JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 11, 2018
© Mary Ann Liebert, Inc.
Pp. 1247–1256
DOI: 10.1089/cmb.2017.0230

Sequential Integration of Fuzzy Clustering and Expectation Maximization for Transcription Factor Binding Site Identification

ALI YOUSEFIAN-JAZI[1] and JINWOOK CHOI[2]

ABSTRACT

The identification of transcription factor binding sites (TFBSs) is a problem for which computational methods offer great hope. Thus far, the expectation maximization (EM) technique has been successfully utilized in finding TFBSs in DNA sequences, but inappropriate initialization of EM has yielded poor performance or running time scalability under a given data set. In this study, we used a sequential integration approach that defined the final solution as the set of solutions acquired from solving objectives in a cascade manner to integrate the fuzzy *C*-means and the EM approaches to DNA motif discovery. The new method is explained in detail and tested on the chromatin immunoprecipitation sequencing (ChIP-seq) data sets for different transcription factors (TFs) with various motif patterns. The proposed algorithm also suggests an efficient process for analyzing motif similarity to known motifs as well as finding a target motif. A comparison of results with those of the well-known motif-finding tool, MEME-ChIP, shows the advantages of our proposed framework over this existing tool. Experimental results show that we were able to find the true motifs for all TFs, and that the motifs found by our proposed algorithm were more similar to JASPAR-known motifs for the STAT1, GATA1, and JUN TFs than those found by MEME-ChIP.

Keywords: chromatin immunoprecipitation sequencing, expectation maximization, fuzzy C-means, motif discovery.

1. INTRODUCTION

Transcription factor binding sites (TFBSs), motifs, which generally reside in the noncoding regions of the DNA, play important roles in regulating the gene expression. One of the major challenges biologists face today is the identification of such binding sites. High-throughput assays generate significant amounts of information on the sequence preferences of transcription factors (TFs). Chromatin immunoprecipitation sequencing (ChIP-seq) technology (Johnson et al., 2007) identifies ‘‘peaks’’ of (direct or indirect) binding by the ChIP-ed factors. Moreover, ENCODE (Consortium, 2012) is a database that has

[1]Interdisciplinary Program in Bioengineering, Graduate School, Seoul National University, Seoul, Korea.
[2]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, Korea.

made binding site sequences available for hundreds of different TFs. Because ChIP-seq experiments are slow, expensive, and unable to find exact binding site locations, they cannot find patterns that are common across all of the positive binding sites to give insight into why TFs bind to those locations (Chen et al., 2010). On the contrary, good computational methods can potentially provide high-quality prediction of binding sites and reduce the time needed for experimental verification. Thus, there is a need for large-scale computational methods for accurate binding site classification and clear TFBS pattern representation.

The problem of de novo identification of TFBSs has been widely studied, and motif discovery algorithms have been categorized as either search based or probabilistic. Search-based algorithms infer motifs as consensus sequences, while algorithms based on probability infer motifs through position frequency matrices (PFMs), which specify the frequency of nucleotides for each position in the binding site. While the search-based methods are appropriate for finding short motifs, PFMs provide more information on the TF's binding specificity (Das and Dai, 2007). Bailey and Elkan (1994) presented a probabilistic motif discovery program (MEME) that uses the expectation maximization (EM) technique to infer PFMs. It uses a probabilistic model of the input data set and selects parameter values for the model that maximize the likelihood value. Since its initial introduction in 1994, it has gone through several versions.

The MEME-ChIP tool (Machanick and Bailey, 2011), which is built to analyze ChIP-seq peak regions, attempts to utilize MEME and DREME as two motif discovery algorithms and compare the results with a database of known TF motifs using the TOMTOM (Gupta et al., 2007) algorithm. MEME scales poorly with large data sets, and DREME is able to detect only very short motifs as a search-based algorithm. EXTREME (Quang and Xie, 2014) is also one of the latest versions that uses an online EM algorithm to improve MEME's performance in dealing with large data sets. EXTREME can discover motifs in a practical amount of time without discarding any sequences that could be a benefit for infrequent motif discovery. Recently, Ibrikci and Karabulut (2010) proposed fuzzy *C*-means (FCM) for motif discovery and tested their algorithm against DNA sequences extracted from the genome of *Saccharomyces cerevisiae* and *Escherichia coli*.

