Hybrid Approach Using SVM and MM2 in Splice Site Junction Identification

Srabanti Maji and Deepak Garg*

Department of Computer Science and Engineering, Thapar University, Patiala-147004, India

Abstract: Prediction of coding region from genomic DNA sequence is the foremost step in the quest of gene identification. In the eukaryotic organism, the gene structure consists of promoter, intron, start codon, exon and stop codon, etc. In the prediction of splice site, which is the separation between exons and introns, the accuracy is lower than 90% even when the sequences adjacent to the splice sites have a high conservation. Therefore, the algorithms used in the splice sites identification must be improved in order to recover the prediction accuracy. Hence, an efficient method, MM2F-SVM is proposed through this article, which consists of three stages – initial stage, in which a second order Markov Model (MM2) is used, i.e. feature extraction; intermediate, or the second stage in which principal feature analysis (PFA) is done, i.e. feature selection; and the final or the third stage, in which a support vector machine (SVM) with Gaussian kernel is used for final classification. While comparing this proposed MM2F-SVM model with the other existing splice site prediction programs, superior performance for the former has been noticed.

Keywords: Gene identification, markov models, principal feature analysis, splicing site, support vector machine.

1. INTRODUCTION

Gene identification is one of the main objectives in genome sequencing. In the past few years, an elevated increase in the genomic primary sequence data for a broad range of organisms has been noticed [1]. The translation of data into knowledge is the key for future biological research and a great challenge as well. Watson and Crick [2] in 1953 discovered the double-helical structure of DNA, and within short period, researchers achieved a detailed understanding of the molecular methodology involved in gene replication and expression. In 1970s, direct access to the sequence of gene became possible through the invention of DNA sequencing and cloning. An essential characteristic for gene finding in genome sequencing projects is the occurrence of splice sites in the gene sequences. In sequencing of known structural elements, the signals observed are explored by the latest available computational techniques. The key aspect in the systematic study of eukaryotic genes is the accurate prediction of the partition between exons and introns, i.e. the splice sites, which further depends largely on exactly locating the splice sites.

The basic structural and resourceful unit of all living organisms is called the cell, which may be classified into two types – eukaryotic and prokaryotic. The prokaryote cell is simpler and smaller than a eukaryote cell. The nucleus is present in eukaryote, not in prokaryote. In eukaryotic organism, the gene structure consists of promoter, intron, start codon, exon and stop codon. Identification of the coding region is done by the presence of exon and that of non-coding region by the presence of intron. The size of intron sequences is in the range of 80-10000 nucleotides or more. In protein synthesis, introns are removed from the sequence during the process of transcription and translation. In all known intron sequences, the consensus sequences at both ends of an intron are almost the same. From DNA, pre-mRNA is produced through transcription process, which contains all the necessary information of the gene sequence, but only before it is fully converted (or processed) into mRNA. In the process called splicing, introns are removed and exons are retained in the mRNA and the reactions in splicing process are catalyzed by spliceosome. Within the intron, an acceptor site (3' end of the intron) and a donor site (5' end of the intron) are essential for splicing. The splice donor site includes invariant sequence GT at the 5' end of the intron with a larger and less preserved region. The splice acceptor site at the 3’ end of the intron terminates the intron with nearly invariant AG sequence (Fig. **1**). This is known as GT–AG law [3].

Dissimilar modification of a protein can arise when single exon is bounced or if only one out of two splice sites is used from an exon. This is called alternative splicing [4]. Essential mechanism for splice site selection in alternative splicing is the changeability in signal strength [2].

Identifying the presence of splice site within DNA sequence is the initial step in the accurate prediction of gene structure. Biology researchers have extensively studied the laboratory procedures such as PCR on cDNA libraries, northern blot, sequencing, etc. for accurately identifying the gene structure; but, due to presence of large number of hidden genes, it is almost impossible to describe all of them by using laboratory experiments only. Therefore, lab experiments are combined with bioinformatics approaches in the modern researches [5-9]. Numerous bioinformatics and computational approaches have been applied for gene prediction with the help of gene splicing. Some of the examples include probabilistic approaches, support vector machine and neural network approaches, discriminant analysis and the information theoretic approaches.

*Address correspondence to this author at the Department of Computer Science and Engineering, Thapar University, Patiala 147004, India; Tel: +91-175-2393007; Fax: +91-175-2393005; E-mail: dgarg@thapar.edu

2212-392X/14 $58.00+.00

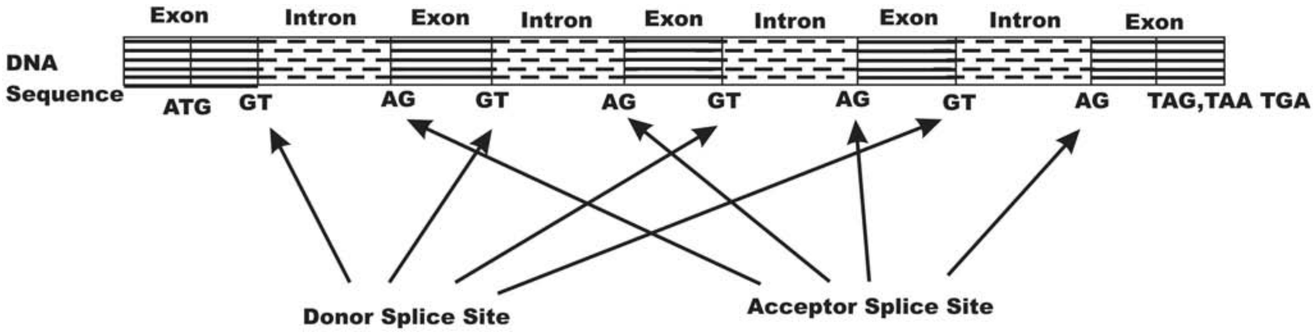

Fig. (1). Acceptor and donor splice sites in DNA sequence.

