Pacific Symposium on Biocomputing 4:171-181 (1999)

Clustering and detection of 5' splice sites of mRNA by k weight-matrices model

K. Murakami
Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan

T. Takagi
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

We propose a novel method to detect 5' splice sites of eukaryotic mRNA. We have grouped the 5' splice splice sites into various classes. The clustered sites are represented by a set of PWMs. The clustering algorithm is similar to k-means clustering algorithm but the distance definition and the training score function were arranged. The clustered PWMs were applied to 5' splice site detection. The results showed an improvement in comparison with traditional single PWM. The result of the clusters suggests there are new motifs of 5' splice sites.

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 RNAs. Such signal sequences are widely known, but it is still difficult to detect these sites correctly from genomic DNA data alone [1,2,3], because such signals are vaguely defined. And though similar sequences are ubiquitous, only a part of them are actually recognized and work *in vivo* or *in vitro*. Such signal sequences are traditionally characterized by positional weight matrix (PWM) introduce by Staden [4]. The PWM method was extended to weight array matrix (WAM) by M. Q. Zhan [5] and S. L. Salzberg [3]. WAM model is based on di-nucleotide positional statistics while the PWM is based on single-nucleotide positional statistics. There are other approaches to the recognition of splice sites with a large size window, such as linear discriminant analysis [6] and neural networks [7,8]. In recognizing splicing site of eukaryotic mRNA, coding/non-coding potential (content analysis) is often used simultaneously in either an explicit or an implicit way and has succeeded to some extent. In biological systems, however, coding biases are unlikely to be used in recognition of splice sites, especially 5' splice sites. Thus, studying artificial recognition without content analysis is important to understanding biological systems. Recently C. Burge has introduced the maximal dependence decomposition (MDD) model for detection of 5' splice sites, which does not consider the coding biases[9]. The MDD model was used with coding biases in the gene finding system GENSCAN. This work suggested that there are strong correlations between some specific two or three positions with base constraints. For instance, there is a compensation relation between +5:G and -1:G in 5' splice sites. Here, (+5:G) represents a base G in the position +5. This suggests that the 5' splice site recognition is performed *in vivo* through a combination of several rules which are still unknown. The current study is not sufficient to elucidate the actual recognition mechanisms. Complete recognition of the splice site from a DNA/mRNA sequence has not been accomplished. Provided that there are some different recognition rules in a recognition mechanism, development of elucidation methods are needed.

There was an approach to divide the data set for parameter estimation of PWMs in Tsunoda's recent work[10]. However, they estimated the cut-off values of PWMs for transcription factor binding sites, not the weights of the PWMs. Moreover, they used a single PWM for each binding site, while we set the weights of several PWMs for the specific DNA sequence signal.

In this work, we have grouped the 5' splice splice site of eukaryotic mRNA into various classes. The clustered sites are represented by a set of PWM. The clustered results suggest that there are other motifs of the 5' splice sites. Furthermore, the clustered PWM are applied to 5' splice site recognition. The accuracy was better than the traditional method using single PWM, and comparable to WAM model, but inferior to MDD model.

2 Data and Methods

2.1 Data

The data we used was obtained from GENSCAN training/test sets collected for the gene-finding system as described in the references[9,11]. The data sets were arranged versions of other data sets, which were collected by D. Kulp (University of California at Santa Cruz) and M.G. Reese (LBNL, USA) from GenBank (Release 89, August, 1995) [ftp://ftp.cse.ucsc.edu/pub/dna/genes] and (GenBank Rel. 95, June, 1996) [ftp://www-hgc.lbl.gov/inf/genesets.html]. In the original data set, some sequences in the test set which were similar to ones in the training set were discarded from the test set by C. Burge.

