Computational Biology and Chemistry 73 (2018) 159–170

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Markovian encoding models in human splice site recognition using SVM

Elham Pashaei*, Nizamettin Aydin

Department of Computer Engineering, Yildiz Technical University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 27 May 2017
Received in revised form 4 February 2018
Accepted 5 February 2018
Available online 14 February 2018

Keywords:
Markovian model
Splice sites
Machine learning
DNA encoding method
MMSVM

ABSTRACT

Splice site recognition is among the most significant and challenging tasks in bioinformatics due to its key role in gene annotation. Effective prediction of splice site requires nucleotide encoding methods that reveal the characteristics of DNA sequences to provide appropriate features to serve as input of machine learning classifiers. Markovian models are the most influential encoding methods that highly used for pattern recognition in biological data. However, a direct performance comparison of these methods in splice site domain has not been assessed yet. This study compares various Markovian encoding models for splice site prediction utilizing support vector machine, as the most outstanding learning method in the domain, and conducts a new precise evaluation of Markovian approaches that corrects this limitation. Moreover, a novel sequence encoding approach based on third order Markov model (MM3) is proposed. The experimental results show that the proposed method, namely MM3-SVM, performs significantly better than thirteen best known state-of-the-art algorithms, while tested on HS3D dataset considering several performance criteria. Further, it achieved higher prediction accuracy than several well-known tools like NNsplice, MEM, MM1, WMM, and GeneID, using an independent test set of 50 genes. We also developed MMSVM, a web tool to predict splice sites in any human sequence using the proposed approach. The MMSVM web server can be assessed at https://pashaei.shinyapps.io/mmsvm.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

A eukaryotic gene consists of exons and introns. Exons are regions that contain the information required to produce protein, while introns are regions that are required to be discarded before translation process (Nassa et al., 2013). Boundaries between exons and introns are called splice sites (Roca et al., 2008). There is a highly conserved dimer pattern at both ends of introns (GT-AG rule) that is recognized during initial steps of transcription process (Leo et al., 2002). Acceptor splice sites (3′ss) are boundaries between intron-exon and conserved with dimer AG, while donor splice sites (5′ss) are boundaries between exon-intron and conserved with dimer GT (Sonnenburg et al., 2007) (see Fig 1). As GT and AG dimers represent splice sites, each GT and AG dimer in the sequence of a gene is a candidate for splice site. However, only 0.1% of those candidate sites are real splice sites (Li et al., 2012). So, it is very hard to identify whether a candidate GT/AG dimer in the gene is a real splice site or a decoy site (Li et al., 2012; Meher et al., 2016b).

Except for those canonical splice site patterns (GT-AG rule), there are non-canonical GC-AG and AT-AC patterns as consensus regions for splice site detection. While the GT-AG pattern comprises over 98% of all human splice site, the GC-AG and the AT-AC patterns contain ∼0.9% and ∼0.09% of the human splice sites, respectively (Parada et al., 2014).

RNA splicing is one of the most important biological processes in eukaryotic gene expression. Introns are excised from pre-mRNA in order to produce mature mRNAs for protein translation (Tang et al., 2016). The precise detection of intron boundaries is critical to this process (Jian et al., 2014b). Alternative splicing diversifies gene expression, and mutations at splice sites may also change splicing process in a detrimental way (Tang et al., 2016). The DNA variations that disrupt normal splicing comprise a considerable proportion of genetic disease etiology (Flanagan et al., 2013; Jian et al., 2014b). Therefore, identification of splicing signals is important in both new drug discovery and the biomedical research (Xu et al., 2017). Moreover, with the rapid growth of the sequenced genome utilizing advanced sequencing technologies like RNA-seq, analyzing these sequences to identify location and structure of their genes is one of the main issues in bioinformatics. Splice site

* Corresponding author.
E-mail addresses: elham.pashaei@std.yildiz.edu (E. Pashaei), naydin@yildiz.edu (N. Aydin).

https://doi.org/10.1016/j.compbiolchem.2018.02.005
1476-9271/© 2018 Elsevier Ltd. All rights reserved.



Fig. 1. Representation of acceptor and donor splice site in DNA sequence.

identification is the main part of the genome annotation (Nassa et al., 2013).

Splice site detection is very expensive and time consuming based only on experimental techniques like qPRC technology (Xu et al., 2017). Recently developed RNA-Seq technology has provided an unprecedented approach to obtain various information from transcriptome by mapping million short reads to the reference genome (Costa et al., 2013). However, multiple potential matches of the short read on the reference genome make sequence mapping prone to errors (Li and Dewe, 2011). Splice site identification through new computational methods based on machine learning is an alternative approach for medical geneticists which provides more economical, and quicker way of recognizing the splicing signals (Xu et al., 2017). On the other hand, the main goal of using splice site prediction tools has shifted from the recognition of possible splice sites to the prediction of the effect of a genetic variant (mutation) within splicing consensus regions (GT/AG dinucleotides and their close vicinity) in the human genome project era (Jian et al., 2014a,b). A thorough performance comparison of the most common in silico splice site prediction tools for splicing defect prediction has been presented in (Jian et al., 2014a) and (Desmet et al., 2010). In addition, several dedicated tools such as SPANR (Xiong et al., 2015), Skippy (Woolfe et al., 2010), and Human Splicing Finder (Desmet et al., 2009) have been developed to predict the impact of mutations on splicing too.

Various computational methods have been developed for splice site recognition. These methods are based on probabilistic approach (Burge and Karlin, 1997; Pertea et al., 2001; Staden, 1984; Yeo and Burge, 2004; Zhang and Marr, 1993), hidden Markov model (Maji and Garg, 2013; Zhang et al., 2010), decision trees (Lopes et al., 2007), support vector machine (SVM) (Baten et al., 2006; Baten et al., 2008; Huang et al., 2006; Li et al., 2012; Maji and Garg, 2014; Wei et al., 2013; Wei et al., 2012), Bayesian network (Chen et al., 2005), random forest (Meher et al., 2016a; Pashaei et al., 2017), artificial neural network (NN) (Bin and Jing, 2014; Loi and Rajapakse, 2003; Nassa et al., 2013) and so on. Among them, SVM-based classification technique produces more accurate result than the other machine learning approaches and probabilistic approaches for predicting splice sites (Bari et al., 2012; Bari et al., 2014; Kamath et al., 2014). However, due to existence of enormous false splice sites compared to true sites in genomes, splice site prediction problem is still one of the most important and challenging task in bioinformatics (Baten et al., 2008; Li et al., 2012; Sonnenburg et al., 2007). Therefore, it is essential to develop new methods or improve existing methods to predict splice sites more accurately (Bin and Jing, 2014; Meher et al., 2016b).

Feature extraction, namely nucleotide encoding, is a fundamental issue in the prediction of splice sites using machine learning techniques (Huang et al., 2006; Wei et al., 2012). Encoding approaches try to reveal characteristics, dependencies and relationship of nucleotides in the splice sites and the surrounding region well enough to provide informative features, which are then used as input in machine learning classifiers for prediction (Bari et al., 2014; Baten et al., 2006). Markovian models are regarded as an efficient technique that highly used for pattern recognition in biological data (Baten et al., 2006; Baten et al., 2008; Fink, 2014). However, to the best of our knowledge, there is no detailed discussion and categorization of the methods with regards to using true or both true/false DNA sequences.

This paper provides a new precise evaluation and categorization of Markovian encoding models utilizing the most widely used SVM classifier, and proposes a novel efficient approach for human splice site recognition based on the hybrid of the third order Markovian encoding method (MM3) and SVM classifier. The results showed that utilizing both true and false DNA sequences produce a better model for splice site prediction than only considering the true sequences. It was also found that the proposed MM3-SVM method shows a superior performance compared with a lot of current best methods and existing well-known tools on human splice site prediction.

2. Related work

The splice site recognition process is subdivided into two separate classification problems, i.e. acceptor site prediction and donor site prediction. These are binary classification problems in which the process involves first transforming sequence data into numeric feature vectors over which an underlying classifier operates (Kamath et al., 2014). The learned model is then applied to unknown sequences to predict real splice site from decoy one.

To facilitate the discussion of results in the comparison of our proposed method with state-of-the-art methods and tools, we categorized our review below into separate parts and provided a summary of representative methods in each category.

2.1. Markovian encoding base methods for splice site prediction

Baten et al. (2006) proposed the MM1 as an effective preprocessing scheme for the poly SVM, using only true splice sites (MM1-SVM). Later, the author improved performance of the previous work using Fisher score feature ranking method and RBF kernel (Reduced MM1-SVM) (Baten et al., 2008). In Maji and Garg (2014), a three-stage algorithm was proposed which is a combination of the MM2 encoding method, principle feature analysis feature selection method and SVM classifier for solving the splice site problem (MM2F-SVM). Interestingly, it has not been mentioned clearly how the MM2 model was constructed utilizing both sites. Loi and Rajapakse (2002), Loi and Rajapakse (2003) proposed MCM, which is hybrid of MM1 and MM2, as an efficient encoding method and fed its produced features to the neural network. Later, Goel et al. (2015) improved the performance of the previous approach employing SVM classifier (MCM-SVM). Accurate segmentation of DNA sequence is a key step in the performance of the MCM encoding method, which is not easy work (Pashaei et al., 2017). Inspired by the approach published in Wei et al. (2013), we developed DMM2-SVM (Double MM2-SVM) methods which consider a new hybrid of the MM2 model utilizing both MCC and false DNA sequences (Pashaei et al., 2016a).

This study puts forward a novel method by the hybrid of MM3 and SVM to predict human splice sites more accurately (MM3-SVM). Moreover, a new evaluation of Markovian encoding models is provided.

2.2. Current state-of-the-arts methods for splice site prediction

To improve accuracy and reduce time complexity of splice site detection approaches, Zhang et al. (2006) used linear SVM with a Bayes kernel (B-SVM). Huang et al. (2006) proposed an efficient DNA encoding method by considering pairwise nucleotides (PN) and calculating their frequency difference between true and false

sites (FDTF). Then, the SVM classifier was applied to predict splice sites (PN FDTF-SVM). In Meher et al. (2016a), the PN-FDTF encoding method was merged with RF (P-RF), which showed an improvement in prediction of donor splice sites. Zhang et al. (2010), proposed length variable Markov model (LVMM2/OLVWMM2) in which the second order Markov model (MM2) is used for features extraction and some threshold for feature selection and classification. Although the method produces good accuracy finding the optimal value for thresholds of the method is a difficult task. Three DNA encoding approaches which were orthogonal encoding, codon encoding, and sequential information were proposed in Wei et al. (2012) and a linear SVM was employed as the classifier (ECS-LSVM). In Wei et al. (2013), the distribution of tri-nucleotides was combined with MM1 encoding method using both true and false sites (DM). Then, F-score feature ranking method was used for feature selection and the SVM classifier for classification (DM-SVM). Li et al. (2012) proposed a high-throughput method, MSC + Pos + APR-SVM, which is one of the leading methods in human splice site detection considering its high accuracy. The method serves a hybrid of position features (Pos), adjacent position relationship (APR) features, and multiscale sequence component (MSC) features as input into SVM. The method requires high computational cost due to so many useless features that are produced by the MSC encoding method (Li et al., 2012). In Kamath et al. (2012), an evolutionary algorithm was utilized to generate complex (local and non-local) features (GF-EA) from DNA sequence to predict splice sites using SVM classifier. In the recent paper by Kamath et al. (2014), an interesting algorithmic framework (EFFECT) has been proposed to automate the process of extracting discriminatory features for detecting functional signals in DNA sequences. The EFFECT approach utilizes an evolutionary algorithm for both feature construction and feature selection stages.

