Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Robust Mixture Model Clustering of DNA Binding Sites

Sheng Liu, Qing Song, Aize Cao, Xulei Yang, Yilei Wu

Abstract— **Nucleotide sequences contain motifs that preserved through evolution because they are important to the structure or function of the molecules. DNA binding site analysis is an important issue in biology experiments as well as in computational methods. To find DNA binding sites that bind to specific transcription factors, we develop a Robust Mixed Effect Mixture Model (RMEMM). The DNA sequences are represented as mixed effect model of position specific frequency, considering the relationship of frequency between positions. The results show that the mean effect is similar to position-specific scoring matrices (PSSM), providing a new view of the sequence. This model is robust to outliers or data with a bit large tails on distribution.**

I. INTRODUCTION

The huge amount of biological sequence data, especially involved in the progress of genome projects, demand the development of computational methods and tools to annotate, cluster, or explain their sequences, structures and functions corresponding to living process such as transcriptions, regulatory factors, translations, etc. Sequence-specific DNA-binding proteins involved in transcriptional regulation (transcription factors) play a central role in many biological processes. [6] surmised that the distribution of amino acids around bases found in 130 protein-DNA complexes in the protein data bank could be used to derive empirical interaction potentials, thus to predict DNA target sites for DNA-binding proteins. A strict sequence correspondence between amino acids and nucleotide bases was not found in protein-DNA complexes, while they have preferences with each other. That makes it possible to identify and analyze DNA binding sites through computational methods from primary sequences. An important issue in biology is to discriminate different classes or clusters of DNA binding sites according to different transcription factors on their functions. This include two subproblems. The first is to find DNA binding sites given a large set of related sequences, even whole genome sequence. Many methods for this identification of DNA binding sites problem have been proposed. [16] compared three different algorithms for this motif-finding problem, YMF [14], [15], MEME [1], AlignACE [12]. The second subproblem is, having found a set of DNA binding sites, to find what transcription factors they are related to and their relationships. [8] recently summarized and compared some basic widely used methods, including consensus sequence [4], position-specific scoring matrices (PSSM) [17], [2], [3], and Centroid [8] method from representation of binding sites point of view. These methods are all based on the assumption that the bases in each binding site sequence are independent from each other. Actually it is how the sequences are organized that makes the sequences different from each other in biological meaning. So the bases in a sequence are correlated. In this paper, we proposed a mixed effect model to partially solve this problem. In addition, the representation of binding sites in these methods requires prior information on the relationship between sequences. When sequence data is not in a large amount, the prior information provided may be limited. Due to limitation of number of available binding sites data, the binding sites corresponding to a transcriptional factor may not represent all the possible binding sites of that transcriptional factor. The PSSM or other parameters related to the properties of all binding sites of the transcriptional factor may not be accurate based only on the observed binding sites. We propose a Robust Mixed Effect Mixture Model (RMEMM) to tackle this problem. We just use it for clustering on the second subproblem. This is especially important when unknown transcription factors are involved or when we study the structure of regulatory networks from transcription factors. In this paper, we cluster the binding sites as each cluster is related to transcription factor. The rest of this paper is organized as follows, we first introduce representation of DNA binding site sequences. Then we evaluate the proposed method. Finally we give a brief discussion.

II. METHODS

A. Data Set

The data sets are from database of [11] which contains 59 transcription factors with their DNA binding sites experimentally determined. As in [10], 43 transcription factors corresponding to 357 binding sites are selected for analysis. Each transcript factor has more than 2 binding sites. They are set to at the same length of 15 bases. Unlike [10] producing additional simulated data based on the PSSM, we just use this 357 binding sites directly for the analysis.

B. Sequence representation

Many effective clustering algorithms are based on probabilistic framework and work well. Similar to PSSM, each sequence is represented by the frequency of occurrence at each position. Let $Y(y_1,\ldots,y_n)$, denote n DNA sequences, $Y_i(t_{ij})$ denotes value of the ith sequence at position t_{ij}, for $i = 1,\ldots,n$, $j = 1,\ldots,m$, where n is the number of sequences, m is the width of sequence. For a single sequence, Y_i, the 'frequency' of occurrence, f^i_{jN}, is 1 if each base N(A,C,G,T) is present at this position j, 0 otherwise. Considering the background frequency of occurrence in the data set, $f^0_{jN} = \frac{n_{jN}+1}{n+4}$, where n_{jN} is the number of occurrence in the data set at position j. We get

$$Y_{iN}(t_{ij}) = f^i_{jN} - f^0_{jN} \tag{1}$$

1-4244-0033-3/06/$20.00 ©2006 IEEE.

