Commun. Theor. Phys. (Beijing, China) **39** (2003) pp. 631–634

Vol. 39, No. 5, May 15, 2003

Genomic Signal Enhancement by Clustering*

ZHENG Wei-Mou

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

Beijing Genomics Institute, Beijing 101300, China

(Received November 8, 2002)

Abstract *Weight matrix models for signal sequence motif are simple. A main limitation of the models is the assumption of independence between positions. Signal enhancement is achieved by taking the total likelihood as the objective function for maximization to cluster sequences into groups with different patterns. As an example, the initial and terminal signals for translation in rice genome are examined.*

PACS numbers: 87.10.+e, 87.14.Gg, 89.75.Kd
Key words: genomic signals, cluster analysis

1 Introduction

Raw genomic sequences offer little. The annotation of genomes provides a perspective and overview of the entire genome. Methods of gene identification may be classified as similarity search, content search, and signal search. The similarity search is a homology-based approach by comparison with samples in databases, belonging to the so-called r-nearest neighbor discriminant analysis. The content search, e.g. codon/triplet usage, is based on the assumption that a genome may be divided into uniform regions with considerably different statistical characteristics. The signal search is based on position-specific inhomogeneous models, regarding the uniform regions as "noise".

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. Biochemical binding sites on DNA, or corresponding mRNA and pre-mRNA play a key role in transcription, splicing or translation. We know some "words" in the dictionary of signals, but they may also occur almost anywhere irrelevant to any signals. Most common signals are of a word, and can be aligned according to the given word. At each aligned position each base appears with a specific frequency, which describes the positional preference of signals. There exist signals without any unique word for alignment. Here we consider only signals with a word for alignment.

Simple types of probabilistic models for signal sequences assign a probability to each possible DNA sequence of some fixed length l.[1] For signals of variable length, hidden Markov models,[2] models with an enlarged alphabet (including indel), models with linear multiple patterns[3] can be used.

A general independence model takes the position-specific biases into account. The probability for the particular signal sequence $S = s_1 s_2 \dots s_l$ is given by

$$p(S) = f_1(s_1) f_2(s_2) \cdots f_l(s_l) = \prod_{i=1}^{l} f_i(s_i)\,, \qquad (1)$$

where $f_i(s)$ is the probability of generating nucleotide s at position i. Relation (1) may be viewed as the leading term of the Lazarsfeld–Bahadur expansion.[4,5] Biological signal models of this type are called weight matrix models (WMMs).[6,7] The assumption of independence between positions is the main limitation of WMMs. A natural generalization is inhomogeneous Markov model (IMM) 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.[8] Here we propose a simple way to enhance signals by clustering signal sequences.

2 Methods

Our method is based on WMMs. Instead of single weight matrix, we classify signal sequences into several groups or patterns, and assign each pattern a specific weight matrix.

The information provided by a given signal motif is measured by the relative entropy or Kullback–Leibler (KL) distance[9,10,11]

$$D(P,Q) = \sum_j P_j \log(P_j/Q_j)\,, \qquad (2)$$

where P_j and Q_j are the probabilities of observing sequence j as motif and as background, respectively. $D(P,Q)$ corresponds to the likelihood ratio of motif to background. Taking Q_j independent of sequences makes D correspond to the likelihood. The KL distance can also be used for distinguishing a signal site from a noise site. When aligning signals and their neighboring background sites, we may calculate distribution of bases at

*The project supported in part by the Special Funds for Major National Basic Research Projects, National Natural Science Foundation of China and Research Project 248 of Beijing

each aligned position. At both ends, the noise backgrounds are a region where the KL distance among the corresponding distributions of any two sites is small. By taking the noise distribution as the average distribution of noise sites, a signal region is a region where the KL distance of the distribution at each site to that of noise is large. The KL distance $D(p,q)$ 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

$$d = \sum_i 2(p_i - q_i)^2/(p_i + q_i)\,, \tag{3}$$

where the summation is taken over those i with both p_i and q_i not vanishing. This distance is the leading term of the KL distance when expanding the latter with respect to p_i around $p_i = q_i$.