Today, with the development of new techniques and algorithms, new approaches to making algorithms, such as integrating existing algorithms, have become more popular. The *global objective* and *sequential* approach are the two frameworks exploiting existing integration approaches. In the global objective approach, a single global objective is defined using a combination of multiple objectives, and then the solution is defined as an equilibrium point of the single objective (Gennert and Yuille, 1988). However, sometimes merging the different objectives into a global objective and computing the actual weights are hard to achieve, problematic, or computationally expensive. In the sequential approach (Xi and Fenton, 1993), the final solution is defined as the set of solutions acquired from solving the objectives in a cascade manner. In this way, each module is permitted to influence other modules and take advantage of the benefits of each module, which means a multiple-objectives problem is transformed into a stage-wise single-objective problem.

The rationale behind the fuzzy clustering lies in the reality that an object or data point could be assigned to different classes by degree of membership (Asyali and Alci, 2004). On the contrary, it has been demonstrated that initializing the parameters of EM with different values has a significant effect on performance (Chen, 2006). In this article, the sequential integration of fuzzy clustering and EM techniques (SIFEM) has been proposed as a de novo motif discovery algorithm. SIFEM identifies novel sequence patterns (motifs) in the ChIP-seq regions that may represent direct or indirect TFBSs. We utilize a thorough quantitative methodology to examine the potential values and limitations of the proposed method. In motif enrichment analysis, we consider both standard and central enrichments of discovered motifs for finding the direct DNA-binding affinity of the ChIP-ed TF, and in motif identification we compare the ab initio motifs to known TF binding motifs as well. As a result, our algorithm is tested with five different ChIP-seq experiment data sets and had some improvements over the previous existing work. The source codes along with some ChIP-seq data used in this study can be found in the Supplementary Material.

2. METHODOLOGY

FCM and EM are the most popular clustering algorithms for solving several problems in computational biology (Do and Batzoglou, 2008; Jin and Wang, 2009). Both methods include iterative steps repeated until a satisfactory objective is reached. In FCM, a membership value for each input is first calculated, and then, the center of each cluster with the given vectors is updated by using the membership values as weighting factors in the next step. Likewise, the EM technique iteratively improves the parameters in the E-step and M-step. A sequential integration method is proposed here to take advantage of the benefits of both methods, casting motif identification as a stage-wise single-objective problem. Let $Y=(Y_1, Y_2,\ldots,Y_N)$ denote the data set of ChIP-seq peaks, where N is the number of sequences in the data set. The proposed method divides the data set into n (overlapping) subsequences of length W, then converts each subsequence to a $W\times 4$ binary array.

We refer to this new data set $X=(X_1, X_2,\ldots,X_n)$ as the input vectors, and $c=\{1,\ldots,C'\}$ represents the cluster centers. The membership of each input X_n to each cluster C' in $(t+1)$th iteration is given in Equation (1), and can be considered the posterior probability by adding the following constraint on the total membership values of an individual input vector (Asyali and Alci, 2004):

$$\xi_{ij}^{(t+1)} = \frac{\left(\frac{1}{(d_{ij}^{(t)})^2}\right)^{1/(q-1)}}{\sum_{j=1}^{C'}\left(\frac{1}{(d_{ij}^{(t)})^2}\right)^{1/(q-1)}}, \sum_{j=1}^{C'} \xi_{ij}=1, t \epsilon T= \{1, 2, \ldots, t_1\} \tag{1}$$

where d_{ij} corresponds to the distance between the input and the cluster center, and q denotes the fuzziness value. In this problem, dissimilarity between the DNA subsequences and position probability matrices (PPMs) is given as follows:

$$d_{ij}^{(t)} = 1 - \sum_{k\in \mathrm{A}} \sum_{s=1}^{w} I(x_{i,s}, k)\omega_{j,s,k}, A= (A, G, C, T) \tag{2}$$

and

$$I(k, a)= \begin{cases} 1 & if\ a = k \\ 0 & otherwise \end{cases}$$

where $\omega_{j,s,k}$ is probability of nucleotide k at position s of cluster j in the PPM. Once the membership matrix has been constructed, the PPMs can be updated according to Equation (3). Ibrikci and Karabulut (2010) updated each cluster center with a selected group of subsequences by selecting some elements from the membership matrix using a certain threshold rather than with all the subsequences, resulting in better predictive performance. Thus, the function $Sel(x_i)$ is utilized in updating the cluster centers:

