MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **71** (2014) 241-258

ISSN 0340 - 6253

Effective DNA Encoding for Splice Site Prediction Using SVM

A. T. M. Golam Bari, M. Rokeya Reaz, Byeong-Soo Jeong*

Department of Computer Engineering, Kyung Hee University

1-Seocheon-dong, Gyeonggi-do, Yongin-si 446-701, Republic of Korea

{bari, rokeya, jeong}@khu.ac.kr

(Received May 13, 2013)

Abstract

Splice site prediction in the pre-mRNA is a very important task for understanding gene structure and its function. To predict splice sites, SVM (support vector machine)-based classification technique is frequently used because of its classification accuracy. High performance of SVM largely depends on DNA encoding method. However, existing encoding approaches do not reveal the characteristics of DNA sequences very well enough to provide as much information as sequences have. In this paper, we propose new effective DNA encoding method for feature extraction which can give more information of DNA sequence. Our encoding method can provide density information of each nucleotide along with positional information and chemical property. Extensive performance study shows that the proposed method can provide better performance than existing encoding methods based on several performance criteria such as classification accuracy, sensitivity, specificity and auROC (area under receiver operating characteristicscurve).

*Corresponding author
Byeong-Soo Jeong
Email: jeong@khu.ac.kr

1. Introduction

As more whole-genome sequences are increasingly generated with the continued development of new high-throughput methods for DNA sequencing, gene identification becomes one of the important tasks for computational biology. In order to understand how the genome works, we need to identify a set of coding fragments known as *exons*, which are separated by non-coding intervening fragments known as *introns*. As shown in Figure 1, the boundaries between exons and introns are called *splice sites*. The vast majority of all splice sites are characterized by the presence of specific dimers: GT for donor and AG for acceptor sites. However, only about 0.1%~1% of all GT and AG occurrences in the genome represents true splicing sites. Thus accurate prediction of splice site is naturally required for a systematic study of eukaryotic genes.

For this reason, there have been a lot of research works for predicting gene's structure and its function. Several machine learning algorithms have been developed for splice site prediction such as Bayesian networks, ANN (artificial neural network), discriminant analysis, and SVMs. Among them, SVMs and related kernel methods are most frequently used for solving such problems [1-10] due to their high accuracy and capability to deal with high-dimensional large data sets.

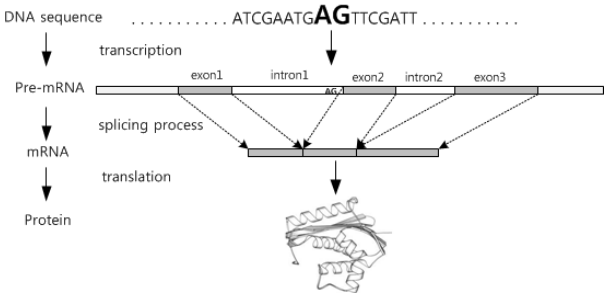


Figure1. Central dogma and splice sites

When we use SVM based classification technique, the feature extraction is a very important step for better classification accuracy. For feature extraction, DNA encoding method has the advantage of simple process. It can also provide the characteristics of DNA sequences to transform splice site sequence to a feature vector. However, existing encoding approaches do not reveal the characteristics of DNA sequence very well enough to provide as much informa-

tion as DNA sequences have. In this paper, we propose a new effective DNA encoding method which can give more information of DNA sequence. Our encoding method can provide density information of each nucleotide along with their positional information and chemical property.

The paper is organized as follows: In Section 2, we briefly survey related work about splice site prediction. We describe our proposed encoding method in detail in Section 3. We explain experimental environment and analyze its results in Section 4. Section 5 discusses some of the properties of the proposed encoding. Finally, Section 6 presents our conclusion.

2. Related Work

A large number of computational methods have been applied for solving biological sequence (e.g. DNA, protein) analysis, finding gene regulation, protein-protein interaction and so on. Of them, splice site prediction which is known as an important component of computational gene finder has received much attention to biological scientists. Their methods are mainly based on HMM (hidden Markov model), NN (neural network), and several statistical analysis. Even though a large number of splice site prediction tools are publicly available, they still need to improve their performance for high prediction accuracy and capability of handling a large scale DNA sequences. Table 1 summarizes the characteristics of representative prediction tools [11].

Table 1. Expert systems for splice site recognition

Tools name	Organism for training data set	Learning model*
GENESPLICER	Arabidopsis, human	HMM + MDD
NETPLANTGENE	Arabidopsis	NN
NETGENE2	Human, Arabidopsis	NN + HMM
NNSPLICE0.9	Drosophila, human	NN
SPLICEDETECTOR	Arabidopsis,maize	Logit linear models
BCM-SPL	Human,Drosophila,yeast,plant,C.elegans	LDA
SPLICEVIEW	Eukaryotes	Score with consensus

*MDD:maximal dependence decomposition, LDA: linear discriminant analysis

SPLICEVIEW [12] searches for a match with a consensus sequenceon a set of aligned functional sites considering the correlations between nucleotides of those sites. SPLICEDETECTOR [13] also uses the same approach as SPLICEVIEW does with some additional information. PWM (positional weight matrix) determines the appearance probability

of a given base at each position of the signal which can also be optimized by a neural network method, as proposed in NETPLANETGENE[14], NETGENE2[15] and NNSPLICE[16].