Similarly, for a cluster of sequence of the same transcriptional site, we can use the frequency representation, f^c_{jN}, for cth cluster, using the same formulation.

For simplicity, frequency representation of Y_i for base A is added after that of base C. In the same manner, G after C, T after G. That is $Y_i = [f_{1A}, \ldots, f_{mA}, f_{1C}, \ldots, f_{mC}, f_{1G}, \ldots, f_{mG}, f_{1T}, \ldots, f_{mT}]$, the width of Y_i is 4*m;

C. Mixed effect mixture model with t-distribution

Linear mixed-effects models [5] have become a popular tool in many areas such as biology. To model the frequency of occurrence of each sequence and cluster frequency, we use a mixed effect model, fixed effect that represent the cluster mean frequency and random effect that models the frequency difference between cluster frequency and each sequence's frequency. $Y_i(t_{ij})$ consisting two parts. The first part is the cluster frequency, f^c_{jN}. The second part is the difference between each sequence's frequency and cluster frequency, d^c_{jN}. So Y can also be written as:

$$Y = f^c + d^c + \varepsilon \tag{2}$$

where ε is the within-object error which is independent distributed. We assume that these n sequences are from g different clusters indexed by $c = 1, \ldots, g$. Let $Z(z_1, \ldots, z_n)$ denote cluster labels defining cluster of origin of $y_1, \ldots, y_n$, respectively. z_{ci} is 1 if yi belongs to the cth cluster, 0 elsewhere. According to [7], with B-spline basis, Y can be modeled as:

$$Y_i(t_{ij}) = \left(\sum_{l=1}^{p} \beta_l^{(c)} \bar{B}_l(t_{ij}) \right) + \sum_{l=1}^{q} \psi_{il} B_l(t_{ij}) + \varepsilon_{ij} \tag{3}$$

Where $\bar{B} = \{\bar{B}_l(), l = 1, \ldots, p\}$ and $B = \{B_l(), l = 1, \ldots, q\}$ are basis for spline function on [0, m]. $\beta_l^{(c)}$ is coefficient corresponding to cluster c. ψ_{il} is the random effect coefficient with mean 0, and covariance matrix $Cov(\psi_i) = \Psi$. The first part of right part of (3) represents the mean of the cth cluster, the second part represents the deviation from the cluster mean, the third part represents the uncorrelated errors with mean 0, and covariance matrix σ. As these sequences are from g different clusters, the n sequences Y are realized with g-component normal mixture probability density function(p.d.f).

$$f(y; \pi, \mu, \Sigma) = \sum_{c=1}^{g} \pi_c \phi(y; \mu_c, \Sigma_c), \tag{4}$$

where the mixing proportions π_c are nonnegative and sum to one. $\phi(y; \mu_c, \Sigma_c)$ denotes the multivariate normal p.d.f. with mean μ_c, covariance matrix Σ_c. To add robustness to model (4) to noise or outlier, we introduce a new random variable $U(u_1, \ldots, u_n)$ with gamma distribution to scale normal p.d.f. as mentioned by [9],

$$U \sim \gamma(\frac{1}{2}\nu, \frac{1}{2}\nu), \tag{5}$$

where ν is the degree of freedom of U, according to [9] , Y is distributed as t-distribution:

$$Y_i \sim t(\mu, \Sigma, \nu_i) \tag{6}$$

and

$$Y_i | u_i, z_{ci} = 1 \sim N(\mu_c, \Sigma_c / u_i) \tag{7}$$

We want to combine (3), (4) and (6) models into one to have improved performance in terms of clustering, we get,

$$f(y; \beta, \psi, \varepsilon, \nu) = \frac{\Gamma\left(\frac{\nu+m}{2}\right) |(B\Psi B' + \sigma^2 I)|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{m}{2}} \Gamma(\frac{\nu}{2}) \{1 + \frac{\delta(y;\beta,\psi,\varepsilon)}{\nu}\}^{\frac{1}{2}(\nu+m)}} \tag{8}$$

where $(B\Psi B' + \sigma^2 I)$ is the covariance matrix of $(\psi B + \varepsilon)$, and

$$\delta(y; \beta, \psi, \varepsilon) = (y - \beta\bar{B})^T (B\Psi B' + \sigma^2 I)^{-1} (y - \beta\bar{B}) \tag{9}$$