2.3. Popular web tools for splice site prediction

The position weight matrix method (WMM) (Staden, 1984; Tavares et al., 2009) is the earliest and popular method that calculates the frequency of each nucleotide at each position to score splice sites motifs and utilizes a threshold to predict true sites. The Position weight array method (WAM) also known as the MM1 (Salzberg, 1997; Zhang and Marr, 1993), is a position weight matrix where every position is a Markov chain instead of a simple nucleotide distribution. The NNsplice (Reese et al., 1997a) is a tool based on the combination of sparse encoding methods (A = 0001, C = 0010, G = 0100 and T = 1000) and the NN for splice site recognition. The MEM (Yeo and Burge, 2004) employs Maximum Entropy distribution consistent with reliably estimated constraints, for approximating the short sequence motif distributions and the GeneID (Blanco et al., 2007) predicts splice sites by using the WAM (Salzberg, 1997; Zhang and Marr, 1993). The SplicePort (Dogan et al., 2007) utilizes feature generation algorithm (FGA), one of the most successful feature extraction and selection method, for automatically providing features as input to the SVM classifier for splice site prediction.

3. Methodology

The proposed MM3-SVM framework consists of two stages. In the first stage, a third order Markov model (MM3) is used for feature extraction from DNA sequences. These features are then fed to the second stage, where SVM with RBF kernel is used for classification. A schematic of the framework has been given in Fig. 2.

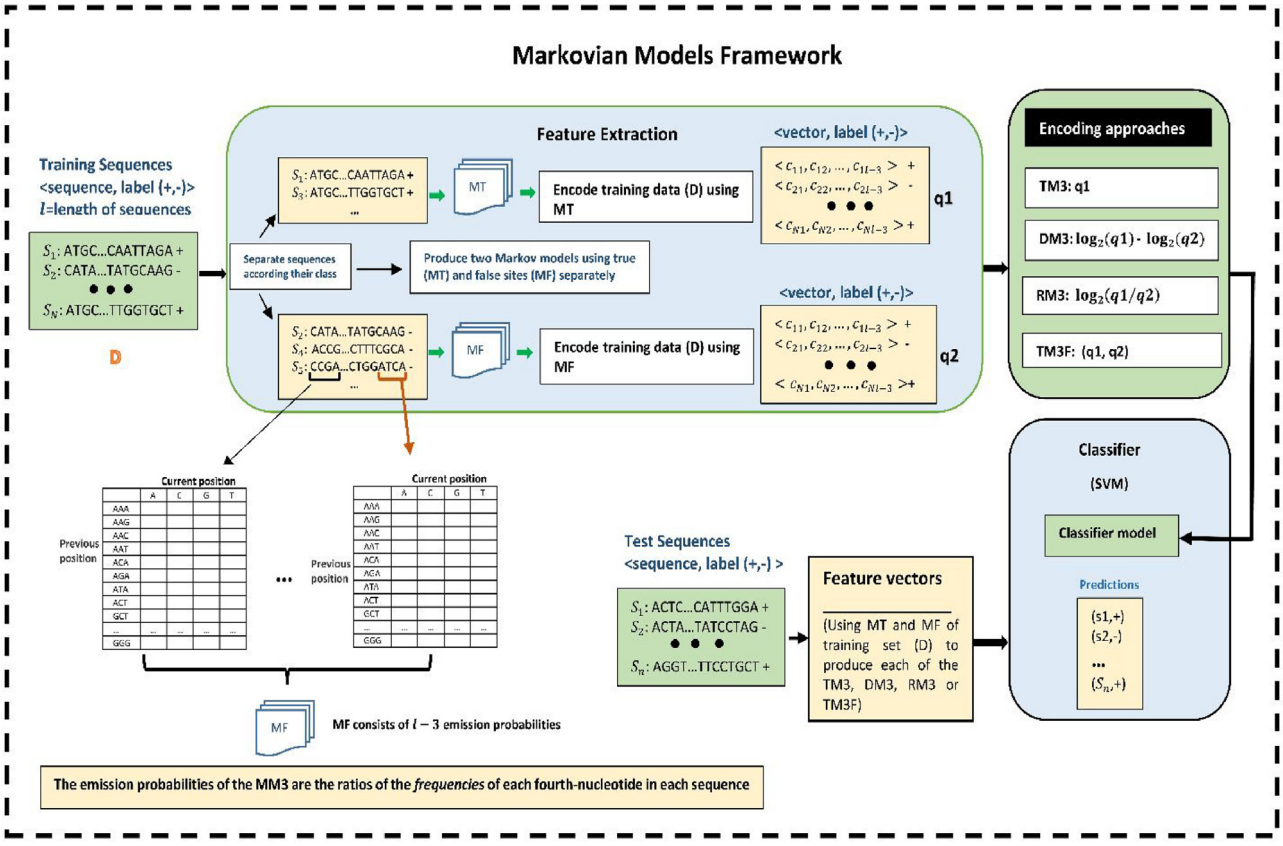


Fig. 2. The MM3-SVM framework consists of two algorithms, MM3 and SVM.

3.1. Encoding of sequence data using Markovian models (feature extraction)

A Markov chain is represented as a sequence of states, where the probability distribution for each state depends on a fixed number of previous states (Salzberg et al., 1999). In the context of DNA sequence analysis, each nucleotide represents a state in the Markov chain (Baten et al., 2008). The probability distribution of a base is calculated by examining the frequency of a fixed number of bases just prior to that base in the sequence. For example, an MM1, MM2, and MM3 predict a base by looking at the one, two and three previous bases, respectively (Salzberg et al., 1999).

To learn the Markov model, some position specific probabilistic parameters (emission probabilities) should be calculated from training sequences. We note that all the training sequences were aligned with respect to the dinucleotide consensus AG/GT before. To create Markovian models, we subdivide training sequences into two parts using a class label. So, true training sequences (positive samples) were utilized to create "true" Markovian model (TM), while false training sequences (negative samples) were employed to build "false" Markovian model (FM). For a sequence of length$l$, there are $l-k$ emission probabilities for each of Markovian models, where em > k is the order of Markov chain. The emission probabilities of the Markovian models are the ratios of the frequencies of each $k+1$ nucleotide substring in each sequence position as shown in the following equation (Baten et al., 2006; Baten et al., 2008),

$$P_i(s_i) = P(s_i|s_{i-1}, s_{i-2}, \ldots, s_{i-k}) = N(s_{i-k}, \ldots, s_{i-2}, s_{i-1}, s_i)/N(s_{i-k}, \ldots, s_{i-2}, s_{i-1}) \quad (1)$$

where $N(s_{i-k}, \ldots, s_{i-2}, s_{i-1}, s_i)$ means the occurrence number of $(s_{i-k}, \ldots, s_{i-2}, s_{i-1}, s_i)$ substring in the training sequence at position $i \in (1, 2, l)$, and, $s_i \in (A, C, G, T)$ is DNA nucleotide. In this paper $k = 1, 2, 3$.

The Markovian models (MT and MF) were used for encoding the training sequences. We called encoded training set q_1 and q_2 utilizing MT and MF, respectively. Then, the final encoded training set was obtained by one of the following approaces:

- Utilizing only true Markovian model, our final encoded training set is equal to q_1, named as TM
- Utilizing both Markovian models, ratio of $\ln(q_1)$ and $\ln(q_2)$ is calculated to construct the final encoded training set, named as RM
- Utilizing both Markovian models, difference of $\ln(q_1)$ and $\ln(q_2)$ is computed to build the final encoded training set, named as DM
- Utilizing both Markovian models, hybrid of q_1 and q_2 is considered to construct the final encoded training set, named as TMF

The first approach was proposed by Baten in Baten et al. (2006), while others were investigated in this study for the first time.

3.2. Classifier

We employed SVM as our classifier to predict splice sites, due to its superior performance. Several reasons describe its excellent performance in many domains, including the ability to deal with high dimensional data, the ability to learn complex classification function efficiently, fast convergence and utilizing powerful regularization principles to avoid overfitting (Statnikov et al., 2008). The underlying idea of SVM classifier is to separate two classes of data by constructing a maximal margin hyperplane. To learn linearly inseparable data, the SVM transforms data into higher dimensional space via a kernel function, where a separating hyperplane is found. The binary SVM classifies unknown samples according to the side of the hyperplane they belong to (Statnikov et al., 2011).

The "e1071" R package was used for implementation of SVM and the radial basis function (RBF) kernel was employed to learn the models. A grid-based search was utilized to tuning the parameters C and γ.

3.3. Datasets, performance measurements and experimental setting for analysis

3.3.1. Evaluation datasets

The experiments were performed on the well-known benchmark Homo Sapiens Splice Site Dataset (HS3D) (Pollastro and Rampone, 2002), which is available at URL: http://www.sci.unisannio.it/docenti/rampone/ To assess the performance of the proposed methods, the dataset was split into an acceptor subset and a donor subset, and evaluation was done separately on each subset. The length of each DNA sequence in each subset is 140 nucleotide. The Acceptor set consists of 2880 true and 329374 false sequences and dinucleotide AG is conserved at positions 69 and 70 of the sequences. The donor set consists of 2796 true and 271937 false sequences and dinucleotide GT is conserved at positions 71 and 72 of the sequences. We constructed two datasets for each acceptor and donor sets, which are known as balanced and imbalanced datasets. The ratio between the size of the positive and negative classes is 1:1 for the balanced dataset, while the ratio is 1:10 for the imbalanced dataset. In other words, we choose all the positive samples and randomly selected the same and 10 times more negative samples to construct the balanced and unbalanced datasets, respectively. This dataset has also been utilized in earlier studies (Maji and Garg, 2014; Meher et al., 2016a; Meher et al., 2016b; Wei et al., 2013; Zhang et al., 2010).

To find optimal length of flanking region of splice sites, we extended the length of sequences in the HS3D dataset from 140 to 202 nucleotide (nt) (100 +AG/GT+100) for both balanced (1:1) donor and acceptor sites. For this purpose, we employed BLAST and a subroutine in R to obtain genomic features of the sequences. Then, we extracted longer sequences from NCBI RefSeq collection of human pre-mRNA sequences (version GRCh38) (https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml) using those genomic features.

