Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 71(2) 2017 159–166

A Non-parametric Regression based Computational Approach for Prediction of Donor Splice Sites*

Prabina Kumar Meher and A.R. Rao
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 31 August 2016; Accepted 20 June 2017

SUMMARY

Identification of splice sites with higher accuracy is vital for systematic study of gene structures in eukaryotes. In this paper, an attempt has been made to develop a kernel regression based probabilistic approach for prediction of donor splice sites. The proposed method achieved an estimate of area under receiving operating characteristics curve of 93.75±0.56 and 93.50±0.56 and area under precision-recall curve of 96.13±0.43 and 96.13±0.43 on cattle (*Bos Taurus*) and fish (*Danio rario*) datasets respectively. The prediction accuracy of the developed approach was also found comparable with the existing probabilistic approaches viz., SAE, MEM, MDD, MM1 and WMM, while tested by using an independent splice site dataset. Thus, we believe that the proposed approach will supplement the existing approaches for prediction of donor splice sites.

Keywords: Kernel regression, Indicator variable, Conditional expectation, Nucleotide dependencies.

1. INTRODUCTION

Identification of genes in genomic DNA is one of the important steps to understand the genome of any species, once sequencing of the species is completed. *Ab initio* gene finding programs are mainly based on signal detection. In particular, determining splicing signal (splice site) is an important component of computational gene finding methods (Golam Bari *et al*. 2014). Genes present in the genomic DNA are not continuous but consist of several exons separated by non-coding introns (Degroeve *et al.* 2005). The exon-intron boundaries are called donor (5′) splice sites, whereas intron-exon boundaries are called acceptor (3′) splice sites (Sonnenburg *et al.* 2007). Majority of donor and acceptor splice sites are characterized by the presence of dimers GT and AG respectively. However, the presence of such dimmers is only a necessary but not sufficient to declare a candidate sequence as a true donor or acceptor splice sites (Golam Bari *et al.* 2014). Since splice sites play a key role for predicting the gene structure, development of efficient analytical methods for splice site prediction is vital (Meher *et al.* 2016a).

Several computational approaches have been proposed by the researchers for the identification of splice sites. These approaches can be broadly categorized into two classes i.e. (i) probabilistic approaches and (ii) machine learning based approaches (Wei *et al.* 2013). In probabilistic approaches, the likelihood of the candidate sequences are computed by using position specific probabilities of nucleotides i.e., weighted matrix model (WMM; Staden 1984), or 1st order positional dependencies i.e. first order Markov model (MM1; Zhang and Marr 1993), maximal dependency decomposition (MDD; Burge and Karlin 1997), maximum entropy model (MEM; Yeo and Burge 2004), sum of absolute error (SAE; Meher *et al.* 2014) or second order positional dependencies i.e. length variable Markov model (LVMM; Zhang *et al.* 2010). Among these probabilistic approaches, LVMM is computationally expensive because of higher ordered Markov model. Furthermore, in the

Corresponding author: Prabina Kumar Meher
E-mail address: meherprabin@yahoo.com

* Paper presented in Dr. G.R. Seth Memorial Young Scientist Award Session on November 21, 2016 during the 70th Annual (International) Conference of the Indian Society of Agricultural Statistics held at ICRISAT, Patancheru, Hyderabad, Telangana during November 21-23, 2016.

probabilistic approach like MEM determining a threshold value required to distinguish true splice sites from false ones is difficult (Wei *et al.* 2013). In case of machine learning based approaches, the true and false splice site sequences are mapped onto numeric feature vectors, which are then used as input in binary machine learning classifiers (Golam Bari *et al.* 2014; Meher *et al.* 2016b; Meher *et al.* 2016c). Moreover, the process of encoding the sequence data requires one step more towards the prediction of splice sites, which may need additional memory allocation. Besides, the prediction accuracies also vary with different sequence encoding approaches (Meher *et al*. 2014).

In this study, we made an attempt to develop a kernel regression based probabilistic approach for the prediction of splice sites in eukaryotes. The developed approach was employed for the prediction of donor splice sites in *Bos taurus* and *Danio rario* and achieved high prediction accuracy. The approach was also found comparable with other probabilistic approaches, while compared using the bench mark NN269 (Reese *et al.* 1997) dataset.

