Finding Signals for Plant Promoters

Weimou Zheng
Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China; Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China

The strongest signal of plant promoter is searched with the model of single motif with two types. It turns out that the dominant type is the TATA-box. The other type may be called TATA-less signal, and may be used in gene finders for promoter recognition. While the TATA signals are very close for the monocot and the dicot, their TATA-less signals are significantly different. A general and flexible multi-motif model is also proposed for promoter analysis based on dynamic programming. By extending the Gibbs sampler to the dynamic programming and introducing temperature, an efficient algorithm is developed for searching signals in plant promoters.

Introduction

Methods for gene recognition are based on either homology analysis, on content search, or on signal search. Signals are short sequence segments with a definite structure. The signal search tries to recognize the location in genome where the gene expression machinery interacts with the nucleic acid. Signals as biochemical binding sites on DNA, or corresponding mRNA and pre-mRNA play a key role in transcription, splicing or translation. Promoter is the most important regulatory region which controls the initiation of transcription. Promoter prediction is crucial for gene annotation. In eukaryotes, a promoter, encompassing a gene's transcription start site (TSS), contains aggregates of transcription factor binding sites (TFBSs). Several ubiquitous and cell-specific regulatory factors work together to achieve a combinatorial control. TFBSs can appear in different combinations on different promoters. The order of TFBSs in promoters varies, and relative distances of TFBSs in various promoters differ. Promoter is undoubtedly extremely complex. Efficient gene hunting using promoter recognition is still impossible. For example, GenScan uses a very simplified model for promoter: a 15 bp TATA-box weight matrix model (WMM), a 14-20 bp intergenic-null model of spacer, and then a 8 bp cap site WMM (*1*). About 30% of eukaryotic promoters lack an apparent TATA signal. TATA-less promoters are modelled simply as intergenic-null regions of 40 bp in length.

Weight matrix can be used to describe a signal as pattern of a multiple sequence alignment, and is good for modelling certain TFBSs. This simple type of probabilistic models for signals assigns a probability to each position for signal sequence of some fixed length l (*2*, *3*, *4*). The assumption of independence between positions is the main limitation of WMMs. A natural generalization is an inhomogeneous Markov model and its modification is called windowed weight array model, replacing the independent probabilities with conditional probabilities. To reliably capture the most significant dependencies between positions, the maximal dependence decomposition (MDD) model has been developed (*5*). We have proposed a simple way to enhance signals by clustering signal sequences (*6*).

To discover novel motif sites, multiple sequence alignment methods are useful. Some statistical methods, e.g. expectation-maximization (EM) or Gibbs sampling algorithm for independent block model (*7*, *8*) or hidden Markov model (*9*), have been developed for finding patterns in unaligned sequences (see reviews, e.g. (*10, 11, 12*)).

There are 71 monocot and 220 dicot promoter sequences for plants available from the web (http://www.softberry.com/). The sequences are taken at [−200, +51] with respect to the TSS. We shall search signals with the simplest model of a single motif in a noise background for each sequence. Then, we shall propose a flexible multi-motif model to cope with the complicated combination of TFBSs based on dynamic programming.

Single Motif Model

We align the 71 monocot sequences according to their TSSs, and calculate base frequencies at each position.

E-mail: zhang@itp.ac.cn

We estimate the 5' and 3' noises by taking an average over 30 bases at the two ends, $[-200, -171]$ and $[+22, +51]$, respectively. To compare signal with noise, we need a measure for the distance between two distributions. The most often used distance is the relative entropy or Kullback-Leibler (KL) distance (*13, 14, 15*)

$$D(p,q) = \sum_i p_i \log(p_i/q_i), \qquad (1)$$

where $\{p_i\}$ and $\{q_i\}$ are the two probability distributions. $D(p,q)$ corresponds to a likelihood ratio. $D(p, q)$ is not convenient when some p_i or q_i is close to zero, which is often the case for signals. We introduce the following modified χ_2 distance

$$d = \sum_i 2(p_i - q_i)^2/(p_i + q_i), \qquad (2)$$

where the summation is taken over those i with either p_i or q_i not vanishing. This distance is the leading term of the KL distance when expanding the latter with respect to p_i around $p_i = q_i$. The KL distance can be used for distinguishing a signal site from a noise site.