For model training of the provided web tool, we utilized the latest release of GENCODE annotation (version 27) (https://www.gencodegenes.org) for extracting splice-site sequences from the complete human genome assembly (version GRCh38.p10) (Harrow et al., 2012). We wrote a script in R, in which "Biostrings", "GenomicRanges" "GenomicFeatures", and "Rsamtools" R Bioconductors was utilized to extract splice-sites sequences based on the GTF file annotation. We broke down the genome assembly into chromosomes and considered chr1 to chr22 as well as X and Y chromosomes for sequence analysis. We count the borders of all exon start (acceptor) and end (donor) positions whose sites followed the AG-GT dinucleotide consensus. Negative sequences were constructed by randomly selecting AG or GT locations that were not annotated as splice sites. For providing non-canonical splice-site sequences, we followed the same steps, as we did for canonical sites. The provided sequences have 70 nt in length from downstream and upstream of the annotated splice sites (70+AG/GT +70). The Detailed information about the provided dataset, called GCHSS, is summarized in Table 1.

The performance of the proposed method was further evaluated by using an independent test set of 50 genes, which was extracted from Burset and Guigo test dataset (Burset and Guigó, 1996). This benchmark test dataset has been used earlier in Meher et al. (2016b) to assess the efficiency of the proposed

Table 1
Characteristics of the provided splice site dataset using GENCODE annotation.

Type	Dinucleotide	Genome (%)	Extract from	Utilized Region	Utilized Strand	Extracted True sites	Extracted Decoy sites
canonical splice sites	AG-GT	~ 98%	All chromosome	CDS	"+"	87,350	200,000
non-canonical splice sites	AG-GC	~ 0.9%	All chromosome	CDS	"+" and "-"	5879	30,000
	AC-AT	~ 0.09%	All chromosome	Exon	"+" and "-"	174	1000

method, which is publicly available at http://genome.crg.es/datasets/genomics96/ An imbalanced dataset with a ratio of 1:5 was constructed, by choosing all the true site and randomly selected 5 times more false site than true ones, for both donor and acceptor sites separately.

3.4. Performance measurements

To evaluate the classification performance, seven most popular measures namely sensitivity (S_n), specificity (S_p), global accuracy (Q^9), Matthew's correlation coefficients (Mcc) (Li et al., 2012), area under ROC curve (AUC) (Baten et al., 2008), area under precision-recall curve (auPRC) (Sonnenburg et al., 2007) and F-measure (Meher et al., 2016a; Zhang et al., 2006) were taken. The Q^9 measure is defined as follow:

$$q^9 = \begin{cases} Q^9 = (1+q^9)/2 \\ \dfrac{(TN-FP)}{(TN+FP)} if\ \ TP+FN=0 \\ \dfrac{(TP-FN)}{(TP+FN)} if\ \ TN+FP=0 \\ 1-\sqrt{2\left[\left(\dfrac{FN}{TP+FN}\right)^2+\left(\dfrac{FP}{TN+FP}\right)^2\right]} if TP+FN\neq 0 \\ \quad and\ \ TN+FP\neq 0 \end{cases} \quad (2)$$

where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives, respectively (Wei et al., 2013; Zhang et al., 2010). The larger values of all measures show better performance of a classifier. The AUC is effective and widely used criteria in binary classification problems (Bari et al., 2014; Meher et al., 2016b). However, the Mcc and auPRC are better measurements for assessing the performance of the classifiers when the class distribution is heavily unbalanced (Kamath et al., 2014; Saito and Rehmsmeier, 2015; Sonnenburg et al., 2007).

Ten-fold cross-validation (CV) procedure was conducted to evaluate the performance of the classifiers. The idea behind CV is to divide the dataset into 10 equally parts (folds), train on 9/10 folds, and test on the remaining fold. This process is repeated 10 times and averaged over 10 testing sets. The final performance was reported in terms of the evaluation criteria described above. Mann-Whitney *U* test was used to measure statistical significance at 95% confidence intervals, considering F-measure criteria. The "stats" R package was used to implement the *U* test.

It should be noted that due to the large imbalance between positive and negative samples of the unbalanced dataset, the evaluating indicator tends to be biased towards the majority class (Lin and Che, 2012). So we used "CV includes sampling" technique (Blagus and Lusa, 2015) to assess performance of imbalanced datasets, in which a random under-sampling technique (Longadge et al., 2013) was applied to the training set (9 folds) and a reduced (balanced) training set was obtained at each iteration of CV for training phase, while the test set (remaining fold) was not modified.

We should note that different models were trained for classification of canonical and non-canonical splice sites in our web tool. Considering canonical 5′ss and 3′ss, we trained the model on the imbalanced HS3D with the ratio of 1:10, choosing 100% of positive samples (for comparison), while we trained the model on the imbalanced GCHSS dataset with the ratio of 1:5, selecting 50% of positive samples (embedded in server). Considering non-canonical splice sites, the GCHSS with the ratio of 1:2 and 1:5 was used to train the models for the AT-AC and GC-AG patterns respectively, in which 100% of positive samples were chosen.

3.4.1. Performance evaluation and categorization of Markovian Encoding models using HS3D dataset

There are three main Markovian-based encoding methods on splice site prediction in the literature, which are MM1 (Baten et al., 2006; Baten et al., 2008), MM2 (Maji and Garg, 2014) and MCM (Goel et al., 2015; Loi and Rajapakse, 2003). The contribution of our study is to introduce MM3 for splice site prediction. Therefore, we grouped Markovian encoding methods into four categories: MM1, MM2, MCM and proposed MM3. To provide a rigorous evaluation of Markovian encoding methods, we examine the methods both in the setting when only true samples are used and when both true and false samples are used. So, four different cases have been considered for each category except MCM, namely TM, DM, RM, and TMF (see Section 3.1). The TM case was proposed by Baten (Baten et al., 2006; Baten et al., 2008), while the TMF case was proposed in our previous study (Pashaei et al., 2016a), inspired by the approach published in Wei et al. (2013). The DM and RM cases have never been considered in the literature when a machine learning classifier is used for prediction rather than a threshold value.

We also evaluated three different cases of MCM encoding methods by using true sites (TM) and hybrid of true and false sites (TMF) in model's region part. More details about structure of this encoding method has been given in Goel et al. (2015), Loi and Rajapakse (2002, 2003); Pashaei et al. (2017).

Totally, fifteen various Markovian models were combined with SVM classifier to fulfill the evaluation. A 10-fold cross validation was performed on both balanced and unbalanced acceptor and donor HS3D datasets. Parameters of SVM classifier for each Markovian model have been provided in supplement.

3.4.2. Comparison with existing best state-of-the-arts methods using HS3D dataset

In order to assess the efficiency of the proposed method, we compared it with existing best current state-of-arts methods: MM1-SVM (Baten et al., 2006), Reduced MM1-SVM (Baten et al., 2008), B-LSVM (Zhang et al., 2006), MM2F-SVM (Maji and Garg, 2014), PN FDTF-SVM (Huang et al., 2006), LVMM (Zhang et al., 2010), ECS-LSVM (Dan Wei et al., 2012), Sparse-SVM (Sun et al., 2003), DM-SVM (Wei et al., 2013), MCM-SVM (Goel et al., 2015), DMM2-SVM (Pashaei et al., 2016a), SCMM1-AdaBoost (Pashaei et al., 2016b), P1-RF (Meher et al., 2016a) and High accuracy MSC-Pos-APR-SVM (Li et al., 2012), using both balanced and unbalanced of acceptor and donor HS3D datasets. The performances of the methods were measured in all the terms mentioned in the previous section, while averaged over 10-fold cross validation. The Mann-Whitney *U* test was used to show the significance of the results, considering F-measure criteria. For this purpose, the results of the MM3-SVM (10-fold-CV) were compared with results of other methods (10-fold-CV) separately. All the methods were

implemented in R programming language, while it has been tried to find the best value for their parameters in both encoding and classification parts.

3.4.3. Comparison with existing tools using an independent test dataset

The performance of the MMSVM tool was compared with several well-known splice site predictor tools: WMM (Staden, 1984), NNsplice (Reese et al., 1997b), MM1(Durbin et al., 1998), MEM (Yeo and Burge, 2004), GeneID (Blanco et al., 2007), and SplicePort (Dogan et al., 2007), using an independent test dataset of 50 genes. Web link of the WMM, MM1, and MEM is http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html The links for NNsplice and the GeneID are http://www.fruitfly.org/seq_tools/splice.html and http://genome.crg.es/software/geneid/geneid.html respectively, while the SplicePort web tool can be accessed at http://spliceport.cbcb.umd.edu/. We wrote a subroutine in R programming language for analyzing text output of each tool, whereas the "PRROC" R package was used for calculating the values of the ROC and auPRC metrics. We also set the threshold parameter (for classification) to 3 for all the WMM, MM1 and MEM methods, while the value of the threshold was adjusted to 0.4 for the NNsplice and 0.0 for the GeneID and SplicPort tools. Besides, each tool utilizes the different length of the flanking region. The WMM, MM1, and MEM use very short flanking sequences, i.e. 23 bps for acceptor and 9 bps for donor splice sites. The NNSplice uses 42 bps for acceptor, and 17 bps for donor sites. The GeneID uses 10 bps for both sites (Parra et al., 2000), while SplicePort uses 80 bps on each side of the consensus dinucleotide.

4. Results

4.1. Performance analysis of Markovian Encoding models using HS3D dataset

We summarized the performance of the Markovian encoding methods (MM1, MM2, proposed MM3) under each of the four different models (TM, DM, RM and TMF), using SVM classifier, as mentioned in Subsection 3.1. The results are shown in Table 2 and Table 3 for acceptor HS3D dataset (both balanced and imbalanced) and donor HS3D dataset (both balanced and imbalanced), respectively. Moreover, the performance of the MCM encoding method was evaluated to facilitate comparison between Markovian encoding based methods.

From Table 2, it is observed that the DM, RM, and TMF models perform better than TM model in all MM1, MM2, and MM3 encoding methods. Therefore, it can be said that utilizing both true and false sites for extracting Markovian modeling is more efficient than considering only true sites by taking into account the accuracy. It is further observed that the proposed MM3 encoding method outperforms other encoding methods in all the terms for both balanced and imbalanced acceptor HS3D dataset. Besides, the TM3F method has higher accuracy than the other models in the MM3 method. We note that DM and RM approximately produce the same results when the parameter of the SVM is identical. Table 3 for donor dataset exhibits the same results. Again, it shows that MM3 performs better than MM1, MM2, and MCM, while the TM3F achieved higher values of metrics than other models.

Table 2
Performance comparison of Markovian models using SVM on predicting ACCEPTOR splice sites of both balanced and unbalanced HS3D datasets.