For simplicity, we shall consider only constant $Q_j = 4^{-l}$. From Eqs. (1) and (2), we have

$$D = \sum_{i=1}^{l} \sum_{\alpha} f_i(\alpha) \log[4 f_i(\alpha)]\,, \tag{4}$$

where $\alpha \in \{A, C, G, T\}$.

We determine the patterns of a fixed number, say 3, as follows. Suppose that we have divided the signal sequence set into 3 groups or subsets. We may estimate a weight matrix for each subset, and then calculate the likelihood function (1) for each sequence in the subset. The total likelihood function of the whole set is the product of the likelihood function of all sequences. The optimal patterns correspond to the weight matrices estimated from a subset partition which gives the maximal total likelihood function. That is, the pattern determination is an optimization with the objective function for maximization being the total likelihood. Denote by $w(k,i,\alpha)$ the score or logarithmic probability for nucleotide α to be at position i of pattern k. Equivalently, we may use the following objective function for optimal clustering:

$$F = \sum_k \sum_i w(\nu_k, i, b_{ki})\,, \tag{5}$$

where ν_k is the pattern index of sequence k, and the first summation is taken over all the sequences in the set. For given score matrices w, the probability for sequence k belong to pattern j is proportional to $U_k(j) = \exp[\sum_i w(j,i,b_{ki}]$. More generally, we may introduce a temperature τ to replace $U_k(j)$ with $[U_k(j)]^{1/\tau}$. By taking the limit $\tau \to 0$, sequence k is assigned to the pattern with the largest probability. Once the objective function is chosen, many algorithms can be adopted. Among common algorithms are the simulated annealing, expectation-maximization (EM) algorithm,[12] and Gibbs sampler.[3]

Let us explain the procedure of clustering with the simple greedy algorithm although generally such an algorithm should be used in practice with care due to trapping by local optimum. Assume again that the total number of clusters is 3. We first randomly assign each sequence to one of the 3 clusters. In this way we divide the sequence data set into 3 subsets. We may then estimate 3 weight matrices for the 3 subsets. With these 3 matrices, we obtain 3 values of likelihood for each sequence. We update the assignment of a sequence to one of the clusters, simply depending on which likelihood value is the largest. That is, we make the inference that the sequence belongs to the cluster whose weight matrix gives the largest likelihood. This finishes one iteration. Three new weight matrices can be estimated from the updated 3 clusters of sequences. Further iterations converge when clustering does not change.

3 Results

We use 129 gene sequences obtained from the rice genome[13] by aligning cDNAs with the genome sequence. Although the method can well be applied to the splicing signals, we shall focus on the less discussed initiation and termination signals. If we align the 129 sequences according to the first codon ATG (being sites +1, +2 and +3), the frequencies of bases at positions from −11 to +11 are listed in Table 1, where we have multiply frequencies by 4. The 5′-noise is obtained by averaging over 42 sites from −45 to −4. The sites with a positive index are in the coding region for not too short first exons. For the coding region we calculate 3 3′-noises according to the 3 codon positions, which are obtained by averaging over 14 sites taken every other 2 sites from site +8, +9 and +10 for codon position 1, 2, and 3, respectively. In the table the 3′-noises are arranged in the order of codon positions. The χ^2 distances to noise are calculated for each site and its corresponding noise. The distances tell us that it is reasonable to take signal zone from site −6 to +6. This agrees with GenScan's Kozak signal region.[14] The Kozak consensus for the sequence set is GNRGCSatgGCG.

Table 1 Position-specific base distribution near the start codon. χ^2 distances of ATG neighboring sites from their corresponding noises are given in the last row.