2. MATERIALS AND METHODS

2.1 Collection and Processing of Splice Sites

The splice site datasets of cattle and fish were used to evaluate the performance of the proposed prediction approach. Besides, the bench mark NN269 splice site dataset was used to compare the performance of the proposed approach with that of other probabilistic approaches.

Initially, the exon and intron sequences of *Bos taurus* and *Danio rario* were collected from UCSC genome browser (https://genome.ucsc.edu/). The exon and intron sequences were then processed and the true splice site sequences of 20 base pairs (bp) long (10nt bp at exon-end and 10bp at intron-start) were extracted keeping GT at 11th and 12th positions respectively. Further, the false splice site sequences of same length were randomly extracted from exonic and intronic regions keeping G and T at 11th and 12th positions respectively. In both the species, 5,000 sequences of true and 5,000 sequences of false splice sites were collected.

The NN269 dataset is available at http://cs.gmu.edu/~ashehu/sites/default/files/ tools/ EFFECT_2013/ data.html, which has been extracted from 269 human genes. It consists of 1,324 true donor and 4,922 false donor splice site sequences, where each sequence is of 15bp long with GT at 9th and 10th positions respectively. This dataset has been partitioned into training and test sets consisting of 5256 (1116 true + 4140 false) and 990 (208 true +782 false) sequences respectively. To maintain uniformity, the length of splice sites for *Bos taurus* and *Danio rario* were also restricted to 15bp long, which is same as the length of the sequences in NN269 dataset.

2.2 Proposed Prediction Approach

For the position-wise aligned true and false splice sites sequences of training datasets, the occurrence of the bases at different positions can be described by an indicator variable. Let $I^{Tss}(x_{ij}=s)$ be the indicator variable that represents the occurrence of nucleotide base s at j^{th} position of i^{th} sequence corresponding to the training dataset of true splice sites (*Tss*), where, $i=1,2,...,N$, $j=1,2,...,P$. Similarly, let $I^{Fss}(x_{ij}=s)$ be the indicator variable for false splice sites (*Fss*) training dataset, where $i=1,2,...,M$. Further, let $I^{Ts}(z_j=s)$ be the indicator variable that represents the occurrence of base s at j^{th} position in any test sequence (*Ts*), where $j=1,2,...,P$. Then, without loss of generality we can write

$$I^{Tss}(x_{ij}=s)=\begin{cases}1, \textit{ if s occurs}\\ 0, \textit{ otherwise}\end{cases}, s\in\{A,T,G,C\}; i=1,2,...,N; j=1,2,...,P,$$

$$I^{Fss}(x_{ij}=s)=\begin{cases}1, \textit{ if s occurs}\\ 0, \textit{ otherwise}\end{cases}, s\in\{A,T,G,C\}; i=1,2,...,M; j=1,2,...,P$$

and

$$I^{Ts}(z_j=s)=\begin{cases}1, \textit{ if s occurs}\\ 0, \textit{ otherwise}\end{cases}, s\in\{A,T,G,C\}; j=1,2,...,P.$$

Now, the prediction error for any j^{th} position of test sequence by assuming it as a *Tss* ($e_j^{Ts|Tss}$) can be computed as

$$e_j^{Ts|Tss}=\frac{\sum_{i=1}^{N}\sum_{s}\left\{I^{Ts}(z_j=s)-E\left[I^{Ts}(z_j=s)\,|\,I^{Tss}(x_{ij'(\neq j)}=s)\right]\right\}}{N},$$

where $E\left[I^{Ts}(z_j=s)\,|\,I^{Tss}(x_{ij'(\neq j)}=s)\right]$ is the expectation of base s occurring at j^{th} position given the bases at other positions by assuming *Ts* as a *Tss*. Then, the average prediction error of the test sequence ($\overline{e}^{Ts|Tss}$) over all the P positions can be computed as

$$\overline{e}^{Ts|Tss}=\frac{\sum_{j=1}^{P}e_j^{Ts|Tss}}{P}.$$

Similarly, the average prediction error of the test sequence by assuming it as *Fss* ($\bar{e}^{Ts|Fss}$) can be computed as

$$\bar{e}^{Ts|Fss} = \frac{\sum_{j=1}^{P} e_j^{Ts|Fss}}{P},$$

where

$$e_j^{Ts|Fss} = \frac{\sum_{i=1}^{M}\sum_{s}\left\{I^{Ts}(z_j = s) - E\left[I^{Ts}(z_j = s) \mid I^{Fss}(x_{ij'(\neq j)} = s)\right]\right\}}{M}$$

and $E\left[I^{Ts}(z_j = s) \mid I^{Fss}(x_{ij'(\neq j)} = s)\right]$ is the expectation of base *s* occurring at j^{th} position given the bases at other positions by assuming *Ts* as a *Fss*. The value of the conditional expectation can be estimated by using non-parametric kernel regression approach explained as follows:

