J. Plant Biochem. Biotechnol. (October–December 2015) 24(4):385–392
DOI 10.1007/s13562-014-0286-2

ORIGINAL ARTICLE

Determination of window size and identification of suitable method for prediction of donor splice sites in rice (*Oryza sativa*) genome

Prabina Kumar Meher · Tanmaya Kumar Sahu · A. R. Rao · S. D. Wahi

Received: 24 February 2014 / Accepted: 28 July 2014 / Published online: 18 November 2014

Abstract Accurate prediction of the gene structure depends upon the accurate prediction of splice sites. The conserved feature in splicing junction has been successfully used for the prediction of eukaryotic splice sites. In eukaryotes, though the di-nucleotide GT is conserved at 5′ splice sites, the pattern surrounding the conserved di-nucleotide varies from species to species. Most of the work related to splice site analysis has been extensively done in *Homo sapiens* and *Arabidopsis thaliana*. However, such works are yet to be fully explored in *Oryza sativa* and other species of grass family. In this study, statistical techniques have been applied to discriminate the real splice sites from pseudo splice sites in rice, maize and barley genomes and based on this a suitable window size is determined for the prediction of donor splice sites. Depending upon the determined window size, appropriate methods for predicting donor splice sites in rice have been considered and compared in terms of prediction accuracy. The results revealed that a window size of 9 base pair (3 bp at the exon end and 6 bp at the intron start including the conserved di-nucleotide GT at the beginning of intron) is an effective window size in all the three species of grass family for the prediction of donor splice sites. Further, the Maximum Entropy Model based method is found as best among the short sequence based prediction methods for donor splice sites with the 9 base pair window size.

Keywords Splice sites · Prediction accuracy · Window size · Short sequence motif

Abbreviations

MLAs	Machine Learning Approaches
MEM	Maximum Entropy Modeling
MDD	Maximal Dependency Decomposition
MM1	Markov Model of 1st order
WMM	Weighted Matrix Method

Splice sites are the junctions of the exon-intron boundaries. There are two type of splice sites namely acceptor (3′) splice sites and donor (5′) splice sites. Donor splice sites, with GT, correspond to the beginning of introns and acceptor splice sites, with AG, corresponds to the end of introns, together known as canonical splice sites (GT-AG type). These canonical splice sites are abundant among different types of splice sites in eukaryotes (Sheth et al. 2006). The accurate prediction of splice sites always leads to accurate gene structure prediction, thus splice sites are vital from genome annotation point of view.

Most of the splice site prediction methods are based on Machine Learning Approaches (MLAs) *i.e.,* Classification trees (Burge and Karlin 1997), Artificial Neural Networks (ANN; Ho and Rajapakse 2003) and Support Vector Machines (SVM; Sun et al. 2003; Sonnenburg et al. 2007).

Electronic supplementary material The online version of this article (doi:10.1007/s13562-014-0286-2) contains supplementary material, which is available to authorized users.

P. K. Meher · S. D. Wahi
Division of Statistical Genetics, Indian Agricultural Statistics Research Institute, New Delhi 110012, India

P. K. Meher
e-mail: pkmeher@iasri.res.in

S. D. Wahi
e-mail: sdwahi@iasri.res.in

T. K. Sahu · A. R. Rao (✉)
Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi 110012, India
e-mail: rao.cshl.work@gmail.com