On the other hand, SVM which is a powerful pattern recognition technique is successfully applied for splice site prediction problem because of its high classification accuracy and capability of handling large-scale DNA sequences. In order to apply SVM, effective DNA encoding approach is necessary for transforming raw DNA sequences into vectors of feature space. On this account, several encoding methods are proposed and analyzed in many ways.

Salekdehet. al. [2] proposed an encoding method which can consider the positional probability of each nucleotide while introducing another 4 ambiguous values to represent possibility of occurrence of some other nucleotides. However, it cannot distinguish distance value between ACC-AGG and ACC-ATT in the case of one matched and two unmatched nucleotides even though AGG is a more important sequence than ATT for splice site prediction.

In [1], for extracting more information from splice site sequences, they utilize three approaches, orthogonal encoding, codon usage, and sequential information. Huang et. al. [9] proposed four different encoding approaches and compared their performance. They are MN (mono nucleotide) encoding which maps each of 4 DNA bases into an integer number, PN (pairwise nucleotide) encoding which maps 16 possible pairs into an integer number, and combining frequency difference between the true and false sites (FDTF) encoding. Their experimentation indicates that PN with FDTF method produces the best accuracy. In [17], they classified 4 DNA bases as 4 different coordinates from the knowledge of biology which is based on nucleotide classification. Their experimental results show that the 4D representation provides good performance in measuring the evolutionary relationship among different species.

Zhang et. al. [6] used weight matrix model for DNA encoding. The problem of this type of weight matrix is that it assumes each position is equally important and therefore each attribute (i.e. nucleotide) is independent. But attributes are not always independent and some positions may be essential while others may be trivial in the area of splice site prediction.

AKMA Batenet. al. [7] produced their best result in reduced Markov encoding where the conditional probability of a nucleotide at any location depends on its immediate predecessor. But the correlation between adjacent nucleotides does not reveal the global feature of splice sites. Markovian probability becomes complex and unrealistic when we consider high order Markov model for global feature.

However, none of the above methods consider density information of each nucleotide in DNA sequences which may be desirable information for splice site prediction. In this paper, we propose a new effective DNA encoding method which can give more information about DNA sequences. Our encoding method includes density information of each nucleotide along with positional information and chemical property.The extensive performance study shows that our method can provide better performance than existing encoding methods based on several performance criteria such as classification accuracy, sensitivity, specificity and auROC.

3. DNA Encoding

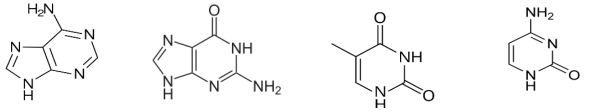
Encoding has its own advantages while extracting features from functional sites (e.g. splice site, promoter site, translation site, etc.) as well as in the visualization and similarity analysis of those sequences. As for DNA encoding, it might be a very simple approach to assign two binary digits to each nucleotide (A=00, G=01, C=10, T=11). However, it cannot give any characteristics of DNA sequences even though it has the advantage of simplicity. Another simple approach is to assign four binary digits which have only single '1' value among 4 digits (A=0001, G=0010, C=0100, T=1000). This sparse encodingmight give benefit when we compress DNA sequences, but this encoding treats the four nucleotides equally and failed to consider the probability of natural mutation in DNA sequences.

Besides, many graphical representations have been proposed for visualizing DNA sequences while mapping each nucleotide or dinucleotide into coordinates of 2D [18-21] or 3D [22]. They are efficient to recognize DNA sequences at a single glance, however, not enough to show sequence characteristics. Lastly, several statistical methods were proposed for analyzing DNA sequences. But these approaches are very hard to encode because of their numerical complexity.

Generally, in the case of DNA sequence encoding, two characteristics (i.e. degeneracy and uniqueness) are required to avoid loss of information, and guarantee unique mapping between DNA sequence and its Cartesian graph. DNA encoding for splice site prediction should be satisfied with these two criteria. In this paper, we propose new encoding approach for splice site prediction. Our approach considers chemical characteristics of DNA and density information of each nucleotide in its every position while satisfying the above criteria.

3.1 Chemical property of each nucleotide

DNA isnucleic acid that contains genetic instruction used in the development and functioning of all known living organism. It is a polymer whose monomer units are nucleotides.There are mainly four different types (i.e. Adenine, Guanine, Cytosine, Thymine) of nucleotides found so far. Each nucleotide has different chemical structure andchemical binding between them as shown in Figure 2 and 3.Depending on chemical property (i.e. ringstructure,hydrogen bond and functionality) of each nucleotide, DNA sequence might represent different biological characteristics.