As per Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964), the estimate of the conditional expectation of any random variable *Y* relative to a random variable *X* can be computed as

$$E(Y|X) = \sum_{i=1}^{n} K_h(x - x_i) y_i \Big/ \sum_{i=1}^{n} K_h(x - x_i),$$

where *K* is a kernel with band width *h* and *n* is the number of observations. However, for multivariate *X*, the conditional expectation can be obtained as

$$E(Y|\mathbf{X}) = \sum_{i=1}^{n} y_i \prod_{j=1}^{P} K_{h_j}(x_j - x_{ij}) \Big/ \sum_{i=1}^{n}\prod_{j=1}^{P} K_{h_j}(x_j - x_{ij}).$$

In the present context, the estimates of conditional expectation can be computed as follows:

$$E\left[I^{Ts}(z_j = s) \mid I^{Tss}(x_{ij'(\neq j)} = s)\right] = \frac{\sum_{i=1}^{N} I^{Tss}(x_{ij}) \prod_{j'\neq j} K(z_j, x_{ij'}, \lambda_j)}{\sum_{i=1}^{N}\prod_{j'\neq j} K(z_j, x_{ij'}, \lambda_j)}$$

$$E\left[I^{Ts}(z_j = s) \mid I^{Fss}(x_{ij'(\neq j)} = s)\right] = \frac{\sum_{i=1}^{M} I^{Fss}(x_{ij}) \prod_{j'\neq j} K(z_j, x_{ij'}, \lambda_j)}{\sum_{i=1}^{M}\prod_{j'\neq j} K(z_j, x_{ij'}, \lambda_j)},$$

where the kernel $K(z_j, x_{ij'}, \lambda_j)$ is the bi-weight function represented as

$$K(z_j, x_{ij'}, \lambda_j) = \begin{cases} \lambda_j, & if\ z_j = x_{ij'} \\ (1-\lambda_j)/3, & if\ z_j \neq x_{ij'};\ 0 < \lambda_j < 1 \end{cases}$$

In other words, the weight λ_j was assigned when the j^{th} base of the test sequence matched with the j'^{th} base of i^{th} training sequence and the weight *(1- λ_j)/3* was assigned for the mismatch. In this study $\lambda_j = \lambda$ for all *j=1,2,...,P* was considered by giving equal importance to all positions.

2.3 Prediction for Test Sequence

If the test sequence (*Ts*) actually belongs to the *Tss* or *Fss* category, the average prediction error will be less by assuming it as *Tss* or *Fss* respectively. Keeping this assumption in mind the following criterion is developed for discrimination of true and false splice sites i.e.,

$$Ts \in \begin{cases} Tss, & If\ (\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}) < \varepsilon \\ Fss, & If\ (\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}) \geq \varepsilon \end{cases}$$

where the threshold value ε can be determined through cross validation technique (Henderson 1996).

2.4 Performance Measure

Area under receiving operating characteristic curve (AUC-ROC; Baten *et al.* 2006) and area under precision-recall curve (AUC-PR; Davis and Goadrich 2006) were used to measure the prediction accuracy of the classifier. The false positive rate (α) and true positive rate (1-β) were extracted for different threshold values between 0 and 1. The AUC-ROC was then computed as

$$\sum_i \left\{(1-\beta_i.\Delta\alpha) + (1/2)\left[\Delta(1-\beta).\Delta\alpha\right]\right\},$$

where $\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1})$, $\Delta\alpha = \alpha_i - \alpha_{i-1}$ and *i*=1,2, …, *m* (number of test instances) (Bradley 1997). For the unbalanced class distribution, AUC-PR provides a better measure for assessing the performance of the classifiers as compared to AUC-ROC (Sonnenburg *et al.* 2007). Thus, AUC-PR was also computed following Davis-Goadrich approach (Davis and Goadrich 2006). Moreover, for comparison among AUC-ROC and AUC-PR the standard errors were computed as

$$SE = \sqrt{[\theta(1-\theta) + (m^{Tss}-1)(q_1 - \theta^2) + (m^{Fss}-1)(q_2 - \theta^2)] / m^{Tss}.m^{Fss}},$$

where $q_1 = \theta/(2-\theta)$ and $q_2 = 2.\theta^2/(1+\theta)$; m^{Tss}, m^{Fss} and θ are the number of positive instances (*Tss*), number of negative instances (*Fss*) and estimate of AUC-ROC (AUC-PR) for the test dataset respectively.