A. R. Rao
e-mail: arrao@iasri.res.in

T. K. Sahu
e-mail: tanmayabioinfo@gmail.com

In splice site prediction using MLAs, selection of proper window size is crucial. Most of the criteria used for window size selection are based on the pilot studies, which involves the use of smaller samples for testing the accuracy of prediction methods with varying window sizes. Subsequently, the final prediction is made with the most favorable window size. Hebsgaard et al. (1996) tested six different window sizes *i.e.,* 101, 151, 201, 251, 301, 351 and 401 nucleotides to train the ANN model with different units in the hidden layer for splice site prediction in *Arabidopsis thaliana* and observed that the best network has a window of 201 nucleotides with 15 hidden units. Also, Degroeve et al. (2002) used five different combinations of window sizes *i.e.,* 20, 40, 60, 80 and 100 nucleotides in both upstream and downstream region of the conserved di-nucleotide GT for donor splice sites prediction in *Arabidopsis*. They observed that consideration of 50 nucleotides on both sides of GT exhibit better performance over other combinations. In addition, they also stated that the optimization of window size for different genomes is crucial for the induction of accurate species-specific splice site prediction model. Huang et al. (2006) examined a series of sequence lengths ranging from 32 to 50 nucleotide bases around the splice junction by keeping the left and right flanking regions symmetrical to get the optimum window size for predicting splice sites using SVM. Thus, from the above it is evident that the determination of suitable window length prior to the application of any prediction method is essential in terms of cost and time. Moreover, the use of short window size with maximum information in splice site prediction may save computational time and memory space. To our limited knowledge, determination of suitable window size in grass family has not yet been fully explored. Also, there exist different methods for splice site prediction with different window sizes. However, identification of an efficient splice site prediction method with an appropriate window size is yet to be explored. Thus, the present study is conducted with two objectives: (i) to determine suitable window size in rice, maize and barley of grass family (ii) to identify appropriate method for donor splice site prediction using the determined window size in rice genome.

Material and methods

Collection and processing of data

For the present study, exon and intron sequences of *Oryza sativa* were collected from the FTP site of rice Genome Annotation Project (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/). Further, real or true splice sites (TSS) having 100 nucleotides at exon end and 102 nucleotides from intron start (including the conserved di-nucleotide GT at intron start) were extracted from the collected exon and intron sequences thorough a developed Perl program. Also, one more Perl script was developed to extract the pseudo or false splice sites (FSS) of length 202 solely within the exonic/intronic sequences having G, T at 101st, 102nd position respectively. Keeping in view the distribution of TSS and FSS on genome and availability of computational resources at hand, a sample dataset with 16,330 TSS and 51,994 FSS was considered in case of rice for the present investigation. In case of maize and barley, the gene sequences were collected from NCBI (http://www.ncbi.nlm.nih.gov/) and splice sites were extracted using a Perl script. False splice sites were extracted in a similar way as followed in rice. A final dataset of 500 TSS and 2,500 FSS, both for maize and barley, was considered for the determination of suitable window size.

Position weight matrix

The splice site motif sequences of both TSS as well as FSS were aligned separately using the di-nucleotide GT as the anchor. The alignment was used to calculate the probabilities of nucleotides at each position on the motif to determine the conservedness in TSS and FSS.

From a given set of N aligned sequences each of length P, $S_k=(x_{1k},x_{2k},\ldots,x_{Pk})$, where $x_{ik}\in\{A,T,G,C\}\,;\forall i=1,2,\ldots,P$; $k=1,2,\ldots,N$, the Position Weight Matrix (PWM) was computed as

$$p_{ib}=\frac{1}{N}\sum_{k=1}^{N} I(x_{ik})\ ,\ b\in\{A,\ C,\ G,\ T\}\ and\ i=1,2,\ \ldots,\ P$$

$$where\ I(x_{ik})=\begin{cases}1, & if\ x_{ik}=b\\ 0, & otherwise\ .\end{cases}$$

The PWM was constructed, with four rows one $_1$ each for A, C, G, and T and 20 columns with 10 from the exon side and 10 from the intron side (excluding GT at the beginning of intron), separately for TSS and FSS and independently for rice, maize and barley. Further, variance in the difference of PWM (VDPWM) between TSS and FSS was computed to visualize the difference in variability pattern between TSS and FSS in all the three species.

Kullback Leibler divergence

Kullback Leibler Divergence (KLD; Kullback and Leibler 1951) was used to measure the positional variation in terms of distribution of nucleotide bases between TSS and FSS. Here, the position wise aligned sequence data of TSS and FSS were used to obtain KLD for 20 positions (10 positions at the exon end and 10 positions from the intron start excluding the conserved di-nucleotide GT at the beginning of intron). The KLD using TSS and FSS was calculated as follows;

In general, for two multinomial populations p_1 and p_2, each having K classes with probability $(p_1(1),p_1(2),\ldots,p_1(K))$ and probability $(p_2(1),p_2(2),\ldots,p_2(K))$, the KLD between the two populations is given by

$$KLD(p_1,p_2) = -\sum_{i=1}^{K} p_1(i).\log_2\left(\frac{p_1(i)}{p_2(i)}\right).$$

