed that the original MEME program cannot process more than 2000 sequences within one day, hence we also used the GPU-accelerated version of MEME, CUDA-MEME[19] (run on two Intel X5670 CPUs and two Fermi M2050 GPUs with 48GB RAM). SEME was run as normal CPU program. SEME is still around 60 times faster than CUDA-MEME which runs on the highly parallelized GPU system. In addition, SEME can process up to 10000 sequences (a typical dataset size for ChIP-Seq experiments) in 1 hour while the CUDA-MEME took more than one day to process 6000 sequences.

3.2 Comparing TF motif finding in large scale real datasets

We compare the performance of SEME with other existing motif-finding programs on two large scale TF binding site data. We also study the ability of SEME in uncovering the hidden position and/or sequence rank preferences in the input dataset when they are present.

The Metazoan Compendium datasets. The first benchmark is a metazoan compendium dataset published by Linhart et.al[17]; consisting of 32 datasets based on experimental data from microarray, ChIP-chip, ChIP-DSL, and DamID as well as Gene Ontology data[1]. A list of the promoter sequences of many target genes (1000bp upstream and 200bp downstream the Transcription Start Site (TSS)) are used as the positive input for each motif-finding program and promoter sequences of other non-target genes are used as background sequences. The performance of six existing motif-finding programs, namely AlignACE [25], MEME [3], YMF [29], Trawler [7], Weeder [22], and Amadeus [17], were compared in the original benchmark study [17]. Each program's predicted PWMs are evaluated by the PWM divergence. Only PWMs with medium and strong matching with the known motifs (PWM divergence < 0.18) are considered to be successfully detected [17].

The result of this comparison is shown in Figure 5(a). We find that SEME successfully detected the correct motifs in 21 datasets whereas the second best program, Amadeus, succeeded in 18. Weeder and Trawler found correct PWMs in 11 and 12 datasets, respectively. SEME also found more accurate motifs than the rest; it found 12 motifs with PWM divergence < 0.12. SEME further detected a significant position preference for the correct motifs for many datasets in this benchmark: most of them tend to bind nearer to the TSS position (see the Supplementary section 1.4 for details).

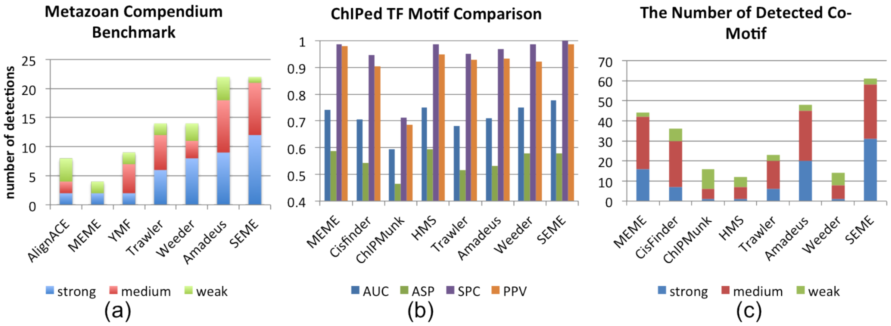


Fig. 5. The performance of SEME compared to existing motif finding programs from large scale real data (a) Comparison result on the metazoan compendium datasets. Four PWM motifs returned by each motif finding program are then compared to the known Transfac motifs using PWM divergence (PD) (as in [17]) and further classified into three matching categories (strong, medium, weak) corresponding to different PD cut-offs (0.12,0.18,0.24). (b) Comparison result on 164 ChIP-Seq libraries over four different measurements? AUC, PPV (Positive Predictive Value), ASP (Average Site Performance) and SPC (Specificity). The result shows that most motif finders perform similarly well in detecting ChIPed TF (but SEME is consistently better than all of them). (c) Comparison result for Co-TF motif finding on 15 ChIP-Seq libraries. The quality of reported PWMs is classified into three categories (strong, medium, weak) corresponding to different STAMP p-value cut-offs (0.0001, 0.01, 0.05). SEME reported the most number of co-TF's motif which match the known PWM with STAMP p-value $\leq$ 0.0001 (strong match, blue bar). Overall, SEME also found the most number of co-TF motif (61) as compared to the second best program, Amadeus (48).