$$\gamma_j^{(t+1)} = \frac{\sum_{i=1}^{n} (\xi_{ij}^{(t+1)})^q Sel(x_i)}{\sum_{i=1}^{n} (\xi_{ij}^{(t+1)})^q} \tag{3}$$

$$Sel(x_i)= \left\{ \begin{matrix} \xi_{ij} \in top(sort(\xi_{ij}), B) \to \xi_{ij} \\ otherwise \to 0 \end{matrix} \right\}$$

where B stands for the number of top subsequences to be considered and $top(Z,B)$ shows a subset of the first B samples of Z. In this algorithm, ξ and γ are iteratively improved until the change (Euclidean distance) in ξ_{ij} falls below a threshold ($\delta = 10^{-6}$) or reaches the maximum number of iterations ($t_1 = 100$).

In the next step, the expected number of times nucleotide k appears at position s in cluster j is calculated, and then, the position probability matrix elements are updated as follows:

$$c_{j,s,k}^{(t+1)} = \sum_{i=1}^{n} \xi_{ij}^{(t)} I(X_{i,s}, k), fork \in \mathrm{A}, s=1, 2, \ldots, W, t \epsilon T' = \{t_1, t_1 + 1, \ldots, t_2\} \tag{4}$$

$$f_{j,s,k}^{(t+1)} = \frac{c_{j,s,k}^{(t+1)} + \beta_k}{\sum_{k\in A} (c_{j,s,k}^{(t+1)} + \beta_k)} \tag{5}$$

where β_k are pseudocounts to prevent any nucleotide frequency $f_{j,s,k}$ from becoming 0. After calculating the PPMs, the membership values with $\theta_j = (f_{j,1}, f_{j,2},\ldots,f_{j,W})$ are updated.

$$\xi_{ij}^{(t+1)} = \frac{\Pr(X_i|\theta_j^{(t+1)})\lambda_j^{(t+1)}}{\sum_{l=1}^{C'} \Pr(X_i|\theta_l^{(t+1)})\lambda_l^{(t+1)}} \tag{6}$$

$$\lambda_j^{(t+1)} = \sum_{i=1}^{n} \frac{\xi_{ij}^{(t)}}{n} \tag{7}$$

where λ_j parameterizes the probability that any subsequence with the length of W is generated by the PPM of jth cluster.

SIFEM iteratively minimizes the following energy function, E, which sequentially integrates the objective functions for both FCM and EM. The above parameters are iteratively improved until $\|E^{(t+1)} - E^{(t)}\| < \delta$.

$$E^{(t)} = \sum_{j=1}^{C}\sum_{i=1}^{n} \xi_{ij}^{q} d_{ij}^{(t)} - \sum_{j=1}^{C}\sum_{i=1}^{n} \xi_{ij}^{(t)} \log(\Pr(x_i \mid \theta_j^{(t)})\lambda_j^{(t)} \tag{8}$$

where $t \in T = \{1,2,\ldots,t_1\}$ and $t \in T' = \{t_1, t_1 + 1, \ldots, t_2\}$ for the first and second summation. Moreover, $\|T \cap T'\| = 1$ and $\|.\|$ denotes the cardinality. The first and second terms are minimizing the total dissimilarity and log likelihood as the objective functions for FCM and EM, respectively. Table 1 shows the complete scheme of the algorithm.

In conclusion, the procedure and parameters for SIFEM can be summarized in the following steps. First, the number of clusters ($C' = 10$), membership values (ξ), and pattern length (W, with the values shown in Table 2) are initialized. Then, for each subsequence X_i ($i = 1,\ldots, n$), the algorithm calculates the membership values (ξ_{ij}, $j = 1,\ldots, C'$) with the specified fuzziness value ($q = 2$) and allows the inputs to be a member of any class with a certain possibility or ''degree of membership,'' a value between 0 and 1. In the next step, PPMs (γ_j) are updated with a selected group of subsequences based on a certain threshold. Performance improvements were seen by choosing a value of 350 for parameter K that was close to the number of TFBS instances residing within the data sets used in this study. Finally, the PPM elements for each cluster, $f_{j,s,k}$, and the probability of each cluster, (λ_j), are updated, and the above process repeated until the termination criteria are met.