Encoding Methods	Type of Model	Balanced Acceptor splice sites							Imbalanced Acceptor splice sites						
		S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure
MM1	TM1	90.17	87.54	88.65	77.78	95.33	94.19	89.00	91.18	88.28	89.54	58.30	95.76	70.71	59.14
		(2.01)	(2.15)	(1.06)	(2.06)	(0.42)	(0/92)	(1.02)	(2.56)	(0.54)	(1.15)	(1.98)	(0.57)	(2.08)	(1.66)
	DM1	93.37	89.51	91.15	82.97	96.91	95.94	91.60	93.37	89.32	91.07	61.62	96.95	77.31	62.22
		(1.48)	(1.30)	(0.63)	(1.32)	(0.45)	(1.08)	(0.66)	(1.35)	(0.56)	(0.61)	(1.38)	(0.39)	(1.38)	(1.31)
	RM1	93.37	89.48	91.13	82.93	96.91	95.95	91.59	93.33	89.31	91.06	61.59	96.95	77.30	62.19
		(1.48)	(1.25)	(0.60)	(1.28)	(0.46)	(1.09)	(0.65)	(1.41)	(0.58)	(0.63)	(1.41)	(0.39)	(1.37)	(1.34)
	TM1F	92.43	90.38	91.31	82.84	97.06	96.32	91.49	92.60	90.16	91.21	62.81	97.03	78.19	63.64
		(1.19)	(0.97)	(0.68)	(1.42)	(0.31)	(0.67)	(0.71)	(2.10)	(0.45)	(0.79)	(1.00)	(0.41)	(1.31)	(0.81)
MM2	TM2	91.42	87.85	89.43	79.34	95.94	94.86	89.81	91.74	88.85	90.16	59.72	96.23	72.28	60.53
		(1.35)	(1.12)	(0.54)	(1.07)	(0.48)	(0.56)	(0.56)	(1.44)	(0.67)	(0.85)	(1.89)	(0.55)	(2.50)	(1.78)
	DM2	93.78	91.53	92.46	85.36	97.90	97.26	92.74	93.51	90.90	92.06	64.96	97.61	82.51	65.74
		(1.65)	(1.54)	(0.84)	(1.7)	(0.39)	(0.68)	(0.85)	(1.06)	(0.54)	(0.41)	(1.09)	(0.37)	(1.19)	(1.14)
	RM2	93.82	91.53	92.47	85.40	97.90	97.26	92.76	93.51	90.88	92.05	64.92	97.61	82.51	65.70
		(1.66)	(1.54)	(0.83)	(1.68)	(0.39)	(0.68)	(0.84)	(1.06)	(0.53)	(0.43)	(1.12)	(0.36)	(1.19)	(1.16)
	TM2F	94.16	91.91	92.82	86.13	97.86	97.33	93.11	94.06	90.37	91.93	64.18	97.46	80.98	64.80
		(2.01)	(1.07)	(0.79)	(1.78)	(0.36)	(0.64)	(0.93)	(1.74)	(0.45)	(0.63)	(1.35)	(0.39)	(1.85)	(1.21)
MCM	C = TM1	92.08	87.85	89.64	80.04	96.04	95.06	90.17	92.43	88.64	90.31	59.76	96.30	73.64	60.43
	M = TM2	(1.71)	(1.54)	(0.61)	(1.28)	(0.45)	(0.94)	(0.65)	(1.50)	(0.54)	(0.68)	(1.46)	(0.47)	(1.99)	(1.34)
	C = TM1F	92.68	88.54	90.34	81.30	96.56	95.74	90.80	93.20	89.03	90.82	60.96	96.72	75.46	61.55
	M = TM2	(1.31)	(1.46)	(1.00)	(1.94)	(0.38)	(0.88)	(0.95)	(1.66)	(0.60)	(0.66)	(1.51)	(0.37)	(1.89)	(1.41)
	C = TM1F	92.46	90.80	91.56	83.28	97.21	96.57	91.70	92.33	90.92	91.48	64.28	97.26	80.29	65.25
	M = TM2t*	(0.94)	(1.23)	(0.78)	(1.53)	(0.48)	(0.81)	(0.75)	(2.53)	(0.64)	(1.16)	(2.07)	(0.52)	(2.63)	(1.86)
MM3	TM3	92.78	91.15	91.82	83.96	97.27	96.61	92.02	93.13	90.94	91.93	64.82	97.34	80.10	65.67
		(1.92)	(1.51)	(1.04)	(2.22)	(0.51)	(0.75)	(1.12)	(1.28)	(0.55)	(0.70)	(1.61)	(0.45)	(1.97)	(1.55)
	DM3	96.63	95.49	95.89	92.14	99.35	99.00	96.08	94.79	93.06	93.80	70.88	98.60	89.89	71.76
		(1.43)	(1.20)	(0.76)	(1.35)	(0.14)	(0.14)	(0.69)	(1.35)	(0.48)	(0.44)	(1.27)	(0.34)	(1.31)	(1.27)
	RM3	96.60	95.49	95.87	92.11	99.35	99.00	96.06	94.76	93.07	93.79	70.89	98.60	89.89	71.78
		(1.44)	(1.20)	(0.75)	(1.34)	(0.14)	(0.14)	(0.69)	(1.35)	(0.49)	(0.44)	(1.30)	(0.34)	(1.31)	(1.30)
	TM3F	**97.19**	**95.42**	**96.09**	**92.64**	**99.28**	**98.94**	**96.33**	**95.07**	**93.38**	**94.12**	**71.93**	**98.50**	**88.42**	**72.82**
		(1.43)	(1.21)	(0.91)	(1.71)	(0.23)	(0.25)	(0.86)	(1.22)	(0.63)	(0.59)	(1.78)	(0.34)	(1.58)	(1.80)

The value inside the brackets () are the standard errors.
Bold printed values reflect criteria with highest scores.
* Using TM2 model twice (TM2t) for coding upstream and downstream segments.

Table 3
Performance comparison of Markovian models using SVM on predicting DONOR splice sites of both balanced and unbalanced HS3D datasets.

Encoding Methods	Type of Models	Balanced Donor splice sites							Imbalanced Donor splice sites						
		S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure
MM1	TM1	93.24	90.99	91.90	84.24	97.12	96.24	92.20	93.10	91.08	91.96	65.10	97.15	79.05	65.98
		(1.88)	(1.77)	(1.01)	(2.07)	(0.57)	(0.92)	(1.04)	(1.89)	(0.48)	(0.90)	(1.66)	(0.35)	(2.35)	(1.50)
	DM1	96.03	91.34	93.21	87.49	97.97	97.12	93.83	95.71	91.98	93.54	68.81	98.12	85.10	69.41
		(1.04)	(1.15)	(0.62)	(0.95)	(0.46)	(0.76)	(0.48)	(1.00)	(0.60)	(0.55)	(1.68)	(0.24)	(1.99)	(1.68)
	RM1	96.03	91.34	93.21	87.49	97.97	97.12	93.83	95.67	92.00	93.54	68.83	98.12	85.10	69.44
		(1.04)	(1.18)	(0.64)	(0.97)	(0.46)	(0.76)	(0.48)	(1.05)	(0.60)	(0.55)	(1.68)	(0.24)	(2.00)	(1.68)
	TM1F	95.03	93.03	93.87	88.09	97.98	97.16	94.09	94.92	92.83	93.72	70.41	98.10	84.59	71.24
		(1.15)	(1.49)	(0.88)	(1.62)	(0.63)	(0.98)	(0.80)	(1.60)	(0.64)	(0.77)	(1.88)	(0.23)	(1.91)	(1.85)
MM2	TM2	93.60	91.56	92.42	85.20	97.29	96.31	92.65	93.42	91.42	92.30	66.04	97.37	80.27	66.92
		(1.51)	(1.42)	(0.76)	(1.46)	(0.58)	(1.08)	(0.74)	(1.72)	(0.40)	(0.80)	(1.48)	(0.31)	(2.98)	(1.33)
	DM2	95.96	93.53	94.56	89.52	98.58	97.98	94.80	95.64	93.28	94.31	72.01	98.53	88.59	72.80
		(0.96)	(0.95)	(0.58)	(1.09)	(0.45)	(0.57)	(0.55)	(1.06)	(0.46)	(0.54)	(1.49)	(0.24)	(1.96)	(1.46)
	RM2	95.96	93.53	94.56	89.52	98.58	97.98	94.80	95.64	93.30	94.32	72.04	98.53	88.60	72.84
		(0.96)	(0.95)	(0.58)	(1.09)	(0.45)	(0.57)	(0.55)	(1.06)	(0.46)	(0.54)	(1.50)	(0.24)	(1.96)	(1.47)
	TM2F	95.96	94.06	94.85	90.05	98.54	97.99	95.06	95.14	93.94	94.47	73.48	98.48	87.78	74.43
		(1.37)	(1.23)	(0.97)	(1.98)	(0.54)	(0.71)	(0.99)	(0.86)	(0.57)	(0.36)	(1.47)	(0.24)	(2.11)	(1.55)
MCM	C(TM1)	94.49	91.67	92.83	86.22	97.76	96.91	93.17	94.49	92.00	93.09	68.09	97.71	82.09	68.87
	M(TM2)	(1.64)	(1.49)	(0.91)	(1.66)	(0.47)	(0.90)	(0.85)	(1.33)	(0.58)	(0.63)	(1.55)	(0.31)	(2.58)	(1.54)
	C(TM1F)	95.96	93.31	94.34	89.33	98.17	97.28	94.70	95.10	93.18	94.02	71.40	98.18	84.48	72.25
	M(TM2)	(1.73)	(1.27)	(0.83)	(1.76)	(0.56)	(1.00)	(0.89)	(1.01)	(0.56)	(0.41)	(1.33)	(0.22)	(2.14)	(1.40)
	C(TM1F)	96.39	93.56	94.73	90.00	98.52	97.91	95.04	95.75	94.24	94.88	74.71	98.66	89.01	75.61
	M(TM2t)*	(1.13)	(0.94)	(0.69)	(1.45)	(0.57)	(0.81)	(0.72)	(1.45)	(0.45)	(0.68)	(1.42)	(0.25)	(1.45)	(1.39)
MM3	TM3	94.71	92.56	93.50	87.30	98.09	97.41	93.70	94.49	92.72	93.50	69.84	98.19	86.22	70.73
		(1.13)	(1.23)	(0.84)	(1.57)	(0.48)	(0.79)	(0.78)	(1.37)	(0.47)	(0.64)	(1.47)	(0.22)	(2.12)	(1.42)
	DM3	97.03	95.74	96.25	92.79	99.30	98.90	96.41	96.64	94.48	95.40	75.98	98.96	92.46	76.78
		(1.12)	(0.97)	(0.67)	(1.34)	(0.27)	(0.29)	(0.68)	(0.81)	(0.53)	(0.45)	(1.71)	(0.25)	(1.29)	(1.73)
	RM3	97.03	95.74	96.25	92.79	99.30	98.90	96.41	96.64	94.47	95.40	75.95	98.96	92.46	76.75
		(1.12)	(0.97)	(0.67)	(1.34)	(0.27)	(0.29)	(0.68)	(0.81)	(0.55)	(0.46)	(1.75)	(0.25)	(1.30)	(1.76)
	TM3F	**97.09**	**95.76**	**96.28**	**92.80**	**99.30**	**98.88**	**96.45**	**96.60**	**94.94**	**95.70**	**77.38**	**98.99**	**91.57**	**78.21**
		(1.34)	(1.19)	(1.03)	(2.09)	(0.35)	(0.42)	(1.05)	(0.92)	(0.58)	(0.42)	(1.76)	(0.26)	(1.99)	(1.80)

The value inside the brackets () are the standard errors.
Bold printed values reflect criteria with highest scores.