ChIP-Seq experimental datasets: Discovery of the ChIPed TF motif from ChIP-Seq data. The second benchmark is a collection of large scale ChIP-Seq experimental data which consists of 164 published ChIP-Seq libraries from the ENCODE project[8] and our lab over different cell-lines and TFs[6, 33, 14]. ChIP-Seq usually reports more than 10000 target sequences with narrower target regions (100bp). We compute the Area Under ROC Curve , Positive Predictive Value, Average Site Performance and Specificity scores of each program's predicted PWM. The formula for the above scores are given in the Supplementary section 4. From each library, the 100bp sequences around the top 10000 ChIP-Seq peaks were selected (sorted by ChIP intensity) as our input data. MEME and Weeder only use the top 2000 peaks due to their long running time. Peaks with odd numbered ranks were used for training while the even numbered peaks were used as positive testing data. The negative dataset is generated a 1^{st}-order Markov model trained using the same number of 100bp random sequences extracted from the regions 1000bp away from the ChIP-Seq peaks.

We compared SEME with 7 popular *de novo* motif finding programs for ChIP data: MEME, Weeder, Cisfinder, Trawler, Amadeus, ChIPMunk and HMS. Each program's top 5 motifs are evaluated using the four statistics measurements on the test data. For each scoring, the best of the 5 motifs will be used to represent

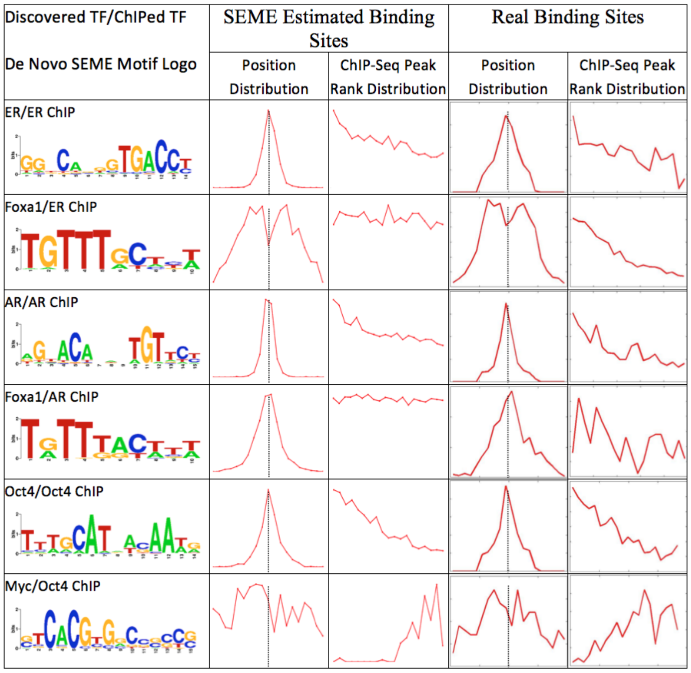


Fig. 6. Automatic learning of the position and sequence rank preference from the input data. Instead of requiring the user to input the expected co-TF motif preference distribution (position and/or sequence rank distribution), SEME learns such distributions directly from the input data. We show that most of the time, SEME can learn the correct distributions of each TF (as compared to real binding sites distribution in the rightmost column, defined by the ChIP-Seq and the known PWM of the TF). For position distribution, the x-axis is +/-200bp from ChIP-seq peak summit (the black dash line), and the y-axis is the fraction of binding sites in a given position. For rank distribution, the x-axis is the rank of ChIP-seq peak (left : high ChIP intensity, right : low ChIP intensity), and the y-axis is the fraction of binding sites in a given rank. The ChIP-seq peak rank distributions (MCF7 ER ChIP, LNCaP AR ChIP) of FoxA1 and the position distribution of Myc are tested to be insignificant by SEME.

the performance of a program. Figure 5(b) shows the average performances of the motif finders. Again, we find that SEME is consistently better than all other programs (1^{st} rank in Area Under ROC Curve , Positive Predictive Value and Specificity, and 3^{rd} rank in Average Site Performance).

Discovery of co-TF motifs from ChIP-Seq data. We note that most motif finders show good performance in finding the ChIPed TF motifs. This is expected since the ChIPed TFs are highly enriched[33]. Compared to finding ChIPed TF motifs in ChIP-Seq datasets, the problem of finding co-TF motifs in the ChIP-Seq datasets is much more challenging. The co-TF motif instances are less abundant and most are not located exactly at the ChIP-Seq peaks. Nevertheless, finding the co-TF(s) could potentially uncover previously unknown TF-TF interaction.

For co-TF motif comparison, we used 15 ChIP-Seq libraries whose co-TFs have been characterized (the list of co-TFs for each ChIP-Seq is in Supplementary section 2.5). We extracted 400bp sequences around the ChIP-Seq peaks and compared the top 20 *de novo* motifs of each program to the known co-TF motifs in the JASPAR and Transfac database; we cannot use the previous statistical measurements since co-TFs may not occur in all ChIP-Seq peaks. Furthermore, the ChIPed TF binding sites need to be masked before we start the co-TF motif finding. SEME and ChIPMunk can do this automatically and, for other programs without auto-masking mode, the input sequences were masked by the top 2 motifs reported from their ChIPed motif finding results.