Similarly, from (8) and (3),

$$Y_i | u_i, z_{ci} = 1 \sim N(\beta_c \bar{B}, (B\Psi B' + \sigma^2 I)/u_i) \tag{10}$$

for $i = 1, \ldots, n$, where n is number of data, c is number of cluster.

From (8) and (3),

$$\Sigma_c = B\Psi B' + \sigma^2 I \tag{11}$$

$$\mu = \beta\bar{B} \tag{12}$$

D. ML estimation via Expectation Maximization

In EM framework of ML estimation of t-distribution,considering ψ, U and Z as missing data, the complete-data log likelihood is:

$$\begin{aligned}
&L_{com}(\pi, \beta, \psi, \varepsilon, \nu) \\
&= \sum_{c=1}^{g} \sum_{i=1}^{n} z_{ci} \log \pi_c + \sum_{c=1}^{g} \sum_{i=1}^{n} z_{ci} \left\{ -\log\Gamma\left(\frac{1}{2}\nu_c\right) \right. \\
&\quad \left. + \frac{1}{2}\nu_c \log\left(\frac{1}{2}\nu_c\right) + \frac{1}{2}\nu_c(\log u_i - u_i) - \log u_i \right\} \\
&\quad + \sum_{c=1}^{g} \sum_{i=1}^{n} z_{ci} \left(-\frac{1}{2}\log|\Psi| - \frac{u_i}{2}\psi_i' \Psi^{-1} \psi_i - \frac{m}{2}\log\sigma^2 \right. \\
&\quad \left. - \frac{u_i}{2\sigma^2} \left\| Y_i - \beta^{(c)}\bar{B} - \psi_i B \right\|^2 \right)
\end{aligned} \tag{13}$$

Where π is the proportion of each cluster among all the clusters, And $\sum_{c=1}^{g} \pi_c = 1$

Details of the E-step and the M-step are given upon request supplement material.

After the convergence of the EM iterations, the parameters $\pi_{c|i}, \Psi, \beta^{(c)}, \sigma^2$ are estimated.

III. Results and Discussion

A. Effectiveness of Robust Mixed Effect Mixture Model

We generate simulated data (SD1) from this model as [7]:

$$Y(j) = (\beta_1^{(c)}+\psi_1)+(\beta_2^{(c)}+\psi_2)\cos(\frac{2\pi j}{m})+(\beta_3^{(c)}+\psi_3)\sin(\frac{2\pi j}{m})$$

where $j = 1, \ldots, m$, β and other parameters are the same as [7]. Results (Table II) shows 180 to 192 data for each cluster were correctly clustered, similar to [7].

Error rate is calculated as follow:

$$Er = \left(1 - \frac{\text{Data correctly clustered}}{\text{Total number of data analyzed}}\right) \times 100$$

B. Robustness to the random noise

The same simulation model was used as previous section. But parameters are changed as follows:

$$\beta = \begin{bmatrix} 0 & 0.5 & 0.87 \\ 0 & -0.81 & 0.59 \\ 0 & 0.5 & -0.87 \\ 0 & -0.71 & -0.71 \end{bmatrix}$$

where 4 rows represent parameters of each cluster. ψ reflects the random effects, with the variances of 0.29, 0.22, 0.19, and pair-wise correlations 0.50, -0.05, and 0.04. For each cluster, 200 data are generated. In addition, we generate 20 random data points with the same length. Robust Mixed Effect Mixture Model (RMEMM) , Deterministic Annealing (DA) and K-means clustering algorithm were used for the 820 data points.