The data sets were derived so that they met the following criteria[11]. Only sequences from homo sapiens were used in this analysis. There are no alternative splicing features. All splice sites of coding sequence (CDS) obey the GT-AG rule (intron start with GT and end with AG). The CDS starts with a start codon and ends with a in-frame stop codon. Database entries with annotations of alternative splicing, partial or putative CDS or ORF designation, viral or mitochondrial origin, were discarded. Furthermore, no two sequences in the data sets are substantially similar (BLAST score ≥ 100) to each other. The number of entries (genes) are 380 for the training set, and 65 entries for test set. The training set also includes some single-exon genes (i.e. intron-less genes).

We extracted both authentic and pseudo 5' splice sites from the data with the GT-AG rule (most of the intron start with GT and end with AG). Negative data of 5' splice sites were extracted from the sequences in such a way that each subsequence has a GT conservative dinucleotides at the starting point of intron (+1,+2), and has a sufficient base-length to be analyzed (3 bases for 5'-side, 6 bases for 3'-side).

We also took the 5' splice sites on 5' untranslated region (UTR) or 3' UTR that satisfy with GT nucleotide according to annotation of mRNA in the training data set. On the other hand, as for the test data set, we took the 5' splice sites only in CDS, so that we could compare our method with other methods described in the above paper[11] in terms of recognition accuracy.

2.2 Clustering using Positional Weight Matrix

Biological signal sequences are traditionally characterized by the PWM introduce by Staden[4]. In this method, given a uncharacterized short sequence $X = x_1, x_2, ..., x_n$, the PWM score $S_p(X)$ under positive model is calculated by the formula:

$$S_p(X) = \sum_{i=1...n, x_i \in A,C,G,T} log(P(i, x_i)), \tag{1}$$

where $P(i, x_i)$ is the probability of generating the nucleotide x_i at position i of the positive model of the signal. Here, a positional weight matrix (PWM) is defined as a matrix which is constructed from $P(i, x_i)$ for all i, x_i. Similarly, a score $S_n(X)$ for a negative model of the signal is defined using pseudo site samples, and the relative score of positive model against the negative model is often used as the effective score defined as:

$$R(X) = S_p(X) - S_n(X). \tag{2}$$

We clustered actual 5' splice sites using the PWM. Fig. 1 shows the overview of the clustering process, which is also the training process of several PWMs. The process is similar to the k-means clustering algorithm[12,13]. The main differences are 1) a training data sample is represented as a sequence,