STAMP program[20] was used to compute the p-value of the match between a predicted co-TF motif against the known co-TF motif. STAMP p-value provides a better match measurement compared to PWM divergence since it removes the motif length bias [20]. We separated the p-value of the PWM matching into three significance levels: (1) weak match ($0.05 \geq$ p-value > 0.01), (2) medium match ($0.01 \geq$ p-value > 0.0001) and (3) strong match (p-value ≤ 0.0001). Figure 5(c) shows the performances for different motif-finding programs in finding the co-TF motifs from the 15 datasets. SEME recovered 61 known co-TF motifs; compared to Amadeus and MEME which find 48 and 44 co-TF motifs, respectively. 31 out of the 61 co-TF motifs of SEME belong to the strong match category (Amadeus only found 20) and another 27 are in the medium match category. This indicates that SEME's predicted co-TF PWMs are highly accurate.

To study the biological significance of the learnt preferences, we further study the output of three datasets, involving the ER, AR, FoxA1, Oct4 and c-Myc TFs, in details (see Figure 6). The real binding site of each TF is defined to be the site around +/-100bp around the TF's ChIP-Seq peak whose known PWM score is better than a cutoff that yields FDR$= 0.01$. If multiple matches occur, only the best scoring site is chosen. Comparison between SEME's learnt distributions (Figure 6, middle columns) and the real binding site distributions (Figure 6, rightmost columns) indicates that SEME is able to learn the correct co-TF position and sequence rank preferences. We also found that the motif positions of FoxA1, a known co-TF of ER, is not enriched exactly at the ER ChIP-Seq peak in the MCF7 data; instead it is found in the flanking regions near the ER peaks. Interestingly, in the LnCAP AR ChIP-Seq dataset (FoxA1 is also a known co-TF of AR), we found that FoxA1 binds very closely to AR—it is enriched at the AR ChIP-Seq peak summits. This observation is consistent with the previous report that FoxA1 can physically interact with AR [10]. This observation also indicates the different roles FoxA1 assumes when working with AR and ER[26]. In the ChIP-Seq data of Oct4 from mouse's ES cell, SEME found the motif of c-Myc enriched within Oct4's low intensity peaks regions. We conjectured that, in these regions, Oct4 indirectly bind the DNA through c-Myc (hence explaining the ChIP-Seq's low intensity). An earlier report showed that Oct4, along with Sox2, Nanog, and Stat3 form an enhancer module while c-Myc along with n-Myc, E2F1 and Zfx form a promoter module in the ES cell[6]. In fact, interaction between these enhancer and promotor modules had also been reported previously[32].

These examples indicate that the position and sequence rank distribution learnt by SEME are reasonably accurate and users could use them to infer the nature of the interaction between the ChIPed TF and the co-TF(s). In this manner, SEME can be used to generate biological hypothesis for further experimental validations. Moreover, the highly diverse preferences that we observe highlight the difficulty for users to provide the correct prior in the first place.

4 Conclusion

This paper developed a novel algorithm called SEME for mining motifs using mixture model and EM algorithm. We presented three important contributions: (1) automatic detection and learning of the position and sequence rank preferences of a candidate motif. (2) ability to estimate the correct TF motif length (with possible gaps within) and (3) using importance sampling for efficiency while still able to estimate the EM parameters unbiasedly. As a result, we showed that SEME is substantially better, both in terms of accuracy and efficiency, compared to the existing motif finding programs.

Moreover, in the task of finding co-TF motif in the ChIP-Seq data, SEME not only report more accurate co-TF motifs than other programs but also correctly estimated the position and sequence rank distribution of each co-TF's motif. We showed that such information provides useful insights on the interaction between the ChIPed TF and the predicted co-TFs. SEME does have a few limitations. Firstly, it assumes that the target motif contains conserved 5-mer region. In cases without such 5-mer, SEME also allows user to provide custom seeds. Secondly, SEME is more suitable for large scale input ($\geq$ 100 sequences) since it needs enough samples to determine whether we should do extension (EEM) or include additional binding preferences (REM).