Error rate of RMEMM is much lower (Table I) than that of DA and K-means, showing the robustness of RMEMM to random outliers.

C. Robustness to outliers

Data are generated as the previous section, in addition, we modify each cluster of the first two data sets by constant value 4 which form outliers (SD3). Also, we add a constant value to first 4 data of each cluster in SD1 so these data become outliers (SD2), as Fig. 1 shows. Results (Table II) demonstrate that RMEMM is robust against such data. This means that RMEMM could deal with outliers that shift a little from normal data on every dimension.

TABLE I
RESULTS OF CLUSTERING OF SIMULATED DATA USING ROBUST MIXED EFFECT MIXTURE MODEL(RMEMM), DETERMINISTIC ANNEALING(DA) AND K-MEANSON DATA WITH RANDOM NOISE

Algorithm	RMEMM	Deterministic Annealing	k-means
Er	16.92(2.1186)	34.67(1.7067)	37.68(8.8020)

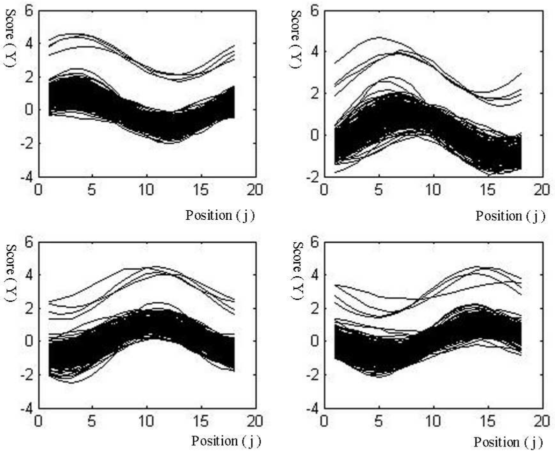

Fig. 1. Each graph represents a cluster of simulated data in SD2 showing abnormal data in each cluster

TABLE II
ERROR RATE(ER) OF CLUSTERING OF SIMULATED DATA USING ROBUST MIXED EFFECT MIXTURE MODEL(RMEMM), DETERMINISTIC ANNEALING(DA) ON DIFFERENT SIMULATED DATA

Algorithm	RMEMM			Deterministic Annealing		
Data	SD1	SD2	SD3	SD1	SD2	SD3
Er	6.77 (0.58)	7.15 (0.96)	18.84 (1.29)	6.24 (0.53)	12.89 (3.11)	26.74 (6.85)

D. Position-specific scoring matrices (PSSM)

After clustering the DNA binding sites, we can have the effect (mean value) of the represented Y. Along the index of position of 4 bases, we plot PSSM and estimated mean value as the RMEMM clustering results (Fig. 2). The two data are very close to each other. That is, at the same time of clustering, Robust Mixed Effect Mixture Model (RMEMM) also get the cluster mean that corresponding to PSSM.

E. Improved clustering results of DNA binding sites

To compare the performance of proposed method, we also cluster DNA binding sites using deterministic annealing (DA) method.

From Table III, RMEMM has a lower error rate than that of DA. As the data set is from a limited data, each cluster corresponding to a transcription factor may contain much more binding sites. So cluster frequency at certain positions may not actually reflect the true cluster frequency, there may be errors on the frequency when we have limited number of biological experiment verified data set. On the other hand, some positions in a cluster of DNA binding sites may have less relation to the cluster property thus vary more than those homologous positions. The distribution of this positions may be large tailed.

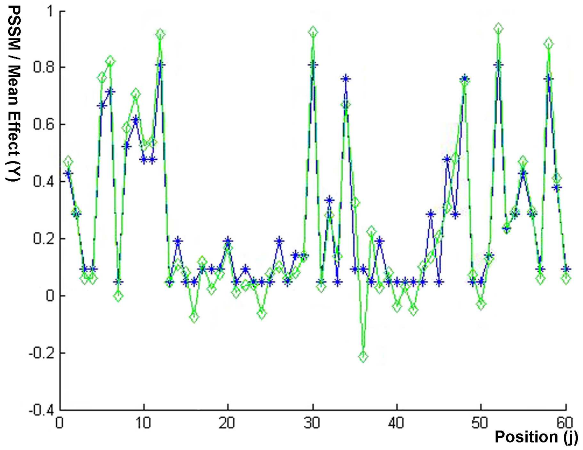

Fig. 2. Graph comparison of PSSM and Mixed Effect mean value. X axis is the position of one DNA sequence analyzed in the order of A,C,G,T as mentioned in the body. Y axis is PSSM(with asterisk) or mixed effect mean value estimated (with diamond sign)