2.5 Optimizing the Parameter λ

The only parameter need to be optimized is lambda (λ). The model was trained using NN269 training dataset with 9 different values of lambda i.e., 0.1 to 0.9 at an interval of 0.1, and prediction was made for the corresponding test set. The value of lambda at which lowest misclassification error occurred was chosen as the optimum one.

2.6 Comparison with Other Methods

The performance of the proposed approach was compared with that of other probabilistic approaches viz., WAM, MM1, MDD, MEM and SAE. The comparison was made using NN269 test dataset that comprises 208 true and 782 false splice site sequences. The scores of WAM, MM1, MDD and MEM were computed by using MaxEntScan tool (http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html), whereas the score for SAE was obtained by using *d*ssPred tool (http://cabgrid.res.in:8080/sspred/). The comparison was made in terms of estimates of AUC-ROC and AUC-PR.

3. RESULTS

3.1 Optimum Value of Lambda

The ROC and PR curves are plotted for all the nine values of λ (Fig. 1). It can be seen that the ROC and PR curves are better for small values of lambda (λ≤0.5) as compared to the higher values of lambda (λ≥0.6). It is also seen that the ROC and PR curves are very close to each other for λ=0.1, 0.2, 0.3, 0.4. For more clarity, the values of ($\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}$) are plotted for these four values of lambda (Fig. 2), where the x-axis represents the test sequence (true and false) and y-axis represents the values of ($\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}$) for the test sequences. It can be seen that the difference ($\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}$) are less than zero for true splice sites, whereas it is greater than zero for false splice site sequences (Fig. 2). The *Tss* and *Fss* sequences having $\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss} > 0$ and $\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss} < 0$ of are said to be misclassified as *Fss* and *Tss* respectively. After looking at Fig. 1 and Fig. 2 it can be said that the misclassification error is lowest at λ=0.3. Thus the optimum value of lambda was determined as 0.3.

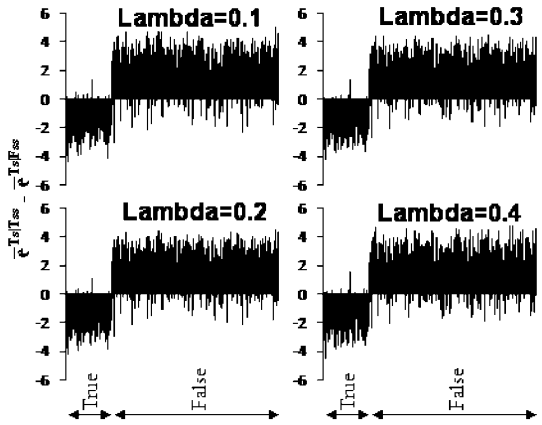

Fig. 2. Barplot of $\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}$ values for 208 true and 782 false splice site sequences of NN269 test dataset. The scores $\bar{e}^{Ts|Tss}$ and $\bar{e}^{Ts|Fss}$ are obtained by using the proposed approach. It can be seen that the difference values $\bar{e}^{Ts|Tss} - \bar{e}^{Ts|Fss}$ for first 208 true sites are mostly less than zero and for the false splice sites the difference values are above zero.