Splice site identification on the basis of models of site recognition and sequence data (supported by experimental confirmations) are described in a multi-agent system namely AMELIE [10]. This system, and the NetPlantGene are the two independent systems devoted to the recognition of splice sites in plant and human genomes respectively [11]. The HMM system, proposed by Salzberg *et al.* [12] is used to predict translation start site and splice site in the eukaryotic genes. They also developed Viterbi exon-intron locator (VAIL) [13], which is also an HMM based eukaryotic gene predictor tool. An effective HMM has been developed by Michael *et al.* [14], which is capable of signifying the consensus and degeneracy features of splicing sites in eukaryotic genes, and is utterly trained by using expectation maximization (EM) algorithm. A 12-fold cross-validation method was also used on this system to calculate its performance. The Feature Subset Selection (FSS), developed by using Estimation of Distribution Algorithm (EDA) is also reported [15], which have shown superior performance in classification of splice sites in case of Arabidopsis thaliana [16]. An FSS, based on wrapper algorithm and united with SVM [17] has also been used for splice site prediction. Another newer technique, EDA based feature ranking, was used for splice site identification in rapid feature selection process [18]. The SVM was also used by Brown *et al.* [19] to calculate a useful function for microarray gene expression by classifying genes and the existing data sets. To trace the signals in ribosome binding sites, splice site and promoter section, one computational technique was developed by Rodger Staden [20]. This method allocates separate values to all the bases at every position of the identification sequence to specify the comparative significance of each base, which is performed by weight matrix models (WMM). Zhang and Marr [21] introduced WAM that oversimplified the conventional WMM and integrated the dependencies between contiguous locations. Splice site can be identified by applying four stochastic regular grammar (SRG) inference algorithms controlled by generalization parameters with 10-way cross-validation to choose the best grammar for each algorithm [22]. To identify Translational Start Sites (TSS) and splice sites junction in eukaryotic mRNA, Salzberg [12] established conditional probability (CP) matrices. Gene finding model GENSCAN was introduced by Burge and Karlin [23, 24], tested on human and vertebrate genes. It has the blend of the double-stranded nature of the model and the ability to deal with inconsistent numbers of genes, and is particularly useful for studying long human genomic sequences. The maximal dependence decomposition (MDD) was also developed by them for modeling useful signals in DNA sequence and recommended that there were strong relationship between some of two or three precise positions with base constraints and probably relate to the splice site recognition. To identify transcription factor binding sites (TFBS) and splice sites, Huang and Zhao [25] applied permuted variable length Markov models (PVLMM) that can confine the potentially important dependencies with locations. For the prediction of splice site region in human pre-mRNA an artificial neural networks (ANN) was also applied [26]. A time-delay neural network model, which is a type of feed-forward neural network, shown their application to promoter annotation in the Drosophila melanogaster genome, and name of the tool was neural network for promoter prediction (NNPP) [27].

Some of the reasons for utilizing SVMs in Bioinformatics are – these have a strong widespread application in machine learning for classification and, these can target the relevant data positions automatically [28]. Other applications of SVM in bioinformatics are – prediction of protein secondary structure, multi-class protein fold recognition, and the prediction of human signal peptide cleavage sites. Till date, the most popular method for splice site recognition are Markov models which need the labor-intensive selection of information resource – SVM, the support vector machine kernels [12, 29-38].

As already mentioned, a large improvement in the recognition of splice sites is possible if a basic model uses hybrid architecture, e.g. WMM, MM1, MDD etc., is combined with other signal methods. GeneSplicer [35] is such type of method, where second order Markov models (MM2) are united with MDD. Probabilistic parameters of first order MM are joined with support vector machine (SVM) to predict splice site [39]. The addition of RNA structure information increases the accuracy of eukaryotic splice site finding, as examined by Markov models of zero to second orders [40]. Rajapakse and Ho *et al.* [41] merged a typical MM2 and back propagation neural networks (BPNN) to establish another splice site predictor.

Here, in our proposed method, MM2 is combined with SVM for the objective of increasing the efficiency and accuracy of splice-site prediction. The model consists of three stages – initial stage, in which a second order Markov Model (MM2) is used, i.e. feature extraction; intermediate, or the second stage in which principal feature analysis (PFA) is done, i.e. feature selection; and the final or the third stage, in which a support vector machine (SVM) with Gaussian kernel is used for final classification. By comparing this proposed MM2F-SVM model with the other existing splice site prediction programs, it was found that the model give superior performance than the other programs.

2. MATERIALS AND METHODOLOGY

2.1. Evaluation Datasets

Three ‘standard’ datasets of splice site were used to evaluate the performance of the proposed algorithms, which are publicly available, and are described in detail below.

HS3D (Homo Sapiens Splice Sites data set) was our first dataset [42], which is a dataset of intron, exons and splice sites. This dataset (of Human genes only) was extracted from Genbank. The length of each splice site sequence was 140bp. There were 2796 true donor sites and 271937 pseudo donor sites which contained “GT” dinucleotides and there were 2880 true acceptor and 329374 pseudo acceptor sites which contained “AG” dinucleotides. In case of donor splice site, GT dinucleotide was conserved at positions -71 and -72 of the sequences, and for acceptor splice site, AG was conserved at positions -69 and -70 of the sequences. The ratio between the number of true splice site and pseudo splice site was 1:10 and we used this dataset to extract features for modeling further.

The second dataset, DGSplicer [34], which is a true dataset, was created by extracting 2381 real acceptor sites and 2381 real donor sites from 462 annotated multiple-exon human genes [43]. Two donor splice sites and one acceptor splice site were excluded from the collection to form a set of 2380 real acceptor sites and 2379 real donor sites because those three splice sites contained the symbols other than A, C, G, and T. From 462 annotated human genes, a large collection of 400314 pseudo acceptor sites and 283062 pseudo donor sites were collected and used as the false dataset. The window size for the donor splice site was 18 nucleotides {-9 to +9} with consensus GT at positions +1 and +2, which included the last 9 bases of the exon and first 9 bases of the succeeding intron. The acceptor splice sites have a window of 36 nucleotides {-27 to +9} with consensus AG at positions -26 and -27, which includes the last 27 nucleotides of the intron and first 9 nucleotides of the succeeding exon.

In order to verify the effectiveness of our method, we performed additional evaluation on the third dataset namely NN269 [44]. It consisted of 1324 confirmed true acceptor sites, 1324 confirmed true donor sites, 5552 pseudo acceptor sites and 4922 pseudo donor sites; collected from 269 human genes. The window size of donor splice sites was 15 nucleotides {-7 to +8} with consensus GT at positions +1 and +2. This includes the last 9 bases of the exon and first 6 bases of the succeeding intron. The acceptor splice site have a window size of 90 nucleotides {-70 to +20} with consensus AG at positions -69 and -70. This includes the last 70 nucleotides of the intron and first 20 nucleotides of the succeeding exon, which is available at [45]. This data set was split into a training set and a testing set. The training dataset contained 1116 true acceptor, 1116 true donor, 4672 pseudo acceptor, and 4140 pseudo donor sites. The test data set contained 208 true acceptor sites, 208 true donor sites, 881 false acceptor sites, and 782 false donor sites. In NN269 donor splice sites, GT was conserved in positions -8 and -9 of the sequences and for acceptor dataset; AG was conserved in positions -69 and -70 of the sequences.

