

A flexible model for promoter motifs

Wei-Mou Zheng

Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China

and

Beijing Genomics Institute, Beijing 101300, China

Abstract

Transcription factor binding sites (TFBS) 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. A general and flexible multi-motif model is proposed for promoter motif analysis based on dynamic programming. In the model, motifs are described with weight matrices, all possible arrangement of motifs are examined, and the total probability of training sequence set is maximized for determination of parameters. By extending the Gibbs sampler to the dynamic programming and introducing temperature, an efficient algorithm is developed for searching motifs in promoters. The algorithm is tested with plant promoters.

1 Introduction

Methods for gene recognition either are based on homology analysis, or on content search, or on signal search. Signals are short sequence segments with a definite structure. 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 in GenScan.

Weight matrix can be used to describe a signal as a 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 inhomogeneous Markov model and its modification 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 [13]. 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 the complicated combination of TFBSs based on dynamic programming.

2 Single motif model

We align the 71 monocot sequences according to their TSSs, and calculate base frequencies at each position. 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 [14, 15, 16]

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

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

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

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 less large 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 are shown in Fig. 1.

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 by examining the distance from distributions of motif and its nearby bases to that of noise background. The results for monocot and dicot are listed in Figs. 2 and 3, respectively. The TATA signals of the monocot and dicot are very similar except for one base shift, while their TATA-less signals significantly different. The average start positions of the former for the monocot and dicot are -49 and -59 , respectively. The average start position of the TATA-less signals are -128 and -92 , respectively. The monocot and dicot are also different in the GC content of their noises and TATA-less signals. Only 23 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.

3 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 , $i = 1, 2, \dots, 6$, respectively. Here, $\sum_0^6 \pi_i = 1$. After a noise

or a motif is selected, another set of probabilities $p(0, 0, \alpha)$ and $p(i, j, \alpha), i = 1, \dots, 6; j = 0, 1, \dots, 7; \alpha \in \{A, C, G, T\}$ is then used to generate individual bases, where i is the index of motif type and j that of the position in a motif. Under the statistical model, the probability to observe the sequence $S_{0;n} = b_0 b_1 \dots b_n$, or the partition function, can be calculated by considering all possible ways to arrange motifs and noise on the sequence. The partition function $Z(S_{0;k})$ will satisfy the recursion relations:

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

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

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

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 $q_k \in \{0, 1, \dots, 6\}$ indicates the state of b_k being noise or belonging to one of the 6 motifs. We call a path the possible state assignment of each base in the sequence. For our model, in a path any non-zero q_k must appear successively in a multiple of 8. The path with the greatest probability may be determined by the following Viterbi algorithm. Replacing the summation in (5) by maximum selection, 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 determined, 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 two 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 the motifs) 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 $Z(S_{0;k}|q_k), q_k = 0, 1, \dots, 6$ to weights, we sample a state q_k for b_k . We keep doing this until reach b_n , then we can trace base states back to obtain a full path, which may be called a Gibbs path. After finding 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 whole region $[-200, +51]$ of the 71 monocot promoter sequences. The 6 motifs and the noise found by the Gibbs algorithm and EM algorithm are listed in Fig. 4, where the noise pattern and the motif probabilities π_i are also given. One of the 6 motifs (motif 5) fits well the TATA pattern found in last section. The TATA-less pattern for the monocot is more or less associated with motif 2. We further calculate the position distribution of motifs. It is found that the TATA-motif located by the Viterbi algorithm concentrates around site -32 while there are no prominent peaks in position distribution for other motifs. As an example, we show the motif counts at every three sites for motifs 4 and 5 in Fig. 5. We also investigate the model at 8 motifs. The motifs found are shown in Fig. 6. By comparing the Figs. 4 and 6, it is easily recognized that there is a correspondence between some motifs of the two models. From the model of 8 motifs to that of 6 motifs, the correspondence written in motif indices is: $5 \rightarrow 1, 3 \rightarrow 5, 1 \rightarrow 4,$ and $2 \rightarrow 8$ (with one site shifted). The average number \bar{m} of motifs per sequence may be estimated as follows. Since the number of noise bases is $(251 - 8\bar{m})$, we have the relation $\pi_0 = (251 - 8\bar{m})/(251 - 7\bar{m})$, which, for $\pi_0 = 0.887$ at the model of 6 motifs, leads to $\bar{m} = 16.6$. For the model of 8 motifs, $\pi_0 = 0.888$ leads to $\bar{m} = 15.8$. The increase of motif number results in the sharpening of motif patterns, and the reduce of the total number of identified motifs.