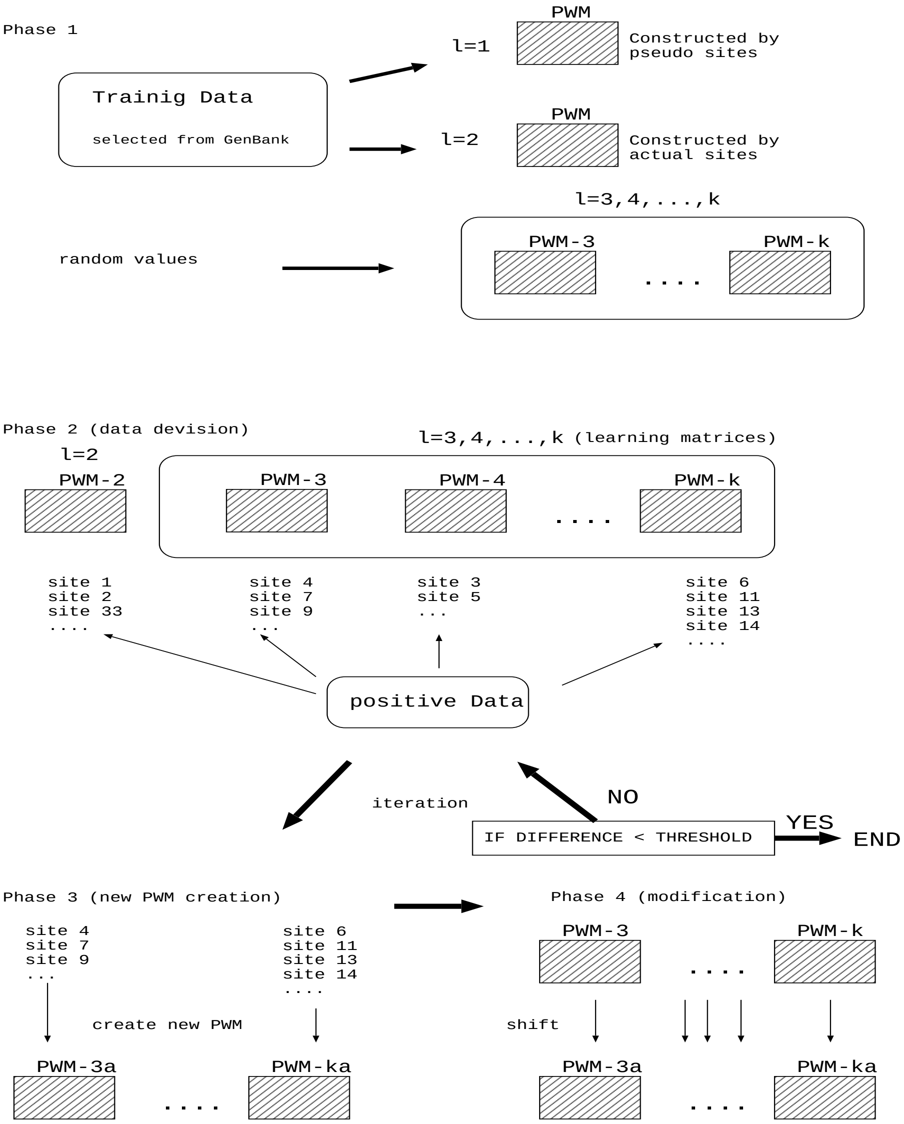

Figure 1: Overview of the process for the classification and PWM training.

not a point in a multi-dimensional space, and 2) a 'centroid' is not a vector but a PWM, and 3) the distance between the training data sample and the centroid is defined by the score (the probability of the sequence generated from the PWM). The training procedure is comprised of four phases. At the first phase, two PWM were constructed from the training data for both the positive data (subsequences around the actual 5' splice sites) and the pseudo sites (subsequences with 'GT' dinucleotides but not annotated as 5' splice sites in the database). The constructed PWM are called PWM-1 and PWM-2. In addition, some other PWM were created by random numbers. As the result, k PWMs are created in total. The random matrices are trained in the following phases.

The procedures from phase 2 to phase 4 are iterated until a convergence condition (described below) is satisfied. At phase 2, all positive subsequences are distributed to one of the PWM in such a way that the distributed PWM gives the highest score to the sequence. This score (generating probability) corresponds to the distance of famous k-means clustering algorithm.

The phase 3 creates new PWMs using assigned training sequences for each class. These new PWM are called PWM-3a, PWM-4a, and PWM-ka. These PWM become new goals for PWM-3, PWM-4, and PWM-k, respectively. PWM-1a and PWM-2a are not created because both existing models are already concrete models. PWM-1 corresponds to a pseudo model, and PWM-2 corresponds to a positive model which represents the overall features of the 5' splice sites. One of our aims is to extract new motifs (base biases) by producing some PWMs which are different from the consensus sequence and embedded in a single PWM. Therefore PWM-1 and PWM-2 are necessary as references in phase 2. Each $P^l(i,j)$ in PWM-l is updated at phase 4 by

$$P^l(i,j) \leftarrow P^l(i,j)(1-\alpha) + P^l_{mod}(i,j) * \alpha, \tag{3}$$

where $P^l(i,j)$ is the probability of the matrix PWM-l (l=3,4,..., k), $P^l_{mod}(i,j)$ is the probability of the matrix PWM-la (l=3,4,..., k), and α is a small real constant. After the modification (equation 3) is applied, $\sum_{j=A,C,G,T} P^l(i,j) = 1$ is kept, because $\sum_{j=A,C,G,T} P^l_{mod}(i,j) = 1$. The convergence condition is that the differences between PWM-l and PWM-la in terms of all probabilities are less than a threshold value θ. When we set θ to 0.01, then all the matrices were observed to converge within a few dozens iterations. Equation 3 realizes the maximization of the score function T:

$$\begin{aligned} T &= \textstyle\sum_{l=2}^{k} \sum_{m=1}^{N} \delta_{l,m} S_p(X_m) \\ &= \textstyle\sum_{l=2}^{k} \sum_{m=1}^{N} \sum_{i=1, j \in A,C,G,T}^{n} \delta_{l,m} \; log(P^l(i,j)), \end{aligned} \tag{4}$$