2.2. Overview of the Projected Model

The proposed model MM2F-SVM includes a number of separate modules and sub modules that were anticipated to capture properties of DNA and specially designed to identify splice site. As splice site corresponds to the donor splice site and acceptor splice site, so splice site categorization process is subdivided into two classification modules – donor splice site classification and acceptor splice site classification process. Further, for the recognition of acceptor splice sites and donor splice sites, two different models are assembled which consist of three phases (or sub modules). The model employs several important aspects; these are (1) appropriate features encoding scheme, (2) feature selection or ranking method, and (3) parameters optimization. The basic subsequent processing steps are outlined in the following:

1. *Feature extraction*: Positional probabilistic descriptions of different orders are constructed and a pool of candidate features is generated.
2. *Feature selection*: The discriminative power of each feature is assessed and the most informative features are selected using PFA.
3. *Classification step*: The SVM classifier is trained on the probabilistic parameters.

The proposed model architecture is described in Fig. (**2**).

2.3. Feature Extraction

In Markov process, the probability of the given condition in the given instant is likely to be presumed from information about the previous conditions [46]. A Markov chain represents a statistical system that undergoes transitions from one state to another between a limited or unlimited number of possible states and indicates next state depends only on the present state. A simple and existing Markov Model for DNA sequence is shown in Fig. (**3**). The systems which follow this specific type of characteristic are called Markov property, and the behavior of Markov chains are described by transition probability matrix. Every element of the matrix signifies probability of passage from a specific condition to a next state. In Markov model, we require a learning set of sequences on which these probabilities will be predictable. By using this technique, we can simply calculate the likelihood of the sequence, i.e. the probability that the sequence has been produced in accordance with this model.

In a DNA sequence, every nucleotide corresponds to a state in the Markov chain, where the observed state variables are derived from the symbol $X_{DNA} = \{A, T, G, C\}$. If length of the MM is L , then this probabilistic model describes the probability distribution of sequences of states $S_1, S_2 ..., S_L$ through transition probabilities, where transition probability $P(S_I = q \mid S_{I-1} = p)$ describes the probability of state $S_I = q$ (given, state $S_{I-1} = p$). A Markov model is used to capture

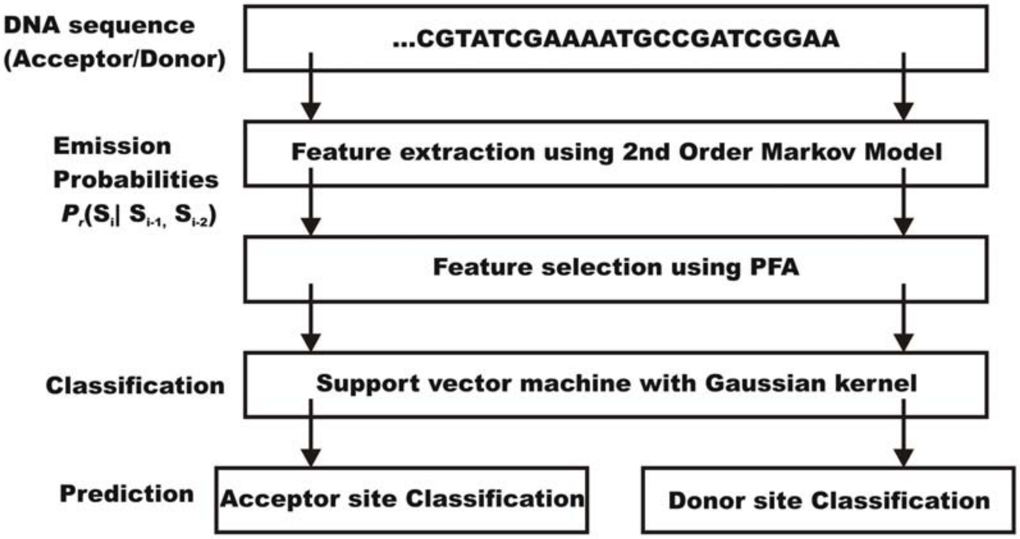
DNA sequence
(Acceptor/Donor)
...CGTATCGAAAATGCCGATCGGAA
Emission
Probabilities
$P_r(S_i| S_{i-1}, S_{i-2})$
Feature extraction using 2nd Order Markov Model
Feature selection using PFA
Classification
Support vector machine with Gaussian kernel
Prediction
Acceptor site Classification
Donor site Classification

Fig. (2). The MM2F-SVM Model. The input DNA sequence is preprocessed by 2nd order MM, PFA based feature selection. An SVM with Gaussian kernel function takes these parameters as its input for the splice site prediction.

the inter-dependencies among successive states in order to extract a set of probabilistic features [47]. If K is the order of MM, then likelihood of a sequence in this model is shown below:

$$P(S_1, S_2, ..., S_L) = \prod_{I=1}^{L} P_I(S_I \mid S_{I-1}) \qquad (1)$$

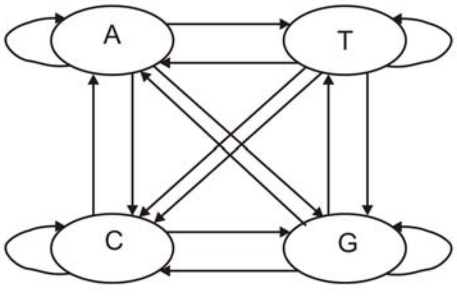

Fig. (3). Markov model for DNA sequence.

The ensuing model allocates different transition probabilities for each position. In the proposed method, MM2F-SVM employs 2nd order Markov model (MM2) to build the probabilistic feature set, which has transition probabilities of the format $P(S_I = q \mid S_{I-1} = p, S_{I-2} = v)$, and can be described by the collection of parameters:

$$\left\{P(S_I \mid S_{I-1}, S_{I-2}) : S_I, S_{I-1}, S_{I-2} \in X_{DNA}, I = 1, 2, .., L\right\} \qquad (2)$$

2.4. Feature Selection

The feature selection areas includes text processing of internet documents, gene expression array analysis, and combinatorial chemistry to improve prediction performance providing faster and more cost-effective predictors and better understanding of the underlying process that generated the data. For pattern classification, feature selection plays a very crucial role in the preprocessing step, which aims to deal with the storage space, dimensionality reduction problem and classification time, to improve the understanding of the problem as well as result interpretation [48, 49].

Feature selection process is useful to provide the necessary mechanism that clean out redundant features and provide some biological interpretation of the incorporated features. In this framework, feature selection in biological data can be employed in two different ways:

- By using Positional Feature Selection (PFS) [50] technique that identifies the best-fitting dependency length based on the discriminative power of each feature.
- Reducing feature set by selecting a subset of the original features that contains most of the essential information, using the same criteria as the PCA, the method name is principal feature analysis (PFA) [51].