TABLE III

RESULTS OF CLUSTERING OF DNA BINDING SITE MOTIFS USING ROBUST MIXED EFFECT MIXTURE MODEL(RMEMM) AND DETERMINISTIC ANNEALING(DA)

Algorithm	RMEMM	Deterministic Annealing
Er	35.85	41.51

NOTATION LIST

i	ith sequence of total number of n sequences
j	jth base position in a sequence of length m
c	cth cluster of total number of g clusters
Y	DNA sequences with length m each
N	One base of a set of four bases (A,T,C,G)
f^i_{jN}	Frequency of occurrence of base N at position j for ith sequence
f^0_{jN}	Background frequency of occurrence
Z	Cluster Labels, where z_{ci} is 1 if y_i belongs to the cth cluster
$\bar{B}$	Basis for spline function for fixed effect
B	Basis for spline function for random effect
β	Fixed effect coefficient
ψ	Random effect coefficient
Ψ	Covariance matrix of ψ
ε	Uncorrelated error with mean 0 and covariance matrix σ
π_c	Component proportion of cluster c
$\phi(y;\mu,\Sigma)$	Normal probability density function with mean μ and covariance matrix Σ
U	Random variable with gamma distribution to scale normal p.d.f.
L_{com}	Complete data log likelihood
ν	Degree of freedom of U
Φ	Parameters of $\beta,\psi,\varepsilon,\nu$
$Pr(.)$	Probability
Er	Error rate

REFERENCES

[1] Bailey T.L. and Elkan C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994, pp. 28-36.

[2] Berg, O. and von Hippel, P. , Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, vol. 193, 1987, pp. 723-750.

[3] Berg, O. and von Hippel, P. , Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *Journal of Molecular Biology*, vol. 200, 1988, pp. 709-723.

[4] Day, W. H. and McMorris, F., Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Research*, vol. 20, 1992, pp. 1093-1099.

[5] Hartley, H. O. and Rao, J. N. K., Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika*, vol 54, 1967, pp. 93-108.

[6] Kono, H., and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins., *Proteins*, vol. 35(1), 1999, pp. 114-121.

[7] Yihui Luan and Hongzhe Li., Clustering of time-course gene expression data using a mixed-effects model with B-splines., *Bioinformatics*, vol. 19, 2003, pp. 474-482.

[8] Osada, R., Zaslavsky E., et al., Comparative analysis of methods for representing and searching for transcription factor binding sites., *Bioinformatics*, vol. 20, 2004, pp. 3516-3525.

[9] Peel, D., McLachlan, G.J., Robust mixture modeling using the t distribution., *Statistics and Computing*, vol. 10, 2000, pp. 339-348.

[10] Qin ZS., McCue LA., Thompson W., Mayerhofer L., Lawrence CE., Liu JS., Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology*, vol. 21(4), 2003, pp. 435-9.

[11] Robison, K. and McGuire, A. M. and Church, G. M., A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 Genome., *Journal of Molecular Biology*, vol. 284, 1998, pp. 241-254.

[12] Roth F.P., Hughes J.D., Estep P.W., and Church G. M., Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, vol. 16, 1998, pp. 939-945.

[13] Shoham, S., Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition*, vol. 35(5), 2002, pp. 1127-1142.

[14] Sinha S. and Tompa M., A statistical method for finding transcription factor binding sites. In *proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 344-354.

[15] Sinha S. and Tompa M., Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, vol. 30(24), 2002, pp. 549-5560.

[16] Sinha S. and Tompa M., Performance Comparison of Algorithms for Finding Transcription Factor Binding Sites. *Third IEEE Symposium on Bioinformatics and Bioengineering (BIBE)*, 2003.

[17] Staden, R., Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, vol. 12, 1984, pp. 505-519.