Vol 17 No 12, December 2008
1674-1056/2008/17(12)/4396-05

Chinese Physics B

© 2008 Chin. Phys. Soc.
and IOP Publishing Ltd

A combined statistical model for multiple motifs search*

Gao Li-Feng(高丽锋)[a], Liu Xin(刘 鑫)[b], and Guan Shan(官 山)[c]†

[a] *Chinese Academy of Agriculture Science, Beijing* 100081, *China*
[b] *Institute of Theoretical Physics, Beijing* 100080, *China*
[c] *Physics Science and Technology Department, Yangzhou University, Yangzhou* 225009, *China*

(Received 3 January 2008; revised manuscript received 13 February 2008)

Transcription factor binding sites (TFBS) play key roles in genebior 6.8 wavelet expression and regulation. They are short sequence segments with definite structure and can be recognized by the corresponding transcription factors correctly. From the viewpoint of statistics, the candidates of TFBS should be quite different from the segments that are randomly combined together by nucleotide. This paper proposes a combined statistical model for finding over-represented short sequence segments in different kinds of data set. While the over-represented short sequence segment is described by position weight matrix, the nucleotide distribution at most sites of the segment should be far from the background nucleotide distribution. The central idea of this approach is to search for such kind of signals. This algorithm is tested on 3 data sets, including binding sites data set of cyclic AMP receptor protein in E.coli, PlantProm DB which is a non-redundant collection of proximal promoter sequences from different species, collection of the intergenic sequences of the whole genome of E.Coli. Even though the complexity of these three data sets is quite different, the results show that this model is rather general and sensible.

Keywords: transcription factor binding sites, motif, position weight matrix
PACC: 0500, 0590

1. Introduction

As more and more genome are sequenced, one major challenge in the 'post-genome' era is to understand biological functions with the help of genome sequence in which secret of life is encoded. One of the first important steps in this process is to identify evolutionarily conserved fragments in the promoter regions of genes. These DNA segments are called motif and play key roles in the procedure of gene's regulation, transcription and expression. Most of them act as the transcriptional factor binding sites (TFBS) that are bond by some proteins to activate or inhibit the expression of corresponding genes.

Traditionally, two main computational models have been used to describe the experimentally determined TFBS: consensus sequence and position weight matrix (PWM). Consensus sequence is essentially the most likely sequence of the binding sites, and the variability at each site has been ignored. PWM model, on the other hand, tried to capture the inherent variability at each sites, it assigns a probability distribution of the four nucleotide to each position of the binding sites. Therefore, consensus sequence is just a particular case of the PWM, and the PWM is more general than consensus to describe the TFBS.

Algorithms that based on consensus model usually adopt the strategy of enumeration,[1−4] and oligonucleotides with statistical significance are regarded as potential TFBS. The limitation of those approaches is that the computational effort grows exponentially in the length of the site. The other kinds of algorithms that based on the PWM model can be further categorized into three different classes: (i) consensus building greedy algorithm,[5] the well known free software package of this algorithm is CONSENSUS (http://ural.wustl.edu/∼jhc1/consensus/); (ii) Expectation-Maximization (EM) algorithm,[6−8] MEME (http://meme.sdsc.edu/meme/website/) is one of the free software packages of this algorithm; (iii) Gibbs sampling algorithm,[9,10] Gibbs motif sampler (http://bayesweb.wadsworth.org/gibbs/gibbs.html/) is a realization of this algorithm. The difficulty of such kind of motif search algorithms is that both the positions where motifs are located in the DNA sequences and the parameters of the model describing the motifs need to be determined simultaneously. Without any prior knowledge, these algorithms adopt

*Project supported by the National Natural Science Foundation of China (Grant No 70671089), and the Key Important Project (No 10635040).
†Corresponding author. E-mail: dhq2gs@yahoo.com.cn

different strategy to determine the parameters of the model by optimizing an objective function, i.e., the original difficult problem is now transformed to an optimization problem. Since it is very difficult to find a global optimization solution, the first common feature of these algorithms is to initialize a model randomly at first and then to improve it iteratively. The second common feature of these algorithms is to restrict them on a data set with relatively high signal-noise ratio and to search for those common and over-represented sequence patterns. The data set usually consists of the upstream regulatory regions of co-regulated genes (by clustering genes on the basis of their expression profiles or functional annotation). With the advent of whole genome sequencing, an alternative approach is to compare the regulatory regions of orthologous genes in different species to identify functionally conserved sequence motifs.[11,12]

In this paper, we report our model which is used to search multiple motifs simultaneously, the results show that it not only works well on carefully prepared data set, but also works well on the collection of intergenic sequences of the whole genome of E.Coli.