***Positional Feature Selection (PFS)*:** PFS recognizes the best-fitting dependency length by distinguishing each feature. It selects the optimal feature among those describing a specific position by comparing their discriminative power; from a set of positional models of different lengths. It is applicable to solve binary classification problems [50]. The central model uses the F-score value as a selection criterion for the best-suitable feature per splice site [52].

***Principal Feature Analysis (PFA)*:** It is very much similar with the methods such as principal component analysis (PCA). One variant is possible by choosing a subset of the original feature vector that retains the underlying discriminative information by using the same optimality criteria as in PCA. Instead of identifying a projection of all features included to the original feature space to a lower dimensional space, PFA utilize the properties of the primary components to select a subset of the original features [51]. PFA considers the mutual information among the selected features. In this case, the source features are the second-order MMs and the outcome is the principal feature subset that competently characterizes the initial group of probabilistic parameters. The extracted components are separately studied for their statistical importance by performing the Wilcoxon rank sum test ($p < 0.05$).

In our proposed model we have selected Principal Feature Analysis (PFA) for feature selection process due to its better sensitivity as compared to Positional Feature Selection (PFS).

2.5. Classification

The fundamentals of Support Vector Machines (SVM) were studied extensively by Vapnik [53-56]. The formulation of SVM uses the Structural Risk Minimization (SRM) principle [57], which is more superior to the conventional Empirical Risk Minimization (ERM) principle used with usual neural networks. An SVM builds one or multiple hyperplane in a high dimensional space. Better partition can be accomplished by using the hyperplane that has the largest distance to the nearest training data point of any class (functional margin). The basic rule in SVM classifier is that the generalization error gets reduced when the margin is high [55, 58, 59]. Fig. (**4**) shows the SVM with hyperplane and margin.

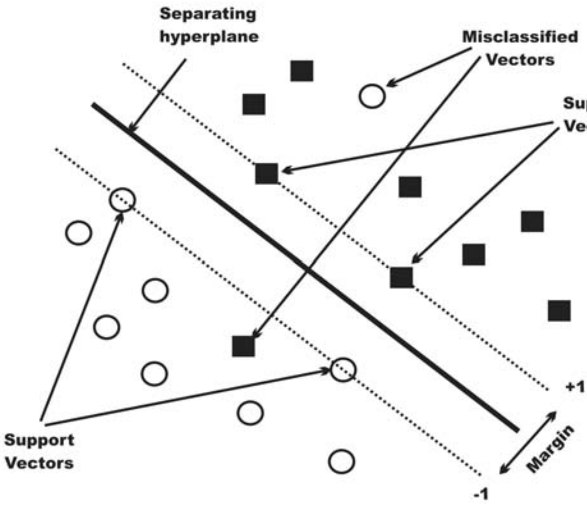

Fig. (4). Support Vector Machine (SVM) with hyperplane and margin.

SVM uses hypothetical space of linear function in high dimensional feature space trained with a learning algorithm. Using the method of Lagrange multipliers, we can obtain the dual formulation, which is expressed in terms of variables α_i. To solve the optimization problem, SVM classification is given by:

$$Maximize\ f(\alpha) = \sum\nolimits_{I=1}^{N} \alpha_I - \frac{1}{2}\sum\nolimits_{I=1}^{N}\sum\nolimits_{J=1}^{N} \alpha_I \alpha_J Y_I Y_J K(X_I, X_J), \quad (3)$$

Subject to $\sum\nolimits_{I=1}^{N} \alpha_I Y_J = 0$, $0 \le \alpha_I \le C$, $I = 1,...N$

In the above equation N is the number of training data, X is input vectors, Y defines class value that can be either -1 or 1 and C is trade off parameter for generalization performance. The dual formulation leads to an expansion of the weight vector in terms of the input examples:

$$w = \sum\nolimits_{I=1}^{N} Y_I \alpha_I X_I \quad (4)$$

Different data points of X_I for which $\alpha_I > 0$ are those points that are on the margin or within the margin when a soft-margin SVM is used. These are the so-called support vectors.

Assuming a query DNA segment is D, the trained SVM classifies based on the decision function:

$$o(D) = sign\left[\sum_{I=T} \alpha_I y_I K(X_I, D)\right] \quad (5)$$

where set of support vectors are represented by T.

For classification purpose we have used Gaussian RBF kernel with width $\sigma = 1$, where σ controls the flexibility of the resulting classifier. Therefore equation becomes:

$$K_\sigma^{GaussianRBF}(X, D) = \exp(-\frac{1}{\sigma} \| X - D \|^2) \quad (6)$$

After expanding, this equation becomes

$$K_\sigma^{GaussianRBF}(X, D) = \exp(-\frac{1}{\sigma}\left[\sum_{I=1}^{N}\left(X_I - D_I\right)^2\right] \quad (7)$$

where N is the number of dimensions in vectors X and D , correspondingly Ith element in vectors X and D are X_I and D_I . After substituting equation (7) into equation (5), the output O(D) becomes Gaussian kernel with width $\sigma = 2$,

While D is a vector of conditional probabilities of a sequence of length L:

$$D = \left[P(S_2 \mid S_1), P(S_3 \mid S_2), P(S_4 \mid S_3), ..., P(S_L \mid S_{L-1})\right] \quad (8)$$

Therefore a SVM classifier with the Gaussian kernel function can approximate higher order Markov model.

2.6. Model Design

The splice site identification process is divided into two sub modules; these are donor splice site identification and the acceptor splice site identification. For each module, separate models are created, e.g., for HS3D donor data-set, one MM2F-SVM model is created and trained with HS3D donor training dataset. To estimate the classification performance of this model, the HS3D donor test dataset is used. Likewise, a separate MM2F-SVM model is trained and tested with HS3D acceptor training and acceptor test dataset. Similarly, DGSplice and NN269, donor and acceptor dataset's are trained and tested.

2.7. Model Learning

Training of the proposed model was conducted in three stages: the MM2 parameters estimation, feature selection using PFA and the SVM with Gaussian kernel training having width of 20 (for classification). True and false splice site training sequences are used to create the second order markov model. Depending upon the true and false splice site class label, the desired output level was set to +1 and -1. We used MATLAB [60] for implementation of the support vector machine.

2.8. Model Comparison

To validate the usefulness of our proposed MM2F-SVM method and to compare its performance with others, we have selected other accepted methods those are closely related to the proposed method. We used another preprocessing scheme that is zero order markov model (MM0) with SVM and compare their preprocessing performance with our proposed model.