(a) Adenine (b) Guanine (c) Cytosine (d) Thymine

Figure 2. Chemical structure of each nucleotide

As shown in Figure 2, Adenine and Guanine have two rings (i.e. hexagon and pentagon). On the contrary, Cytosine and Thymine have only one ring structure(i.e. hexagon only). Thus, they belong to different groups in terms of ring.In the same way, as shown in Figure 3, Guanine and Cytosine show strong hydrogen bondcompared with Adenine and Thymine.Lastly, they are grouped on different chemical functionality i.e. amino group and keto group.

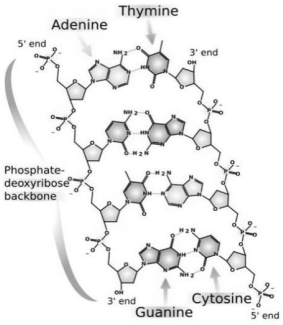


Figure 3. Complementary DNA binding

Table 2 summarizes such chemical property of each nucleotide.

T

able 2. Chemical property of each nucleotide

Chemical property	*Class*	*Nucleotides*
Ring Structure	Purine	{A, G}
	Pyrimidine	{C, T}
Functional Group	Amino	{A, G}
	Keto	{G, T}
Hydrogen Bond	Strong-H	{C, G}
	Weak-H	{A, T}

In order to include such chemical property in DNA encoding, we put 3 coordinates (x, y, z) to represent three chemical group and assign 1 or 0 values. Each nucleotide, s_i= (x_i, y_i, z_i) is represented according to following formula.

$$x_i = \begin{cases} 1 & if\ s_i \in \{A, G\} \\ 0 & if\ s_i \in \{C, T\} \end{cases} \qquad y_i = \begin{cases} 1 & if\ s_i \in \{A, C\} \\ 0 & if\ s_i \in \{G, T\} \end{cases} \qquad z_i = \begin{cases} 1 & if\ s_i \in \{A, T\} \\ 0 & if\ s_i \in \{C, G\} \end{cases}$$

That is, coordinate value of each nucleotide is determined by their chemical property of the nucleotide. Purine {A, G} and pyrimidine {C, T} both have rings – purinehas two rings, pyrimidine has one. So, they will fall into same coordinate (here, x coordinate). Similarly amino {A, C} and keto {G, T} group fall into y coordinate because they have same functionality. Eventually, $z$coordinate is plotted by strong-H {C, G} and weak-H {A, T} group because they possess same Hydrogen bond.

3.2 Nucleotide density

Another desirable feature of DNA sequence is the occurrence of each nucleotide in DNA sequence. For this, we add another coordinate d_i in (x_i, y_i, z_i). Inthis added coordinate d_i, we represent frequency information and also distribution of each nucleotide in DNA sequence. For this purpose, we define d_i value which represents density of each nucleotide by following formula.

Let$\sum = \{A, T, C, G\}$ and $S = \{s_1, s_2, s_3, \dots, s_l\}$is a DNA sequence of length $l$where $s_i \in \sum$ and $i = 1, 2, 3, \dots, l$. We further define that |S| represents the length of the string S and |S_i| is the length of substring$[1, i]$. Then, the density d_i of any nucleotide s_i in the position$i$can be formally derived as

$$d_i = \frac{1}{|S_i|} \sum_{i=1}^{l} f(s_i),$$

where $f(q) = \begin{cases} 1 & if s_i = q \\ 0 & otherwise \end{cases}$, $i = 1, 2, 3, \ldots, l$ and $q \in \sum$.

The $d_i$acts as nucleotide's positional weight within a sequence. For example, consider a sequence "ATAGTCATAA". The density of 'A' is 1, 0.67, 0.43, 0.44, and 0.50 in the position 1, 3, 7, 9 and 10 respectively, 'T' is 0.5, 0.40, and 0.37 in the position 2, 5 and 8 respectively, 'C' is 0.17 in the position 6 and 'G' is 0.25 in the position 4.The following table clearly explains density information for the above example sequence.

Table 3. Density information of example sequence "ATAGTCATAA"

i	s_i	$f(s_i)$	$\|S_i\|$	$d_i = f(s_i)/\|S_i\|$
1	A	1	1	1.00
2	T	1	2	0.50
3	A	2	3	0.67
4	G	1	4	0.25
5	T	2	5	0.40
6	C	1	6	0.17
7	A	3	7	0.43
8	T	3	8	0.375
9	A	4	9	0.44
10	A	5	10	0.50

On the other hand, A can be represented as (1, 1, 1), T can be represented by (0, 0, 1) as per the equation described in section 3.1. Similarly, G and C's representation in (x, y, z) format is (1, 0, 0) and (0, 1, 0) respectively. We add density as a fourth dimension to represent each nucleotide in (x, y, z, d) format.