2. Method description

Suppose that there is a data set O which consists of N DNA sequences, $O = \{o_i, i = 1, \ldots, N\}$, and the length of each sequence is L_i. These sequences are obtained either from the upstream regions of co-regulated genes, or from orthologous genes of different species, or from the intergenic sequences of a genome. The basic model of motif search algorithm assumes that one or more motifs are hidden among the sequences, and they are regarded as signals to be detected, the other part of the sequences except motifs are background noises. Therefore, for each sequence in O, there exists a hidden state sequence which assigns each base in the sequence either a state of motif or a state of noise, this state sequence is usually called a path. Motifs are described by PWM, and noise is described by single nucleotide probability distribution.

For CONSENSUS algorithm,[5] searching for motifs which hidden in O is realized by maximizing the objective function

$$I = \sum_i \sum_b f_{bi} \log \frac{f_{bi}}{p_b} \tag{1}$$

where b is the index of four nucleotide, i represents the position of a site in the motif, p_b is the probability of nucleotide b under background, f_{bi} is an element of the PWM, it represents the probability of nucleotide b appearing at position i in the motif. If the objective function I is greater than a threshold, the PWM is accepted as a potential motif.

For Gibbs sampling algorithm,[10] it first assume that there is only one motif in O, and then this motif is detected by maximizing objective function I,

$$I = \sum_{i=1}^{N} \sum_k \log \frac{f_{o_i(a_i+k),k}}{p_{o_i(a_i+k)}}, \tag{2}$$

where i is the index of sequences in O, k represents the position of a site in the motif, a_i is the start site position of the motif in sequence o_i, therefor $f_{oi(ai+k),k}$ represents the probability distribution of nucleotide $o_i(a_i + k)$ at position k in the motif, and $p_{oi(ai+k)}$ is the background probability distribution of nucleotide $o_i(a_i + k)$.

For MEME algorithm,[8] data set O is first broken up conceptually into a new data set X which consists of subsequences of length w (this is the length of the motif to be detected), these subsequences are cut from the sequences in O one by one in overlapping manner. Then MEME algorithm transforms the target of searching for motifs hidden in O to the problem of optimizing the classification of X, i.e. to classify X into different groups by maximizing the likelihood function

$$L(X; \Theta) = \prod_{i=1}^{n} P(x_i \,|\, \Theta) = \prod_{i=1}^{n} \sum_g P(x_i, g \,|\, \Theta) \tag{3}$$

where x_i is a sequence in X, Θ represents the parametrized model, and is defined by (λ_0, $p(b|0)$, λ_i, $p(b|j, i)$), where λ_0 is the probability that a sequence in X is generated from background, λ_i is the probability of the sequence that is generated from motif i, $p(b|0)$ is the background probability distribution of nucleotide b, $p(b|j, i)$ is the probability of nucleotide b appearing at position j in motif i; g represents one of the possible group schemes for sequences in X; the summation covers all possible group schemes.

Since it is very difficult to find global maxima in optimization problem, especially when the solution space is complex, reducing the dimension of solution space will be helpful for finding optimization solution. In our algorithm, the parametrized model is reduced as

$$\Theta = \{p(b),\ p(b|j, i),\ i = 1, 2, ..., K; \\ j = 1, 2, ..., w;\ b \in (A, C, G, T)\}, \tag{4}$$

$$\sum_{b\in(A,C,G,T)} p(b) = 1, \quad \sum_{b\in(A,C,G,T)} p(b\,|j,i) = 1\,, \quad (5)$$

where K is the number of motifs assumed in O, w is the length of the motif, $p(b|j,i)$ represents the probability of nucleotide b at position j of motif i; the noise background, which refers to the remainder of the sequences except motifs, is described by $p(b)$, $b \in (A,C,G,T)$. Under such a model, the likelihood function which needs to be maximized is