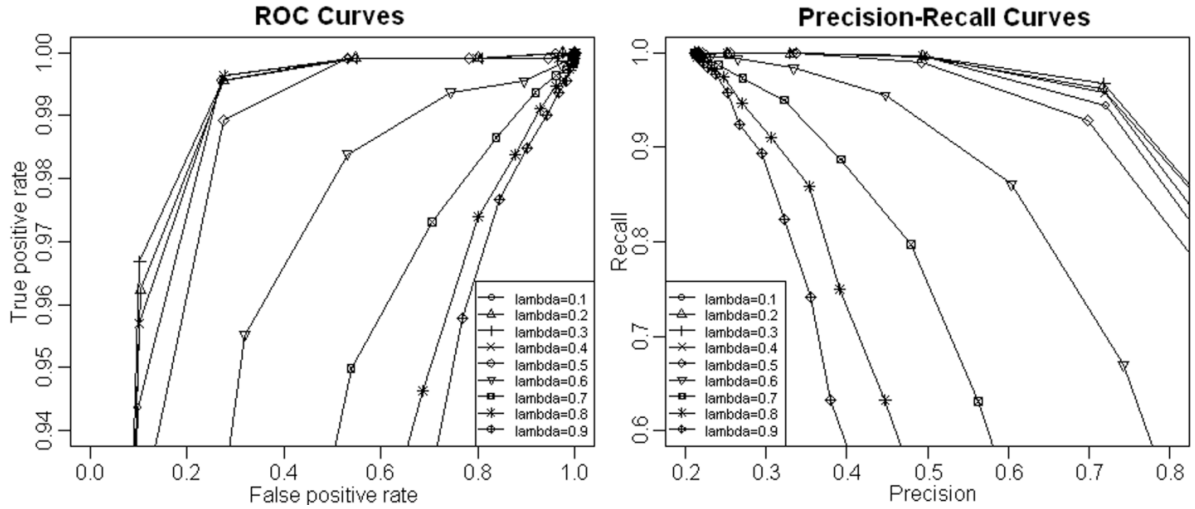

Fig. 1. ROC and PR curves for nine different values of λ. The ROC and PR curves are drawn based on the scores obtained from the proposed approach by using NN269 dataset. It can be seen that the ROC and PR curves covered maximum area at λ=0.3 as compared to the other values of λ.

3.2 Performance Analysis of the Proposed Approach

The prediction accuracy of the proposed approach was assessed through 5-fold cross validation technique. In cattle and fish data, the estimates of AUC-ROC and AUC-PR of the proposed approach for all the folds are given in Table 1. It is observed that except for the first fold the estimates of AUC-ROC and AUC-PR are consistent over rest of the folds. It is also seen that the average values of AUC-ROC are 93.75% and 96.13%, whereas AUC-PR are 93.50% and 96.13% for fish and cattle dataset respectively (Table 1). It is also noticed that the estimates AUC-ROC and AUC-PR for cattle are almost same with that of fish across all the five folds of the cross validation.

Table 1. Estimates of AUC-ROC and AUC-PR of the proposed approach for all the five folds of the cross validation corresponding to both fish and cattle splice site datasets.

Cross Validation	Fish		Cattle	
	AUC-ROC±SE	AUC-PR±SE	AUC-ROC±SE	AUC-PR±SE
CV-1	95.31±0.48	97.60±0.34	95.28±0.48	97.58±0.34
CV-2	93.12±0.59	95.50±0.47	93.12±0.59	95.51±0.47
CV-3	92.96±0.59	95.93±0.45	92.70±0.60	95.93±0.45
CV-4	93.49±0.57	95.05±0.50	92.50±0.61	95.05±0.50
CV-5	93.87±0.55	96.58±0.41	93.91±0.55	96.58±0.41
Average	93.75±0.56	96.13±0.43	93.50±0.56	96.13±0.43

3.3 Comparative Analysis

The scores obtained from different probabilistic approaches were used to estimate the values of AUC-ROC and AUC-PR. The scores obtained under SAE and proposed approach are observed to be highly correlated (Fig. 3) and at the same time the AUC-ROC and AUC-PR are seen to be similar (Table 2). On the other hand, though the scores of proposed approach is found to be highly correlated with that of WMM as compared with the others (Fig. 3), its estimates of AUC-ROC and AUC-PR are ~1% and ~2% respectively higher than that of WMM (Table 2). It is also seen that the estimates of AUC-ROC and AUC-PR of MEM and MM1 are almost equal i.e., ~96% and ~92% and higher than that of MDD, WMM, SAE and proposed approach. Further, it can be seen that the estimates of AUC-ROC of MDD and proposed approach are almost same (~95.5%), whereas AUC-PR of proposed approach is ~11% higher than that of MDD. Though the scores of WMM is observed to be highly correlated (>0.9) with the remaining approaches (Fig. 3), its estimates of AUC-ROC is lowest among

