Clustering and Detection of 5' Splice Sites of mRNA by *k* Weight Matrix Model

Katsuhiko Murakami [12] **Toshihisa Takagi** [1]
katsu@ims.u-tokyo.ac.jp takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2] Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

1 Introduction

Biological signals on DNA sequences such as TATA box, GC box, CAAT box, the Shine-Dalgarno sequence in the promoter regions, and splice sites (donor/acceptor sites) in eukaryotic mRNA are of considerable interest because they play numerous crucial roles in binding with proteins, or RNA. Recently it has been suggested that the 5' splice site recognition is performed *in vivo* through a combination of several rules which are still unclear [1].

2 Materials and Methods

The data we used was obtained from the GENSCAN training/test sets as described in reference [1]. We extracted both the actual and pseudo 5' splice sites from the data with the GT-AG rule (most of the introns start with GT and end with AG).

Biological signal sequences are traditionally characterized by the positional weight matrices (PWMs) introduced by Staden [2]. In this method, given an uncharacterized short sequence $X = x_1, x_2, ..., x_n$, the score of the sequence is calculated by the formula: $S_p(X) = \sum_{i=1...n, x_i \in A,C,G,T} \log(P(i, x_i))$, where $P(i, x_i)$ is the probability of generating the nucleotide x_i at position i for the site model under consideration. Here, a PWM is defined as a matrix which is constructed from $P(i, x_i)$ for all i, x_i.

In this work, we clustered the actual 5' splice sites using the PWM. The process of clustering sites is similar to the k-means clustering algorithm [3]. The first step in our clustering algorithm is to create two PWMs, which are constructed from all the positive data (authentic sites) and all the negative data (pseudo sites) respectively. Some PWMs (called random PWMs) are also constructed from random weights for each class, except the above two classes. Second, all positive data is distributed to the classes in such a way that the PWM of the class generates the distributed sequences with the highest probability among the classes (data distribution). Third, a new PWM is constructed for each class with the distributed data. Fourth, the each random PWM is modified slightly so that the random PWM is closer to the target PWM in the class (training). The data distribution procedure and training procedures of the PWMs are iterated. During the iteration, the score function T is maximized at each step, where T is defined as:

$$\begin{aligned} T &= \sum_{l=2}^{k} \sum_{m=1}^{N} \delta_{l,m} S_p(X_m) \\ &= \sum_{l=2}^{k} \sum_{m=1}^{N} \sum_{i=1, x_i \in A,C,G,T}^{n} \delta_{l,m} \ \log(P^l(i, x_i)), \end{aligned}$$

with the constraints $\sum_{x_i = A,C,G,T} P(i, x_i) = 1$ for all position i. X_m is a sequence of the training data, and $\delta_{l,m}$ is the delta function:

$$\delta_{l,m} = \begin{cases} 1 & \text{if } X_m \text{ belongs to the class } l \\ 0 & \text{otherwise.} \end{cases}$$

3 Results and Discussion

We have clustered the 5' splice sites into several classes. Each class was represented by the PWMs. Table 1 shows the motifs extracted from the PWMs. The characters are typed in upper case if the probability is more than 50%. If the probability is more than 35%, they are typed in lower case. The '-' indicates the exon-intron boundary. The motifs of class 2 to class 5 are new and different from the traditional consensus sequence (C1: a/cAG-GTa/gAGt).

Table 1: Consensus sequences extracted from the PWMs in the classes.

class	-3	-2	-1	—	+1	+2	+3	+4	+5	+6
C1	a/c	A	G	—	G	T	a/g	A	G	t
C2	c	C	a/t	—	G	T	a/G	A	G	T
C3	A	A	G	—	G	T	A	A	G	T
C4	D	g/T	G	—	G	T	a/G	A	G	T
C5	c	A	G	—	G	T	A	A	G	a/g

We applied the extracted PWMs for splice site detection. The specificity at different sensitivity levels are calculated for the test set. The specificity (SP) is defined using True Positive (TP) and False Positive (FP) by $SP = \frac{TP}{TP+FP}$. At the threshold level in which 20% of actual sites are detected, the specificity of the traditional PWM was only 50%, while that of our method was 52%. We observed especially high specificities for classes 2, 3, and 4 (59%, 57%, and 57%).

Acknowledgements

This work is partially supported by Grant-in-Aid for Scientific Research on Priority Areas, "Genome informatics" from the Ministry of Education, Science, Sports and Culture, Japan.

References

[1] Burge, C. and Karlin, S., Prediction of complete gene structures in Human genomic DNA, *J. Mol. Biol.*, 268:78–94, 1997.

[2] Staden, R., Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes, *Nucleic Acids Res.*, 12:551–567, 1984.

[3] Ripley, B.D., *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.