To compare monocot with dicot, we have also examined the 220 dicot sequences at the whole region $[-200, +51]$ with the model of 8 motifs. The 8 motifs found are shown in Fig. 7. The TATA motif (d5, letter ‘d’ added to indicate ‘dicot’) is rather similar to the TATA motif (m1) of the monocot with one site shifted. Motifs d7 and d1 of the dicot bear some similarity with motifs m8 and m7 of the monocot, respectively. Other motifs and noises of the monocot and dicot are quite different.

After motifs have been located on sequences by the Viterbi algorithm, we may inspect the correlation between motifs. Adding an virtual termination motif at the both ends of each sequence, and ignoring the distance in between, we count motif pairs. Denote the number of motif i , the total number of motifs and the number of pairs of nearby motifs i and j by M_i , M and M_{ij} , respectively. The correlation between motifs i and j may be described by $r_{ij} = (M_{ij} - M_i M_j / M) / \sqrt{M_i M_j / M}$. The results are listed in Tables 1 and 2 for the monocot and dicot, respectively. For the monocot the autocorrelation of the TATA motif is negative while most motifs exhibit a positive autocorrelation. The strongest cross correlation is found between TATA motif and motif 7 with $r_{17} = 3.38$. This means that a TATA motif is likely to be followed by motif 7. The rather large positive values of r_{21} and r_{31} imply that a TATA motif is also likely to be led by motifs 2 and 3. Almost all motifs of the dicot, including TATA motif, have a positive autocorrelation. The TATA motif of the dicot does not show as strong cross correlation as in the monocot.

4 Discussions

In the above 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 model can be used for discrimination of promoters. 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 are small, we should join the two motifs into one.

The method is rather general. The use of the method for splicing signal and poly-A signal analysis near 3'UTR will be discussed elsewhere.

This work was supported in part by the Special Funds for Major National Basic Research Projects, the National Natural Science Foundation of China.

References

- [1] C. Burge and S. Karlin, *J. Mol. Biol.* **268** (1997) 78-94.

- [2] M. S. Gelfand, *J. Comput. Biol.*, **2** (1995) 87-115.
- [3] R. Staden, *Nucleic Acids Res.*, **12** (1984) 505-519.
- [4] G.D. Stormo, T.D. Schneider, L. Gold and A. Ehrenfeucht, *Nucleic Acids Res.*, **10** (1982) 2997-3011.
- [5] S.L. Salzberg, D.B. Searls and S. Kasif (eds.), *Computational Methods in Molecular Biology* (Elsevier, Amsterdam, 1998).
- [6] W.M. Zheng, Genomic signal enhancement by clustering, ITPAS-preprint (2002).
- [7] C.E. Lawrence and A.A. Reilly, *Proteins* **7** (1990) 41-51.
- [8] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald and J.C. Wootton, *Science* **262** (1993) 208.
- [9] L.R. Rabiner, *Proc. IEEE*, **77** (1989) 257.
- [10] A. Vanet, L. Marsan, and M.-F. Sagot, Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.* **150** (1999) 779-799.
- [11] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert, Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.* **5** (1998) 279-305.
- [12] M.Q. Zhang, Computational methods for promoter recognition, in *Current Topics in Computational Molecular Biology* (Tsinghua University Press, Beijing, 2002).
- [13] <http://www.softberry.com/>
- [14] Kullback,S., Keegel,J.C. and Kullback,J.H. (1959) *Information Theory and Statistics*, Wiley, New York.
- [15] Kullback,S. (1987) *Topics in Statistical Information Theory*, Springer, Berlin.
- [16] Sakamoto,T., Ishiguro,M. and Kitagawa,G. (1986) *Akaike Information Criterion Statistics*, KTK Scientific, Tokyo.

Table 1. Correlation coefficients r_{ij} between motifs in the monocot.

	1	2	3	4	5	6	7	8
1	-1.52	0.64	-0.64	-0.02	1.56	0.05	3.38	-1.93
2	2.09	-0.00	-0.06	0.42	-0.55	1.68	-1.32	-0.04
3	2.24	2.91	-0.20	0.36	-1.02	-0.49	-0.56	-0.24
4	-0.68	0.42	-0.39	2.68	0.24	-0.54	0.16	-0.18
5	1.26	0.05	0.64	-1.37	1.32	-0.70	-0.32	0.52
6	-0.19	-0.60	0.32	-0.54	0.45	3.09	-2.02	1.09
7	0.95	-0.60	0.23	-1.21	-0.08	-2.02	6.21	-1.41
8	-1.18	-0.30	-0.66	2.17	-1.29	0.46	-1.11	2.25

Table 2. Correlation coefficients r_{ij} between motifs in the dicot.