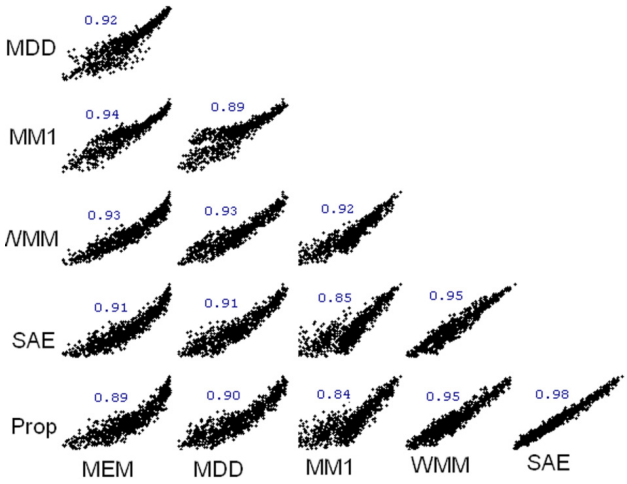

Fig. 3. Scatter plots of the scores between all possible pair of splice site prediction approaches. The scores of the approaches are highly correlated with each other.

Table 2. Estimates of AUC-ROC and AUC-PR of different probabilistic approaches based on NN269 splice site dataset.

Approaches	Performance measure	
	AUC-ROC±SE	AUC-PR±SE
MEM	96.30±0.81	92.25±1.13
MDD	95.89±0.84	80.65±1.68
MM1	96.25±0.81	92.64±1.12
WMM	94.61±0.96	89.68±1.30
SAE	95.53±0.86	92.08±1.16
Proposed	95.52±0.87	92.66±1.16

Table 3. P-values of the Mann-Whitney U statistic for testing the significant difference among the prediction accuracies of different probabilistic approaches.

	AUC-ROC					AUC-PR				
	SAE	MEM	MDD	MM1	WMM	SAE	MEM	MDD	MM1	WMM
MEM	0.92					0.58				
MDD	0.17	0.13				0.04	0.04			
MM1	0.63	0.59	0.13			0.58	0.99	0.04		
WMM	0.31	0.24	0.39	0.09		0.03	0.15	0.04	0.18	
Proposed	0.96	0.94	0.13	0.70	0.31	0.95	0.89	0.06	0.97	0.31

all (Table 2). Similarly, MDD is found to be highly correlated with MEM, WMM and proposed approach but its AUC-PR is ~10% less than the remaining approaches. Though, MEM and MM1 and SAE seem to perform better, their accuracies are not significantly higher than that of proposed approach (Table 3).

3.4 Determining the Threshold Value

Determining a threshold value is essential to distinguish the true splice sites from false ones. Though, MEM seems to be better as compared to the others (Table 2), choosing a threshold value in MEM is difficult (Golam Bari *et al*. 2014). In our approach, the TPR and TNR are plotted across a range of threshold values in all the five folds of cross validation for balanced splice site dataset, in both cattle and fish (Fig. 4). It is seen that the values of true positive rate (TPR) and true negative rate (TNR) are almost equal at $\varepsilon = 0$. Besides, from the box plot of the scores of true and false splice sites (Fig. 5), it is observed that the scores of false splice sites are mostly lying above zero whereas less than zero for the true splice sites. So, it can be said that the threshold value of 0 can be used to differentiate the true splice sites from false ones. At this threshold, the values of different performance metrics viz., TPR, TNR, Classification accuracy and Matthew's correlation coefficient (MCC) were also computed for the NN269 test dataset (Table 4) and observed that the accuracy is ~95% and MCC is ~81% (Table 4), which indicates the better predictive ability of the proposed approach.

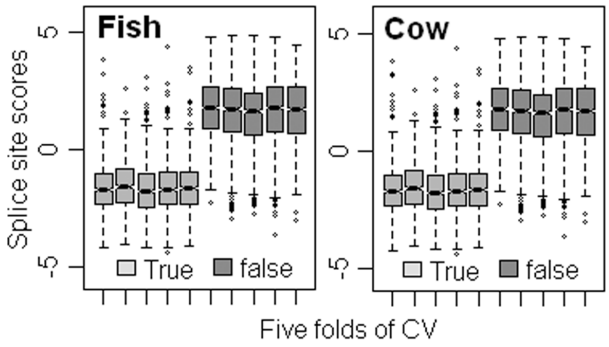

Fig. 5. Box plots of the scores obtained by using the proposed method in both fish and cattle. It can be noticed that the scores for the true splice sites are mostly less than zero, whereas the scores are mostly above zero for the false splice site sequences in both the species, across all the five folds of the cross validation.