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

$$\delta_{l,m} = \begin{cases} 1 & \text{if } X_m \text{ belongs to the class } l \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Here $P^1(i,j)$ were not taken into account because it is a negative model constructed from the pseudo sites.

2.3 Detection of 5' splice sites with k-PWMs

We describe how a sequence is scored by this method using k-PWMs. Let $S_p(X|l)$ be a score of PWM-l calculated as described above. The sequence is classified to a class which gives the best score to the sequence, and the class l is defined by a specific PWM-l. We define here a score $S_{path}(X)$ as:

$$S_{path}(X) = \Sigma_{l=2}^{k} S_p(X|l)P(l) \tag{6}$$

where $P(l)$ is a prior probability that a sequence is generated from class l PWM. Therefore $S_{path}(X)$ is interpreted as a sum of probabilities throughout all the positive classes. The final score of a sequence was defined as:

$$R(X) = S_{path}(X) - S_p(X|1), \tag{7}$$

where $S_p(X|1)$ was a score calculated by the negative model. We tested for various k ranging from 3 (the least relevant value for this method) to 6. If $k = 2$, it is the same model as the single PWM model with the training of negative data. Five tests were performed for each k. The best set of k-PWMs will be shown in the next section.

3 Results and Discussion

3.1 PWM and motifs

Table 1 shows the PWM of each class extracted by this method. Class 1 is the same as the traditional PWM. The consensus sequence is a/cAG-GTa/gAGt. In the expression, 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 standard consensus is visualized in Fig 2 using sequence logo[14].

C. Burge indicated in the MDD model that there is a compensation relation between +5:G and -1:G in 5' splice sites[9]. This effect was seen in classes 2, 3,

Table 1: The extracted PWM by the k-PWMs method (k=6). The PWM-1 corresponds to a pseudo site model and are not shown. PWM-2 (class 1) corresponds to the traditional PWM. Class 2 to class 5 are trained and produced. Each cell shows the percentage of the base for the position.

position	A	C	G	T	position	A	C	G	T
		class 1					class 2		
-3	31.4	37.6	18.7	12.3	-3	24.4	31.3	27.8	16.6
-2	60.4	13.0	13.8	12.8	-2	31.2	52.4	16.3	0.1
-1	8.2	3.5	81.1	7.1	-1	30.8	10.8	24.5	33.9
1	0.0	0.0	100.0	0.0	1	0.0	0.0	100.0	0.0
2	0.0	0.0	0.0	100.0	2	0.0	0.0	0.0	100.0
3	49.0	3.0	45.1	2.9	3	42.6	1.4	56.0	0.0
4	71.3	7.7	12.4	8.6	4	96.9	0.5	1.8	0.9
5	6.4	4.9	84.1	4.6	5	0.0	0.0	100.0	0.0
6	16.0	17.7	20.0	46.4	6	0.3	16.4	1.3	82.0
		class 3					class 4		
-3	36.9	41.3	15.1	6.8	-3	32.6	29.4	0.0	38.0
-2	80.9	2.8	10.6	5.7	-2	0.5	0.0	40.9	58.6
-1	1.9	0.0	97.2	1.0	-1	0.0	4.4	90.7	4.8
1	0.0	0.0	100.0	0.0	1	0.0	0.0	100.0	0.0
2	0.0	0.0	0.0	100.0	2	0.0	0.0	0.0	100.0
3	73.9	3.5	15.1	7.5	3	48.5	0.0	50.3	1.2
4	9.7	0.1	0.0	90.2	4	87.7	1.5	7.9	2.9
5	3.8	6.3	69.8	20.1	5	0.0	1.8	98.2	0.0
6	14.2	14.1	13.9	57.7	6	12.5	17.6	12.4	57.5
		class 5							
-3	34.2	47.5	16.4	1.8					
-2	98.4	0.2	1.4	0.0					
-1	0.4	0.0	99.6	0.0					
1	0.0	0.0	100.0	0.0					
2	0.0	0.0	0.0	100.0					
3	81.3	6.7	12.0	0.0					
4	57.8	21.0	21.2	0.0					
5	23.3	13.8	54.7	8.2					
6	34.1	22.2	36.1	7.6					