	5′noise	−11	−10	−9	−8	−7	−6	−5
A	0.90	1.21	0.87	0.43	0.96	1.15	0.56	0.87
C	1.03	1.09	0.78	1.30	1.24	0.78	0.68	1.55
G	1.15	0.96	1.43	1.36	0.99	1.43	2.05	0.81
T	0.93	0.74	0.93	0.90	0.81	0.65	0.71	0.78
Distance		0.04	0.03	0.11	0.02	0.07	0.23	0.09
−4	−3	−2	−1	+1	+2	+3	+4	+5
1.49	0.99	1.40	0.74	4.00	0.00	0.00	0.53	0.71
0.96	0.43	1.67	1.49	0.00	0.00	0.00	0.31	2.08
1.18	2.14	0.53	1.55	0.00	0.00	4.00	2.76	0.84
0.37	0.43	0.40	0.22	0.00	4.00	0.00	0.40	0.37
0.19	0.39	0.35	0.33	2.63	2.31	1.75	0.38	0.29
+6	+7	+8	+9	+10	+11		3′noises	
0.37	1.09	0.62	0.28	0.96	0.90	0.83	0.86	0.33
0.87	0.87	1.49	1.46	0.71	1.15	0.88	1.20	1.63
2.17	1.30	1.21	2.05	1.58	1.02	1.59	0.87	1.56
0.59	0.74	0.68	0.22	0.74	0.93	0.71	1.07	0.48
0.17	0.03	0.10	0.09	0.02	0.01			

Table 2 Weight matrices for the best clustering of 129 sequences of start codon signal into 3 groups.

Pattern 1: GMSGAGatgGGG with size of 49												
A	0.24	0.90	0.73	0.65	1.88	0.82	4.00	0.00	0.00	1.39	1.31	0.33
C	0.90	1.55	1.31	0.33	1.55	1.22	0.00	0.00	0.00	0.41	0.00	1.14
G	2.12	0.73	1.63	3.02	0.33	1.88	0.00	0.00	4.00	2.20	1.80	2.37
T	0.73	0.82	0.33	0.00	0.24	0.08	0.00	4.00	0.00	0.00	0.90	0.16
Pattern 2: GNGAAGatgGCG with size of 39												
A	1.23	1.44	1.13	1.85	1.74	0.51	4.00	0.00	0.00	0.10	0.72	0.41
C	0.62	0.92	0.41	0.82	0.31	0.92	0.00	0.00	0.00	0.00	3.18	0.10
G	1.74	0.62	1.74	0.82	0.92	2.26	0.00	0.00	4.00	2.87	0.00	2.26
T	0.41	1.03	0.72	0.51	1.03	0.31	0.00	4.00	0.00	1.03	0.10	1.23
Pattern 3: GCAGCCatgGCG with size of 41												
A	0.20	0.29	2.73	0.59	0.49	0.88	4.00	0.00	0.00	0.00	0.00	0.39
C	0.49	2.15	1.07	0.10	3.12	2.44	0.00	0.00	0.00	0.39	3.41	1.27
G	2.34	1.07	0.10	2.44	0.39	0.49	0.00	0.00	4.00	3.32	0.59	1.85
T	0.98	0.49	0.10	0.88	0.00	0.20	0.00	4.00	0.00	0.29	0.00	0.49

Table 3 Weight matrix for the data set of 129 sequences of stop codon signal and weight matrices for its best clustering into 3 groups.