3. RESULTS

3.1. Experimental setup

We used five TF ChIP-seq experiments from the ENCODE project (Consortium, 2012) to benchmark the performance of our model. The ChIP-Seq data included analyses of STAT1, GATA1, JUN, NFYA, and

Table 1. Pseudocode for SIFEM Algorithm

SIFEM algorithm

```
Input: $C'$, $\xi$, $X$
Output: $\gamma$
repeat
  if $t < t_1$
    updating membership values, $\xi_{ij}$, in view of Eq. (1)
    updating the distance measure, $d_{ij}$, in view of Eq. (2)
    updating PPM for each cluster, $\gamma_j$, in view of Eq. (3)
  else
    updating $c_{j,s,k}$ in view of Eq. (4)
    updating PPM elements, $f_{j,s,k}$, in view of Eq. (5)
    updating membership values, $\xi_{ij}$, in view of Eq. (6)
    updating $\lambda_j$ in view of Eq. (7)
  end
until $\|E^{(t+1)} - E^{(t)}\| < \delta$ or $t > t_2$
```

PPM, position probability matrix.

Table 2. Data Sets Utilized in This Study Extracted from the ENCODE Project and JASPAR Database

Cell line	*Factor*	*Pattern length*	*Data set size (sequences)*
K562	STAT1	15	1476
K562	GATA1	11	1704
K562	JUN	14	1708
K562	NFYA	18	7647
K562	REST	21	6406

REST performed on the K562 cell line, and peaks were already called and organized into BED files. The details of the data sets used for this study are given in Table 2. The motifs are detected as simple or complex patterns that include short or long motifs with highly or slightly conserved cores (Das and Dai, 2007). Therefore, the selection of motif patterns from a variety of TF families of different lengths within data sets of different sizes allows us to measure the performance of the proposed algorithm. We processed the data using the 101 bp sequences centered at the point source called for each ChIP-seq peak and using shuffled sequences with matching dinucleotide composition as the background set. This is the same data preprocessing procedure used in most motif discovery work (Quang and Xie, 2014; Alipanahi et al., 2015).

3.2. *Comparison methodologies*

As MEME-ChIP is one of the most popular packages used to analyze ChIP-seq ''peak regions'' for evaluation in recent motif discovery studies (Weirauch et al., 2013), we exploited it for the evaluation and comparison of our results. It has been a valuable tool still in use for the ongoing challenge of identifying regulatory elements and runs two complementary motif discovery algorithms on the input data: MEME and DREME. In this article, both quantitative and visual measures are utilized for result evaluation and comparison. *p* Values are used as quantitative measures to show PPM enrichment and similarity. Moreover, sequence logo and Two Sample Logo are also two visual measures analyzed in this section for evaluation of results.

In SIFEM, we trim positions for each discovered motif that show low levels of probability in the start and end of the PPM before calculating its *p* value. The *p* values are calculated using the AME package from the MEME-suite (Bailey et al., 2009). This motif enrichment significance measure is the number of motifs with the same width and number of occurrences that could have generated an equal or higher log likelihood ratio if the data set had been generated according to the background model. The usual way is to assume the binding affinity of the most highly ''enriched'' motif as the direct DNA-binding affinity of the ChIP-ed TF (Bailey and Elkan, 1994; Quang and Xie, 2014); this can be problematic in some cases, including instances of poor ChIP-seq data quality due to poor antibody performance or sample preparation issues, as well as highly enriched cofactor binding sites. ChIP-seq data sets are known to contain binding sites for unrelated TFs (Teytelman et al., 2013). Our main criterion for choosing a ChIP-based model was that its motif detectors should respond to binding sites for the TF of interest, and not respond to binding sites for unrelated or cofactor TFs. To deal with this challenge, we also used the central motif enrichment analysis package, CentriMo (Bailey and Machanick, 2012), to consider all discovered motifs and find candidates for the direct ChIP-ed TF binding motif.

Figure 1 shows the comparison results between both measures for motif enrichment. When we applied AME (McLeay and Bailey, 2010) to the (100 bp) STAT1 and REST ChIP-seq regions, the consensus sequence of the motifs found by SIFEM, HASTTTCRKTTYCYB and GCTGTCCAWGGTGCTGAA, respectively, rank second in terms of enrichment among the discovered motifs (Fig. 1a, e). However, GCTGTCCAWGGTGCTGAA has a worse central enrichment *p* value (4.2e-291) than the other motifs discovered by SIFEM in the REST ChIP-seq regions; it has a unimodal curve with a narrower region of maximum central enrichment ($w = 36$ bp). Moreover, it achieves a higher maximum site probability in the center of the ChIP-seq regions. In Figure 1d, the width of the region of maximum enrichment ($w = 202$) is quite large, suggesting that the resolution of the ChIP-seq peaks is not as good as in typical ChIP-seq experiments. In addition, the relative number of ChIP-seq regions containing the discovered motif is also rather small (1045/