Accordingly, for the splice site motifs, the divergence between the i^{th} position of TSS (t_i) and the i^{th} position of FSS (f_i) is computed as

$$KLD(t_i, f_i) = -\sum_{b\in\{A,T,G,C\}} p_t^b \log_2\left(\frac{p_t^b}{p_f^b}\right)$$

$$= p_t^A \log_2\left(\frac{p_t^A}{p_f^A}\right) + p_t^T \log_2\left(\frac{p_t^T}{p_f^T}\right) + p_t^G \log_2\left(\frac{p_t^G}{p_f^G}\right) + p_t^C \log_2\left(\frac{p_t^C}{p_f^C}\right)$$

where, p_t^b is the probability of occurrence of base b at i^{th} position in TSS and p_f^b is the corresponding value in the FSS. The probabilities p_t^b and p_f^b can be obtained from the PWM of TSS and FSS respectively.

Pearson Chi-square

The difference in the distributions of four nucleotide bases *i.e., A, T, G, C* corresponding to a given position between TSS and FSS was computed using a Chi-square statistic, which was obtained from the 2×4 contingency table (Table 1). The Chi-square statistic is given by

$$\chi^2 = \frac{N^2}{R_t \times R_f}\left[\sum_{b\in\{A,T,G,C\}}\left(\frac{\left(n_t^b\right)^2}{C_b} - \frac{R_t^{\,2}}{N}\right)\right].$$

In order to observe the difference in the distribution pattern between TSS and FSS, a bar diagram was plotted by taking 20 different positions (10 positions at the exon end and 10 positions from the intron start excluding the conserved di-nucleotide GT at the beginning of intron) on the X-axis and the calculated Chi-square on the Y-axis.

Table 1 2×4 Contingency table showing the frequency distribution of bases of TSS and FSS for a given position. Here, $n_t^A, n_t^T, n_t^G, n_t^C$ and $n_f^A, n_f^T, n_f^G, n_f^C$ are the frequency of bases *A, T, G, C* respectively in TSS and FSS corresponding to the same position

Bases Sites	A	T	G	C	Total
TSS	n_t^A	n_t^T	n_t^G	n_t^C	R_t
FSS	n_f^A	n_f^T	n_f^G	n_f^C	R_f
Total	C_A	C_T	C_G	C_C	N

Cramer's V coefficient

Cramer's V coefficient (CVC; Cramér 1946) was used for finding associations among different positions in TSS and FSS motifs, described as below:

Let n_i^b and $n_j^{\,b}$ be the frequencies of nucleotide base b at positions i and j respectively. Similarly $n_{ij}^{b\times b}$ is the frequency of pair of nucleotides $b\times b$ together corresponding to the position (i, j), where $i,j=1,2,\ldots,P\ andi<j$; $b\in\{A,T,G,C\}$; $b\times b\in\{AA,AT,AG,AC,\ldots,CC\}$. Then a 4×4 contingency table (Table 2) was prepared for computing the association between any two positions in respect of distribution of four nucleotide bases. Using this contingency table, the Pearson Chi-square value was computed as

$$\chi^2 = \sum_b\sum_b \frac{\left[n_{ij}^{b\times b} - \left(\frac{n_i^b.n_j^b}{N}\right)\right]^2}{\left(\frac{n_i^b.n_j^b}{N}\right)}$$

Then, the CVC was computed using the formula $\phi_c = \sqrt{\frac{\chi^2}{N(m-1)}}$, where χ^2 is the Pearson Chi-square obtained from a 4×4 contingency table (Table 2) and m=min(4,4). The range of CVC varies from 0 (no association) to 1 (complete association).

CVCs were calculated for all possible pairs of positions, by taking 20 positions (10 positions at the exon end and 10 positions from the intron start excluding GT at the beginning of intron), separately both for TSS and FSS. CVCs were plotted in a line graph by taking different positions in the X-axis and their associations in the Y-axis.

