ibc Interdisciplinary Bio Central

IBC 2012;4:14, 1-6 • DOI: 10.4051/ibc.2012.4.4.0014

Open Access

REVIEW

Survey on Nucleotide Encoding Techniques and SVM Kernel Design for Human Splice Site Prediction

A.T.M. Golam Bari[1], Mst. Rokeya Reaz[1], Ho-Jin Choi[2] and Byeong-Soo Jeong[1]*

[1]Department of Computer Engineering, Kyung Hee University, Suwon, Korea
[2]Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

Subject areas; Bioinformatics/Computational biology/Molecular modeling, Omics (Physiomics/metabolomics/proteomics/genomics)

Author contribution; The first author (A.T.M.G.B.) studied extensively the domain area, conglomerated the knowledge of recent works and summarized. The co-author (Mst.R.R.) analyzed some statistical data got from various papers as well as some future research direction. The third author (H.J.C.) is the co-supervisor and the last author (B.S.J.) is our supervisor. Both of them reviewed the paper and gave us direction to improve it. This work is undone unless their sincere suggestion and cooperation.

***Correspondence** and requests for materials should be addressed to B.S.J. (jeong@khu.ac.kr).

Editor; Keun Woo Lee, Gyeongsang National University, Korea

Received December 12, 2012
Revised December 30, 2012
Accepted December 31, 2012
Published December 31, 2012

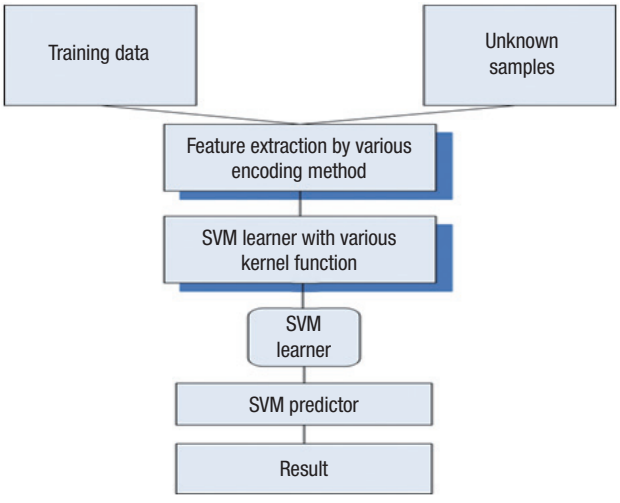
Citation; Bari, A.T.M. G., et al. Survey on Nucleotide Encoding Techniques and SVM Kernel Design for Human Splice Site Prediction. IBC 2012, 4:14, 1-6.
doi: 10.4051/ibc.2012.4.4.0014

Funding; NIPA (National IT Industry Promotion Agency); Fostering Global IT Human Resource Project and the National Research Foundation (NRF) grant (No. 2011-0018264) of Ministry of Education, Science and Technology (MEST) of Korea.

Competing interest; All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

Splice site prediction in DNA sequence is a basic search problem for finding exon/intron and intron/exon boundaries. Removing introns and then joining the exons together forms the mRNA sequence. These sequences are the input of the translation process. It is a necessary step in the central dogma of molecular biology. The main task of splice site prediction is to find out the exact GT and AG ended sequences. Then it identifies the true and false GT and AG ended sequences among those candidate sequences. In this paper, we survey research works on splice site prediction based on support vector machine (SVM). The basic difference between these research works is nucleotide encoding technique and SVM kernel selection. Some methods encode the DNA sequence in a sparse way whereas others encode in a probabilistic manner. The encoded sequences serve as input of SVM. The task of SVM is to classify them using its learning model. The accuracy of classification largely depends on the proper kernel selection for sequence data as well as a selection of kernel parameter. We observe each encoding technique and classify them according to their similarity. Then we discuss about kernel and their parameter selection. Our survey paper provides a basic understanding of encoding approaches and proper kernel selection of SVM for splice site prediction.