$$L(O;\Theta) = \sum_{S} \prod_{b\in(A,C,G,T)} p(b)^{n_0(b)} \prod_{i=1}^{K}\prod_{j=1}^{w} \times \prod_{b\in(A,C,G,T)} p(b\,\Big|j,i)^{n_{ij}(b)}\,, \quad (6)$$

where S represents one of the possible paths of O, $n_0(b)$ is the number of times that nucleotide b appearing as background noise under path S, $n_{ij}(b)$ is the number of times that nucleotide b appearing at position j in the motif i under path S. To maximize the likelihood function with EM algorithm, we actually need to optimize $p(b)$ and $p(b|j,i)$. Comparing the four algorithms mentioned above, we find that our algorithm adopts both the parametrized model used in CONSENSUS and Gibbs sampling algorithm and on the optimization method used in MEME. The following test results also show that this approach is rather sensible and general for different data sets.

3. Test results

Our algorithm is tested on 3 data sets with different complexity. The first one we used to test our algorithm is 18 DNA fragments of length 105 each known to contain cyclic AMP receptor protein (CRP) binding sites.[5,8] This is a small data set with strong signal-noise ratio, and all the three methods mentioned above work well to find the signal. In our model, while we set the length of the motif to be 20 bps and the number of motifs to be 5, the most strong signal is just the CRP binding sites whose logo (http://weblogo.berkeley.edu/) is shown in Fig.1.



Fig.1. Sequence logo of CRP binding site.

The second data set we used is selected from PlantProm DB which is downloaded from http://www.softberry.com/. PlantProm DB[13] is an annotated, non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site(s), it contains 305 entries, including 71 from monocot, 220 from dicot and 14 from other plants. All stored promoter sequences in this database are of the same length, 251 bp, and position 201 corresponds to the TSS (Transcriptional Start Site). The strongest signal in both the monocot and dicot data sets is the TATA box motif. For the 220 promoter sequences of dicot, In our model, TATA box was found 199 times totally in 144 promoter sequences, and 46% of them , which is 91 TATA boxes, are located between TSS+60bp and TSS+40bp. Figure 2 is the logo of TATA box we have found, it is a very strong signal in comparison with its environment. The strongest signal of this data set found by Gibbs sampler and MEME are show in Figs.3 and 4 respectively. Obviously, both of them fail to find the strongest signal.



Fig.2. Logo of TATA box in dicot.

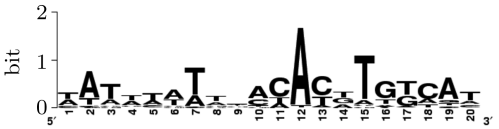


Fig.3. Logo of motif found by Gibbs sampler in dicot.

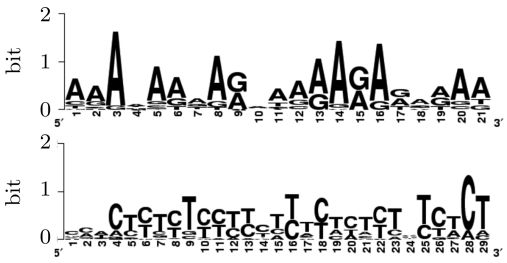


Fig.4. Logo of motifs found by MEME in dicot.

For the 71 promoter sequences of monocot, Gibbs sampler, MEME and our method all found a signal, whose logo is shown in Figs.5, 6 and 7 respectively. Clearly, both MEME and our method found TATA

box as the strongest signal, while Gibbs sampler found something else. In this case, TATA box was found 66 times in promoter sequences in our method, and 61% of them, which is 41 TATA boxes, are located between TSS+60bp and TSS+40bp. Again, we can see that TATA box is the most strong signal in comparison with its environment (Fig.7).



Fig.5. Logo of TATA box found by Gibbs sampler in monocot.



Fig.6. Logo of TATA box found by MEME in monocot.



Fig.7. Logo of TATA box found by MEME in monocot.