	3′noises		−4	−3	−2	−1	+1	+2	+3	+4	+5	+6	+7	+8	5′noise	
Single pattern: SNNCtrrNNNNN with size of 129																
A	1.03	1.24	0.58	0.56	1.05	1.30	0.50	0.00	1.95	3.04	1.27	0.78	1.30	1.30	0.87	1.07
C	0.82	1.06	1.48	1.43	0.78	0.96	1.74	0.00	0.00	0.00	0.65	1.12	0.84	0.65	1.15	0.86
G	1.43	0.77	1.13	1.30	1.02	0.78	0.90	0.00	2.05	0.96	1.12	0.78	0.84	0.96	1.02	0.89
T	0.72	0.93	0.81	0.71	1.15	0.96	0.87	4.00	0.00	0.00	0.96	1.33	1.02	1.09	0.96	1.19
Distance				0.01	0.09	0.00	0.03	2.79	1.37	1.98	0.05	0.05	0.02	0.03	0.05	
Pattern 1: NGCTtraKTNNW with size of 43																
A	1.15	1.27	0.57	0.74	0.37	0.65	0.74	0.00	1.95	4.00	0.65	1.49	1.30	1.40	1.12	1.11
C	0.75	1.12	1.40	1.12	1.02	1.86	1.30	0.00	0.00	0.00	0.19	0.19	0.65	0.93	0.56	0.86
G	1.37	0.67	1.25	0.93	1.67	0.84	0.09	0.00	2.05	0.00	1.58	0.37	1.12	1.02	0.84	0.82
T	0.72	0.94	0.78	1.21	0.93	0.65	1.86	4.00	0.00	0.00	1.58	1.95	0.93	0.65	1.49	1.21
Distance			0.10	0.25	0.23	0.73	2.78	1.45	3.00	0.42	0.42	0.06	0.11	0.05		
Pattern 2: GTACtagATNNN with size of 37																
A	0.93	1.14	0.66	0.32	0.54	1.51	0.22	0.00	4.00	0.65	1.95	0.22	1.30	1.41	0.54	1.02
C	0.90	1.11	1.45	1.51	0.43	0.65	2.05	0.00	0.00	0.00	1.19	1.19	0.86	0.76	1.19	0.83
G	1.41	0.91	1.03	1.84	1.30	0.76	0.97	0.00	0.00	3.35	0.76	0.76	0.86	0.97	1.19	0.90
T	0.77	0.84	0.86	0.32	1.73	1.08	0.76	4.00	0.00	0.00	0.11	1.84	0.97	0.86	1.08	1.25
Distance				0.29	0.32	0.11	0.17	2.72	2.23	1.77	0.66	0.36	0.03	0.07	0.13	
Pattern 3: CAACtgaNCNTC with size of 49																
A	1.00	1.29	0.53	0.57	2.04	1.71	0.49	0.00	0.41	4.00	1.31	0.57	1.31	1.14	0.90	1.06
C	0.82	0.98	1.58	1.63	0.82	0.41	1.88	0.00	0.00	0.00	0.65	1.88	0.98	0.33	1.63	0.87
G	1.51	0.73	1.10	1.22	0.24	0.73	1.55	0.00	3.59	0.00	0.98	1.14	0.57	0.90	1.06	0.95
T	0.67	1.00	0.78	0.57	0.90	1.14	0.08	4.00	0.00	0.00	1.06	0.41	1.14	1.63	0.41	1.12
Distance				0.02	0.65	0.15	0.34	2.85	2.16	3.06	0.03	0.43	0.06	0.17	0.29	

After having fixed the signal region, the optimal clustering divides the signal sequence set into 3 groups of sizes 49, 39, and 41. The weight matrices for these 3 patterns are listed in Table 2. Long-range correlation is clearly seen. The 3 patterns exhibit a significant difference in their GC-content: 0.68, 0.57, and 0.75 (without counting ATG). The GC-content (0.78, 0.76, and 0.86) of noise at the third codon position also shows the same tendency.

The total logarithmic likelihood, or the sum of scores for the whole set of signal sequences, obtained by using the single weight matrix model is 761. The value of the first-order Markov model is 810, while that of the above 3-pattern model is 967. We have examined the discriminant power of different models. 7911 pseudo-signal sequences are extracted from the gene sequence data set. Both the Markov model and multi-pattern model are superior to the single weight matrix model. For FN below 2% the Markov model gives smaller FP than the multi-pattern model, while for FN above 2% the multi-pattern model performs better. Using likelihood ratio to the single base

distribution of the pseudo-signal sequences always gives better discrimination than using likelihood.

Table 4 Nucleotide contents in the 3 codon usage tables found by maximizing the total likelihood function.

cluster	A	C	G	T
1	0.273	0.207	0.250	0.270
2	0.177	0.347	0.333	0.143
3	0.223	0.283	0.293	0.203

The termination signal of translation is rather weak. From the 129 termination signal sequences we determine the location of the signal zone to be the sites from -3 to $+8$ with the starting site of stop codons marked as site $+1$. The total logarithmic likelihood for the whole set of signal sequences has been calculated on sites from -4 to $+8$, which is of the same width 12 as the start codon signal. The value obtained by using the single weight matrix model is only 415. The optimal clustering divides the sequence set into 3 groups of sizes 43, 37, and 49. The likelihood value found for the 3 patterns is 629. The 31 sequences with stop codon TAG almost form the group 2, while group 3 mainly belongs to stop codon TGA. The position-specific distributions for single and multiple pattern models are listed in Table 3 together with distributions of noises. From the table we see that after clustering the signal is indeed enhanced and its width increased by one. Comparing GC content of the coding noises at each codon position among the 3 clusters, we see their different low(L)-high(H) patterns: LLH (0.53, 0.45, 0.66), HHL (0.58, 0.51, 0.62) and HLH (0.58, 0.43, 0.67).