2.9. Performance Measures

The proposed hybrid method's classification performance is estimated on the ROC curves, which gives a measure of the tradeoff between the true positive rate TPR and false positive rate FPR. Sensitivity, S_n is the percentage of correct prediction of true splice sites and specificity, S_p is the percentage of correct prediction of pseudo splice sites. S_N (or TPR) is the percentage of correct prediction of true sites and S_P is the percentage of correct prediction of false sites as defined below:

$$Sensitivity(S_N) = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity(S_P) = \frac{TN}{TN + FP} \tag{10}$$

$$FPR = 1 - S_p = \frac{FP}{FP + TN} \tag{11}$$

$$precision = \frac{TP}{TP + FP} \tag{12}$$

A true positive is a true donor (true acceptor, respectively) site that is also classified as a true donor (true acceptor, respectively) site. A false positive is a false donor (false acceptor, respectively) site that is wrongly predicted as a true donor (true acceptor, respectively) site. A true negative is a false donor (false acceptor, respectively) site that is also classified as a false donor (false acceptor, respectively) site. A false negative is a true donor (true acceptor, respectively) site that is wrongly classified as a false donor (false acceptor, respectively) site given in Table **1**.

Table 1. Definitions of TP, TN, FP and FN

	Predicted Positive	Predicted Negative
True positive	True positives, TP	False negatives, FN
True negative	False positives, FP	True negatives, TN

Accuracy (ACC) is the proportion of the candidate sites in the test data set that are classified correctly [61], which tells how well the proposed MM2F-SVM system can assign true sites and false sites into the right categories; it was calculated by the following formula:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \tag{13}$$

Matthews's correlation coefficient (MCC) is used as a comprehensive classification performance metric incorporating both sensitivity and specificity measures defined by the following formula [61].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where MCC ranges from -1 to 1, and completely well trained classifiers are denoted by 1.

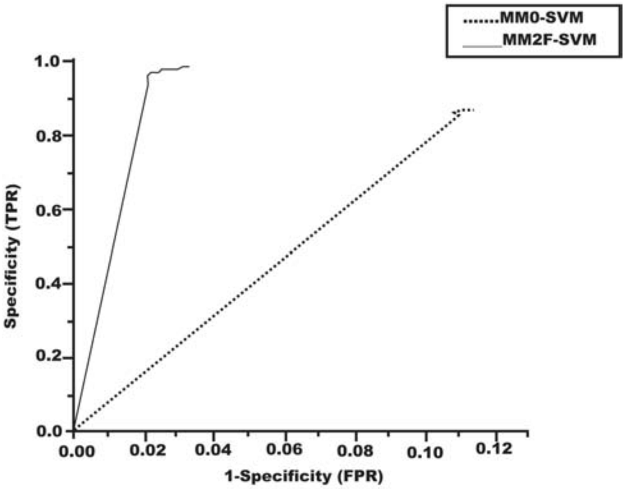

Fig. (5). ROC curve showing the comparison of performance between MM0-SVM and MM2F-SVM using HS3D donor dataset.

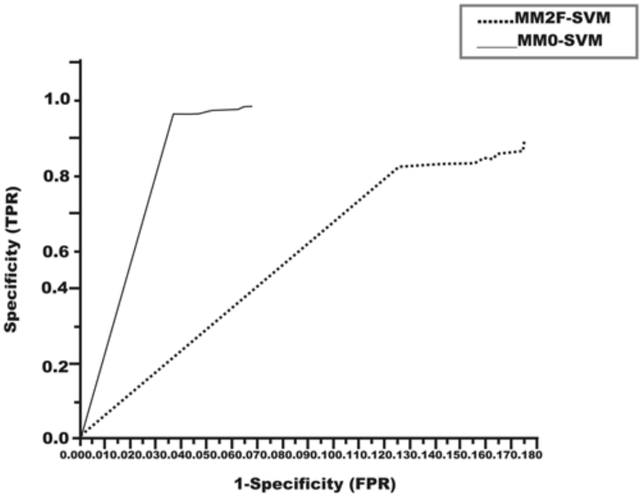

Fig. (6). ROC curve showing the comparison of performance between MM0-SVM and MM2F-SVM using HS3D acceptor dataset.

2.10. ROC Analysis

Receiver operator curve (ROC) analysis is an effective and widely used method of assessing the performance of models [62]. It is a graphical representation of sensitivity and specificity of a classification model. To approximate the best possible FPR and TPR pair, we used the Euclidean metric. Specially, the best sensitivity, specificity tradeoff is defined by the coordinates of each point on the ROC curve with the minimum distance from a perfectly well-trained classifier. When the ROC is created from the TPR (on the y-axis) and FPR (on the x-axis) of the model, the closer a curve approaches the (0,0) point, the more accurate the model is (refer to Figs. **5-12**).

2.11. Cross Validation

A twelve fold cross validation (CV) technique is applied to identify the MM2F-SVM splice site prediction accuracy and to compare their performance with the other available methods [63]. Here cross validation is performed by splitting the data into twelve independent subsets, in which every subset does not share any repeating sequences. Each model was trained by selecting eleven of the subsets (training data) and tested on twelfth unused subset (test data). We calculated the average of twelve prediction accuracies as the final prediction performance of the model, because CV is used for estimating the efficiency of the model.

3. RESULTS AND DISCUSSION

3.1. Selection of the Best Preprocessing Method

For preprocessing method selection, we have used methods like MM0 and MM2F with SVM classifiers for splice site prediction. We took HS3D donor and acceptor dataset for predictive accuracy comparison of MM0-SVM and MM2F-SVM methods. The ROC analysis of the models MM0-SVM and MM2F-SVM are shown in Figs. (**5**, **6**) respectively. After observing their performance, the MM2F-SVM model was used primarily for splice site identification.

3.2. Comparison in the Predictive Performance

The proposed model's 12-fold cross validation results - sensitivity (S_n), specificity (S_p), FPR and MCC for donor MM2F-SVM and the acceptor MM2F-SVM using HS3D are shown in Tables **2** and **3** respectively.

Table 2. Performance of Donor MM2F-SVM with Gaussian Kernel Width 20 for Identifying Donor (5' Splice) Sites

S. No.	No of True Donor	No of Pseudo Donor	TP	FP	TN	FN	Sensitivity (S_n)	Specificity (S_p)	FPR	MCC
1	233	22670	224	480	22190	9	0.96137	0.97882	0.0211	0.5466
2	235	22700	227	490	22210	8	0.96595	0.97841	0.0215	0.5464
3	240	23001	232	501	22500	8	0.96666	0.97821	0.0217	0.5465
4	235	22800	228	515	22285	7	0.97021	0.97741	0.0225	0.5389
5	241	22850	233	532	22318	8	0.96680	0.97671	0.0232	0.5357
6	238	23004	230	553	22451	8	0.96638	0.97596	0.0240	0.5258
7	237	22760	232	567	22193	5	0.97890	0.97508	0.0249	0.5261
8	236	22850	231	589	22261	5	0.97881	0.97422	0.0257	0.5179
9	238	23577	233	613	22964	5	0.97899	0.97400	0.026	0.5121
10	235	22890	230	674	22216	5	0.97872	0.97055	0.0294	0.4912
11	238	22779	234	691	22088	4	0.98319	0.96966	0.0303	0.4907
12	237	22547	233	720	21827	4	0.98312	0.96806	0.0319	0.4820
	Average						**0.97326**	**0.97476**	**0.0252**	**0.5217**