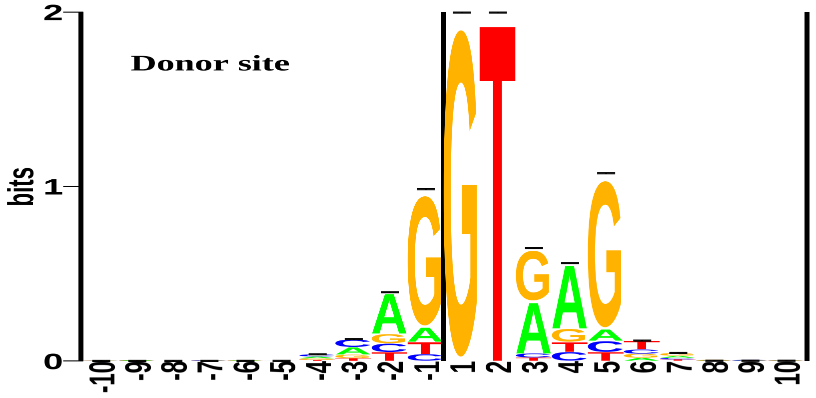

Figure 2: The sequence logo of standard consensus of 5' splice site. Y-axis indicates the strength of base composition bias based on information content.

and 5. Class 2 and 3 prefer +5:G and class 5 prefers -1:G. The class 2 showed a high specificity and has a sufficient amount of samples (18% $\simeq$ 1/5), and has a strong bias of (-2:C, -1:W, +4:A, +5:G, +6:T) in comparison with the standard bias (class 1). W indicates A or T. Class 3 shows (-2:A, -1:G, +3:A, +4:T) bias, although the sample size of the class is small. Class 4 shows (-3:D, -2:T) bias. Finally, Class 5 shows (-2:A, -1:G, +3:A, +5:A, 6:a/G) bias. Except class 1, the classes are apparently different from the standard consensus (class 1 PWM).

3.2 Performance of the signal detection using k-PWMs

Table 3 shows the specificity for different sensitivity levels for the test set. Specificity is defined using True Positive (TP) and False Positive (FP) by $SP = \frac{TP}{TP+FP}$. It should be noted that it is easy to filter out more than 85 % of the negative data, when traditional PWM detects 95% of the true sites. Due to the huge amount of pseudo sites, even about 10% of the negative data is comparable to the number of all positive data. Therefore the usual specificity defined by $SP = \frac{TN}{TN+FN}$ was not used here. In Table 3, the other two results of other existing approaches for detection of 5' splice sites are also shown. The WAM model is based on di-nucleotide positional statistics while the PWM is based on single nucleotide positional statistics. The MDD model is a elaborated approach, but it still employs the PWM in the methodology.

Table 2: Consensus sequences of extracted PWMs. PWM-2 (class 1) corresponds to the traditional PWM. Class 2 to class 5 are trained and produced. 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. D indicates A,G,or T. '—' indicates exon-intron boundary.

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

There are other approaches, such as neural networks[7,8] or linear discriminant analysis[6]. However we do not compare our method with them, because they need a large window size to discriminate true sites from false sites and deal with content information of coding/non-coding regions. Therefore they cannot be compared with our methods directly. In this study, we focus on the 5' splice site information itself, not the content information. From Table 3, the total accuracy of this method is better than the results when PWM only was used, although it did not reached the accuracy of MDD. In the model 'All', the current method treats every class equally.