4 Discussion

Weight matrix models have the advantages that they are simple, and use a small number of parameters. A main limitation of WMMs is the assumption of independence between positions. Correlations between positions weakened the information contained in weight matrix since subpatterns are mixed together by averaging. Clustering is an efficient way to deal with correlations. Cluster WMMs is simple and flexible. A parameter in cluster WMMs is the total number G of clusters. From the Jensen theorem for convex functions,[15,16] the object function L of the total likelihood increases with increasing G. However, involving the well-known issue of model selection, the optimal G is not determined directly by L,[11] and is strongly dependent on the size of data set.

The method of maximal dependence decomposition (MDD) is efficient in dealing with correlation. The MDD may be viewed as clustering according to a binary tree based on signal consensus. The general clustering algorithm may find a likelihood higher than MDD. We have analyzed 620 donor splice sequences extracted from the 129 full gene sequences. The total score of MDD is extremely close to that of the optimal clustering. As is seen from the above analysis for initial and terminal signals for translation, the multi-pattern model can be used even when signal consensus is very weak. Furthermore, the main idea of clustering can be applied to dealing with inhomogeneity in codon usage. We catenate coding sequences in the data set of 129 full sequences to get a sequence, and divide it into 1013 non-overlapping windows of width 120. By maximizing the total likelihood estimated from 3 codon usage tables, we divide the windows into 3 clusters, and at the same time obtain optimal 3 codon usage tables. The contrast among these 3 codon usage tables is rather strong. Their nucleotide contents are shown in Table 4, where the GC content difference is remarkable.

Cluster WMMs are useful not only for DNA sequences, but also for protein sequences. Protein sequences have an alphabet of size 20, which is 5 times higher than that of DNA sequences. Cluster WMMs provide a simple way to coarse-grain symbolic sequences with least cost in increasing the number of parameters. The use of cluster WMMs for protein sequences will be discussed elsewhere.

References

[1] M.S. Gelfand, J. Comput. Biol. **2** (1995) 87.

[2] L.R. Rabiner, Proc. IEEE **77** (1989) 257.

[3] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton, Science **262** (1993) 208.

[4] M. Goldstein and W.R. Dillon, *Discrete Discriminant Analysis*, John Willey & Sons, New York (1978).

[5] H. Solomon (ed.), *Studies in Item Analysis and Prediction*, Stanford Univerdity Press, Stanford (1961).

[6] R. Staden, Nucleic Acids Res. **12** (1984) 505.

[7] G.D. Stormo, T.D. Schneider, L. Gold, and A. Ehrenfeucht, Nucleic Acids Res. **10** (1982) 2997.

[8] S.L. Salzberg, D.B. Searls, and S. Kasif (eds.), *Computational Methods in Molecular Biology*, Elsevier, Amsterdam (1998).

[9] S. Kullback, J.C. Keegel, and J.H. Kullback, *Information Theory and Statistics*, Wiley, New York (1959).

[10] S. Kullback, *Topics in Statistical Information Theory*, Springer, Berlin (1987).

[11] T. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics*, KTK Scientific, Tokyo (1986).

[12] C.E. Lawrence and A.A. Reilly, Proteins **7** (1990) 41.

[13] J. Yu, *et al.*, Science **296** (2002) 79.

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

[15] T.M. Rassias, *Survey on Classical Inequalities*, Kluwer Academic, Dordrecht (2000).

[16] T.M. Rassias, and H.M. Srivastava, *Analytic and Geometric Inequalities and Applications*, Kluwer Academic, Dordrecht (1999).