Finally, the example sequence "ATAGTCATAA" is represented by {(1, 1, 1, 1), (0, 0, 1, 0.5), (1, 1, 1, 0.67), (1, 0, 0, 0.25), (0, 0, 1, 0.4), (0, 1, 0, 0.17), (1, 1, 1, 0.43), (0, 0, 1, 0.37), (1, 1, 1, 0.43), (1, 1, 1, 0.50)}where (, , ,) stands for a single nucleotide with their x, y, z and d values as described.

3.3 Classifier model construction

A binary SVM is adopted to classify NN269 [23] sequences into two classes, true sites and false sites. Let$S = \{s_1, s_2, s_3, \ldots, s_l\}$denoteasplice site of NN269 datasets of length l and$R = \{r_1, r_2, r_3, \ldots, r_l\}$is the input feature vector, where $r_k = \{x_i, y_i, z_i, d_i\} \in \mathbb{R}$, $k = 1, 2, 3, \ldots, l$ is the coordinate value ofa nucleotide (s_k). The classification of DNA sequence S finds an optimal mapping from $\mathbb{R}^{4l}$the space of coordinate values into $\{+1, -1\}$ where $+1$ corresponds to true splice site and -1 to false splice site respectively.

Let some target function $\hat{f}: \mathbb{R}^{4l} \rightarrow \{+1, -1\}$ and $D = \{(R_j,\ y_j) \mid j = 1, 2, 3, \ldots, N\} \subseteq \mathbb{R}^{4l}$ denotes the set of training examples, where $y_j = \hat{f}(R_j)$ denotes the desired class, true site or false site, for the input feature vector R_j of all coordinate values of sequence S_j; N denotes the number of training sequences. We need to compute a model $\hat{f}: \mathbb{R}^{4l} \rightarrow \{+1, -1\}$ by D.

SVM first transforms the input to a higher dimensional space with a kernel function K and then linearly combines them with a weight vector $\boldsymbol{w}$ to obtain the output. The binary SVM is trained to classify the input vectors to correct the class of splice sites.

For this purpose, SVM constructs a discriminate function by solving the following optimization problem:

Minimize

$$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma \sum_{j=1}^{N} \xi_j$$

subject to the constraint

$$y_j\left(\boldsymbol{w}^T\boldsymbol{\phi}(R_j) + b\right) \geq 1 - \xi_j \; and \; \xi_j \geq 0,$$

where slack variables ξ_j represent the magnitude of error in the classification, $\boldsymbol{\phi}$ represents the mapping function to a higher dimension, b is the bias used to classify samples, and $\gamma > 0$ is the sensitivity parameter that decides the trade-off between the training error and the margin of separation [29-30].

The minimization of the above optimization problem is equivalent to maximizing the following quadratic function:

$$\max_{\alpha} \sum_{j=1}^{N} \alpha_j - \frac{1}{2} \sum_{j,i=1}^{N} \alpha_j \alpha_i y_j y_i \, K(R_j, R_i)$$

subject to $0 \leq \alpha_j \leq \gamma$ and $\sum_{j=1}^{N} \alpha_j y_j = 0$.

Function $K(R_j, R_i) = \boldsymbol{\phi}(R_j)^T \boldsymbol{\phi}(R_i)$ is the kernel function and the weight vector $\boldsymbol{w} = \sum_{j=1}^{N} y_j \alpha_j \boldsymbol{\phi}(R_j)$. Once the parameters α_j are obtained from the optimization, the model function for any input pattern R_i, $\hat{f}(R_i)$ is given by:

$$\hat{f}(R_i) = sgn\left(\sum_{j=1}^{N} \alpha_j y_j K(R_j, R_i) + b\right),$$

where $sgn(k) = \begin{cases} +1 \ if k \geq 0 \\ -1 \ if k < 0 \end{cases}$.

4. Experimental Results

4.1 Dataset, experimental environment and performance metric

To evaluate the performance of our proposed encoding method, weconducted several experiments on NN269 [23] dataset. Table 4 shows the different important characteristics of the dataset. The dataset was created to compare different splice site models. The donor datasets have 7 base pairs (bp) of the exon and 8bp of the following intron (starting with GT).The acceptor data sets have 70 bp in the intron (ending with AG) and 20 bp of the following exon.

Table 4. Dataset characteristics

Dataset	Acceptor			Donor		
	True	False	Total	True	False	Total
Training	1116	4672	5788	1116	4140	5256
Testing	208	881	1089	208	782	990
Total	1324	5553	6877	1324	4922	6246

The existing Reduced MM1-SVM [7] outperforms the other methods like Information Content (IC Shapiro) and MM1-SVM. On the other hand, MM1-SVM/WMM1-SVM [8] performs better than Loi-Rajapakse [24], NNSplice [16] and GeneSplicer [25].So, we compare the performance of our algorithm with Reduced MM1-SVM only. At first, we show the overall performance of the proposed encoding ondifferent SVM kernels. After that, we evaluate the performance of sparse encoding in the same experimental environment. Then we add density in sparse encoding and reevaluate the performance to show the importance of density in already existing model. Finally, we compare our classification accuracy with [7] for performance comparison.