Table 3. Performance of Acceptor MM2F-SVM with Gaussian Kernel Width 20 for Identifying Acceptor (3' Splice) Sites

S. No.	No of True Acceptor	No of Pseudo Acceptor	TP	FP	TN	FN	Sensitivity (S_n)	Specificity (S_p)	FPR	MCC
1	240	27500	229	1005	26495	11	0.95416	0.96345	0.0365	0.4121
2	245	27820	234	1259	26561	11	0.95510	0.95474	0.0452	0.3771
3	239	27450	229	1308	26142	10	0.95815	0.95234	0.0476	0.3678
4	250	28000	240	1392	26608	10	0.96	0.95028	0.0497	0.3654
5	239	27780	230	1399	26381	9	0.96234	0.94964	0.0503	0.3584
6	256	28300	247	1487	26813	9	0.9648	0.9474	0.0525	0.3600
7	248	28670	240	1698	26972	8	0.96774	0.94077	0.0592	0.3350
8	253	27600	245	1729	25871	8	0.96837	0.93735	0.0626	0.3348
9	244	27676	237	1745	25931	7	0.97131	0.93694	0.0630	0.3291
10	254	27855	247	1759	26096	7	0.97244	0.93685	0.0631	0.3342
11	249	27650	243	1789	25861	6	0.97590	0.93529	0.0647	0.3297
12	241	27554	236	1861	25693	5	0.97925	0.93245	0.0675	0.3200
	Average						**0.96580**	**0.9448**	**0.0551**	**0.3520**

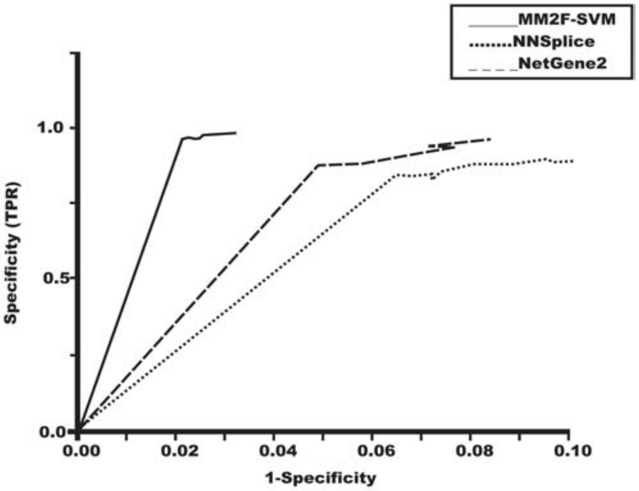

Fig. (7). ROC curve showing the comparison of performance between NNSplice, NetGene2 and MM2F-SVM using HS3D donor dataset.

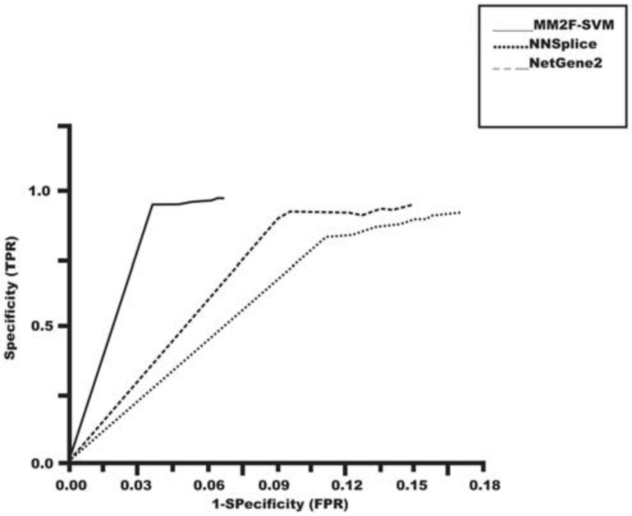

Fig. (8). ROC curve showing the comparison of performance between NNSplice, NetGene2 and MM2F-SVM using HS3D acceptor dataset.

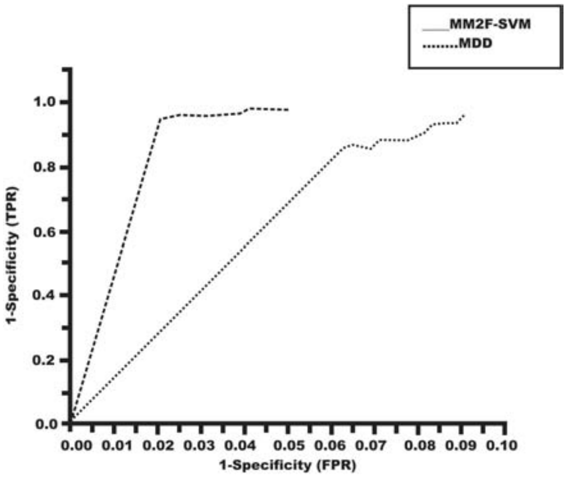

Fig. (9). ROC curve showing the comparison of performance between MDD and MM2F-SVM using DGSplicer donor dataset.

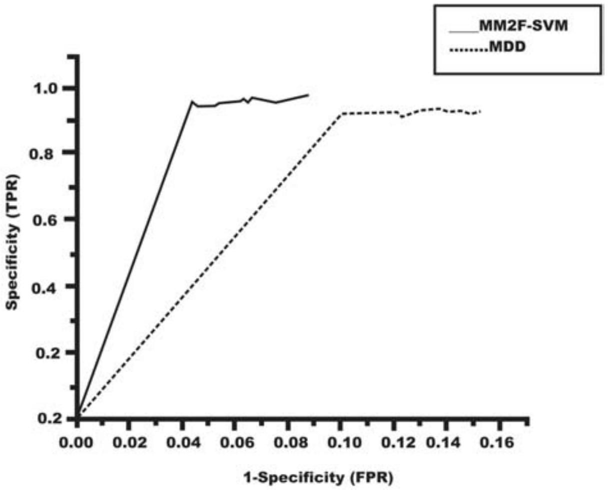

Fig. (10). ROC curve showing the comparison of performance between MDD and MM2F-SVM using DGSplicer acceptor dataset.