	1	2	3	4	5	6	7	8
1	6.67	-1.63	-1.81	1.07	1.27	0.36	-4.18	-0.61
2	-0.74	1.86	0.77	0.69	1.00	0.26	-0.22	-0.46
3	-1.13	0.37	4.92	-1.10	1.62	-0.13	-0.65	-0.72
4	1.47	-0.13	-0.62	-0.01	-0.87	1.18	1.48	0.23
5	1.27	-0.17	0.71	0.19	2.20	0.01	0.82	-0.16
6	2.13	-0.29	-0.26	1.84	-0.89	0.70	0.09	-0.04
7	-3.63	2.02	-1.34	0.54	-0.77	0.23	5.47	1.04
8	0.75	-1.50	0.35	0.44	-0.16	-0.04	0.01	2.89

Fig. 1 Distances of nucleotide distribution at each site from that of the noise.

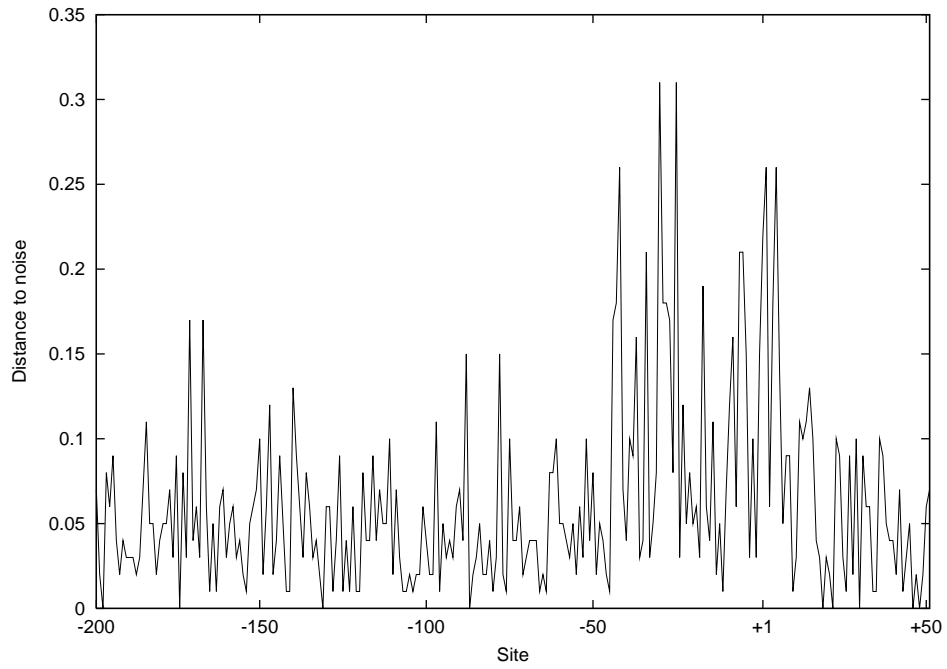


Fig. 2 TATA and TATA-less motifs for the monocot found by the best alignment of single motif with two types. The background noise distribution is also shown at the top. Values above the motif indices indicate the proportion of sequences identified for the motif types.



Fig. 3 TATA and TATA-less motifs for the dicot found by the best alignment of single motif with two types. The background noise distribution is also shown at the top. Values above the motif indices indicate the proportion of sequences identified for the motif types.



Fig. 4 Motifs for the monocot found by the flexible model of 6 motifs based on dynamic programming. The background noise distribution is also shown at the top. Values above the motif indices indicate the probabilities of the corresponding motifs.

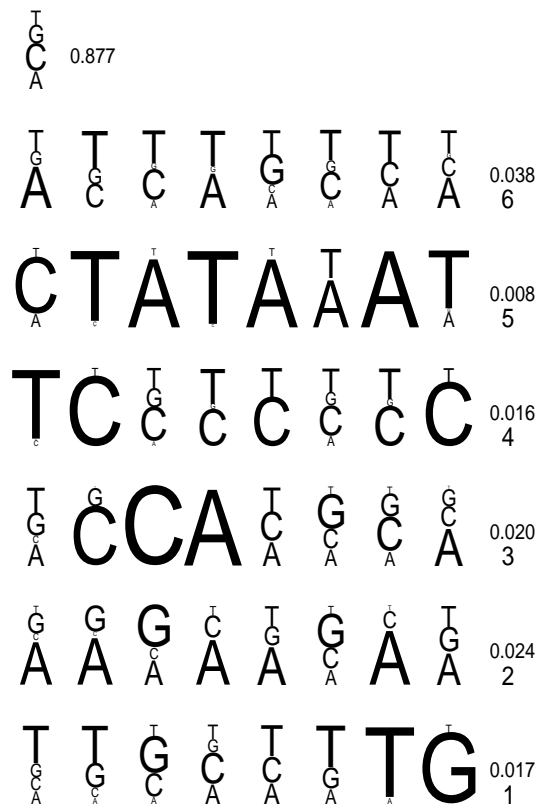


Fig. 5 Motif counts at every 3 sites for motifs 3 and 5 of the monocot located by the Viterbi algorithm using the flexible model of 6 motifs based on dynamic programming.