Determination of window size

Initially, the window sizes under KLD, CVC, Pearson Chi-square and VDPWM were determined by taking those positions in the splice site motifs where the respective values were

Table 2 4×4 Contingency table depicting the distribution of nucleotide bases for two different positions either in TSS or FSS motifs

i *j*	A	T	G	C	Total
A	n_{ij}^{AA}	n_{ij}^{AT}	n_{ij}^{AG}	n_{ij}^{AC}	n_i^A
T	n_{ij}^{TA}	n_{ij}^{TT}	n_{ij}^{TG}	n_{ij}^{TC}	n_i^T
G	n_{ij}^{GA}	n_{ij}^{GT}	n_{ij}^{GG}	n_{ij}^{GC}	n_i^G
C	n_{ij}^{CA}	n_{ij}^{CT}	n_{ij}^{CG}	n_{ij}^{CC}	n_i^C
Total	n_j^A	n_j^T	n_j^G	n_j^C	N

higher as compared to the other positions. Subsequently, the final window size was determined on a consensus basis.

Prediction of donor splice sites

Prediction of donor splice sites in rice was made by considering four existing splice site prediction methods with the determined window size. In this study, four different scoring based approaches, those are capable of predicting splice sites using short sequence motifs were considered for the prediction purposes. These methods are based on Maximum Entropy Modeling (MEM; Yeo and Burge 2004) score, Maximal Dependency Decomposition (MDD; Burge and Karlin 1997) score, Markov Model of 1st order (MM1) score and Weighted Matrix Method (WMM; Staden 1984) score. All these scores were obtained from the URL: http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html. The performance metrics were further computed for comparison of the prediction methods.

Performance metrics

The metrics such as Classification Accuracy (CA), True Positive Rate (TPR), True Negative Rate (TNR), Precision, F-measure, Weighted Accuracy (WA) and Matthew's Correlation Coefficient (MCC; Matthews 1975), all of which are the function of confusion matrix, were used to evaluate the performance of different prediction methods. The confusion matrix contains information about actual and predicted classes. Supplementary Fig. 1 shows the confusion matrix for a binary classifier. The different performance metrics used for assessing the prediction accuracy are as follows;

$$\mathrm{CA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}},\ \text{TPR or Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\ (\text{Same as recall for binary classification}),\ \text{TNR or Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\ \text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\ \text{F-measure}^{(\alpha)} = \frac{(1+\alpha)\times \text{recall}\times\text{precision}}{(\alpha\times\text{recall})+\text{precision}},\ \text{F-measure}^{(\beta)} = \frac{\left(1+\beta^2\right)\times \text{recall}\times\text{precision}}{\left(\beta^2\times\text{recall}\right)+\text{precision}},$$

$$\mathrm{WA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)\ \text{and}\ \mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})\text{-}(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})\left(\mathrm{TP}+\mathrm{FN}\right)(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