The third data set on which we tried our algorithm is the collection of intergenic sequences of one strand of E.Coli. The length of intergenic sequence in E.Coli is different. For simplicity, sequences with length longer than 200 bps and shorter than 1000 bps are remained, and if there is sequence whose length is longer than 1000 bps, then only the forward 1000 bps is cut out and accepted. This data set contains 959 entries. To save time, this data set is divided into 3 subsets, each contains ∼300 sequences. In this case, both Gibbs sampler and MEME fail to find signals from the data sets, and our method found 4 strong motifs which are most probably some biological functions. The logo is shown in Figs.8, 9, 10 and 11, respectively. The first motif shown in Fig.8 is a palindrome, it appeared 44 times in the intergenic region of the positive strand, and while we search the pattern GCCGGATNNGGCGNNNNCGCC in the intergenic region of the whole genome in web server (http://genolist.pasteur.fr/Colibri/), we found 111 matches. This palindrome is also found by Li,[14] and in their web server, it is named matrix 11 and does not match with any known factor.



Fig.8. logo of the first motif found in the intergenic region of E.Coli.



Fig.9. Logo of the second motif found in the intergenic region of E.Coli.



Fig.10. Logo of the third motif found in the intergenic region of E.Coli.

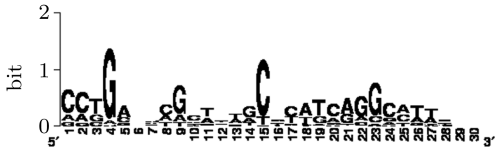


Fig.11. Logo of the fourth motif found in the intergenic region of E.Coli.

The second motif shown in Fig.9 appeared 15 times in the intergenic region of the positive strand. To search the pattern GTAGGCNNGATAAGACGC in (http://genolist.pasteur.fr/Colibri/), we found 104 matches in intergenic region of the whole genome.

The third motif shown in Fig.10 appeared 58 times in the intergenic region of the positive strand, it has a weak palindrome, CCCCGCNNNNGCGGGG , and this pattern has 50 matches in the intergenic region of the whole genome.

The fourth motif shown in Fig.11 appeared 32 times in the intergenic region of the positive strand, it also is a weak palindrome, CCTGA(Nx)TCAGG. When the interval between CCTGA and TCAGG is allowed to change from 0 to 14, there is 494 matches in the intergenic region of the whole genome.

4. Discussion

It is well known that the structure of the promoter sequence is rather complicated, and many transcriptional factors usually need to bind with promoter sequence and work together to activate or inhibit the expression of the gene, therefore there often exist several correlated TFBs in the promoter sequence. To find these signals out simultaneously and furthermore to uncover the correlation relationship among them is the key to for us to understand the interaction between genes. Based on CONSENSUS, Gibbs sampling and MEME algorithm, we supposed a combined approach for multiple motifs search, and testing results show that our approach still works well after we reduced the model parameter space.

References

[1] Bussemaker H J, Li H and Siggia ED 2000 *Proc. Natl. Acad. Sci. USA* **97**(18) 10096

[2] Bussemaker H J, Li H and Siggia E D 2000 *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8** 67

[3] Helden J V, Andre B and Collado-Vides J 1998 *J. Mol. Biol.* **281** 827

[4] Sinha S and Tompa M 2002 *Nucleic Acids Res.* **30** 5549

[5] Stormo G D and Hartzell G W 1989 *Proc. Natl. Acad. Sci. USA* **86**(4) 1183

[6] Dempster A P, Laird N M and Rubin D B 1977 *J. Roy. Statist. Soc.* Ser. B **39** 1

[7] Lawrence C and Reilly A 1990 *Proteins* **7** 41

[8] Bailey T L and Elkan C 1994 *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2** 28

[9] Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, and Wootton J 1993 *Science* **262**(5131) 208

[10] Liu J S, Neuwald A F and Lawrence C E 1995 *J. Amer. Statist. Assoc.* **90** 1156

[11] Mcguire A M and Church G M 2000 *Nucleic Acids Res.* **28** 4523

[12] McCue L A, Thompson W, Carmack C S, Ryan M P, Liu J S, Derbyshire V and Lawrence C E 2001 *Nucleic Acids Res.* **29** 774

[13] Shahmuradov I A, Gammerman A J, Hancock J M, Bramley P M and Solovyev V V 2003 *Nucleic Acids Res.* **31**(1) 114

[14] Li H, Rhodius V, Gross C and Siggia E D 2002 *Proc. Natl. Acad. Sci. USA* **99**(18) 11772