Table 4
Performance comparison of the proposed methods with others state of art methods on ACCEPTOR splice sites for both balanced and unbalanced datasets.

Methods	Balanced Acceptor sites							p-value*	Unbalanced Acceptor sites							p-value*
	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	Fmeasure		S_n	S_p	Q^9	*Mcc*	AUC	auPRC	Fmeasure	
MM1-SVM	90.17	87.54	88.65	77.78	95.33	94.19	89.00	0.005062	91.18	88.28	89.54	58.30	95.76	70.71	59.14	0.005062
	(2.01)	(2.15)	(1.06)	(2.06)	(0.42)	(0.92)	(1.02)		(2.56)	(0.54)	(1.15)	(1.99)	(0.57)	(2.08)	(1.66)	
Reduced MM1-SVM	90.83	88.02	89.27	78.91	95.48	94.36	89.57	0.005062	90.72	88.23	89.35	58.01	95.66	70.00	58.90	0.005062
	(1.49)	(1.42)	(0.78)	(1.47)	(0.51)	(1.13)	(0.75)		(1.45)	(0.88)	(0.67)	(1.59)	(0.33)	(2.41)	(1.62)	
B-LSVM	90.90	88.16	89.34	79.13	95.88	94.85	89.67	0.005034	91.81	88.20	89.82	58.58	96.04	71.10	59.30	0.005062
	(1.97)	(1.68)	(0.99)	(2.09)	(0.40)	(0.86)	(1.06)		(1.40)	(0.72)	(0.82)	(1.79)	(0.40)	(1.73)	(1.69)	
MM2F-SVM	91.84	88.16	89.76	80.08	96.02	95.08	90.17	0.005062	92.22	88.81	90.33	59.93	96.23	72.33	60.66	0.005034
	(1.77)	(1.00)	(0.66)	(1.57)	(0.49)	(1.01)	(0.82)		(1.43)	(0.55)	(0.72)	(1.52)	(0.55)	(2.57)	(1.41)	
PN FDTF-SVM	92.78	88.54	90.37	81.41	96.53	95.42	90.85	0.005062	92.29	88.84	90.35	60.07	96.40	73.47	60.79	0.005062
	(1.25)	(1.16)	(0.52)	(0.92)	(0.58)	(1.24)	(0.47)		(1.69)	(0.93)	(0.77)	(1.83)	(0.42)	(2.15)	(1.85)	
LVMM2	88.92	90.11	89.41	79.20	90.34	86.98	89.70	0.005062	88.99	90.01	89.45	60.04	90.43	44.73	61.41	0.005062
	(1.70)	(1.08)	(1.24)	(1.48)	(0.24)	(1.34)	(1.29)		(1.56)	(0.62)	(0.32)	(2.04)	(0.32)	(1.96)	(1.78)	
ECS-LSVM	91.74	90.00	90.78	81.76	96.64	95.73	90.95	0.005062	91.15	89.47	90.22	60.55	96.41	73.94	61.53	0.005062
	(1.26)	(1.67)	(1.09)	(2.10)	(0.49)	(0.98)	(1.03)		(1.65)	(0.94)	(0.85)	(1.80)	(0.34)	(1.36)	(1.84)	
Sparse-SVM	92.71	89.86	91.10	82.62	96.94	96.19	91.41	0.005034	92.64	89.47	90.89	61.48	96.81	76.19	62.22	0.005062
	(1.13)	(1.63)	(0.82)	(1.53)	(0.40)	(0.74)	(0.73)		(1.01)	(0.74)	(0.5)	(1.36)	(0.36)	(1.56)	(1.44)	
DM-SVM	92.64	90.73	91.54	83.41	97.31	96.66	91.76	0.005034	92.81	90.44	91.51	63.56	97.16	78.91	64.39	0.005062
	(1.22)	(1.57)	(0.62)	(1.19)	(0.34)	(0.43)	(0.57)		(1.13)	(0.66)	(0.54)	(1.43)	(0.38)	(1.72)	(1.45)	
MCM-SVM	92.46	90.80	91.56	83.28	97.21	96.57	91.70	0.005062	92.33	90.92	91.48	64.28	97.26	80.29	65.25	0.005062
	(0.94)	(1.23)	(0.78)	(1.53)	(0.48)	(0.81)	(0.75)		(2.53)	(0.64)	(1.16)	(2.07)	(0.52)	(2.63)	(1.86)	
DMM2-SVM	94.16	91.91	92.82	86.13	97.86	97.33	93.11	0.005034	94.06	90.37	91.93	64.18	97.46	80.98	64.80	0.005062
	(2.01)	(1.07)	(0.79)	(1.78)	(0.36)	(0.64)	(0.93)		(1.74)	(0.45)	(0.63)	(1.35)	(0.39)	(1.85)	(1.21)	
SCMM1-AdaBoost	95.07	92.60	93.64	87.72	98.14	97.44	93.91	0.005062	94.86	92.27	93.39	68.95	98.20	84.68	69.88	0.005062
	(1.16)	(1.40)	(0.77)	(1.40)	(0.33)	(0.56)	(0.69)		(1.25)	(0.61)	(0.55)	(1.54)	(0.27)	(1.72)	(1.56)	
MSC+POS-SVM	94.69	92.78	93.61	87.49	98.30	97.72	93.80	0.005062	94.93	92.76	93.68	70.21	98.24	85.96	71.03	0.005062
	(1.42)	(1.09)	(1.00)	(2.11)	(0.29)	(0.47)	(1.05)		(1.63)	(0.57)	(0.69)	(1.38)	(0.32)	(1.62)	(1.38)	
Our Proposed Method	**97.19**	**95.42**	**96.09**	**92.64**	**99.28**	**98.94**	**96.33**	–	**95.07**	**93.38**	**94.12**	**71.93**	**98.50**	**88.42**	**72.82**	**-**
	(1.43)	(1.21)	(0.91)	(1.71)	(0.23)	(0.25)	(0.86)		(1.22)	(0.63)	(0.59)	(1.78)	(0.34)	(1.58)	(1.80)	

The value inside the brackets () are the standard errors.
Bold printed values reflect criteria with highest scores.
* The reported *p*-values have been computed between proposed method and other state of arts classifiers using F-measure criteria.

Table 5
Performance comparison of the proposed methods with others state of art methods on DONOR splice sites for both balanced and unbalanced datasets.

Methods	Balanced Donor sites							p-value*	Unbalanced Donor sites							p-value*
	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure		S_n	S_p	Q^9	*Mcc*	AUC	auPRC	F-measure	
MM1-SVM	93.24	90.99	92.00	84.28	97.12	96.24	92.20	0.005062	93.10	91.08	91.96	65.10	97.15	79.05	65.98	0.005062
	(1.88)	(1.77)	(1.01)	(2.07)	(0.57)	(0.92)	(1.04)		(1.89)	(0.48)	(0.90)	(1.66)	(0.35)	(2.35)	(1.50)	
Reduced MM1-SVM	93.92	91.24	92.36	85.22	97.32	96.41	92.67	0.005062	93.96	91.08	92.34	65.63	97.32	79.81	66.37	0.005062
	(1.88)	(1.05)	(0.81)	(1.70)	(0.74)	(1.19)	(0.89)		(1.22)	(0.49)	(0.52)	(1.21)	(0.36)	(2.87)	(1.19)	
B-LSVM	94.28	90.77	92.30	85.11	97.33	96.36	92.65	0.005062	93.99	90.96	92.23	65.43	97.34	80.22	66.15	0.005062
	(1.70)	(1.29)	(1.20)	(2.59)	(0.66)	(1.06)	(1.28)		(1.89)	(0.80)	(0.70)	(1.59)	(0.28)	(2.10)	(1.63)	
MM2F-SVM	93.70	91.60	92.50	85.34	97.35	96.38	92.72	0.005062	93.21	91.39	92.20	65.85	97.44	80.76	66.75	0.005062
	(1.42)	(1.32)	(0.79)	(1.51)	(0.67)	(1.21)	(0.76)		(1.49)	(0.44)	(0.70)	(1.40)	(0.31)	(2.80)	(1.30)	
ECS-LSVM	95.03	90.91	92.60	86.04	97.70	96.88	93.12	0.005034	94.92	90.86	92.56	65.78	97.69	82.12	66.33	0.005034
	(1.30)	(1.75)	(1.02)	(1.59)	(0.61)	(0.90)	(0.77)		(1.46)	(0.52)	(0.52)	(1.51)	(0.34)	(2.35)	(1.56)	
Sparse-SVM	94.64	91.95	93.07	86.64	97.73	96.77	93.38	0.005062	94.42	91.61	92.84	67.11	97.80	83.28	67.86	0.005034
	(1.65)	(1.15)	(0.80)	(1.68)	(0.55)	(0.92)	(0.86)		(1.21)	(0.46)	(0.50)	(1.27)	(0.29)	(1.86)	(1.24)	
OLVWMM2	94.42	91.60	93.20	86.11	92.72	89.93	93.50	0.005062	94.21	92.50	93.12	67.71	93.85	51.79	68.50	0.005062
	(1.17)	(1.26)	(0.80)	(1.58)	(0.79)	(1.08)	(0.78)		(1.33)	(0.69)	(0.66)	(1.05)	(0.72)	(1.95)	(1.40)	
DM-SVM	95.17	92.13	93.36	87.34	98.05	97.34	93.73	0.005062	94.60	92.34	93.30	68.71	98.00	84.78	69.51	0.005062
	(1.16)	(1.92)	(1.06)	(1.72)	(0.45)	(0.97)	(0.83)		(1.28)	(0.52)	(0.54)	(1.37)	(0.29)	(2.23)	(1.59)	
P1-RF	95.56	92.06	93.52	87.70	97.82	96.90	93.93	0.012515	95.31	92.22	93.55	69.13	97.84	82.82	69.82	0.005062
	(0.93)	(2.08)	(1.37)	(2.37)	(0.75)	(1.01)	(1.14)		(1.01)	(0.61)	(0.61)	(1.81)	(0.34)	(1.83)	(1.80)	
PN FDTF-SVM	95.39	92.74	93.83	88.18	97.87	96.89	94.14	0.00691	94.92	92.63	93.63	69.88	97.95	84.01	70.69	0.005062
	(1.33)	(1.59)	(0.98)	(1.79)	(0.66)	(1.27)	(0.88)		(1.05)	(0.53)	(0.42)	(1.24)	(0.19)	(2.83)	(1.29)	
DMM2-SVM	95.96	94.06	94.85	90.05	98.54	97.99	95.06	0.005034	95.14	93.94	94.47	73.48	98.48	87.78	74.43	0.005062
	(1.37)	(1.23)	(0.97)	(1.98)	(0.54)	(0.71)	(0.99)		(0.86)	(0.57)	(0.36)	(1.47)	(0.24)	(2.11)	(1.55)	
MCM-SVM	96.39	93.56	94.73	90.00	98.52	97.91	95.04	0.009344	95.75	94.24	94.88	74.71	98.66	89.01	75.61	0.00691
	(1.13)	(0.94)	(0.69)	(1.45)	(0.57)	(0.81)	(0.72)		(1.45)	(0.45)	(0.68)	(1.42)	(0.25)	(1.45)	(1.39)	
SCMM1-AdaBoost	97.03	94.28	95.39	91.35	98.81	98.25	95.71	0.092601	96.71	94.54	95.45	76.20	98.86	90.20	77.00	0.009344
	(1.09)	(0.99)	(0.77)	(1.52)	(0.53)	(0.83)	(0.76)		(1.08)	(0.39)	(0.47)	(1.37)	(0.19)	(2.28)	(1.34)	
MSC+POS+APR-SVM	96.92	94.67	95.62	91.63	98.84	98.34	95.85	0.202622	96.78	94.89	95.69	77.28	98.93	91.04	78.11	0.878482
	(0.81)	(1.42)	(1.04)	(1.93)	(0.63)	(0.85)	(0.94)		(0.96)	(0.37)	(0.38)	(1.19)	(0.19)	(2.27)	(1.19)	
Our Proposed Method	**97.09**	**95.76**	**96.28**	**92.80**	**99.30**	**98.88**	**96.45**	–	**96.60**	**94.94**	**95.70**	**77.38**	**98.99**	**91.57**	**78.21**	–
	(1.34)	(1.19)	(1.03)	(2.09)	(0.35)	(0.42)	(1.05)		(0.92)	(0.58)	(0.42)	(1.76)	(0.26)	(1.99)	(1.80)	

The value inside the brackets () are the standard errors.
Bold printed values reflect criteria with highest scores.