The distance between the 5' and 3' noises is very small, only 0.002. We then calculate distances of each base on 5' and 3' sides of the TSS to its corresponding noise. The distances on the 3' side are generally smaller than those on the 5' side. At 19 bases, the distances are over 0.15, and two of them reach 0.31. Two segments of large distance are $[-45, -43]$ and $[-31, -25]$, inside the so-called core promoter region. The cap region $[-1, +6]$ is a region of a smaller distance. Another segment of a moderate distance on 3' side is at $[+12, +16]$. The distributions and their distances to noise for bases around TSS and the 5' and 3' noise distributions are listed in Table 1.

Table 1 Position Weight Matrix around the TSS, Noise Distribution, and Distances between Base Distributions and Noise Distribution (Last Row).

	5'noise	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	3'noise
A	1.08	0.79	1.13	0.90	1.92	0.62	1.52	0.85	1.69	1.75	0.85	1.07
C	1.11	1.30	1.18	1.63	1.01	2.14	0.90	1.69	1.41	0.96	1.58	1.15
G	0.88	0.56	0.62	1.01	0.51	0.56	0.73	0.39	0.62	0.79	0.79	0.91
T	0.93	1.35	1.07	0.45	0.56	0.68	0.85	1.07	0.28	0.51	0.79	0.87
Distance		0.10	0.03	0.15	0.22	0.26	0.06	0.18	0.26	0.14	0.05	

Table 2 Position Weight Matrices of TATA and TATA-Less Signals, Noise Distribution, and Distances between Base Distributions and Noise Distribution (Last Row) for Monocot.

	noise	TATA signal: cctataaatacc											
A	0.96	0.68	0.15	0.00	3.70	0.00	3.62	2.64	4.00	1.06	2.42	0.15	0.60
C	1.20	1.81	3.47	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.30	2.34	2.04
G	0.84	0.75	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.68	0.98	0.83
T	1.01	0.75	0.30	3.92	0.30	4.00	0.38	1.36	0.00	2.87	0.60	0.53	0.53
Distance		0.11	1.35	2.26	2.02	2.39	1.94	1.44	2.46	1.37	0.64	0.56	0.22
		TATA-less signal: aagaaaaaaaca											
A		2.44	1.78	1.11	3.78	2.89	2.67	3.33	2.22	2.89	2.67	0.22	3.78
C		1.56	0.44	0.00	0.00	0.00	0.00	0.00	0.44	0.00	0.00	2.00	0.22
G C		0.00	1.33	2.44	0.00	0.89	1.11	0.67	1.33	0.67	0.00	0.44	0.00
T		0.00	0.44	0.44	0.22	0.22	0.22	0.00	0.00	0.44	1.33	1.33	0.00
Distance		1.27	0.46	1.11	2.11	1.34	1.27	1.77	0.99	1.20	1.45	0.41	2.10

While averaging will generally blur out signals of a variable position, a large distribution distance indicates the existence of signal. To extract the strongest signal, we consider a simple model of a single motif in the noise background. Bearing TATA and TATA-less sequences in mind, we think two types of the motif.

We apply the model to the region $[-200, -1]$, taking the cap region as a separator. The algorithm used for multiple sequence alignment is similar to that described in (*7*, *8*). The main difference is that we now have to determine the position and type of motif at the same time, instead of just position. We fix the length of motif to be 12. The optimal length may be determined from the distance between distributions of motif and its nearby bases and that of noise background. The results for monocot and dicot are listed in Tables 2 and 3, respectively. The TATA signals of the monocot and the dicot are very similar except for one base shift, while their TATA-less signals are significantly different. The average start positions of the former for the monocot and the dicot are -49 and -59. And the average start positions of the TATA-less signals are -128 and -92. The monocot and the dicot are also different in the GC content of their noises and TATA-less signals. Only 18 monocot sequences of the 71 are identified as TATA-less, while 100 dicot sequences of the 220 are TATA-less. From the distance to noise, it seems more appropriate to take the width for TATA-signal to be 11.