Key Words: coding sequence; exon-intron boundary; intron-exon boundary; splice site; support vector machine; translation process

INTRODUCTION

The basic gene structure for higher eukaryotes includes promoter, start codons, introns, exons, and stop codons[1]. Of them, the exon sequences are coding region of the gene. The translation of protein requires removal of introns from DNA sequences. The process for intron removal is called DNA splicing. The name of the position where intron to exon or exon to intron occurs is splicing site. The 5′ boundary or donor site of introns in most eukaryotes usually contains the di-nucleotide GT, while the 3′ boundary or acceptor site contains the di-nucleotide AG. This is called GT-AG rule[2]. Accurate prediction of splice sites is the very first step for a systematic study of eukaryotic genes. It has been recognized that accurate prediction of eukaryotic gene structure largely depends on the ability to pinpoint the exact splice sites within a sequence[3]. The advances in sequencing technologies have resulted in a large amount of DNA sequence information and therefore the size of genetic and genomic database has drastically increased. We cannot determine which regions of the gene would encode for protein unless different regions of the genome and their functions are characterized[4]. This is why, the annotation of the genome sequence within an acceptable timeframe is an important goal for the study of Bioinformatics. Gene expression in eukaryotes starts with the transcription of DNA sequences into pre-mRNA sequences, followed by the processing of pre-mRNAs to mature mRNAs, and the translation of mRNAs to proteins. Splicing is one of the primary post-processing steps of pre-mRNAs in eukaryotes.

Since 1980s, many computational biotechnology has been applied for locating gene-coding regions (exons). Several machine learning approaches have so far been introduced for the prediction of donor and acceptor splice sites and the secondary structure of proteins[5,6]. Here, we only focus on the support vector machine approach for finding splice site. The SVM method, which is a canonical machine learning approach, was initially proposed by Vapnik et al., is a very effective method for general purpose pattern recognition[3,7-12]. In this paper, we compare some research works on splice site prediction that use SVM for identifying true acceptor/donor sites and false acceptor/donor sites. The GT-AG rule of does not always hold[2]. So as discussed in, it is natural to model the prediction of splice sites[4] as a binary classification problem, using DNA sequences with experimentally confirmed splice sites as positive training examples and those DNA sequences with GT-AG structure but confirmed not to be real splice sites as negative training examples.

Successful recognition of splice junction sites of human DNA sequences was achieved via three machine learning approaches. Both unsupervised (Kohonen's Self-Organizing Map, KSOM) and supervised (Back-propagation Neural Network, BNN; and Support Vector Machine, SVM) machine learning techniques were used for classifying sequences of the test set into one of the three categories: transition from exon to intron (E/I boundaries), transition from intron to exon (I/E boundaries), and no transition[9]. The performance of machine learning techniques have accuracy of the comparative as follows SVM > BNN > KSOM, suggesting that SVM is a robust method of identifying unknown splice sites. It is expected that the SVM can provide a powerful computational tool for predicting the splice junction sites of uncharacterized DNA.

This paper analyzes the research works on different encoding approaches for splice site prediction using SVM and suggests some coding techniques along with kernel selection for better prediction accuracies.

PROBLEM DEFINITION

DNA Sequence

A deoxyribonucleic acid (DNA) is composed of four types of bases that are Adenine (A), Cytosine (C), guanine (G) and thymine (T). So, a DNA sequence is a string containing those four alphabets. As for example AGCATACGTACTGAC is a DNA sequence.

The central dogma of molecular biology

The central dogma of molecular biology explains the transfer of sequential information in details. It states that such information cannot be transferred back from protein to either protein or nucleic acid[13]. According to Marshal Nirenberg, the dogma can be defined as precisely "DNA makes RNA makes Protein"[14]. The schematic diagram of the central dogma[4] is depicted in Figure 1.

Splice Junction Sites

DNA splice junction sites (Figure 2) are boundaries where splicing occurs and are found between the regions of DNA that code for gene products (exon) and those that do not (intron)[15].

The GT di-nucleotide is usually referred to as "donor" whereas the AG di-nucleotide is known as "acceptor"[16]. The donor and acceptor are sketched[17] in Figure 3.