4.2. Comparison analysis with existing best state-of-the-arts methods using HS3D dataset

The performance comparison of the proposed approach with thirteen current state-of-the-arts methods is represented for acceptor HS3D dataset (both balanced and imbalanced) over 10-fold cross validation in Table 4. It can be seen that the proposed MM3-SVM approach significantly outperforms all the other methods in all the terms.

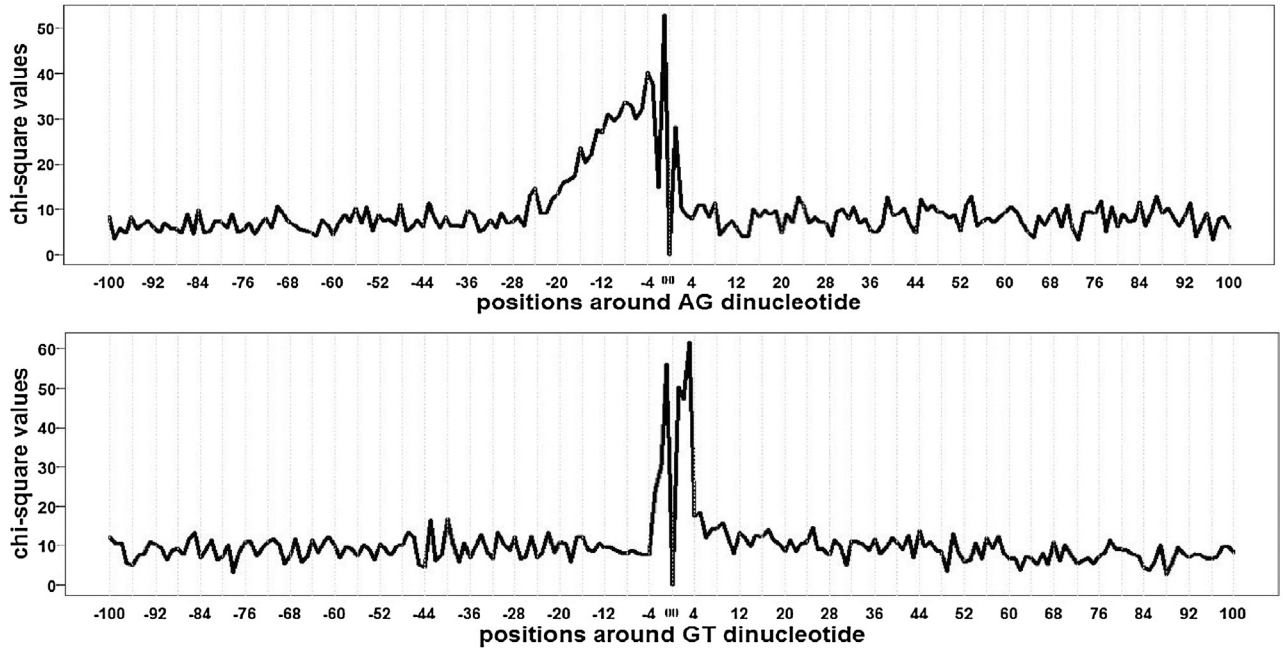


Fig. 3. The chi-square values for flanking regions around splice sites. The values were computed for 202 positions from −100 to 100 nt around the splice sites and given for acceptor (top panel) and the donor sites (down panel).

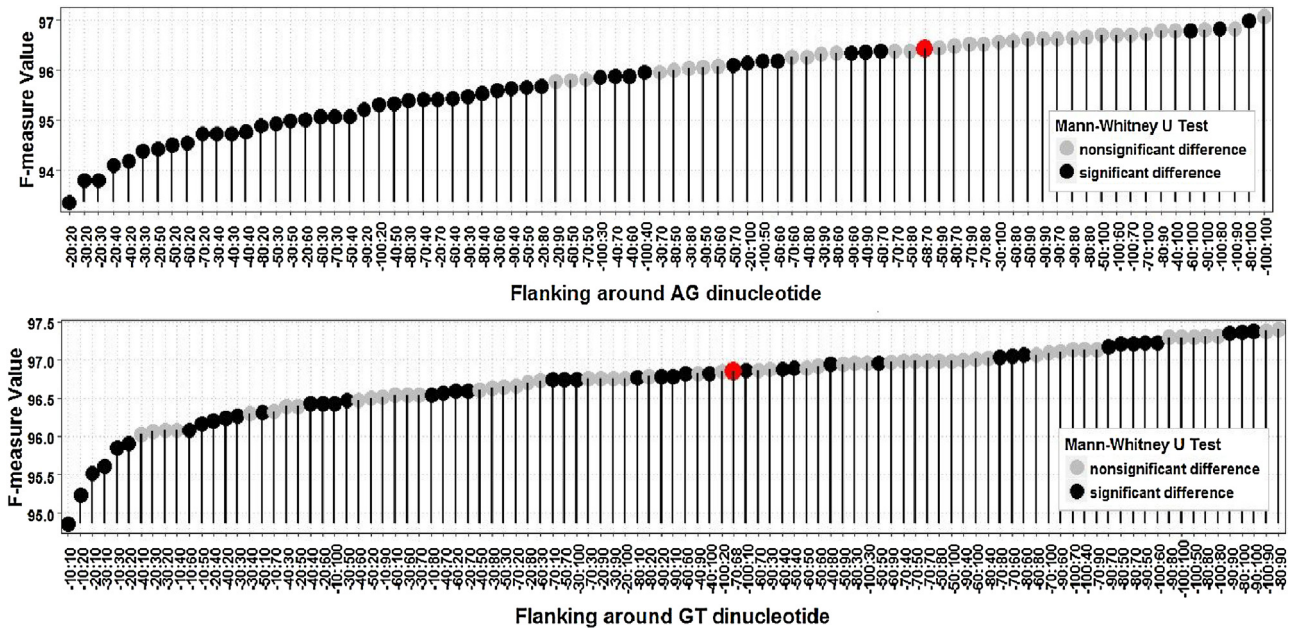


Fig. 4. Classifiers performance for different lengths of flanking region on the balanced HS3D dataset.

Table 5 compares the performance of the proposed method with existing best approaches on donor HS3D dataset (both balanced and imbalanced). It is seen that the proposed method achieved higher accuracies than the other methods. However, the performance of MM3-SVM is almost as high as that of the MSC+Pos +APR-SVM method.

4.3. Impact of sequence length on the model

Length of the flanking sequences around splice sites is important for accurate splice site prediction (Meher et al., 2016b), and chi-square test is usually employed to determine them. The chi-square test calculates the frequency of nucleotides between positive and negative splice-site sequences at each position (Li et al., 2012; Meher et al., 2016b). Considering the balanced HS3D dataset with the length of 200 bps around the splice sites, the chi-square value for each position was calculated. Fig. 3 shows the chi-square value against the position of the flanking regions around the splice sites, where 5′ sites GT and 3′ sites AG have been unified as position 00. We can determine the length of the upstream and downstream sequences by judging whether the chi-square value achieves the significance level of 0.05 at base positions. From Fig. 3, it is observed that the chi-square value is significant for almost all positions at the individual significance level of 0.05. Moreover, it can be seen that the 8 positions (−3 to +5) around GT dinucleotide and 21 positions (−20 to +1) around AG dinucleotide have extremely higher value than other positions. It means that they can play a crucial role in distinguishing the real sites from the decoy sites.

We varied the length of upstream and downstream sequences around the splice sites to find the optimal length of sequences which provide the best classification accuracy. For acceptor splice site, we first extracted 20 bps from both upstream and downstream sequences and increased their length up to 100 bps with the interval of 10, while the initial length of the sequences around donor site was set to 10 bps. Fig 4 shows the performance of the classifiers, in terms of average F-measure on 10-fold-CV, obtained from the variation in the sequence lengths for both balanced donor and acceptor sites.

The red points in the Fig. 4 represent the original length of flanking sequences around splice sites on benchmark HS3D dataset. All results were assessed according to the accuracy of those points. Results on the left side imply a reduction in predictive performance relative to the original length of flanking sequences, while the results on the right side indicate better performance. In order to determine if there is a statistically significant difference between the F-measures of the original length of flanking sequences and others, we used Mann-Whitney *U* test with 95% of confidence level and marked them in a darker tone.

Table 6
Performance comparison of the proposed method with others state of art methods on independent test datasets for prediction canonical splice sites.

Methods	UnBalanced Acceptor sites							Unbalanced Donor sites						
	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	Fmeasure	S_n	S_p	Q^9	*Mcc*	AUC	auPRC	Fmeasure
GeneID	55.69	72.29	63.04	21.75	66.36	30.03	36.43	76.44	89.60	81.79	59.72	82.59	76.23	66.51
WMM	93.67	85.06	88.53	65.31	94.00	81.70	69.81	94.38	84.38	88.26	64.84	94.05	87.67	69.27
MM1	**93.67**	85.82	89.02	66.40	94.11	81.50	70.81	**95.00**	85.38	89.07	66.70	94.78	88.66	70.86
MEM	93.04	85.82	88.83	65.95	94.04	80.55	70.50	93.12	91.12	92.06	74.70	94.65	88.81	78.42
NNSplice	87.97	93.04	90.17	74.84	92.81	83.70	78.98	90.00	94.12	91.80	78.53	93.80	87.00	82.05
SplicePort	88.61	**98.73**	91.89	**89.19**	97.26	**94.90**	**90.91**	89.38	98.12	**92.37**	**87.94**	98.49	**95.62**	**89.94**
MMSVM (HS3D)	90.51	94.05	**92.08**	78.72	95.22	84.49	82.18	88.88	96.46	90.39	82.34	97.21	91.67	85.54
MMSVM (GCHSS)	86.71	97.34	90.42	84.05	**98.25**	90.30	86.71	85.62	**98.12**	89.75	85.50	**98.50**	92.25	87.82

Bold printed values reflect criteria with highest scores.