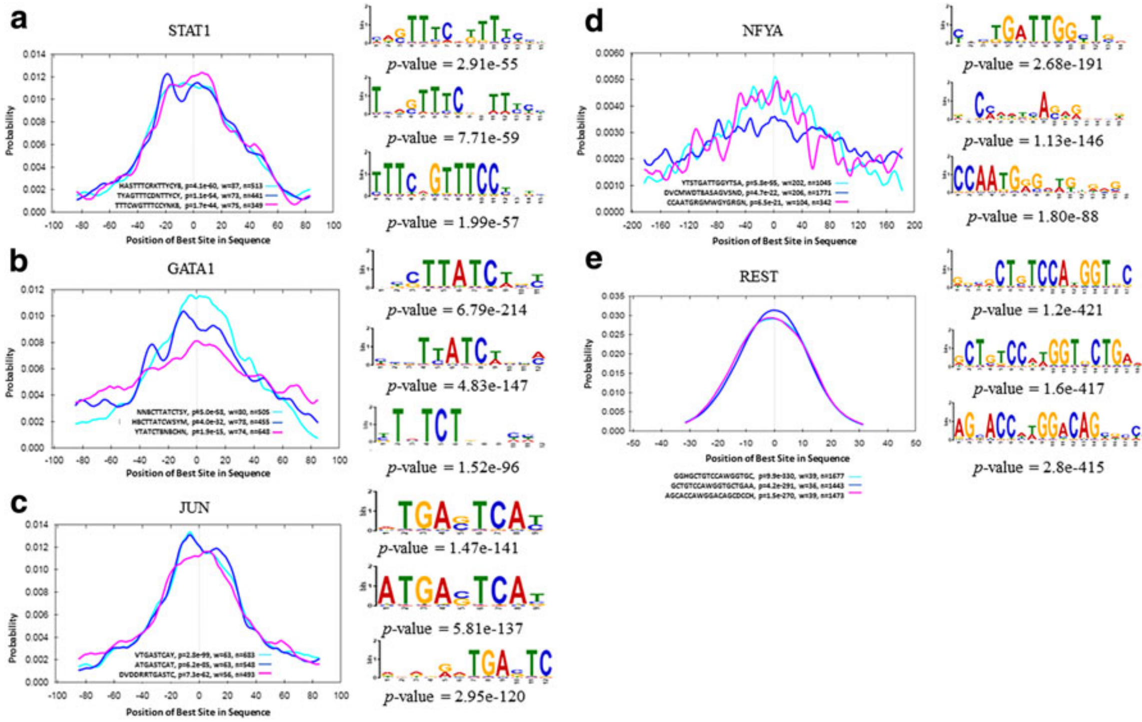


FIG. 1. Comparison of central enrichment and TFBS enrichment results for ChIP-seq of five TFs. On the left is the distribution of the best predicted sites for the three most centrally enriched motifs discovered by SIFEM. Each curve shows the density (averaged over bins of 20-bp width) of the best strong site for the named motif at each position in the ChIP-seq peak regions. The legend shows the consensus, its central enrichment *p* value, the width of the most enriched central region (*w*), and the number of peaks that contain a motif site. On the right are the sequence logos and enrichment *p* values for the binding sites. ChIP-seq, chromatin immunoprecipitation sequencing; TF, transcription factor; TFBS, transcription factor binding site.

7647 = 13.6%), possibly due to an imperfect motif, indirect binding of NFYA to DNA, low ChIP antibody specificity, or other experimental issues.

SIFEM identified several discriminative motifs for each TF, which provide new insights regarding cofactors of STAT1 and JUN. As shown in Figure 2, the found consensus sequences TTTCWGTTTCCYNKB and ATGASTCAT closely match the PPM for IRF1 and NFE2, respectively. This can be also seen in a study by Lehtonen et al. (1997) that characterized the expression of STAT and IFN regulatory factor (IRF) family TFs in primary human blood mononuclear cells, and in a study by Zhang et al. (2006) that concluded hyperactivation of IRF-1 and elevated STAT1 dimer formation may be two general switches that contribute to a much more robust antiviral symphony against virus replication when type I and type II IFNs are coadministered.