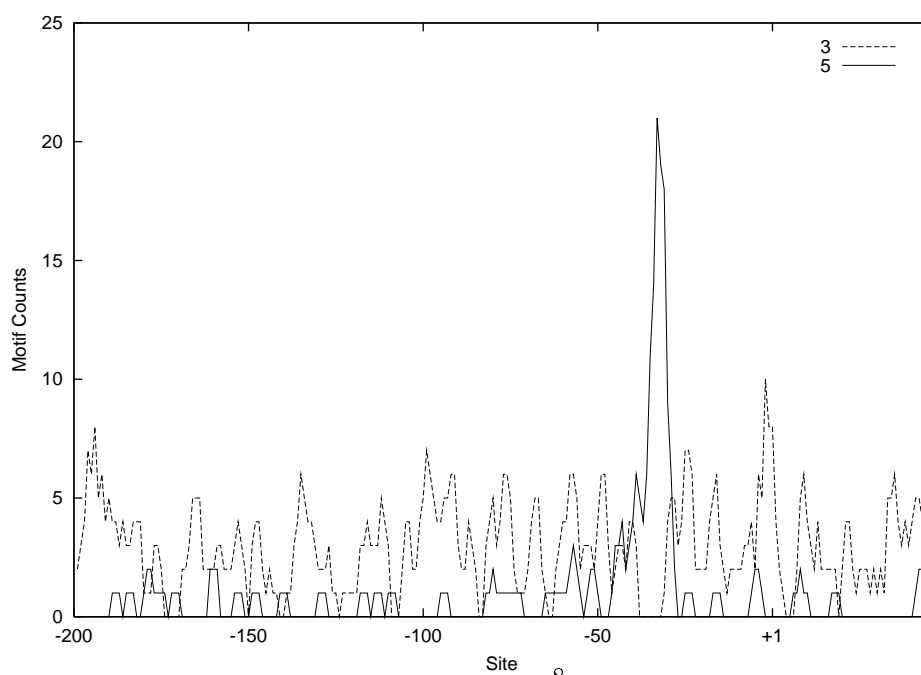


Fig. 6 Motifs for the monocot found by the flexible model of 8 motifs based on dynamic programming. The background noise distribution is also shown at the top. Values above the motif indices indicate the probabilities of the corresponding motifs.

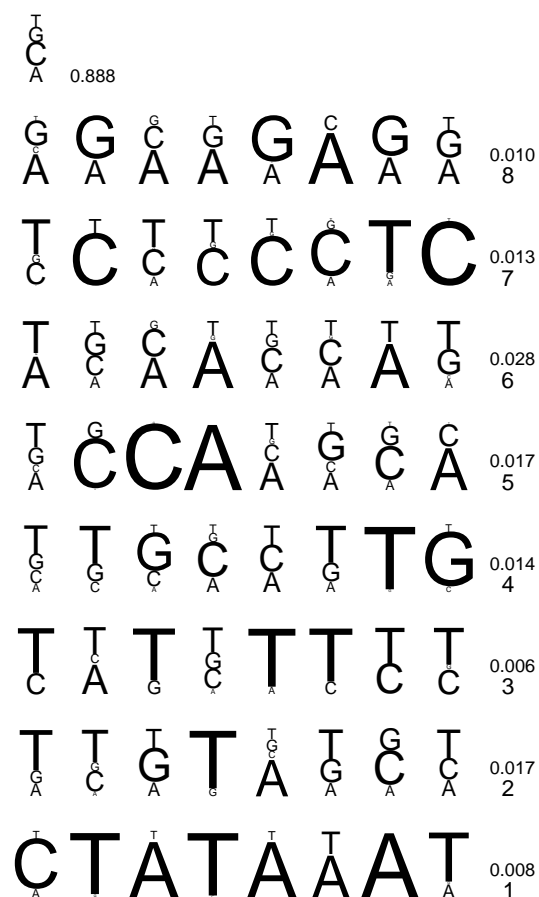


Fig. 7 Motifs for the dicot found by the flexible model of 8 motifs based on dynamic programming. The background noise distribution is also shown at the top. Values above the motif indices indicate the probabilities of the corresponding motifs.