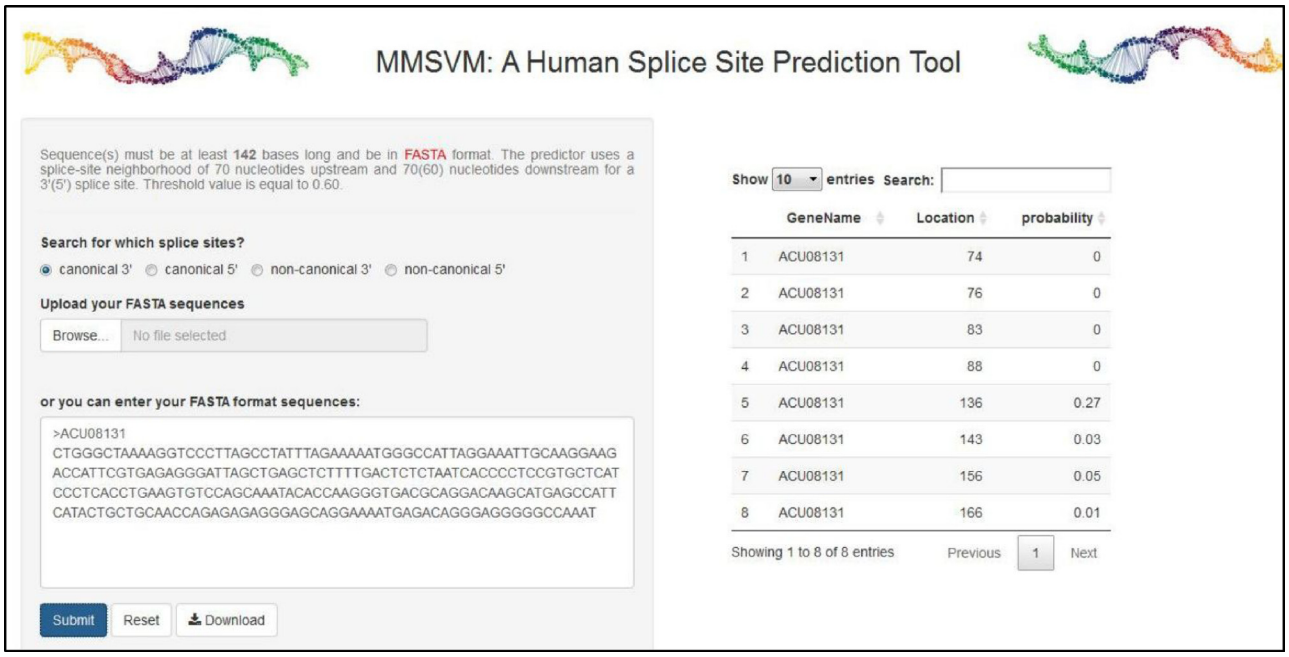


Fig. 5. Image of the prediction server.

Fig. 4 shows that increase in the length of sequences flanking acceptor splice sites can provide a statistically significant difference in classification performance using (−80 to +100), (−100 to +80), and (−60 to +100) flank, when compared with (−68 to +70) flanking. It is worth noting that (−30 to 70) flanking composed of 102 nt is the shortest length that achieved no worse classification performance (with statistical significance) than those obtained by original (−68 to +70) flanking and other longer flanking.

Considering donor splice site, sequences represented by fewer nucleotides allow the construction of classification models that are as accurate as those obtained from original (−70 to 68) flanking or even more accurate. On the other hand, similar to what happens to the acceptor, classification accuracy improves as the length of their sequences increase. It can be seen that 16 flanking have (significantly) better predictive performance than original (black points in right side). The results show that (−50 to 40) flanking composed of 92 nt is the shortest sequence length which provides higher prediction accuracy that original flanking, whereas (−40 to 10), (−20 to 30), (−30 to 20) and (−10 to 40) flanking regions with 50 nt are the shortest length flanking regions which provided not significantly different prediction accuracy than original flanking.

The importance of the above experiment is due to the fact that the greater amount of attributes is produced when we use longer sequences to train our model. It makes the classification process more complex and leads to consuming unfeasible processing time and computational resources (Carvalho et al., 2015). Therefore, to provide a good balance between complexity and high accuracy models, we chose 142 nt (70 + 2 + 70) for acceptor and 132 nt (70 + 2 + 60) for donor sites, respectively.

4.4. Comparison analysis with existing tools using an independent test dataset

The performance metrics of the proposed MM3-SVM method and the considered existing tools which computed by using an independent test dataset for both canonical sites is presented in Table 6. We trained our model using both the benchmark HS3D and the provided GCHSS datasets. It can be seen that utilizing GCHSS dataset enhances the learning process since it contains more real sites than the HS3D dataset. From Table 6, it is observed that performance of the provided model is higher than all the other approaches in terms of *Mcc*, AUC, *auPRC*, and F-measure in both donor and acceptor sites except SplicePort. On the other hand, the S_p criterion of the MM3-SVM is the highest as compared to others (except acceptor SplicePort), while the S_n metric of the proposed method is the lowest among all (except GeneID) for both donor and acceptor sites. Looking at imbalanced metrics and F-measure, it can be said that proposed MM3-SVM performs better than the existing tools although can't show better detection performance than SplicePort.

4.5. Online prediction server

In order to help the biological community for recognizing the human splice sites, an online prediction server based on the proposed approach, named as MMSVM, was developed. Here, the option for predicting acceptor sites and donor sites is given. The server was developed using "shiny" R package in Rstudio, while the developed R code was executed in the background. The web page was designed to facilitate the user for submitting the sequence(s) by both loading FASTA file and pasting the sequences in a text area. The provided tool automatically searches the sequences to find candidate splice sites and reports their locations and their probabilities of being true sites. The model has been trained with the GCHSS dataset and the user has to provide only test sequence (s) of her/his interest to utilize the MMSVM tool. It should be noted that the proposed MMSVM is the only splice site prediction tool that can detect the non-canonical sites. The image of the web server after execution of an example dataset is shown in Fig. 5. The prediction server is freely available at https://pashaei.shinyapps.io/mmsvm.

5. Discussion

Detecting genes accurately is one of the most significant challenges in bioinformatics (Baten et al., 2008). Accuracy in gene detection ultimately depends on addressing various problems, one

of which is splice site recognition (Kamath et al., 2014). The splice site recognition problem is considered as the most important subtask in gene detection (Li and Dewe, 2012). Meanwhile, accurate detection of splicing signals plays a substantial role in new drugs development and investigating DNA functions (Xu et al., 2017). In order to accurately predict splice sites, several methods have been developed and out of them probabilistic approaches (Markov chain) and SVM-based methods are the more popular method (Wei et al., 2013). On the other hand, Markov chain models and SVM classifier do not give satisfactory results when applied individually in the field of splice site prediction (Goel et al., 2015). However, the results can be improved by combining these methods such as MM1-SVM (Baten et al., 2006), Reduced MM1-SVM (Baten et al., 2008), MM2F-SVM (Maji and Garg, 2014) and MCM-SVM (Goel et al., 2015) methods. Moreover, it is desirable to use a higher order Markov model to capture higher and extended sequential relationships in DNA sequences (Baten et al., 2006), to extract more appropriate features to serve as the input of SVM classifier. Also, the earlier splice site prediction approaches have focused only on true splice sites and ignored the false splice sites to train the prediction model, but false splice sites are also essential for the splice sites recognition (Huang et al., 2006; Meher et al., 2016b). Thus, we considered both the true and false sites for splice site recognition.

This study makes a precise performance analysis on all the existing Markovian encoding methods in the literature. For this purpose, we define different models namely TM, DM, RM, and TMF, by considering only true sequences (positive class labels) and both true and false sequences (negative class labels) from the training set. Moreover, this study proposes a novel method for human splice site recognition utilizing hybrid of MM3 and SVM classifier. The proposed method shows superior performance as compared with lower order Markovian methods on both donor and acceptor splice sites (for both balanced and imbalanced HS3D dataset) (Tables 2 and 3). From Tables 2 and 3 it is clear that considering both true and false sites to predict splice sites produces higher accuracy and especially TMF approach is the most efficient one among the others. We compared our proposed method with thirteen best state-of-the-arts methods to judge the efficiency of the proposed method (Tables 4 and 5). The comparison depicts that MM3-SVM outperforms all the methods significantly in terms of several popular criteria for both acceptor and donor splice sites. The performance of the proposed method was further compared with several tools using an independent dataset (Table 6). It can be seen that the proposed methods perform better that the other tools in most of the terms. From above discussion, it is found that the MM3-SVM outperforms most of the leading splice site predictors. We developed a web server, namely MMSVM (Fig. 5), in order to help the researchers to identify splice sites in real-time and easily. Also, the MMSVM was extended to identify non-canonical splice sites. The accuracy of our proposed method suggests that it can be utilized as an analytic solution to more genomic problems such as promoter prediction.

It should be noted that the computational complexity grows exponentially with the increase of the order of Markov chain, due to the fact that a large number of training samples is needed to estimate the Markovian parameters (Baten et al., 2008; Wei et al., 2013).

6. Conclusion

In this paper, we presented a novel approach for human splice site detection, which is the main bottleneck in gene identification. Our simple and easy approach, in which third order Markov model is combined with SVM classifier, is superior to the existing best state-of-the-arts methods and tools. Further, this paper puts forward a new evaluation on Markovian-based encoding methods by considering true splice sites and both true and false sites in different manners. The MMSVM web tool based on the proposed approach is believed to help the biological community to predict human splice sites easily and more accurately. Also, the accuracy of our proposed method suggests that this approach can be successfully included in gene detection systems of large DNA sequences.

References

Bari, A.T.M.G., Reaz, M.R., Choi, H.J., Jeong, B.S., 2012. Survey on nucleotide encoding techniques and SVM Kernel design for human splice site prediction. Interdiscip. Biol. Central 4, 1–6. doi:http://dx.doi.org/10.4051/ibc.2012.4.4.0014.

Bari, A.T.M.G., Reaz, M.R., Jeong, B.-S., 2014. Effective DNA encoding for splice site prediction using SVM. Commun. Math. Comuput. Chem. (MATCH) 71, 241–258.

Baten, A., Chang, B., Halgamuge, S., Li, J., 2006. Splice site identification using probabilistic parameters and SVM classification. BMC Bioinf. 7 (Suppl 5) .

Baten, A., Halgamuge, S., Chang, B., 2008. Fast splice site detection using information content and feature reduction. BMC Bioinf. 9 (Suppl 12) .

Bin, W., Jing, Z., 2014. A novel artificial neural network and an improved particle swarm optimization used in splice site prediction. J. Appl. Comput. Math. 3. doi: http://dx.doi.org/10.4172/2168-9679.1000166.

Blagus, R., Lusa, L., 2015. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinf. 16. doi:http://dx.doi.org/10.1186/s12859-015-0784-9.