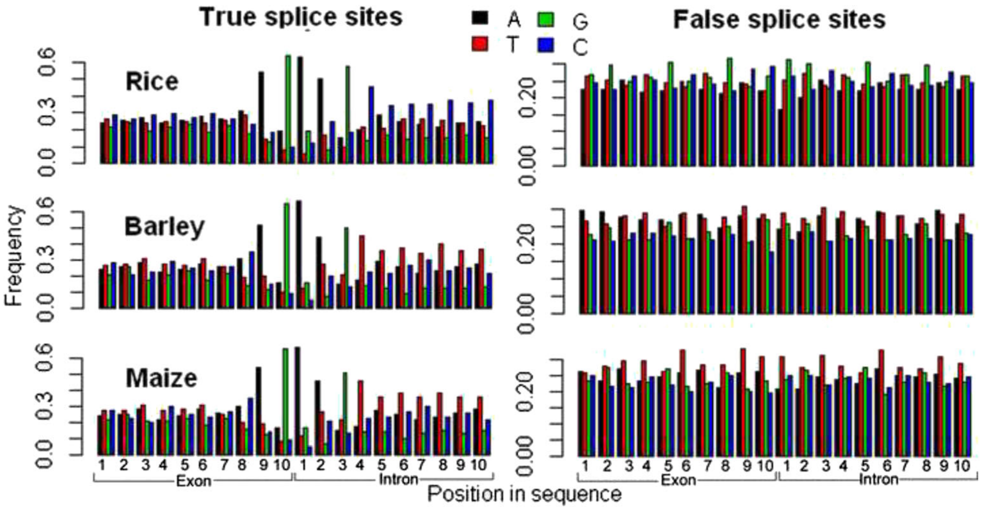
Results and discussion

Identification of protein coding genes requires identification of the start codons, exons, introns and stop codons. The performance of most gene finding systems is greatly influenced by their accuracy at determining the splice sites (Pertea et al. 2001). Therefore, prediction of splice sites plays an important role in predicting the gene structure. The 5′ boundary or donor site is conserved with di-nucleotide GT at intron start and most of the studies in the area of splice site prediction have focused on this conserved feature. Further, in prediction of splice sites, the selection of window size plays an important role as far as computational complexity, memory allocation and feasibility is concerned. Often, the window sizes are determined on pilot study basis, where the prediction method is applied on a sample dataset with varying window sizes and the window size with higher accuracy is considered for final prediction. This takes a lot of time and memory, which can be avoided by selecting window size prior to the application of prediction method. In this study, we explored the application of some of the existing classical statistical techniques like PWM, CVC, KLD and Pearson Chi-square in determining the window size. To demonstrate the techniques in determining the window size, the true and false splice site datasets of rice are used. Besides, to determine a more generalized window size in grass family, the window sizes for two more species *i.e.,* maize and barley are also determined.

Conservedness in donor splice site motifs

After looking at the conservedness of the nucleotides at different positions in splice site motifs, it is inferred that the frequencies of nucleotide bases in the positions 8 to 14 (the two positions at the intron start conserved with GT has not been considered) are not equally likely (Fig. 1). However, the frequencies of nucleotide bases in rest of the positions in the splice site motifs are at par. More specifically, the frequencies of the bases on the exon side are almost equal to each other but on the intron side the occurrence of nucleotide **C** is more frequent as compared to others in case of rice. However, in case of barley and maize, the likelihood of occurrence of nucleotide **T** is more than the other three bases. It is also observed that in the considered three species of grass family the frequencies of nucleotide bases in different positions, in case of FSS motifs are at par (Fig. 1). This indicates the variation among nucleotides is present at the positions surrounding the splicing junction in case of TSS but not in FSS. In other words, in case of TSS, certain nucleotides are partially conserved surrounding the conserved GT at intron start but this has not been observed in case of pseudo splice sites.

Fig. 1 PWM of TSS and FSS in rice, maize and barley. Positions in the splice site motif are represented in X-axis and Y-axis represents the proportion of 4 nucleotide bases for the corresponding positions



Besides, it is also observed that the variability in the difference of true splice sites PWM and false splice sites PWM is higher in these 7 positions (3 from the exon end and 4 from the intron start excluding di-nucleotide GT at the beginning of intron) surrounding the splice junction (Fig. 2). Further, it can be seen that the variability at the intron side is more than that of exon side. From the analysis of PWM and VDPWM, it is inferred that these 9 bp (3 from the exon end and 4 from the intron start including di-nucleotide GT at the beginning of intron) window size plays an important role in discriminating the true splice sites from the false one.

Variations in nucleotides distribution at donor splice sites

The values of the Pearson Chi-square corresponding to the considered 20 positions of splice site motif are plotted in Supplementary Fig. 2. It is observed that the value of calculated Chi-square for each position is larger than 8 and hence greater than the tabulated value of the Chi-square statistic for 3 degrees of freedom (7.815). This indicates the significant difference in nucleotide distribution between TSS and FSS for all the considered positions in splice site motif. Out of 10 positions at the exon end (1–10 positions in Supplementary Fig. 2), the observed value of Chi-square at positions 8, 9 and 10 are higher than rest of the positions. However, out of 10 positions from the intron side (excluding two positions at the beginning of intron), only the first 4 positions show higher chi-square value than the remaining positions. Hence, the window size of length 7 base pair (3 from the exon end and 4 from the intron start excluding di-nucleotide GT at the beginning of intron) is considered for further analysis. Similar to that of VDPWM, it is observed that the variability in the distribution of nucleotides at the intron side is more than the variability of the exon side in all the three species of grass family.