Table 3 Position Weight Matrices of TATA and TATA-Less Signals, Noise Distribution, and Distance between Base Distributions and Noise Distribution (Last Row) for Dicot

	noise	TATA signal: ctataaatabrna											
A	1.33	0.63	0.00	3.60	0.13	3.93	2.17	4.00	1.60	3.50	0.07	1.43	1.63
C	0.79	2.43	0.10	0.00	0.00	0.03	0.00	0.00	0.00	0.20	1.27	1.47	0.83
G	0.61	0.17	0.00	0.00	0.13	0.00	0.07	0.00	0.07	0.00	1.30	0.50	0.53
T	1.27	0.77	3.90	0.40	3.73	0.03	1.77	0.00	2.33	0.30	1.37	0.60	1.00
Distance		0.73	1.91	1.45	1.64	1.88	0.75	2.00	0.78	1.26	0.76	0.23	0.03
	TATA-less signal: ctctcactyctc												
A		0.00	0.48	0.84	0.48	0.24	1.84	0.00	0.00	0.52	0.00	0.68	1.00
C		2.44	1.28	1.80	0.80	3.72	0.24	3.96	1.20	1.40	2.04	1.24	2.36
G		0.44	0.00	1.36	0.48	0.04	0.00	0.00	0.00	0.48	0.72	0.20	0.00
T		1.12	2.24	0.00	2.24	0.00	1.92	0.04	2.80	1.60	1.24	1.88	0.64
Distance		1.11	0.70	1.03	0.34	2.22	0.56	2.61	1.30	0.29	0.95	0.32	0.82

Multi-motif Search by Dynamic Programming

To describe combination of many TFBSs, a general and flexible multi-motif model is proposed based on dynamic programming (*9*). Let us consider the following simple model: 6 motifs of the same width of 8 in the noise background. We introduce the model as a generating model. Suppose that the probability to select a noise base is π_0, and those for motifs are π_i, i= 1, 2,. . . , 6, respectively. Here, $\Sigma_0^6 \pi_i = 1$. After a noise or a motif is selected, another set of probabilities $p(0, 0,\alpha)$ and $p\ (i, j, \alpha), i = 1,\ldots, 6;\ j = 0, 1, \ldots, 7;\ \alpha \in$ {A, C, G, T} is then used to generate specific bases, where i is the index of motif type and j the position in a motif. Under the statistical model, the probability to observe the sequence $S_{0;n} = b_0 b_1 \ldots b_n$, or the partition function, can be calculated by considering all the possible ways to arrange motifs and noise on the sequence. The partition function $Z\ (S_{0;k})$ will satisfy the recursion relations:

$$Z(S_{0;-1}) \equiv 1; \tag{3}$$

$$Z(S_{0;k}) = Z(S_{0;k-1})\pi_0 p(0, 0, b_k),\ 0 \geq k < 7; \tag{4}$$

$$\begin{aligned} Z(S_{0;k}) &= Z(S_{0;k-1})\pi_0 p(0, 0, b_k) \\ &+ \sum_{i=1}^{6} \prod_{j=0}^{7} Z(S_{0;k-8})\pi_i p(i, j, b_{k+j-7}),\ k \geq 7. \end{aligned} \tag{5}$$

For $k \geq 7$, there are always 7 choices of the state for each base, corresponding to the 7 terms in the summation. The terms will be denoted by $Z\ (S_{0;k}|q_k)$, where $q_k \in \{0, 1, \ldots, 6\}$ indicates the state of b_k being noise or belonging to one of the 6 motifs. We call a path the possible assignment state of each base in the sequence. For our model, in a path any nonzero q_k must appear successively in a multiple of 8. The path with the greatest probability may be determined by the following Viterbi algorithm. We record the state of b_k which corresponds to the greatest of

the 7 terms $Z\ (S_{0;k}|q_k)$. Once the state of the last base b_n is etermined, we may trace base states back to get the whole path. We call this 'optimal' path the Viterbi path. After the Viterbi path is identified for each sequence in the sequence data set, we may estimate the whole probability parameter sets $\{\pi\}$ and $\{p\}$ just by counting. This corresponds to the greedy algorithm.

There are recursion relations for $Z\ (S_{k;n})$ similar to those for $Z\ (S_{0;k})$. The previous ones are called the forward relations, while the other ones the backward. In terms of $Z\ (S_{0;k})$ and $Z\ (S_{k;n})$ the probability for any base b_j to be at state q_j (noise or a certain position in one of motifs 1 to 6) can be calculated. This fuzzy assignment will also lead to an estimation of parameters $\{\pi\}$ and $\{p\}$. It may be called the Baum-Welch or EM algorithm.