Our programs were written in Python 2.7, and run with the Windows XP operating system on a Pentium dual-core 2.13 GHz CPU with 2 GB main memory. We used BioPython 1.60 for sequence parsing, SMO (sequential minimal optimization) algorithm of WEKA 3.6 [26] for SVM classification and LibSVM [28] for kernel parameter selection.

To evaluate the classification performance, we used several evaluation methods such as the sensitivity (S_n), specificity (S_p), accuracy, receiver operating characteristics (ROC) curve,and the auROC as described in the following:

$$S_n = \frac{TP}{TP+FN} \times 100 \quad S_p = \frac{TN}{TN+FP} \times 100 \quad Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \times 100$$

where TP, TN, FP, FN represents true positive, true negative, false positive and false negative respectively. PlottingS_n against 1-S_p produces the ROC [27] curve. ROC analysis is an effective and widely used method to assess the performance of classifiers. The larger values of S_n, S_p, accuracy and auROCindicates the better performance of a classifier.

4.2 Performance of the model

We used density information to construct our feature vector because it indicates the positional weigh of nucleotide in a sequence. This weight depends on nucleotide's frequency. So, we get frequency distribution of nucleotides from density which helps SVM to classify splice sites. An example will clear the use of density and supremacy of the proposed method over others.

Let consider five splice sites of length 10 where +1 stands for true and -1 for false class label. Their positional profile (PP) matrix is given in Fig. 4 (b). The encoded sequence of those splice sites based on positional profile and density information (DI) is shown in Table 5.

ATCTATCAGG, +1
TCTCACACTC, +1
CTACCTTAAA, -1
CAGTGAGTAG, +1
TACTTGTCTG, -1

a) Aligned DNA Sequences

∑ \ p	1	2	3	4	5	6	7	8	9	10
A	0.20	0.40	0.20	0.0	0.40	0.20	0.20	0.40	0.40	0.40
C	0.40	0.20	0.40	0.40	0.20	0.20	0.20	0.40	0.0	0.20
G	0.0	0.0	0.20	0.0	0.20	0.20	0.20	0.0	0.20	0.40
T	0.40	0.40	0.40	0.60	0.20	0.40	0.40	0.20	0.40	0.0

b) Positional profile information table of aligned sequences in (a), p is indicating the positions

Figure 4. Positional profile information of a set of aligned splice sites

From table 5, we see that PP does not reflect nucleotide's frequency distribution in the sequence's feature vector. It just fills up each position in feature vector of any site with a value taken from the range [0, 1/N, 2/N, 3/N, …,N/N] where N is the number of aligned sequence.

Table 5. Splice site encoding in positional profile versus density information

Splice site	Class	PP	DI
ATCTATCAGG	+1	0.2, 0.4, 0.4, 0.6, 0.4, 0.40, 0.2, 0.4, 0.2, 0.4	1, 0.5, 0.33, 0.5, 0.4, 0.5, 0.28, 0.25, 0.11, 0.2
TCTCACACTC	+1	0.4, 0.2, 0.4, 0.4, 0.4, 0.2, 0.2, 0.4, 0.4, 0.2	1, 0.5, 0.67, 0.5, 0.2, 0.5, 0.28, 0.5, 0.33, 0.5
CTACCTTAAA	-1	0.4, 0.4, 0.2, 0.4, 0.2, 0.4, 0.4, 0.4, 0.4, 0.4	1, 0.5, 0.33, 0.5, 0.6, 0.33, 0.43, 0.25, 0.33, 0.4
CAGTGAGTAG	+1	0.4, 0.4, 0.2, 0.6, 0.2, 0.2, 0.2, 0.4, 0.4, 0.4	1, 0.5, 0.33, 0.25, 0.4, 0.33, 0.43, 0.25, 0.33,0.3
TACTTGTCTG	-1	0.4, 0.4, 0.4, 0.6, 0.2, 0.2, 0.4, 0.4, 0.4, 0.4	1, 0.5, 0.33, 0.5, 0.4, 0.5, 0.28, 0.25, 0.11, 0.2

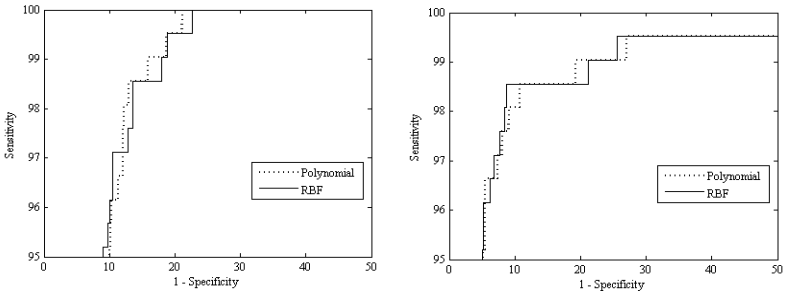
On the other hand, feature vectors extracted from density information shows the appropriate distribution of nucleotides in the sites. This distribution has advantages over others while SVM is used to classify them. Because, we know that splice sites show high frequency of GC content. Furthermore, exon shows a regular pattern based on relative synonymous codon usage (RSCU) but intron does not. Frequency distribution of each nucleotide in site classification based on SVM has added advantages because the dot product operation of two true/false sites will fall into same side of boundary decision most of the time.