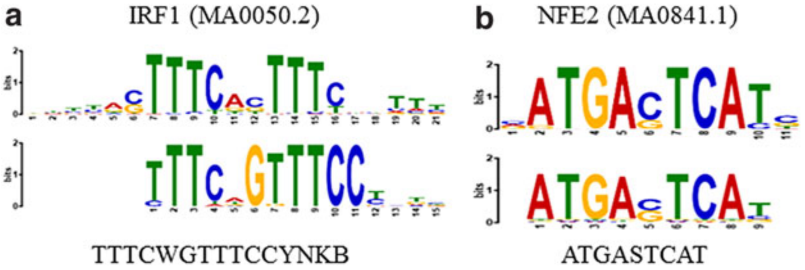


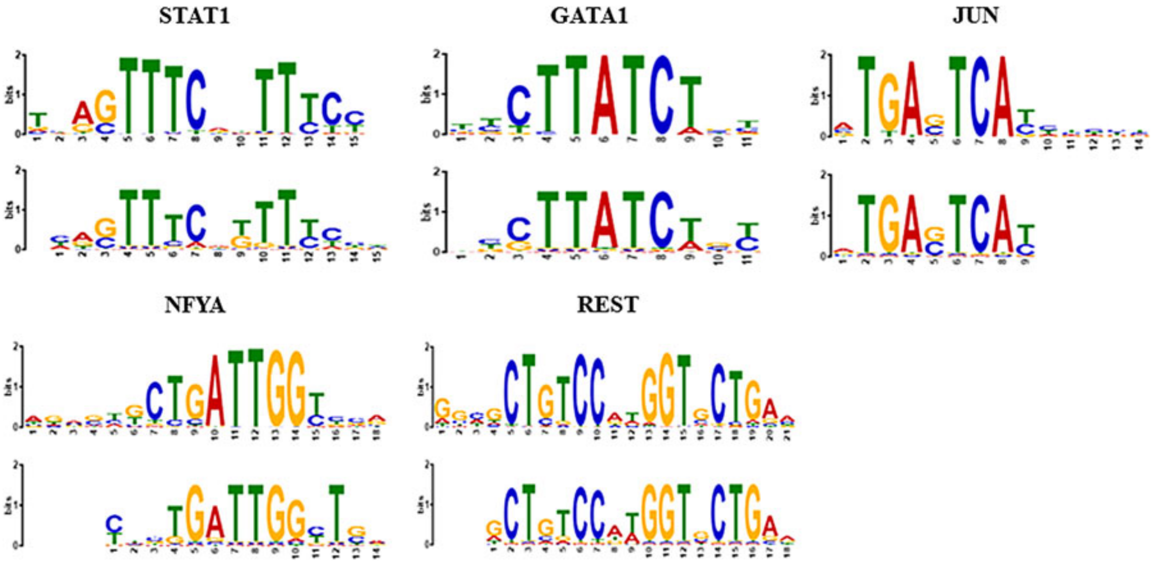
FIG. 2. Unrelated or cofactor TF motifs identified by SIFEM. Bottom: motifs found by SIFEM. Top: matched motifs in JASPAR. **(a)** Unrelated motif found for STAT1, and **(b)** unrelated motif found for JUN in ChIP-seq data.

Table 3. Comparison Results for Five Different Transcription Factors

	Transcription factors				
	STAT1	*GATA1*	*JUN*	*NFYA*	*REST*
Proposed method					
E value (TOMTOM)	**2.59e-11**	**4.27e-08**	**4.07e-08**	8.21e-03	1.36e-14
Overlap	15	11	11	14	21
No. of sites	236	271	354	304	141
MEME-ChIP					
E value (TOMTOM)	2.87e-05	3.89e-07	9.90e-05	**5.00e-03**	**5.52e-17**
Overlap	13	11	13	7	19
No. of sites	232	460	296	337	560