The greedy algorithm would be easily trapped in a rather poor local optimal for a generic initiation. The EM algorithm is not very efficient. We develop an analog of the Gibbs sampler as follows. Converting $Z\ (S_{0;k}|q_k)$, $q_k = 0, 1, \ldots, 6$ to weights, we sample a state q_k for b_k. We keep doing this until reaching b_n, then we can trace base states back to obtain a full path, which may be called a Gibbs path. After finding

Table 4 Position Weight Matrix for Monocot Promoter Motifs and Noise Distribution Obtained for Region [-200, -1] by Dynamic Programming (Motif 1 Fits the Consensus for TATA-Box Signal)

	Noise								
		Motif 1: tataaata(tata) π_1=0.007							
A	1.30	0.73	2.83	0.00	3.80	2.08	3.44	1.88	3.57
C	0.78	0.52	0.84	0.07	0.00	0.04	0.17	0.00	0.00
G	0.64	0.24	0.06	0.00	0.00	0.00	0.13	0.00	0.22
T	1.28	2.51	0.28	3.93	0.20	1.88	0.26	2.12	0.20
		Motif 2: aaaanaaa(a-rich) π_2=0.013							
A		2.47	3.14	3.68	2.30	1.05	2.53	2.62	2.69
C		1.16	0.86	0.17	0.01	1.04	1.33	1.14	0.01
G		0.10	0.00	0.14	0.24	0.94	0.13	0.13	1.11
T		0.27	0.00	0.00	1.45	0.97	0.01	0.12	0.19
		Motif 3: atttttt(t-rich) π_3=0.003							
A		2.85	0.00	1.04	0.00	0.04	0.52	0.04	0.44
C		0.00	0.00	0.15	0.04	0.22	0.00	0.74	0.41
G		0.74	0.00	0.00	0.15	0.00	0.74	0.70	0.15
T		0.41	4.00	2.81	3.81	3.74	2.74	2.52	3.00
		Motif 4: gagatnaa(r-rich) π_4=0.005							
A		0.39	1.65	0.00	2.47	0.39	1.43	2.11	2.13
C		0.00	0.19	0.00	0.34	0.82	0.61	0.32	0.27
G		2.96	0.80	3.64	1.19	1.12	1.04	1.36	1.58
T		0.65	1.36	0.36	0.00	1.67	0.92	0.22	0.02
		Motif 5: ctttyttt(y-rich) π_5=0.018							
A		0.09	1.32	0.12	0.12	0.97	0.81	0.10	1.05
C		3.42	0.05	1.81	1.55	1.33	0.91	2.07	1.20
G		0.16	0.10	0.10	0.18	0.40	0.49	0.00	0.06
T		0.33	2.54	1.96	2.14	1.30	1.79	1.83	1.69
		Motif 6: cacgtgkc(even) π_6=0.008							
A		0.99	2.38	0.46	1.09	0.25	0.31	0.98	0.79
C		1.65	0.02	1.98	0.10	0.00	0.33	0.00	1.83
G		0.71	0.28	0.31	1.64	0.15	3.19	1.59	0.28
T		0.64	1.32	1.24	1.17	3.60	0.17	1.44	1.09

Gibbs path for all sequences, we estimate parameters $\{\pi\}$ and $\{p\}$ by direct counting. This leads to an algorithm which may be called the Gibbs algorithm. Furthermore, we may introduce a temperature τ to raise $Z(S_{0;k}|q_k)$ to the power of $1/\tau$. The temperature adjusts the relative weighting among $Z\ (S_{0;k}|q_k)$. The zero temperature gives the greedy limit. Since the partition function $Z(S_{0;n})$ has the clear meaning being the total probability of observing the sequence set, which provides a standard for comparison of different models, the partition function is taken as the objective function.

Let us examine the region $[-200,-1]$ of the 71 monocot promoter sequences. The 6 motifs and the noise found by the Gibbs algorithm are listed in Table 4. One of the 6 motif fits well the TATA pattern found in last last section. The TATA-less pattern for the monocot cannot be undoubtedly associated with any motifs, and is more or less related to motif 2. We have also examined the 220 dicot sequences in the region $[-200,-1]$ with the same model. As shown in Table 5, the 6 motifs found are: tataaata (tata), aaaanaaa (a-rich), atttttttt (t-rich), gagatnaa (r-rich), ctttyttt (y-rich) and cacgtgkc (even), in comparison with the monocot motifs ctataaat (tata), aaaaawaa (a-rich), cttttrtt (tr-rich), acrtgrws (r-rich), tctcctcc (y-rich) and rccacgsm (c-rich). We see some correspondence between the two sets of motifs, and the tendency of the increasing GC content from dicot to monocot.