We used5-folds cross validation on training dataset. Then the classifier is reevaluated with the test data.

Table 6. Performance evaluation of NN269 acceptor and donor splice site

	TP	*TN*	*FP*	*FN*	S_n	S_p	*Accuracy*	*auROC*	*Kernel*
Acceptor	161	856	25	47	77.40	87.16	93.39	97.90	Poly
	165	857	24	43	79.30	87.28	93.85	97.91	RBF
Donor	183	760	22	25	87.98	97.19	95.25	98.30	Poly
	185	758	24	23	88.94	96.93	95.25	98.24	RBF

Different kernel parameters were applied to obtain the best performance. For donor site, C=0.001 and E=2.0 are used while using polynomial kernel and C=8.0, γ=0.007 for RBF kernel. Similarly, we used C=3.0, E=2.5 in case of acceptor for polynomial kernel and C=8.0,γ=0.004 for RBF kernel. The best values for TP, TN, FP, FN, S_n,S_p, accuracy and auROC are shown for both sites in Table 6.



(a) Acceptor (b) Donor

Figure 4. ROC curves showing the classification performance for NN269 dataset

We drew ROC curves for acceptor and donor splice sites. The curves are drawn in case of thebest performance on both kernels. Figure 4 (a) shows that the model starts being stable considering only 25% FPR (false positive rate) for both kernels. In case of donor, in Figure 4 (b), the difference between auROC is negligible. So, the curves overlap each other.

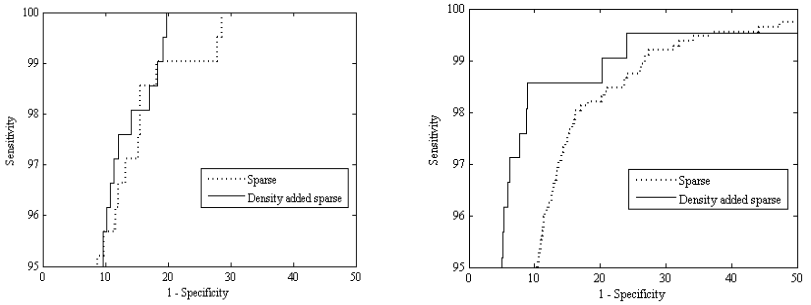
Table 7. Classification performance with and without density

Methods	Splice Site	Without density		With density	
		Accuracy	auROC	Accuracy	auROC
Proposed	Acceptor	-	-	93.85	97.91
	Donor	-	-	95.25	98.30
Sparse	Acceptor	93.57	97.70	93.74	97.80
	Donor	93.79	97.60	95.25	98.30

4.3 Effectiveness of density

In this section, firstly we implement sparse encodingand evaluate the performance. Secondly, we add density as a fifth dimension and reevaluate the classification algorithm. Let define this new encoding as DS (density-based sparse). Thirdly, we compare the performance between them to show the effectiveness of density. Table 7 shows some performance measures of proposed, sparse and DS encoding.

For DS, the accuracy and auROC are increased in case of acceptor sites. The increment of those measures for donor sites is also remarkable. Our proposed method outperforms than sparse and DS encoding. So it is clear that the classifier performs better when density is added in sparse encoding. Figure5 shows the respective ROC curves. For acceptor, in Figure 5(a), the curves differentiate in some places but eventually DS curve stables on less FPR than sparse.



(a) Acceptor (b) Donor

Figure 5. ROC curves comparison between sparse and DS encoding

As shown in Figure 5(b), the difference between sparse and DS encoding is vividly shown. The difference of auROC is 0.70. It can be concluded that the performance of density based sparse encoding is significantly superior to that of sparse encoding.