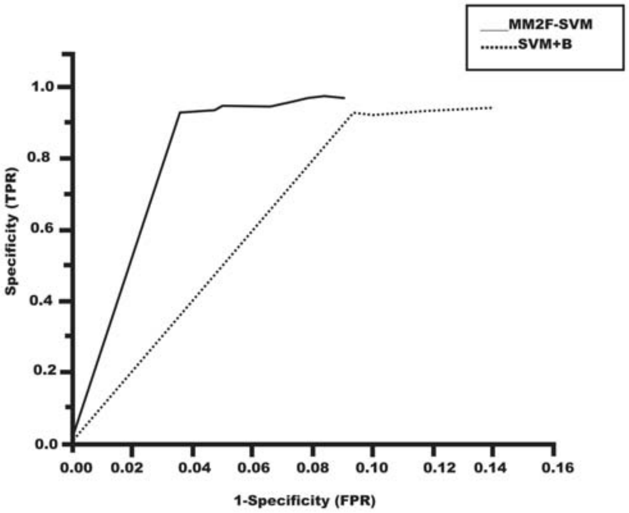

Fig. (11). ROC curve showing the comparison of performance between SVM+B and MM2F-SVM using NN269 donor dataset.

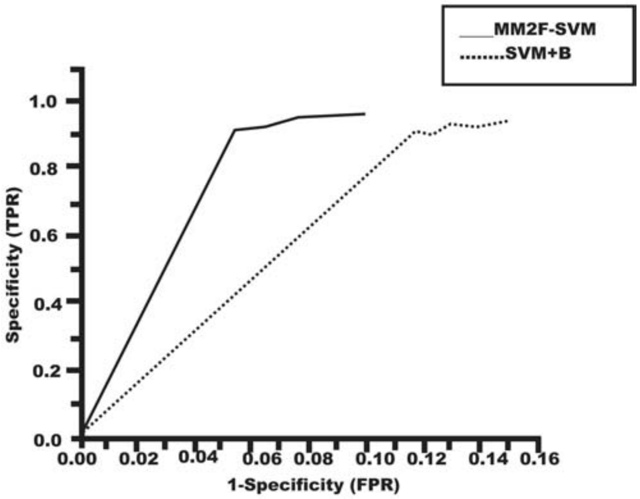

Fig. (12). ROC curve showing the comparison of performance between SVM+B and MM2F-SVM using NN269 acceptor dataset.

The comparison of performance between the MM2F-SVM, NNSplice [44] web site (http://www.frutfly.org/seq_tools/splice.html) and NetGene2 which was trained on human data (http://genome.cbs.dtu.dk/services/NetGene2/) using HS3D dataset is also done. The standard TPR (S_n) and FPR(1- S_p) are employed for this comparison, and the observation is that the MM2F-SVM is the superior model for prediction of donor and acceptor splice site. NetGene2 produced the second best performance as shown in Figs (**7**, **8**). The maximum S_n and S_p values for MM2F-SVM are 98.31% and 97.88% for the donor splice site prediction and 97.92% and 96.34% for acceptor splice site prediction.

To verify the prediction accuracies of the MM2F-SVM method, we used DGSplicer dataset and compared the performance with MDD method [23]. Here, MM2F-SVM showed superior performance as shown in Figs. (**9**, **10**).

To further verify the prediction accuracies of the MM2F-SVM method, we used NN269 dataset and compared the performance with SVM+B method [64], and observed that MM2F-SVM gives superior performance, as shown in Figs. (**11**, **12**).

4. CONCLUSIONS

In this paper, we have proposed a hybrid MM2F-SVM system which is able to choose a specific subset of features to identify a splice site junction according to the ratio of probabilities at every location. We have also used 12-fold cross validation experiment to verify the results. This system is able to correctly identify maximum 98.31% of the true donor sites and 97.88% of the false donor sites; and 97.92% of the true acceptor sites and 96.34% of the false acceptor sites in the test data set. In addition, this method is simpler, more effective and can be used to identify splice site junction on large scale in sequenced genomics.

AUTHORS CONTRIBUTIONS

The concept and design of this study along with implementation of the method(s) and analysis was performed by Srabanti Maji; the interpretation of the results was done by Dr. Deepak Garg. The text writing and thorough revision of the manuscript is contributed by both the authors.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers, Prof. Sanghamitra Bandyopadhyay, Machine Intelligence Unit, Indian Statistical Institute, Kolkata and Prof. Ujjwal Maulik, Department of Computer Science & Engineering, Jadavpur University, Kolkata for their expert and valuable comments, and suggestions in preparing the manuscript perfectly.

REFERENCES

[1] Brenner S. Genomics: The end of the beginning. Science 2000; 287(5461): 2173-2174.

[2] Brenner S. Sequences and consequences. Philos T Roy Soc B 2010; 365(1537): 207-212.

[3] Thanaraj TA, Clark F. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res 2001; 29(12): 2581-2593.

[4] Mersch B, Gepperth A, Suhai S, Hotz-Wagenblatt A. Automatic detection of exonic splicing enhancers (ESEs) using SVMs. BMC Bioinformatics 2008; 9(1): 369.

[5] Jing L, Hua H, Sufang L. Voice Identification Based on HMMs. Trends Appl Sci Res 2006; 1: 79-82.

[6] Maji S, Garg D, Progress in gene prediction: principles and challenges. Curr Bioinform 2013; 8(2): 226-243.

[7] Maji S, Garg D, Gene finding using Hidden Markov Model. J Appl Sci 2012; 12(15): 1518-1525.

[8] Bandyopadhyay S, Maulik U, Roy D, Gene identification: classical and computational intelligence approaches. IEEE Trans Syst Man Cybern Part C Appl Rev 2008; 38(1):55-68.

[9] Maji S, Garg D. Hidden Markov Model for Splicing Junction Sites Identification in DNA Sequences. Curr Bioinform 2013; 8(3):369-379.

[10] Vignal L, Lisacek Fdr, Quinqueton Jl, d'Aubenton-Carafa Y, Thermes C. A multi-agent system simulating human splice site recognition. Comput Chem 1999; 23(3-4): 219-231.

[11] Vignal L, d'Aubenton-Carafa Y, Lisacek F, *et al*. Exon prediction in eucaryotic genomes. Biochimie 1996; 78(5): 327-334.

[12] Sachem SI. A method for identifying splice sites and translational start sites in eukaryotic mRNA. Comput Appl Biosci 1997; 13(4): 365-376.

[13] Henderson J. Finding genes in DNA with a Hidden Markov Model. J Comput Biol 1997; 4(2): 127-141.

[14] Yin MM, Wang JTL. Effective hidden Markov models for detecting splicing junction sites in DNA sequences. Inform Sciences 2001; 139: 139-163.

[15] Inza I, Larranaga P, Etxeberria R, Sierra B. Feature Subset Selection by Bayesian network-based optimization. Artif Intell 2000; 123(1-2): 157-184.