The average number m of motifs per sequence may be estimated as follows. Since the number of noise bases is (200- $8\bar{m}$), we have the relation π_0 = (200-$8\bar{m}$)/(200-$7\bar{m}$), which, for the monocot of $\pi_0 = 0.926$, leads to $\bar{m} = 9.75$. For the dicot, $\pi_0 = 0.946$ leads to $\bar{m} = 7.84$.

Discussion

We have used a multi-type single motif model to find the strongest motif for promoter. The dominant type of the motif turns out to be the known TATA-box. At the same time, a TATA-less signal, as the counterpart of the TATA-box, is determined. This signal may be employed in gene finders to improve the promoter recognition. While the TATA signals for both monocot and dicot are very similar, their TATA-less signals are significantly different. We have proposed a general and flexible multi-motif model based on dynamic programming. By extending the Gibbs sampler to the dynamic programming and introducing temperature, an efficient algorithm has been developed. We have applied the algorithm to analyze plant promoter. The found motifs provide candidates for possible binding sites.

A classification scheme may be proposed. After determination of parameters, the Viterbi path can be identified for each sequence. Sequences can then be grouped according to the motifs appearing in their Viterbi paths. Once the sequence data set has been divided into subsets, the same search algorithm performed on a single subset can help to find more precise patterns for motifs.

The model discussed is still oversimplified. The model can be further refined. The width of motifs need not be the same for each. The tuning of the motif number and width can be done based on the distribution distance. If the distribution distance between an end base of a motif and noise is small, the base should be removed from the motif. On the other hand, if the distribution distance between a base next to a motif and noise is large enough, the base should be included in the motif. If the probability (π_i) of a motif is small, the motif should be removed from the motif list. We may define the distribution distance between two motifs of the same width as the sum of the distribution distance between their bases at each position over the whole width. When the widths of two motifs are different, the distance may be defined as the minimum of the distances obtained when sliding the shorter along the longer and comparing the shorter with substrings of the longer. When the distance of two motifs is small, we should join the two motifs into one. The method is rather general. The use of the method for poly-A signal analysis near 3'UTR will be discussed elsewhere.

References

1. Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78-94.
2. Gelfand, M. S. 1995. Prediction of function in DNA sequence analysis. *J. Comput. Biol.* 2: 87-115.
3. Staden, R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic. Acids. Res.* 12: 505-519.
4. Stormo, G.D., *et al.* 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli. Nucleic Acids Res.* 10: 2997-3011.
5. Salzberg, S.L., *et al.* (eds.) 1998. *Computational*

Methods in Molecular Biology. Elsevier, Amsterdam, Netherlands.

6. Zheng, W.M.. 2002. *Genomic signal enhancement by clustering.* ITPAS-preprint.
7. Lawrence, C.E., and Reilly, A.A. 1990. An expectation maximization (EM) algorithm for the identification and chracterization of common sites in unaligned biopolymer sequences. *Proteins* 7: 41-51.
8. Lawrence, C.E., *et al.* 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262: 208.
9. Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.* 77: 257-285.
10. Vanet, A., *et al.* 1999. Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.* 150: 779-799.
11. Brazma, A., *et al.* 1998. Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.* 5: 279-305.
12. Zhang, M.Q. 2002.*Computational methods for promoter recognition, in Current Topics in Computational Molecular Biology.* Tsinghua University Press, Beijing, China.
13. Kullback, S., *et al.* 1959. *Information Theory and Statistics.* Wiley, New York, USA.
14. Kullback, S. 1987. *Topics in Statistical Information Theory.* Springer, Berlin, Germany.
15. Sakamoto,T., *et al.* 1986. *Akaike Information Criterion Statistics.* KTK Sci-entific, Tokyo, Japan.

This work was supported in part by the Special Funds for Major National Basic Research Projects, the National Natural Science Foundation of China (No. 30170232).

Received: 13 January, 2003
Accepted: 17 January, 2003