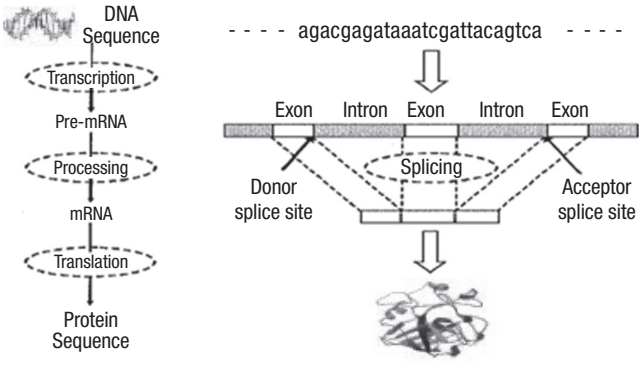


Figure 1. Central dogma of molecular biology.

Support Vector Machine

The SVM is a data driven method for solving bi-nominal classification tasks. Let a dataset T contains *l* instances of x_i (i = 1, ...*l*) with each x_i labeled as y+ or y- indicating a positive or negative instances respectively. The Linear SVM (LSVM) separates the two classes in T with a hyper plane in the feature space such that:

(A) The 'largest' possible fraction of instances of the same class is on the same side of the hyperplane, and
(B) The distance of either class from the hyperplane is maximal.

The prediction of LSVM for an unseen instance Z is 1 (classified as a positive instance) or -1 (classified as a negative instance), given by the decision function[4]

$$\text{Prediction}(z) = \text{sign}(\mathbf{W}\cdot\mathbf{Z} + b) \quad \text{----- (I)}$$

The hyper plane is computed by maximizing a vector of Lagrance multipliers α in

$$W(\alpha) = \sum_{i=1}^{l} \alpha i - 1/2 \sum_{i,j=1}^{l} \alpha i\, \alpha j\, yi\, yj\, K(xi,xj) \quad \text{----- (II)}$$

Constrained to: $0 \le \alpha i \le C$ and $\sum_{i=1}^{l} \alpha i\, yi = 0$ Where C is a parameter set by the user to regulate thoutliers of outliers and noise, i.e., it defines the meaning of the word 'largest' in (A).
For the LSVM this function reduces to the following

$$W = \sum_{i=1}^{l} \alpha i\, xi\, yi \quad \text{----- (III)}$$

All x_i for which α is not zero are called the support vectors. Typically the size of the set of support vectors is much smaller that l.

In this survey, we compare existing methods and also give further direction to improve encoding approaches as well as proper kernel functions and their parameter selection for splice site detection programs[18]. In there are lots of gene prediction programs which are based on Hidden Markov Model (HMM), Maximal Dependence Decomposition (MDD), Neural Networks (NN) and so on. These are summarized in the Table 1.

The main goal of the above programs is to find intron/exon and exon/intron boundaries, not to find the gene structure. SPLICEVIEW is based on prediction of splice signals by classification approaches (a set of consensus). Its two main assumptions are reflection of functional importance from higher frequency of some nucleotides in definite site position and mutual dependence of nucleotides in different site positions. Its accuracy ranges from 95% to 97% of donor and 95% acceptor for different organisms. SPLICEPREDICTOR applied log linear models to find optimal combinations of splice site variables for the purpose of separating true splice sites. Inhomogeneous zero order Markov model per position indicates the probability that a given base appears at each position of the splice signal. This concept is also known as a so called position weight matrix (PWM). NNSPLICE and NETGENE2 optimized PWM weights by neural network (NN) method. The higher Markov model can be used to capture possible dependencies between adjacent positions of a splice signal. This is so called weight array model (WAM) and closely related to position-dependent codon frequency model. Maximal dependence decomposition (MDD)

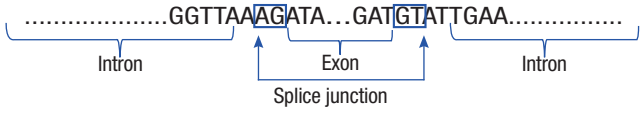


Figure 2. Schematic representation of the splice junction site.