[16] Saeys Y, Degroeve S, Aeyels D, Van De Peer Y, Rouze P. Fast feature selection using a simple estimation of distribution algorithm: A case study on splice site prediction. Bioinformatics 2003; 19(SUPPL 2): ii179-ii188.

[17] Degroeve S, De Baets B, Van de Peer Y, Rouze P. Feature subset selection for splice site prediction. Bioinformatics 2002; 18(suppl 2): S75-S83.

[18] Saeys Y, Degroeve S, Aeyels D, Rouze P, Van de Peer Y. Feature selection for splice site prediction: A new method using EDA-based feature ranking. BMC Bioinformatics 2004; 5(1): 64.

[19] Brown MPS, Grundy WN, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 2000; 97(1): 262-267.

[20] Staden R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res 1984; 12(1 PART2): 505-519.

[21] Zhang MQ, Marr TG. A weight array method for splicing signal analysis. Comput Appl Biosci 1993; 9(5): 499-509.

[22] Kashiwabara AY, Vieira DC, Machado-Lima A, Durham AM. Splice site prediction using stochastic regular grammars. Genet Mol Res 2007; 6(1): 105-115.

[23] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997; 268(1): 78-94.

[24] Burge C. Identification of genes in human genomic DNA. Stanford University 1997.

[25] Zhao X, Huang H, Speed TP. Finding short DNA motifs using permuted markov models. Proceedings of the eighth annual international conference on Research in computational molecular biology. San Diego, California, USA: ACM, 2004.

[26] Brunak S, Enqelbrecht J, Knudsen S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 1991; 220(1): 49-65.

[27] Reese MG. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem 2001; 26(1): 51-56.

[28] Roux B, Winters-Hilt S. Hybrid MM/SVM structural sensors for stochastic sequential data. BMC Bioinformatics 2008; 9(Suppl 9): S12.

[29] Saeys Y, Degroeve S, Aeyels D, Rouze P, Van de Peer Y. Feature selection for splice site prediction: A new method using EDA-based feature ranking. BMC bioinformatics 2004; 5: 64.

[30] Sonnenburg S. New methods for detecting splice junction sites in DNA sequence. Master's Thesis, Humbold University: Germany 2002.

[31] Ratsch G, Sonnenburg S, Schafer C. Learning Interpretable SVMs for Biological Sequence Classification. BMC bioinformatics 2006; 7(Suppl 1): S9.

[32] Sun YF, Fan XD, Li YD. Identifying splicing sites in eukaryotic RNA: Support vector machine approach. Comput Biol Med 2003; 33(1): 17-29.

[33] Zhang XHF, Heller KA, Hefter I, Leslie CS, Chasin LA. Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. Genome Res 2003; 13(12): 2637-2650.

[34] Chen T-M, Lu C-C, Li W-H. Prediction of splice sites with dependency graphs and their expanded bayesian networks. Bioinformatics 2005; 21(4): 471-482.

[35] Pertea M, Lin X, Salzberg SL. GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res 2001; 29(5): 1185-1190.

[36] Marashi SA, Eslahchi C, Pezeshk H, Sadeghi M. Impact of RNA structure on the prediction of donor and acceptor splice sites. BMC Bioinformatics 2006; 7: 297.

[37] Sleator RD. An overview of the current status of eukaryote gene prediction strategies. Gene; 461(1-2): 1-4.

[38] Cai D, Delcher A, Kao B, Kasif S. Modeling splice sites with Bayes networks. Bioinformatics 2000; 16(2): 152-158.

[39] Baten A, Chang B, Halgamuge S, Li J. Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 2006; 7(Suppl 5): S15.

[40] Marashi S-A, Eslahchi C, Pezeshk H, Sadeghi M. Impact of RNA structure on the prediction of donor and acceptor splice sites. BMC Bioinformatics C7 - 297 2006; 7(1): 1-8.

[41] Rajapakse JC, Ho LS. Markov encoding for detecting signals in genomic sequences. IEEE/ACM Trans Comput Biol Bioinform 2005; 2(2): 131-142.

[42] HS3D Dataset [http://www.sci.unisannio.it/docenti/rampone/].

[43] BDGP Data [http://www.fruitfly.org/sequence/human-datasets.html]

[44] Reese MG. Improved splice site detection in Genie. J Comput Biol 1997; 4(3): 311-323.

[45] Genie Dataset [http://www.fruitfly.org/seq_tools/datasets/Human/GENIE_96/]

[46] Shamshad A, Bawadi MA, Wan Hussin WMA, Majid TA, Sanusi SAM. First and second order Markov chain models for synthetic generation of wind speed time series. Energy 2005; 30(5): 693-708.

[47] Durbin RE, SR; Krogh, A; Mitchison, G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press 1998.

[48] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3(3): 1157-1182.

[49] Neumann J, Schnorr C, Steidl G. Combined SVM-based feature selection and classification. Mach Learn 2005; 61(1-3): 129-150.

[50] Malousi A, Chouvarda I, Koutkias V, Kouidou S, Maglaveras N. Variable-length positional modeling for biological sequence classification. In: Proceedings of the American medical informatics association symposium (AMIA); 2008: 91-95.

[51] Cohen I, Xiang QT, Zhou S, Sean X, Thomas Z, Huang TS. Feature selection using principal feature analysis. In: Proceedings of the international conference on image processing 2002.

[52] Chen Y-W, Lin C-J. Combining SVMs with various feature selection strategies. Feature extraction: Stud Fuzz soft comp 2006; 207: 315-324.

[53] Vapnik VN. The Nature of Statistical Learning Theory. Springer, N.Y. 1995.

[54] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and Other Kernel-based Learning Methods. Cambridge University Press 2000.

[55] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20(3): 273-297.

[56] Drucker H, Donghui W, Vapnik VN. Support vector machines for spam categorization. IEEE T Neural Networ 1999; 10(5): 1048-1054.

[57] Haykin SS. Neural Networks: Comprehensive Foundation. Prentice Hall 1999.

[58] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 1998; 2(2): 121-167.

[59] Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G. Support vector machines and kernels for computational biology. PLoS comput biol 2008; 4(10): e1000173.

[60] Rivard SR, Mailloux J-G, Beguenane R, Bui HT. Design of high-performance parallelized gene predictors in MATLAB. BMC Res Notes 2012; 5: 183.

[61] Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405(2): 442 - 451.

[62] Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. J Comput Biol 2004; 11(2-3): 377-394.

[63] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surv; 4: 40-79.

[64] Zhang Y, Chu C-H, Chen Y, Zha H, Ji X. Splice site prediction using support vector machines with a Bayes kernel. Expert Syst Appl 2006; 30(1): 73-81.

Received: December 7, 2012 Revised: June 20, 2013 Accepted: February 9, 2013