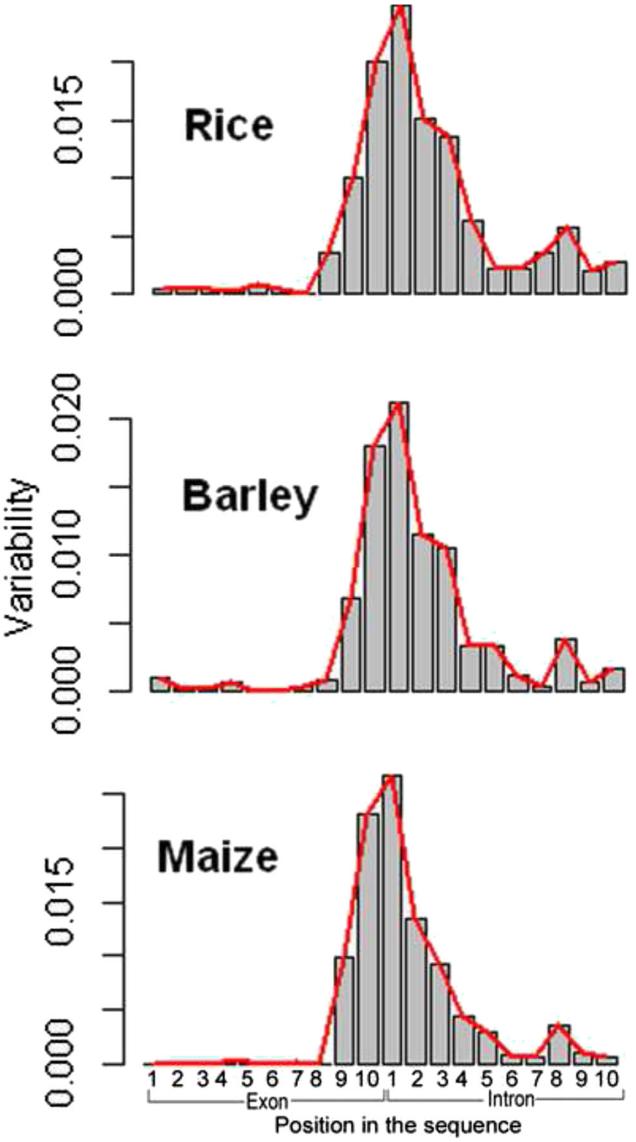


Fig. 2 Plotting of variance of the difference in PWM of TSS and FSS for rice, maize and barley. X-axis represents the position in splice site motif and Y-axis represents the variance of the corresponding positions

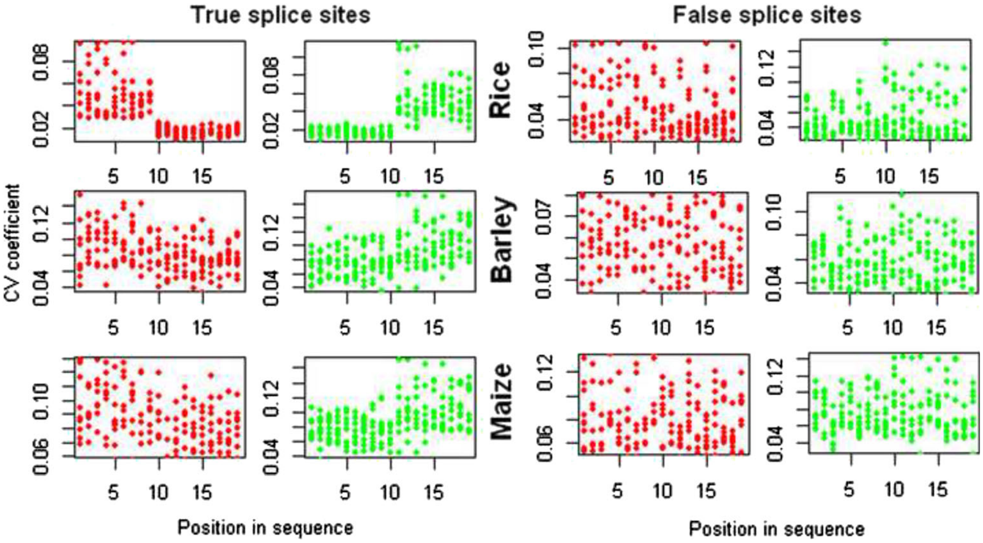


Fig. 3 Graphical representation of the association based on CV coefficients among different positions for TSS and FSS in rice. In case of TSS, the association among different positions both in exon side and intron side are higher compared to the association among the positions of exon and intron. However, there does not exist any such pattern in case of FSS