The motifs in the JASPAR database were used as a reference to benchmark the model performance. The similarity of the discovered motifs in the ChIP-Seq data sets to known motifs was assessed using the TOMTOM package from the MEME suite (Bailey et al., 2009). Table 3 presents *E* values, measures of similarity between the discovered and known motifs, where lower *E* values representing more similar results were found for STAT1, GATA1, and JUN in the proposed model. Moreover, simultaneously finding more TFBSs and also more similar motifs to the reference for JUN and STAT1 shows the superiority of our model in discovering more TFBSs than MEME-ChIP did. In addition to the quantitative evaluation of the proposed method, the sequence logos used for visualization were generated. In the sequence logos, the relatively large letters are the core components of the sought pattern, whereas positions with no distinguished letter indicate that there is ambiguity and any nucleotide can take this specific position. Figure 3 shows the alignments of each direct ChIP-ed TF binding motif to known motifs chosen based on central motif enrichment using TOMTOM. The sequence logos in Figure 3 also prove the proficiency of SIFEM in finding the target motif patterns, since the logos produced by this method closely resemble those of the known motif patterns.

Moreover, the motifs discovered by SIFEM for STAT1 and REST have highly significant matches to known motifs in JASPAR. Two Sample Logo (Vacic et al., 2006) is another interesting visualization tool


STAT1

GATA1

JUN

NFYA

REST

FIG. 3. Comparison of motifs proposed by SIFEM (bottom) versus JASPAR (top) for five different TFs. Motifs are shown as information content in bits.

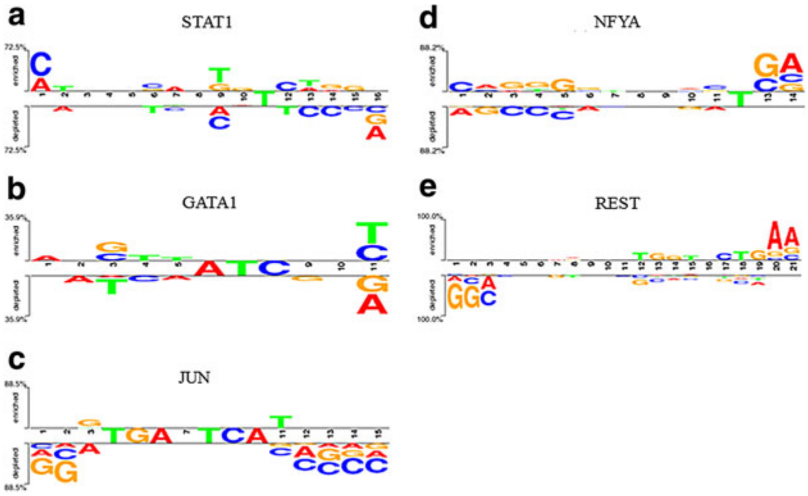


FIG. 4. Two Sample Logo for motifs proposed by SIFEM and MEME-ChIP for five different TFs. The upper and lower sections display sets of symbols that are enriched or depleted, respectively, in the proposed motif, and the middle section displays consensus symbols.

that can concisely show the salient differences between two aligned motifs. In this study, the Two Sample Logo was constructed to present the sequence characteristics of the discovered TFBSs in SIFEM versus MEME-ChIP. As can be seen in Figure 3, there is no significant difference in the highly enriched positions and most differences are in the noninformative positions. In STAT1, we observed that the enriched residues were G and T (position 9 in Fig. 3a), while A and C were depleted. The third position of the GATA1 TFBSs enriches G and C, while T is depleted in this position. Other detailed features are presented in Figure 4.

The overlap and bold values indicate the number of overlapping positions between the discovered motif and reference alignment and the best *E* value for a single TF, respectively.

4. DISCUSSION

In this article, we have proposed SIFEM a new clustering method for finding motifs that infers the motif as a PPM and relies on the overrepresentation of motif instances within the DNA sequence. Unlike current motif discovery algorithms (Alipanahi et al., 2015; Zeng et al., 2016), SIFEM does not require classified input sequences known in advance to contain the motif being sought. Instead, the proposed algorithm, like most de novo motif discovery algorithms, estimates the number of motif appearances from the data. This capability is quite robust, as the results showed that even when only 13.6% of the NFYA ChIP-seq peak regions contain a motif, the motif is still characterized well (Fig. 1d). We also overcame the problem of finding several existing motifs in the ChIP-seq data set in a single pass. Our idea can be considered a sequential integration of FCM and EM methods and demonstrated improvement over the performance of the existing relative method, the MEME-ChIP, for finding motifs in ChIP-seq experiments.