Blanco, E., Parra, G., Guigo, R., 2007. Using geneid to identify genes. Curr. Protocol. Bioinf. doi:http://dx.doi.org/10.1002/0471250953.bi0403s18 (Chapter 4: Unit 4.3).

Burge, C., Karlin, S., 1997. Predictions of complete gene structures in human genomic DNA. J. Mol. Biol. 9, 499–509.

Burset, M., Guigó, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–357.

Carvalho, S.G., GuerraSá, R., Merschmann, L.H., 2015. The impact of sequence length and number of sequences on promoter prediction performance. BMC Bioinf. 16.

Chen, T.M., Lu, C.C., Li, W.H., 2005. Prediction of splice sites with dependency graphs and their expanded Bayesian networks. Bioinformatics 21, 471–482.

Costa, C., Giménez-Capitán, A., Karachaliou, N., Rosell, R., 2013. Comprehensive molecular screening: from the RT-PCR to the RNA-seq. Transl. Lung Cancer Res. 2, 87–91. doi:http://dx.doi.org/10.3978/j.issn.2218-6751.2013.02.05.

Desmet, F., Hamroun, D., Beoud, G.C., Claustres, M., Beoud, C., 2010. Bioinformatics identification of splice site signals and prediction of mutationeffects. In: Mohan, R.M. (Ed.), Research Advances in Nucleic Acids Research. Global Research Network, Kerala, India.

Desmet, F., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., Beroud, C., 2009. Human Splice Finder: an online bioinformatics tool to predictsplicing signals. Nucleic Acids Res. 37.

Dogan, R.I., Getoor, L., Wilbur, W.J., Mount, S.M., 2007. SplicePort—an interactive splice-site analysis tool. Nucleic Acids Res. 35, 285–291. doi:http://dx.doi.org/10.1093/nar/gkm407.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. cambridge university press.

Fink, G.A., 2014. Markov Models for Pattern Recognition. Springer-Verlag, London doi:http://dx.doi.org/10.1007/978-1-4471-6308-4.

Flanagan, S.E., Xie, W., Caswell, R., Damhuis, A., Vianey-Saban, C., Akcay, T., Darendeliler, F., Bas, F., 2013. Next-generation sequencing reveals deep intronic cryptic ABCC8 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation. Am. J. Hum. Genet. 92, 131–136. doi:http://dx.doi.org/10.1016/j.ajhg.2012.11.017.

Goel, N., Singh, S., Aseri, T.C., 2015. An improved method for splice site prediction in DNA sequences using support vector machines. Procedia Comput. Sci. 57, 358–367. doi:http://dx.doi.org/10.1016/j.procs.2015.07.350.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774.

Huang, J., Li, T., Chen, K., Wu, J., 2006. An approach of encoding for prediction of splice sites using SVM. Biochimie 88, 923–929.

Jian, X., Boerwinkle, E., Liu, X., 2014a. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 42, 13534–13544. doi:http://dx.doi.org/10.1093/nar/gku1206.

Jian, X., Boerwinkle, E., Liu, X., 2014b. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet. Med. 16, 497–503. doi:http://dx.doi.org/10.1038/gim.2013.176.

Kamath, U., Compton, J., Dogan, R.I., Jong, D.K.A., Shehu, A., 2012. An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. 9, 1387–1398.

Kamath, U., Jong, K.D., Shehu, A., 2014. Effective automated feature construction and selection for classification of biological sequences. PLoS One 9, e99982. doi: http://dx.doi.org/10.1371/journal.pone.0099982.

Leo, C.P., Hsu, S.Y., Hsueh, A.J.W., 2002. Hormonal genomics. Endocr. Rev. 23, 369–381.

Li, B., Dewe, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 12. doi:http://dx.doi.org/10.1186/1471-2105-12-323.
Li, J., Wang, L., Wang, H., Bai, L., Yuan, Z., 2012. High-accuracy splice sites prediction based on sequence component and position features. Genet. Mol. Res. 11, 3432–3451.
Lin, W.J., Che, J.J., 2012. Class-imbalanced classifiers for high-dimensional data. Brief. Bioinf. 14, 13–26. doi:http://dx.doi.org/10.1093/bib/bbs006.
Loi, H.S., Rajapakse, J.C., 2002. Splice site detection with neural networks/Markov models hybrids. 9th International Conference on Neural Information Processing (ICONIP'OZ) 5.
Loi, H.S., Rajapakse, J.C., 2003. Splice site detection with a higher-order markov model implemented on a neural network. Genome Inf. 14, 64–72.
Longadge, R., Dongre, S.S., Malik, L., 2013. Class imbalance problem in data mining: review. Int. J. Comput. Sci. Netw. (IJCSN) 2, 83–87.
Lopes, H.S., Lima, C.R.E., Murata, N.J., 2007. A configware approach for high-speed parallel analysis of genomic data. J. Circuits Syst. Comput. 16, 527–540.
Maji, S., Garg, D., 2013. Hidden markov model for splicing junction sites identification in DNA sequences. Curr. Bioinf. 8, 369–379.
Maji, S., Garg, D., 2014. Hybrid approach using SVM and MM2 in splice site junction identification. Curr. Bioinf. 9, 76–85.
Meher, P.K., Sahu, T.K., Rao, A.R., 2016a. Prediction of donor splice sites using random forest with a new sequence encoding approach. BioData Mining 9. doi:http://dx.doi.org/10.1186/s13040-016-0086-4.
Meher, P.K., Sahu, T.K., Rao, A.R., Wahi, S.D., 2016b. A computational approach for prediction of donor splice sites with improved accuracy. J. Theor. Biol. 404, 285–294.
Nassa, T., Singh, S., Goel, N., 2013. Splice site detection in DNA sequences using probabilistic neural network. Int. J. Comput. Appl. (IJCA) 76, 1–4.
Parada, G.E., Munita, R., Cerda, C.A., Gysling, K., 2014. A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Res. 42, 10564–10578. doi:http://dx.doi.org/10.1093/nar/gku744.
Parra, G., Blanco, E., Guigó, R., 2000. GeneID in Drosophila. Genome Res. 10, 511–515.
Pashaei, E., Ayilmaz, A., Aydin, N., 2016a. A combined SVM and markov model approach for splice site identification. 6th IEEE International Conference in Computer and Knowledge Enginerring (ICCKE2016) 200–2004, Mashhad, Iran: IEEE doi:http://dx.doi.org/10.1109/ICCKE.2016.7802140.
Pashaei, E., Yilmaz, A., Ozen, M., Aydin, N., 2016b. A novel method for splice sites prediction using sequence component and hidden markov model. In 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC), Florida, USA: IEEE, pp. 3076–3079. doi:http://dx.doi.org/10.1109/EMBC.2016.7591379.
Pashaei, E., Ozen, M., Aydin, N., 2017. Splice site identification in human genome using random forest. Health Technol. 7, 141–152. doi:http://dx.doi.org/10.1007/s12553-016-0157-z.
Pertea, M., Lin, X., Salzberg, S., 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. 29, 1185–1190.
Pollastro, P., Rampone, S., 2002. HS3D, a dataset of Homo sapies splice site regions, and its extraction procedure from a major public database. Int. J. Modern Phys. C13, 1105–1117.
Reese, M., Eeckman, F., Kupl, D., Haussler, D., 1997a. Improved splice site detection in Genie. J. Comput. Biol. 4, 311–324.
Reese, M.G., Eeckman, F.H., Kulp, D., Haussler, D., 1997b. Improved splice site detection in Genie. In First Annual International Conference on Computational Molecular Biology (RECOMB). ACM Press, New York, pp. 232–240.
Roca, X., Olson, A.J., Rao, A.R., Enerly, E., Kristensen, V.N., Børresen-Dale, A.-L., Andresen, B.S., Krainer, A.R., Sachidanandam, R., 2008. Features of 5-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genom. Res. 18, 77–87.
Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10, e0118432. doi:http://dx.doi.org/10.1371/journal.pone.0118432.
Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J., Tettelin, H., 1999. Interpolated Markov models for eukaryotic gene finding. Genomics 59, 24–31.
Salzberg, S., 1997. A method for identifying splice sites and translational start sites in eukaryotic mRNA. Comput. Appl. Biosci. 13, 365–376.
Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., Ratsch, G., 2007. Accurate splice site prediction using support vector machines. BMC Bioinf. 88 (Suppl 10) .
Staden, R., 1984. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12, 505–519.
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinf. 9.
Statnikov, A., Aliferis, C.F., Hardin, P.D., Guyon, I., 2011. A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and Methods, vol 1. World Scientific publishing, USA.
Sun, Y., Fan, X., Li, Y., 2003. Identifying splicing sites in Eukaryotic RNA: support vector machine approach. Comput. Biol. Chem. 33, 17–29.
Tang, R., Prosser, D.O., Love, D.R., 2016. Evaluation of bioinformatic programmes for the analysis of variants within splice site consensus regions. Adv. Bioinf. 2016.
Tavares, L.G., Lopes, H.S., Lima, C.R.E., 2009. Evaluation of weight matrix models in the splice site junction recognition problem. Bioinf. Biomed. Workshop 1, 14–19.
Wei, D., Zhuang, W., Jiang, Q., Wei, Y., 2012. A new classification method for human gene splice site prediction. In: He, J., Liu, X., Krupinski, E., Xu, G. (Eds.), Health Information Science, vol 7231. Springer Berlin Heidelberg, pp. 121–130.
Wei, D., Zhang, H., Wei, Y., Jiang, Q., 2013. A novel splice site prediction method using support vector machine. J. Comput. Inf. Syst. 9, 8053–8060.
Woolfe, A., Mullikin, J., Elnitski, L., 2010. Genomic features defining exonic variants that modulate splicing. Genome Biol. 11. doi:http://dx.doi.org/10.1186/gb-2010-11-2-r20.
Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., Morris, Q., Barash, Y., Krainer, A.R., Jojic, N., Scherer, S.W., Blencowe, B.J., Frey, B.J., 2015. The human splicing code reveals new insights into the genetic determinants of disease. Science 347. doi:http://dx.doi.org/10.1126/science.1254806.
Xu, Z.-C., Wang, P., Qiu, W.-R., Xiao, X., 2017. ISS-PC: identifying splicing sites via physical-Chemical properties using deep sparse auto-encoder. Sci. Rep. 7 doi:http://dx.doi.org/10.1038/s41598-017-08523-8.
Yeo, G., Burge, C., 2004. Maximum entropy modeling of short sequence motifs with application to RNA splicing signals. J. Comput. Biol. 11, 377–394.
Zhang, M., Marr, T., 1993. A weight array method for splicing signal analysis. Comput. Biosci. 9, 499–509.
Zhang, Y., Chu, C.-H., Chen, Y., Zha, H., Ji, X., 2006. Splice site prediction using support vector machines with a Bayes kernel. Expert Syst. Appl. 30, 73–81.
Zhang, Q., Peng, Q., Zhang, Q., Yan, Y., Li, K., Li, J., 2010. Splice site prediction of human genome using Length-variable Markov model and feature selection. Expert Syst. Appl. 37, 2771–2782.