Distance between the positions of TSS and FSS motifs

The distances between the corresponding positions of TSS and FSS motif are computed using KLD measure. It is found that the distances between TSS and FSS corresponding to the first 7 positions are almost close to zero in all the three species of grass family (Supplementary Fig. 3). However, the distances between the positions 8 to 14 are higher as compared to rest of the positions. It is also worth noting that the distances between the corresponding positions of TSS and FSS in intron region is higher than that in the exon region (Supplementary Fig. 3). Taking this distance pattern into account, the positions from 8 to14 (3 from the exon end and 4 from the intron start excluding di-nucleotide GT at the beginning of intron) are considered to be important from discrimination point of view and thus considered as window size for the prediction of donor splice sites for the considered three species of grass family.

Positional associations in donor splice site motifs

The associations among the positions are computed separately for the TSS and FSS motifs for all the considered species. The association is computed using Crammer V measure by taking sequences of 20 bp (10 bp at the exon end and 10 bp at the intron start excluding di-nucleotide GT at the beginning of intron) length. It is observed that in case TSS, the associations among the positions within the exonic region are higher as compared to the associations between the positions of exonic and intronic regions (Fig. 3). Similarly, the associations among the positions within the intronic region are higher as compared to the associations between the positions of intronic and exonic regions (Fig. 3). On the other hand, any such pattern of association among positions in case of FSS does not exist. Further, it can be seen that the association in the three positions at the exon end and four positions at the intron start are higher as compared to other positions in TSS but absence of such feature is found in FSS. Hence, the positions 8–14 are considered as window size for the prediction purpose. It can also be seen that the distinction between the association of exonic and intronic region is clearer in case of rice as compared to barley and maize and this may be due to the fact that the availability of large number of experimentally validated splice sites in case of rice than that of barley and maize. Though, such distinction is not significantly clear in barley and maize, still some indication with respect to the difference in association is present.

Table 3 Performance metrics of four approaches in the prediction of donor splice sites using the sequence data of rice genome with 9 base pair window size

Metrics Methods	CA	TPR	TNR	Precision	F-measure $^{(\alpha=1)}$	F-measure $^{(\beta=2)}$	WA	MCC
MEM	0.852	0.691	0.903	0.691	0.691	0.691	0.797	0.595
MDD	0.846	0.678	0.899	0.678	0.678	0.678	0.788	0.577
MM1	0.846	0.677	0.899	0.677	0.677	0.677	0.788	0.576
WMM	0.840	0.665	0.895	0.665	0.665	0.665	0.780	0.560

Window size on consensus basis

Looking at the results obtained from PWM, VDPWM, KLD, Pearson Chi-square and CVC, the window size of length 7 base pair (3 from the exon end and 4 from the intron start excluding di-nucleotide GT at the beginning of intron) or 9 base pair (3 from the exon end and 6 from the intron start including di-nucleotide GT at the beginning of intron) is considered as the final window size on a consensus basis for the prediction of donor splice sites.