Regarding the comparison results, and as can also be seen in previous studies (Ibrikci and Karabulut, 2010; Zeng et al., 2016), a method may not always produce perfect predictions, possibly due to low information content of the target pattern or a very low number of TFBSs with respect to the sequence in which they reside. However, while the similarity measure for SIFEM was not able to outperform MEME-ChIP results for NFYA and REST TFs, the proposed algorithm found the true motifs for all TFs. On the contrary, SIFEM was able to determine the direct DNA-binding motif of ChIP-ed TFs using the central enrichment method. In conclusion, we have proposed a motif discovery algorithm that can characterize ChIP-seq data with different kinds of pattern characteristics in an efficient manner.

The proposed algorithm is implemented in MATLAB. Each pass through the data set with the FCM and EM algorithm has a time complexity of $O(nWC^2)$ and $O(nW)$, respectively. Therefore, the overall time complexity is proportional to the number of clusters, the width of the motif, and the size of the data set.

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2015R1D1A1A01058075) and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (No. HI14C1277). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M., et al. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838.

Asyali, M.H., and Alci, M. 2004. Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. *Bioinformatics* 21, 644–649.

Bailey, T.L., Boden, M., Buske, F.A., et al. 2009. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res*. 37, 202–208.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Bailey, T.L., and Machanick, P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* 40, e128.

Chen, K., Nimwegen, E.V., Rajewsky, N., et al. 2010. Correlating gene expression variation with cis-regulatory polymorphism in *Saccharomyces cerevisiae*. *Genome Biol. Evol.* 2, 697–707.

Chen, S. 2006. *The Application of the Expectation-Maximization Algorithm to the Identification of Biological Models.* Virginia Polytechnic Institute and State University.

Consortium, E.P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Das, M.K., and Dai, H.-K. 2007. A survey of DNA motif finding algorithms. *BMC Bioinformatics* 8, 1–13.

Do, C.B., and Batzoglou, S. 2008. What is the expectation maximization algorithm? *Nat. Biotechnol.* 26, 897–899.

Gennert, M.A., and Yuille, A.L. 1988. Determining the optimal weights in multiple objective function optimization. Proceedings Second International Conference on Computer Vision, Tampa, FL, 87–89.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., et al. 2007. Quantifying similarity between motifs. *Genome Biol.* 8, R24.

Ibrikci, T., and Karabulut, M. 2010. Employing fuzzy C-means for DNA transcription factor binding site identification. *J. Circuit Syst. Comput.* 19, 15–30.

Jin, Y., and Wang, L. 2009. *Fuzzy Systems in Bioinformatics and Computational Biology*. Springer: Berlin-Heidelberg.

Johnson, D.S., Mortazavi, A., Myers, R.M., et al. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502.

Lehtonen, A., Matikainen, S., and Julkunen, I. 1997. Interferons up-regulate STAT1, STAT2, and IRF family transcription factor gene expression in human peripheral blood mononuclear cells and macrophages. *J. Immunol.* 159, 794–803.

Machanick, P., and Bailey, T.L. 2011. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697.

McLeay, R.C., and Bailey, T.L. 2010. Motif enrichment analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* 11, 1–11.

Quang, D., and Xie, X. 2014. EXTREME: An online EM algorithm for motif discovery. *Bioinformatics* 30, 1667–1673.

Teytelman, L., Thurtle, D.M., Rine, J., et al. 2013. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18602–18607.

Vacic, V., Iakoucheva, L.M., and Radivojac, P. 2006. Two sample logo: A graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537.

Weirauch, M.T., Cote, A., Norel, R., et al. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol*. 31, 126–134.

Xi, R., and Fenton, R.G. 1993. A sequential integration method for inverse dynamic analysis of flexible link manipulators. Proceedings IEEE International Conference on Robotics and Automation, Atlanta, GA, 743–748.

Zeng, H., Edwards, M.D., Liu, G., et al. 2016. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics* 32, i121–i127.

Zhang, X.N., Liu, J.X., Hu, Y.W., et al. 2006. Hyper-activated IRF-1 and STAT1 contribute to enhanced interferon stimulated gene (ISG) expression by interferon alpha and gamma co-treatment in human hepatoma cells. *Biochim. Biophys. Acta* 1759, 417–425.

Address correspondence to:
Dr. Jinwook Choi
Department of Biomedical Engineering
College of Medicine
Seoul National University
103 Daehak-ro
Jongno-gu
Seoul 110-799
Korea

E-mail: jinchoi@snu.ac.kr