Table 4. Values of different performance metrics for proposed approach at threshold value zero.

Performance metrics			
TPR	TNR	Accuracy	MCC
97.59	91.30	94.45	81.22

4. DISCUSSION

The accuracy of gene finding approach depends on addressing many problems and correct identification of splice junction is one of them (Kamath *et al*. 2014). The splice site prediction problem is therefore considered as a primary subtask in gene finding (Kamath *et al*. 2012). This paper presents an innovative approach based on kernel regression for prediction of splice junctions. The approach was employed for the prediction of donor splice sites in cattle, fish and human species.

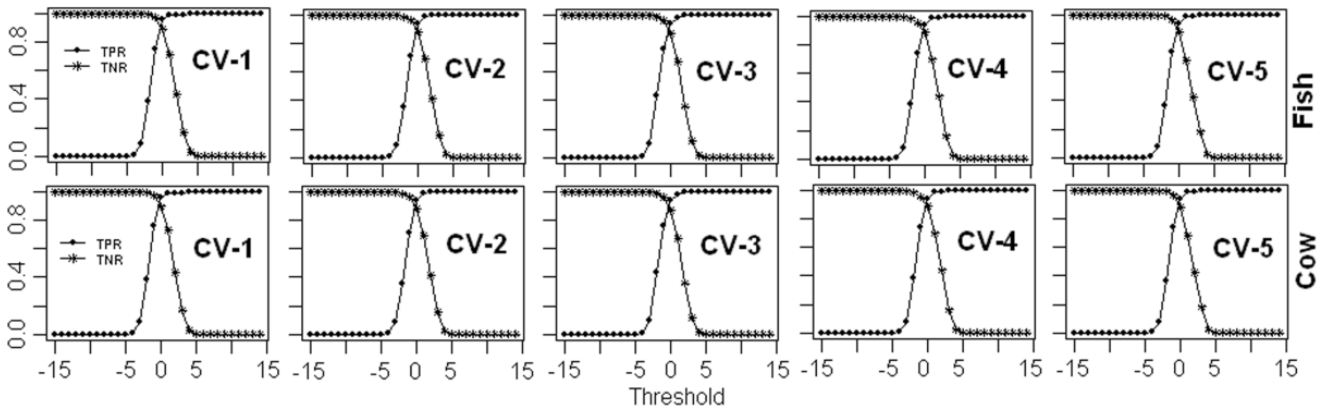

Fig. 4. TPR and TNR across different threshold values for all the five folds of cross validation in both cattle and fish datasets. It can be seen that in both the species, the values of TPR and TNR are almost equal at threshold zero across all the five folds.

In machine learning based approaches, SVM has been more frequently and successfully used in splice site prediction. However, for sub-optimal sequence length it may be less accurate than probabilistic approaches like first order Markov model (Sonnennburg *et al*. 2007). Moreover, by using longer sequence motif the number of parameters need to be estimated may become large. Besides, the parameters of the learning technique also need to be optimized to obtain higher prediction accuracy. In the proposed approach, the only parameter λ was optimized and the value 0.3 was found to be optimum. Also, high prediction accuracy was observed at this value of λ, which may further increase with fine tuning of the parameter λ.

The proposed approach was employed in both balanced (cattle and fish) and unbalanced (human) datasets. It was found that in case of balanced datasets the difference between the estimates of AUC-ROC and AUC-PR is less than that of unbalanced dataset, which is also true for machine learning classifiers (Sonnenburg *et al*. 2007). This may be due to the fact that AUC-ROC is independent of the class distribution, whereas AUC-PR takes into account both TPR and false positive rate (FPR) (Meher *et al*. 2014). Moreover, the difference may further be inflated with increase in the degree of unbalancedness in the datasets.