Predictive analysis

Keeping in view the length of determined window size being small in size, four appropriate splice site prediction approaches *viz.,* MEM, MDD, MM1 and WMM are chosen for the prediction of donor splice sites. All these four approaches required a minimum of 9 base pair window length for prediction of splice sites. It is observed that the performance of MEM approach is higher as compared to the other considered approaches in terms of performance metrics (Table 3). Also, it can be seen that the difference in the performances of MDD and MM1 is negligible. Whereas, the performance of WMM is lowest among the four approaches (Table 3). The WMM is based on the assumption of the positional independence, where the frequency of each nucleotide at each position is computed independently and further used in computing the log-odd score of splice site motifs. Based on the values of log-odd score, a certain threshold value is determined and the prediction of any test instance is made on the basis of this threshold value. However, in MM1, the scores are computed by taking into account the dependencies among the adjacent positions. Hence, the performance of MM1 is better than that of WMM. In case of MDD, the longer distance dependencies are modeled, where MDD splits the training data to fit different WMM and Weighted Array Method (WAM) to suitably define the subsets of the data. Splits are made at the most dependent positions and are chosen by Chi-square test statistic. MEM approximates the short sequence motif distributions with the maximum entropy distribution consistent with low-order marginal constraints estimated from available data that includes dependencies between non-adjacent as well as adjacent positions. Since MEM takes into account both adjacent as well as non-adjacent association, its performance is better than that of other methods considered under the present study and the performance of the prediction methods are in the order of MEM>MDD>MM1>WMM. It is also worth noting that in WMM and MM1, only the TSS are used while training the model in determining the threshold value for prediction. However, FSS are also necessary to train the model and to develop a robust prediction approach (Huang et al. 2006).

The power of statistics has been used to reveal the information present in the genomic sequence data, in particular, for the prediction of functional elements present in the DNA sequence. Splice site is one such important functional element present on DNA sequence. An accurate prediction of splice site is important for accurate prediction of gene structure. The conserved feature present in the exon-intron junction has been fully exploited to get higher prediction accuracy of the donor splice sites. In the present study, statistical techniques have been explored and used to distinguish the TSS from the FSS and the information obtained is further used in determination of window size. The Chi-square statistic is used to determine the effective positions that can help discriminate the TSS from FSS. In a similar way, the KLD measure is used to find the difference in the distribution of four nucleotide bases at each position corresponding to the TSS and FSS. The CVC is used in finding the associations among the positions in the splice site motif. It is concluded from the findings that a window size of 9 bp length, in general, is suitable for the prediction of donor splice sites in rice, maize and barley. Besides, the MEM model can be used for efficient prediction of donor splice site in rice genome. In addition, the reads generated from the next generation technology are shorter in size and recognition of splicing in short reads poses a challenge because they often align to numerous places in a genome, and often lack insufficient sequence specificity on one or both ends of exon-exon junction to accurately define junction (Wu and Nacu 2010). Moreover, to utilize short reads generated from the next generation sequencing technology for transcriptome sequencing and gene structure identification, one need to align accurately the sequence reads over intron boundaries and splice site prediction helps to improve the alignment quality (De Bona et al. 2008) and hence the MEM coupled with the 9 base pair window size is expected to be useful in the prediction of splice variants using short reads generated from next generation sequencing technologies.

Acknowledgements This study is a part of Ph. D. thesis of P. K. Meher, PG School, IARI, New Delhi. Authors acknowledge the INSPIRE fellowship of Department of Science and Technology, New Delhi and IARI Fellowship. The authors also acknowledge the computational facilities of SCGL, developed under NAIP grant NAIP/Comp-4/C4/C-30033/2008-09.

References

Burge C, Karlin S (1997) Prediction of complete gene structure in human genomic DNA. J Comput Biol 268(1):78–94

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton, p 282. ISBN 0-691-08004-6

De Bona F, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal splice alignments of short sequence reads. Bioinformatics 24:174–180

Degroeve S, De Baets B, Van de Peer Y, Rouz P (2002) Feature subset selection for splice site prediction. Bioinformatics 18:S75–S83

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24:3439–3452

Ho LS, Rajapakse JC (2003) Splice site detection with a higher-order Markov model implemented on a neural network. Genome Inf 14: 64–72

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

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

Pertea M, Lin X, Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res 29(5):1185–1190

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R (2006) Comprehensive splice site analysis using comparative genomics. Nucleic Acids Res 34:3955–3967

Sonnenburg S, Schweikert G, Philips P, Behr J, Rätsch G (2007) Accurate splice site prediction using support vector machines. BMC Bioinforma 8(Suppl 10):S7

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

Sun YF, Fan XD, Li YD (2003) Identifying splicing sites in eukaryotic RNA: support vector machine approach. Comput Biol Med 33:17–29

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881

Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11(2–3):377–394