Since almost all classes have a high specificity at low sensitivity level (i.e. 20%) except for class 2 (whose sample size is relatively small), core deviations of the class from the consensus sequence were extracted successfully and the extracted motifs are presumed relevant.

Since each class has different specificity and different sample populations, there is still room to improve the specificity for this method. For instance, classes 2, 4, and 5 show higher specificity than the other models especially at 20% sensitivity level. If these classes are specified, we could detect more true sites more effectively. Most classes contained a sufficient amount of data. In the cases where there is a small amount of data in some classes, the statistics are less reliable and this can lead to failure of both motif extraction and site detection. To avoid this event, one approach would be to underestimate or ignore some classes which have small sample sizes.

4 Conclusion

We have proposed a method to classify biological signal sequences and applied it to 5' splice sites classification. The classes were represented as a set of PWMs

Table 3: Specificity for different sensitivity levels for the test set (k=6). The model ’ALL’ represents the results using class 1 (C1) to class 5 (C5). The model PWM indicates the result when only traditional PWM is used. The results of weight array matrix (WAM) and maximal dependence decomposition (MDD) is referred from Burge’s results.

Model	training true set (%)	test true set (%)	Sensitivity level			
			95%	90%	50%	20%
C1	33	33	2.5	6.6	40.5	49.0
C2	18	22	20.0	21.0	47.5	59.3
C3	9	6	3.8	3.8	17.2	17.2
C4	17	16	8.2	10.8	33.7	57.1
C5	23	22	8.6	9.3	19.9	57.1
ALL	100	100	7.6	10.7	30.3	51.5
PWM	–	100	7.1	9.4	32.0	50.0
WAM	–	100	7.2	12.4	33.0	49.6
MDD	–	100	8.7	13.4	36.0	54.3

and they were applied to 5’ splice site detection. The detection results of k-PWMs model proved better than that of the traditional single PWM model. The k-PWMs model seems comparable to WAM model, but is less specific than the MDD model in terms of site detection. However this approach can extract new motifs as shown in Table 1. This method is applicable to other type of biological signal sequences, such as transcriptional factor binding sites with variations (TATA box, GC box, and Shine-Dalgarno sequence), translation initiation and positional specific motifs in amino acid sequences.

Acknowledgments

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. S. Matis, Y. Xu, M. Shah, X. Guan, J. R. Einstein, R. Mural, and E. C. Uberbacher. Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Computers Chem.*, 20(1):135–140, 1996.

2. M. Q. Zhang. Identification of human gene core promoters *in silico.* *Genome Research*, 8:319–326, 1998.
3. S. L. Salzberg. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Applic. Biosci.*, 13(4):365–376, 1997.
4. R Staden. 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.
5. M.Q. Zhang and T.G. Marr. A weight array method for splicing signal analysis. *Comput. Applic. Biosci.*, 9:499–509, 1993.
6. V. V. Solovyev, A. A. Salamov, and C. B. Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, 22(24):5156–5163, 1994.
7. M. G. Reese and F. H. Eeckman. Improved splice site detection in genie. *J. Comp. Biol.*, 4(3):311–324, 1997.
8. S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, 220:49–65, 1991.
9. C. Burge and S. Karlin. Prediction of complete gene structures in Human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.
10. T. Tsunoda and T. Takagi. A new cut-off estimating algorithm for transcription factor bindability on DNA. In *The First International Conference on Bioinformatics of Genome Regulation and Structure*, volume 1, pages 199–202, Novosibirsk - Altai Mountains, Russia, August 1998.
11. C. Burge. *Identification of complete gene structures in human genomic DNA.* PhD thesis, Stanford University, Stanford, USA, 1997.
12. R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* A JOHN WILEY & SONS, INC., PUBLICATION, Menlo Park, California, 1973.
13. B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.
14. T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.