The performance of the proposed approach was compared with other probabilistic approaches viz., SAE, MEM, MDD, MM1 and WMM. In MEM both 0^{th} and 1^{st} order dependencies are considered (Yeo and Burge 2004), whereas in MDD and MM1 only 1st order dependencies are taken into account. WMM considers only 0^{th} order dependency for scoring the putative splice site motif (Meher *et al*. 2014). In SAE, all possible first order dependencies are taken into consideration. In the proposed approach neither 0th order nor 1^{st} order dependencies were considered. However, certain weights were assigned for match and mismatch of the nucleotides of test sequence with the nucleotides of training sequence and then a normalized score of the test sequence was computed. It was found that by assigning the higher weights ($\lambda \geq 0.5$) for a correct match the classification accuracy was not improved. In terms of AUC-ROC, the performances of different approaches were not found to be significantly different from each other, whereas MDD achieved significantly less accuracy in terms of AUC-PR than the other approaches (except proposed one). For the proposed approach, a threshold value 0 was also determined to distinguish the true splice sites from false splice sites. The earlier methods such as MDD, WMM and MM1 have ignored the false splice site sequences while training the prediction model. However, false ones are also needed to train the prediction model (Huang *et al*. 2006) and thus both true and false splice sites were considered to train the proposed prediction model.

In this study, a kernel regression based probabilistic approach was proposed, which can be used as complementary method to the existing ones for splice site prediction in eukaryotes. Above all, it is an independent thought and further improvement can be made upon it. We have also developed the R-code of the proposed approach that can be used for reproducibility of the work.

ACKNOWLEDGEMENTS

The authors sincerely acknowledged the support received from Director, ICAR-IASRI. The authors also acknowledged the guidance received from Dr. Samsiddhi Bhattacharjee, Assistant Professor, National Institute of Biomedical Genomics, Kalyani, Kolkata.

REFERENCES

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

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern. Recogn.,* **30**, 1145-1159.

Burge, C. and Karlin, S. (1997). Predictions of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78-94.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *In: ML '06: Proceedings of the 23rd international conference on machine learning*. New York, USA, pp 233-240.

Degroeve, S., Saeys, Y., De Baets, B., *et al*. (2005). SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics*, **21(8)**, 1332-1338.

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

Henderson, J., Salzberg, S. and Fasman, K.H. (1996). Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol*., **4**, 127-141.

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

Kamath, U., Compton, J., Islamaj Dogan, R., *et al*. (2012). An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice-site prediction. *IEEE Trans. Comp. Biol. and Bioinf*., **9**, 1387-1398.

Kamath, U., De Jong, K. and Shehu, A. (2014). Effective automated feature construction and selection for classification of biological sequences. *PLoS ONE,* **9(7)**: e99982. doi:10.1371/journal.pone.0099982.

Meher, P.K., Sahu, T.K. and Rao, A.R. (2016a). Prediction of donor splice sites using random forest with a new sequence encoding approach. *BioData Mining*, **9**, 4.

Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016b). Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algor. Mole. Biol.*, **11,** 16.

Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016c). A computational approach for prediction of donor splice sites with improved accuracy. *J. Theo. Biol.*, **404**, 285-294.

Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2014). A statistical approach for 5′ splice site prediction using short sequence motif and without encoding sequence data. *BMC Bioinformatics,* **15**, 362.

Nadaraya, E.A. (1964). On estimating regression. *Theory Pob. Appl*., **9**, 141-142.

Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997). Improved splice site detection in genie. *J. Comput. Biol.,* **43**, 311-323.

Sonnenburg, S., Schweikert, G., Philips, P., *et al*. (2007). Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, **8** (Suppl 10), S7.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*., **12**, 505-519.

Watson, G.S. (1964). Smooth regression analysis. S*ankhya A*, **26**, 359-372.

Wei, D., Zhang, H., Wei, Y. and Jiang, Q. (2013). A novel splice site prediction method using support vector machine. *J. Comput. Inform. Sys.,* **920**, 8053-8060.

Yeo, G. and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol*., **11(2-3)**, 377-394.

Zhang, M.Q. and Marr, T.G. (1993). A weight array method for splicing signal analysis. *Comput. Appl. Biosci*., **9(5)**, 499-509.

Zhang, Q., Peng, Q., Zhang, Q., *et al*. (2010). Splice sites prediction of human genome using length-variable Markov model and feature selection. *Expert Syst. Appl*., **37**, 2771-2782.