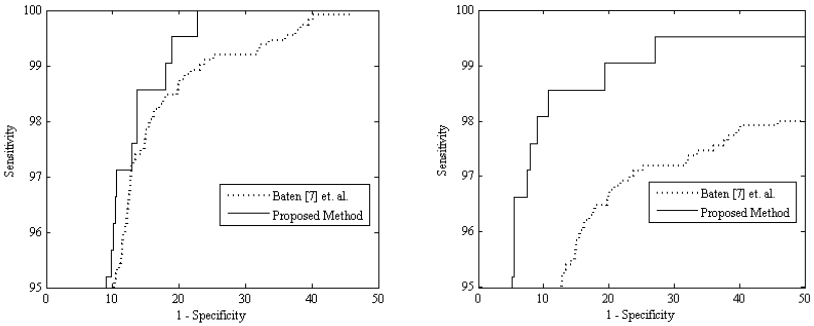
4.4 Classification performance comparison

In this experimental section, we compare the proposed model with Reduced MM1-SVM to verify the practical applicability of the model obtained. Reduced MM1-SVM extracts sequence feature using first order Markov model.It generates some emission probabilities for input sequence to learn the conserved sequence pattern at upstream and downstream regions surrounding the splice site motifs (i.e. GT-AG). Table 8 shows the classification performance of our model with Reduced MM1-SVM.

Table8.auROC comparison of the models

	Proposed SVM Poly	*Proposed SVM RBF*	*Reduced MM1-SVM GRBF*	*Reduced MM1-SVM Poly*	*MM1-SVM Poly*	*IC Shapiro SVM-Poly*
Acceptor	97.90	97.91	97.41	96.96	96.74	96.23
Donor	98.30	98.24	97.90	97.65	97.62	96.66

Compared to MM1-SVM [8], Reduced MM1-SVM and IC Shapiro from [7], our classifier-provides a better performance. For acceptor sites, i) our model gives the best auROC 97.90 (97.91) which is 0.94 (0.50) higher than that of Reduced MM1-SVM poly (Reduced MM1-SVM GRBF), ii) 1.16 higher than that of MM1-SVM poly and iii) 1.67 higher than that of IC Shapiro SVM poly. In case of donor sites, i) our model produces the best auROC 98.30 (98.24) which is 0.65 (0.34) higher than that of Reduced MM1-SVM GRBF (Reduced MM1-SVM poly), ii) 0.68 higher than that of MM1-SVM poly and iii) 1.64 higher than that of IC Shapiro SVM poly.



(a) Acceptor (b) Donor

Figure 6. ROC curves comparison between proposed encoding and Reduced MM1-SVM

Figure 6 shows the comparison of performance between our model and Reduced MM1-SVM. As shown in Figure 6(a) and 6(b), our model is clearly superior for the identification of both acceptor and donor splice sites.

5. Discussion

The proposed encoding has some advantages over sparse and other recent encoding method for splice site prediction. Some of them are discussed in this section.

i) The proposed chemical classification does not treat each nucleotide independently, rather nucleotides are linearly combined. As for example,

$$x_A = x_T + x_C + x_G$$

$$y_A = y_T + y_C + y_G$$

$$z_A = z_T + z_C + z_G.$$

wherex_A, y_A, z_A represents the x, y, z coordinate of Adenine respectively and so on.The linear combination has two way advantages. Firstly, it can detect the natural mutation of DNA sequence that sparse encoding can't. Secondly, the encoded data becomes linearly separable while using SVM.

ii) As the length of splice sites are fixed, the proposed method can also be used for gene classification through homology-based approach. Every nucleotide in a site can be viewed as a four dimensional point. So, a site is a continuous curve connecting those points. We determine the geometric center of those curves and determine the similarity score among them. Let S$= \{P_1,\ P_2, \dots, P_l\}$ where each $P_i = \{x_i, y_i, z_i, d_i\}$ and$i = 1, 2, 3, \dots, l$. Then the geometric center of S is $(X_c,\ Y_c,\ Z_c,\ D_c) = \left(\frac{1}{l}\sum_{i=1}^{l} x_i\ ,\ \frac{1}{l}\sum_{i=1}^{l} y_i\ ,\frac{1}{l}\sum_{i=1}^{l} z_i\ ,\ \frac{1}{l}\sum_{i=1}^{l} d_i\right)$ where X_c, Y_c, Z_c, D_c represents the x, y, z, d coordinate of the geometric center respectively. The importance of geometric center is that it can be used to determine the similarity scoreof splice sites.

6. Conclusion

The accurate prediction of splice site is the key point of gene identification. Several mathematical models and encoding approaches are used for splice site prediction. The accuracy of these approaches depends on their feature extraction method. Our simple and easy approach of density information largely increases the accuracy of the classifiers. We consider the chemical property of nucleotides for encoding approaches.Furthermore, frequency distribution of

each nucleotide in the splice site helps encoded feature vector of those sites correctly classify into high dimensional feature space. The density information directly focuses on the chemical properties (i.e. GC content's high expressiveness and regular exonic pattern) of splice sites that helps SVM to classify them correctly.Our encoding approach with nucleotide density is easy to implement and simple but produce better result than others.

Acknowledgement: "This work was supported by a grant from the Kyung Hee University in 2013."(KHU-20130441)

References

[1] D. Wei, W. Zhuang, Q. Jiang, Y. Wei, A new classification method for human gene splice site prediction, in: J. He, X. Liu, E.A. Krupinski, G. Xu (Eds.), *Health Information Science*, Springer, Berlin, 2012, pp. 121-130.