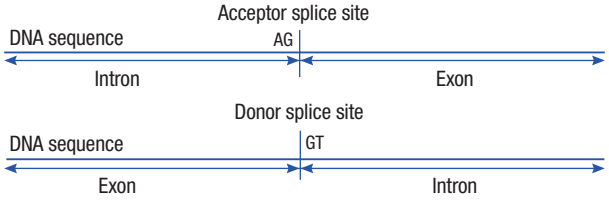


Figure 3. Illustration of acceptor and donor sites.

Table 1. Splice site prediction program

Program	Organism	Method
GeneSplicer[19]	Arabidopsis, human	HMM + MDD
NETPLANTGENE[20] (http://www.cbs.dtu.dk/services/NetPGene/)	Arabidopsis	NN
NETGENE2[21] (http://www.cbs.dtu.dk/services/NetGene2/)	Human, *C.elegans*, Arabidopsis	NN + HMM
SPLICEVIEW[22] (http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html)	Eukaryotes	Score with consensus
NNSPLICE0.9[23] (http://www.fruitly.org/seq_tools/splice.html)	*Drosophila*, human or other	NN
SPLICEPREDICTOR[24,25] (http://bioinformatics.iastate.edu/cgi-bin/sp.cgi	Arabidopsis, maize	Logitlinear models: (i) score with consensus; (ii) local composition
BCM-SPL (http://www.softberry.com/berry.phtml; http://genomic.sanger.ac.uk/gf/gf.html)	Human, Drosophila, *C.elegans*, yeast, plant	Linear discriminant analysis

method captures the most significant dependencies between adjacent as well as non-adjacent positions. The Arabidopsis database is used to evaluate GeneSplicer with NetGene2, NNSplice and SpliceView[19]. But GeneSplicer performed comparatively best alternative in each case in terms of accuracy and sensitivity.

None of the above programs are based on SVM but SVM gives the higher accuracy. In SVM produced 99.09% accuracy whereas the other two methods KSOM (Kohonen's Self Organizing Map) and BNN (Back-propagation Neural Network) estimated 92.72% and 97.27% accuracy respectively[9]. Chao-Hsien Chu et al. showed that SVM out performed better than Naïve Bayes Classifier[3]. They used two data sets, Dsmall and Dlarge and applied their method. SVM gives better performance in terms of CPU time, F-measure and other performance metric. Recently Qingshan Jiang et al. used their new feature extraction method in HS3D (Homo Sapiens Splice Site Dataset) data set and compared their method (ECS_SVM) with other methods like weight matrix model 1 (WMM1) and the first order Markov model (MM1). Both models are based on BPNN and RBFN[26]. Their method with SVM shows better. The accuracy of those methods is summarized in Table 2.

By exploring past and recent works, we can conclude that support vector machine produces more accurate results to predict splice site than other machine learning approaches. Furthermore, we investigated that different encoding approaches have different accuracy for splice site prediction. As a result, we scrutinized splice site prediction using SVM as well as the importance of encoding mechanism to produce better result. Lastly, we described the impact of various kernel functions of SVM for further development in prediction result.

Most of the splice site predictor programs that use SVM as true and false splice site classifier use Figure 4 as a working flow for their main principle.

The training data are experimentally verified true and false splice site. The experiment is done in biotechnology lab. The different algorithm uses different encoding approaches to extract features from training data. Then SVM is trained with those features using various kernel functions. After learning phase, candidate data (unknown samples) are fed into SVM learner. Then SVM predicts result based on its learning. From the above diagram, we can easily say that the accuracy of these problems can be increased by improving two steps – 'feature extraction by various encoding method' and 'SVM learner with various kernel functions'.

In sparse encoding, each nucleotide represents a 4 bit vector. As for example, A is represented as 1000, T is 0001, C is 0100 and G is 0010. FDTF stands for frequency distribution between true site and false site. This method represents each type of splice site in sparse encoding. Then makes a frequency distribution table for the true and false site. After that, the difference of distribution between true and false sites are calculated. This difference is called FDTF. Baye's mapping uses same frequency distribution as FDTF. This method uses concatenation of the false site with true sites, not the difference. Codon is the combination of three nucleotides. There are 64 codons. The sequential information (Information gain) is combined with codon useage[26].