References

1. M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
2. T.L. Bailey. Dreme: Motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653, 2011.
3. T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, volume 2, pages 28–36, 1994.
4. M.F. Berger and M.L. Bulyk. Protein binding microarrays (pbms) for rapid, high-throughput characterization of the sequence specificities of dna binding proteins. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA-*, 338:245, 2006.
5. X. Chen, T.R. Hughes, and Q. Morris. Rankmotif++: a motif-search algorithm that accounts for relative ranks of k-mers in binding transcription factors. *Bioinformatics*, 23(13):i72, 2007.

6. X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.
7. L. Ettwiller, B. Paten, M. Ramialison, E. Birney, and J. Wittbrodt. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods*, 4(7):563–565, 2007.
8. G.M. Euskirchen, J.S. Rozowsky, C.L. Wei, W.H. Lee, Z.D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M.B. Gerstein, et al. Mapping of transcription factor binding regions in mammalian cells by chip: comparison of array-and sequencing-based technologies. *Genome research*, 17(6):898, 2007.
9. M.C. Frith, U. Hansen, J.L. Spouge, and Z. Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32(1):189, 2004.
10. N. Gao, J. Zhang, M.A. Rao, T.C. Case, J. Mirosevich, Y. Wang, R. Jin, A. Gupta, P.S. Rennie, and R.J. Matusik. The role of hepatocyte nuclear factor-3α (forkhead box a1) and androgen receptor in transcriptional regulation of prostatic genes. *Molecular Endocrinology*, 17(8):1484, 2003.
11. P.W. Glynn and D.L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, pages 1367–1392, 1989.
12. M. Hu, J. Yu, J.M.G. Taylor, A.M. Chinnaiyan, and Z.S. Qin. On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic acids research*, 38(7):2154, 2010.
13. J. Keilwagen, J. Grau, I.A. Paponov, S. Posch, M. Strickert, and I. Grosse. De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Computational Biology*, 7(2):e1001070, 2011.
14. S.L. Kong, G. Li, S.L. Loh, W.K. Sung, and E.T. Liu. Cellular reprogramming by the conjoint action of erα, foxa1, and gata3 to a ligand-inducible growth state. *Molecular Systems Biology*, 7(1), 2011.
15. IV Kulakovskiy, VA Boeva, AV Favorov, and VJ Makeev. Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, 26(20):2622, 2010.
16. T.W. Lam, K. Sadakane, W.K. Sung, and S.M. Yiu. A space and time efficient algorithm for constructing compressed suffix arrays. *Computing and Combinatorics*, pages 21–26, 2002.
17. C. Linhart, Y. Halperin, and R. Shamir. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Research*, 18(7):1180, 2008.
18. X.S. Liu, D.L. Brutlag, and J.S. Liu. An algorithm for finding protein–dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, 20(8):835–839, 2002.
19. Y. Liu, B. Schmidt, W. Liu, and D.L. Maskell. CUDA-MEME: Accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units. *Pattern Recognition Letters*, 2009.
20. S. Mahony, P.E. Auron, and P.V. Benos. Dna familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS computational biology*, 3(3):e61, 2007.
21. V. Narang, A. Mittal, and W.K. Sung. Localized motif discovery in gene regulatory sequences. *Bioinformatics*, 26(9):1152, 2010.
22. G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17(Suppl 1):S207–S214, 2001.
23. B. Raphael, L.T. Liu, and G. Varghese. A uniform projection method for motif discovery in dna sequences. *IEEE Transactions on Computational biology and Bioinformatics*, pages 91–94, 2004.

24. J.E. Reid and L. Wernisch. Steme: efficient em to find motifs in large data sets. *Nucleic acids research*, 39(18):e126–e126, 2011.
25. F.P. Roth1JT, J.D. Hughes, P.W. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16:939, 1998.
26. B. Sahu, M. Laakso, K. Ovaska, T. Mirtti, J. Lundin, A. Rannikko, A. Sankila, J.P. Turunen, M. Lundin, J. Konsti, et al. Dual role of foxa1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *The EMBO Journal*, 30(19):3962–3976, 2011.
27. A.A. Sharov and M.S.H. Ko. Exhaustive Search for Over-represented DNA Sequence Motifs with CisFinder. *DNA Research*, 2009.
28. S. Sinha. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14), 2006.
29. S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 344–354, 2000.
30. A. Valouev, D.S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5(9):829, 2008.
31. W.W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.
32. Q. Wu and H.H. Ng. Mark the transition: chromatin modifications and cell fate decision. *Cell Research*, 2011.
33. Z. Zhang, C.W. Chang, W.L. Goh, W.K. Sung, and E. Cheung. Centdist: discovery of co-associated factors by motif distribution. *Nucleic acids research*, 39(suppl 2):W391, 2011.