[2] A. Y. Salekdeh, K. C. Wiese, Improving splice-junctions classification employing a novel encoding schema and decision-tree, IEEE congress on Evolutionary Computation, 2011, pp. 1302-1307.

[3] L. Nanni, A. Lumini, Identifying splice-junction sequences by hierarchical multi classifier, *Pattern Recogn. Lett.* **27** (2006) 1390-1396.

[4] C. Nantasenamat, T. Naenna, C. Isarankura-Na-Ayudhya, V. Prachayasittikul, Recognition of DNA splice junction via machine learning approaches, *EXCLI J.* **4** (2005) 114-129.

[5] Y. Sun, X. Fan, Y. Li, Identifying splicing sites in Eukaryotic RNA: support vector machine approach, *Comput. Biol. Med.* **33** (2003) 17-29.

[6] Y. Zhang, C. H. Chu, Y. Chen, H. Zha, X. Ji, Splice site prediction using support vector machines with a Bayes kernel, *Expert Syst. Appl.* **30** (2006) 73-81.

[7] A. Baten, S. K. Halgamuge, B. Chang, Fast splice site detection using information content and feature reduction, *BMC Bioinformatics* **8** (2008) 1-12.

[8] A. Baten, S. K. Halgamuge, B. Chang, J. Li, Splice site identification using probabilistic parameters and SVM classification, *BMC Bioinformatics* **7** (2006) 1-15.

[9] J. Huang, T. Li, K. Chen, J. Wu, An approach of encoding for prediction of splice sites using SVM, *Biochimie* **88** (2006) 923-929.

[10] Y. Chen, F. Liu, B. Vanschoenwinkel, B. Manderick, Splice site prediction using support vector machines with context-sensitive kernel functions, *J. Univ. Comput. Sci.* **15** (2009) 2528-2546.

[11] C. Mathe, M. F.Sagor, T. Schiex, P. Rouze, Current methods of gene prediction, their strengths and weakness, *Nucl. Acids Res*. **30** (2002) 4103-4117.

[12] I. B. Rogozin, L. Milanesi, Analysis of donor splice signals in different Eukaryotic organisms. *J. Mol. Evol*. **45** (1997) 50-59.

[13] J. Kleffe, K. Hermann, W. Vahrson, B. Wittig, V. Brendel, Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences, *Nucl. Acids Res*. **24** (1996) 4709-4718.

[14] S. M. Hebsgaard, P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouzé, S. Brunak, Splice site prediction in Arabidopsis Thaliana pre-mRNA by combining local and global sequence information, *Nucl. Acids Res.* **24** (1996) 3439-3452.

[15] N. Tolstrup, P. Rouzé, S. Brunak, A branch point consensus fromArabidopsis found by non-circular analysis allows for better prediction of acceptor sites, *Nucl. Acids Res*. **25** (1997) 3159-3163.

[16] M. G. Reese, F. H. Eeckman, D.Kulp, D. Haussler, Improved splice site detection in Genie, First Annual International Conference on Computational Molecular Biology (RECOMB), ACM Press, 1997, New York, pp. 232-240.

[17] B. Liao, M. Tan, K. Ding, A 4D representation of DNA sequences and its application, *Chem. Phys.Lett*. **402** (2005) 380-383.

[18] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, *MATCH Commun. Math. Comput. Chem*. **68** (2012) 611-620.

[19] Y. Li, G. Huang, B. Liao, Z. Liu, H-L curve: A novel 2-D graphical representation of protein sequences, *MATCH Commun. Math. Comput. Chem*. **61** (2009) 519-532.

[20] X. Guo, M. Randić, B. Subhah, A novel 2-D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett*. **350** (2001)106-112.

[21] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem*. **63** (2010) 519-532.

[22] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem*. **56** (2006) 209-216.

[23] M. G. Reese, F. Eeckman, D. Kupl, D. Haussler, Improved splice site detection in Genie, *J. Comput. Biol*. **4** (1997) 311-324.

[24] J. C. Rajapakse, H. S. Loi, Markov encoding for detecting signals in genomic sequences, *IEEE/ACM Trans. Comput. Bio. Bioinf*. **2** (2005) 131-142.

[25] M. Pertea, X. Lin, S. L. Salzberg, GeneSplicer: A new computational method for splice site prediction, *Nucl. Acids Res*. **29** (2001) 1185-1190.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, *SIGKDD Explorations* **11** (2009) 10-18.

[27] T. Fawcett, ROC graphs: notes and practical considerations for data mining researchers, *Technical Report HPL -2003-2004*, HP Laboratories, Palo Alto.

[28] C. C. Chang, C. J Lin, LIBSVM: A library for support vector machines, *ACMTIST* **2** (2011) 1-27.

[29] V. N. Vapnik, *The Nature of Statistical Learning Theory,* Springer, Berlin, 1995.

[30] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.