From the Table 3, we can conclude some facts. Firstly, orthogonal encoding is the best approach as data representation for SVM input. Because it gives more vicinity information among nucleotides and this approach is perfect for representing vector data. Though this type of coding requires more memory than MN and PN encoding, it extracts some more features of DNA sequences. Secondly, sparse encoding with FDTF or Baye's

Table 2. Accuracy comparison between ECS_SVM and other methods

Splice site boundaries	WMM1-BPNN	WMM1-RBFN	MM1-BPNN	MM1-RBFN	ECS_SVM
Exon/intron	0.895	0.893	0.927	0.930	0.948
Intron/Exon	0.880	0.878	0.925	0.921	0.937

Table 3. Research works on splice site prediction using SVM

Coding	Research	Year	Kernel	Parameters	Accuracy	
					IE	EI
Sparse	Virapong[9]	2003	RBF	$C = 2^{0.75}$, $Gama = 2^{-5}$	99.09	97.98
	Ying-Fei[3]	2005	Gaussian, polynomial	S = 1, r = 1, D = 4, Std = 20	93.74-92.04	88.15-88.69
Markov model	Baten[17]	2006	RBF, poly, linear	NA	97-98	96-97
FDTF	T.Li l[1]	2006	RBF	NA	93.7	93.2
Baye's mapping	Chu[4]	2006	Linear, polynomial	d = 2,3 C = 150	86.6-89.8	86.6-89.8
Codon+ sequential	Qingshan jiang[26]	2012	Multiclass SVM	NA	93.7	94.8

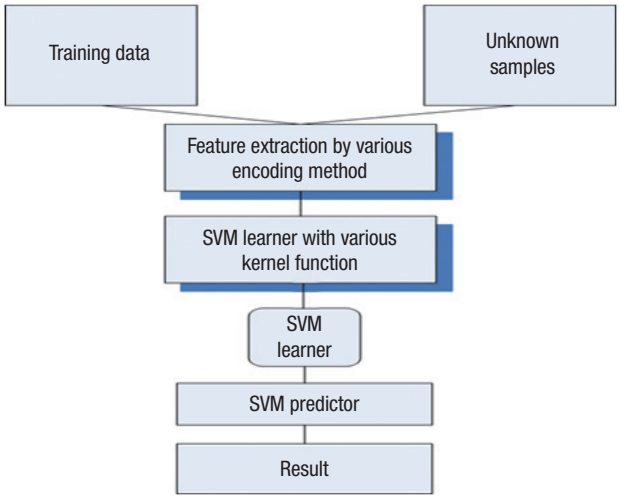


Figure 4. Flow diagram for splice site prediction using SVM.

mapping can tell probability of each nucleotide in each specific position. This probability can help us to determine positions of GC-content of exon sequence. To find out the density of GC-content is very important because GC-content of exon is typically higher than intron. Sparse encoding with FDTF has better accuracy than sparse encoding with Baye's mapping. Baye's mapping uses double length of candidate sequence to represent vector data than FDTF which is computationally expensive. In addition, SVM-kernels have also a great impact on accuracy. Most of the programs use linear, polynomial and Gaussian kernel function for this type of classification.The parameter selection (C and gamma) of kernel function is also a heuristic approach for better performance. Finally, most of the research work thinks this problem as a binary classifier – true and false class. Basically, there is another class which has no transition like intron/exon and exon/intron. We can say this class as 'no transition' class. A new kernel design (*l*- mer content describing kernel, positional information kernel, etc.) for splice site prediction with three class prediction can help to predict better results in future. Finally, TL = LL + RL rule is a brute force approach to deduce the best length of a candidate sequence as a splice site. Determination of LL and RL is also a trial and error method as like as the C and gamma selection of SVM kernel. A statistical approach to determine LL and RL can save time and effort of the researchers.

CONCLUSION

Different works use different types of encoding approaches for a DNA sequence. The encoded sequence acts as input to support vector machine. The performance of splice site prediction mostly depends on encoding approaches as long as the kernel and their parameter selection. In this survey, we discuss the encoding type of different research works and compare them.The better approaches for encoding should take care about information retention and less memory allocation as well as better performance on splice site prediction.

ACKNOWLEDGEMENTS

This work was supported by a grant from the NIPA (National IT Industry Promotion Agency) in 2012. (Fostering Global IT Human Resource Project) and the National Research Foundation (NRF) grant (No. 2011-0018264) of Ministry of Education, Science and Technology (MEST) of Korea.

REFERENCES

1. Huang, J., Li, T., Chen, K., and Wu, J. (2006). An approach of encoding for prediction of splice sites using SVM. *Biochimie* 88, 923-929.
2. Thanaraj, T. A., and Clark, F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res* 29, 2581-2593.
3. Sun, Y. F., Fan, X. D., and Li, Y. D. (2003). Identifying splicing sites in eukaryotic RNA: support vector machine approach. *Comput Biol Med* 33, 17-29.
4. Zhang, Y., Chu, C. H., Chen, Y., Zha, H., and Ji, X. (2006). Splice site prediction using support vector machines with a Bayes kernel. *Expert Syst Appl* 30, 73-81.
5. Hua, S., and Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol* 308, 397-407.
6. Ogura, H., Agata, H., Xie, M., Odaka, T., and Furutani, H. (1997). A study of learning splice sites of DNA sequence by neural networks. *Comput Biol Med* 27, 67-75.
7. Cristianini, N., and Shawe-Taylor, J. (2000). An introduction to support vector machines: and other kernel-based learning methods. New York: Cambridge University Press.
8. Vapnik, V. (1995). The nature of statistical learning theory. Springer-Verlag New York, Inc.
9. Nantasenamat, C., Thanakorn, N., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2005). Recognition of DNA splice junction via machine learning approaches. *Excli Journal* 4, 114-129.
10. Vapnik, V. N. (1998). Statistical learning theory. New York: Wiley.
11. Drucker, H., Wu, D., and Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Trans Neural Network* 10, 1048-1054.
12. Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. Machine Learning. p. 273-297.
13. Wikipedia (2012). Central dogma of molecular biology. Wikipedia, The Free Encyclopedia [cited 2012 Jul 7]. Available from: http://en.wikipedia.org/w/index.php?title=Central_dogma_of_molecular_biology&oldid=522262643.
14. Leavitt, S. A. (2010). Deciphering the Genetic Code: Marshall Nirenberg. Office of NIH History [cited 2012 Jul]. Available from: http://history.nih.gov/exhibits/nirenberg/.
15. Hastings, M. L., and Krainer, A. R. (2001). Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol* 13, 302-309.
16. Snyder, E. E., and Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. *J Mol Biol* 248, 1-18.
17. Baten, A. K., Chang, B. C., Halgamuge, S. K., and Li, J. (2006). Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics* 7 Suppl 5, S15.
18. Mathe, C., Sagot, M. F., Schiex, T., and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* 30, 4103-4117.
19. Pertea, M., Lin, X., and Salzberg, S. L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* 29, 1185-1190.
20. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouze, P., and Brunak, S. (1996). Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* 24, 3439-3452.
21. Tolstrup, N., Rouze, P., and Brunak, S. (1997). A branch point consensus from Arabidopsis found by non-circular analysis allows for better

prediction of acceptor sites. *Nucleic Acids Res* 25, 3159-3163.

22. Rogozin, I., and Milanesi, L. (1997). Analysis of donor splice sites in different eukaryotic organisms. *J Mol Evol* 45, 50-59.
23. Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. Proceedings of the first annual international conference on Computational molecular biology; Santa Fe, New Mexico, United States. USA: ACM. p. 232-240.
24. Brendel, V., Kleffe, J., Carle-Urioste, J. C., and Walbot, V. (1998). Prediction of splice sites in plant pre-mRNA from sequence properties. *J Mol Biol* 276, 85-104.
25. Kleffe, J., Hermann, K., Vahrson, W., Wittig, B., and Brendel, V. (1996). Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucleic Acids Res* 24, 4709-4718.
26. Wei, D., Zhuang, W., Jiang, Q., and Wei, Y. (2012). A new classification method for human gene splice site prediction. Proceedings of the First international conference on Health Information Science; Beijing, China. Springer-Verlag. p. 121-130.