GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

High-accuracy splice site prediction based on sequence component and position features

J.L. Li[1,2]*, L.F. Wang[1]*, H.Y. Wang[3], L.Y. Bai[2] and Z.M. Yuan[1,2]

[1]Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha, China
[2]College of Bio-Safety Science and Technology,
Hunan Agricultural University, Changsha, China
[3]Department of Statistics, Kansas State University, Manhattan, KS, USA

*These authors contributed equally to this study.
Corresponding author: Z.M. Yuan
E-mail: zhmyuan@sina.com

Genet. Mol. Res. 11 (3): 3432-3451 (2012)
Received December 15, 2011
Accepted March 30, 2012
Published September 25, 2012
DOI http://dx.doi.org/10.4238/2012.September.25.12

ABSTRACT. Identification of splice sites plays a key role in the annotation of genes. Consequently, improvement of computational prediction of splice sites would be very useful. We examined the effect of the window size and the number and position of the consensus bases with a chi-square test, and then extracted the sequence multi-scale component features and the position and adjacent position relationship features of consensus sites. Then, we constructed a novel classification model using a support vector machine with the previously selected features and applied it to the *Homo sapiens* splice site dataset. This method greatly improved cross-validation accuracies for training sets with true and spurious splice sites of both equal and different proportions. This method was also applied to the NN269 dataset for further evaluation and independent testing. The results were superior to those obtained with previous methods, and demonstrate the stability and superiority of this method for prediction of splice sites.

Key words: Splice site prediction; Multi-scale component features; Position features; Adjacent position relationship features; Support vector machine

INTRODUCTION

Owing to the tremendous increase in genomic sequence data, there is an urgent demand to improve the efficiency of computational algorithms for gene annotation (Sonnenburg et al., 2007). The accurate identification of splice sites plays a key role in the annotation of genes in eukaryotes (Baten et al., 2007; Rätsch et al., 2007). Most eukaryotic protein-coding genes are split genes that are composed of exons and introns. Introns are the non-protein-coding regions and are removed by RNA splicing in transcription. The border between an exon and an intron is termed a splice site. The splice sites consist of the donor site with an almost invariant dinucleotide GT at the beginning of the intron and the acceptor site with almost invariant dinucleotide AG at the end of the intron, and they are highly conserved consensus regions. Except for those canonical splice sites according to the GT-AG rule, there are very few variant ones with dinucleotide GC and AC as consensus regions, and their number accounts for approximately 1% in total (Burset et al., 2000). There are a large number of GT and AG dinucleotides in eukaryotic genes, but only 0.1% of them are true splice sites (Sonnenburg et al., 2007). How to identify whether or not a GT/AG dinucleotide is a true splice site is always one of the most important and challenging tasks in bioinformatics (Sonnenburg et al., 2007; Baten et al., 2008). In this article, we refer to true splice sites as positives and false ones as negatives.

In the literature, several statistical models have been constructed for splice site prediction. The weight matrix method (WMM) is the earliest and most influential one that uses the position-specific compositional biases (Staden, 1984; Tavares et al., 2009). Subsequently, the pattern recognition algorithms, represented by a Bayesian network (Cai et al., 2000), support vector machine (SVM) (Zhang et al., 2006; Baten et al., 2006; Sonnenburg et al., 2007; Asa et al., 2008; Zhang et al., 2009), hidden Markov model (Baten et al., 2007, 2008; Asa et al., 2008; Zhang et al., 2009, 2010), artificial neural network (Reese et al., 1997; Wang et al., 2009), etc., were successively introduced. A series of special prediction tools were also improved for splice site prediction, such as GeneSplicer (Pertea et al., 2001), DGSplicer (Chen et al., 2005), NNSplice (Reese et al., 1997), SpliceMachine (Kahn et al., 2007), etc. These methods, represented by WMM, construct their splice site statistical models mainly based on splicing signals, including sequence feature information of donor and acceptor splice sites, branch point motifs, protein coding potential of exons, etc. The fusion of splicing signals and RNA secondary structure features (Mareshi et al., 2008) could improve the prediction accuracy of acceptor sites but not so for donor sites. Moreover, it is computationally expensive to extract the features of RNA secondary structures (Zhang et al., 2010). The splicing regulatory elements around splice sites produce an important effect on the splicing process, especially for alternative splicing. These elements are generally short sequence motifs composed of 6-10 bases, including the enhancer and silencer appearing in the exon and intron regions, respectively. Thus, combining the feature information of splicing signals and regulatory elements could effectively improve the level of splice site prediction (Sun et al., 2008).

The existing methods of splice site prediction have achieved an acceptable level of accuracy. However, there are limitations. 1) It is of prime importance to further increase prediction accuracy, especially since the amount of pseudo-splice sites in s genomic sequence is so enormous that even a subtle improvement in prediction accuracy could drastically influence the absolute large number of pseudo-sites in predicted results. 2) Available algorithms are mainly based on Weblogo (Schneider and Stephens, 1990; Crooks et al., 2004), which makes different information content graphs for positives and negatives separately, instead of an integrated graph for positives and negatives. Moreover, the application of these graphs lacks quantitative criteria, such that even with the same datasets, the number and the position of consensus bases determined by different researchers could be different. 3) Considering the protein coding potential of exons and the excavation of regulatory element motifs with unsupervised learning, how to select the length of left and right windows with the splice sites as the center is a problem that researchers must take into deep consideration. 4) The protein coding potential of exons is usually evaluated by the statistical frequency of nucleotide triplets. However, the regulatory elements are mainly composed of 6 nucleotides. Therefore, there is a crucial need to extract the sequence component information in multiple scales. Based on the analysis above, we first quantitatively determined the length of the window and the number and position of the consensus bases by a chi-square test, then extracted the sequence multiscale component (MSC) features, the position (Pos) and adjacent position relationship (APR) features of the consensus sites, and finally constructed an SVM classifier. Satisfactory results showed that our method achieves a high accuracy in the prediction of splice sites.

MATERIAL AND METHODS

Dataset

To construct a reliable prediction model, we used the publicly available HS^{3}D (Pollastro and Rampone, 2002) splice site dataset (http://www.sci.unisannio.it/docenti/rampone) as the model dataset, which was derived from human genes. The dataset contains 2796 confirmed true donor splice sites, 271,937 pseudo-donor sites, 2880 confirmed true acceptor sites, and 329,374 pseudo-acceptor sites. The redundant information has already been removed. Each splice site sequence has the length of 140 bp. For donor splice sites, the GT dinucleotide is conserved in positions 71 and 72 of the sequences, and for acceptor splice sites, AG is conserved in positions 69 and 70 of the sequences. We selected all of the true splice sites and randomly selected the same number of pseudo-sites (2796 donor sites and 2880 acceptor sites) to construct the training set. In this case, the ratio between the number of true splice sites and that of pseudo-splice sites is 1:1. We used this 1:1 dataset to extract features for further modeling, and constructed another 1:10 (true sites:pseudo-sites) dataset to compare the performance of our model with that of Zhang et al. (2010).

To assess the reproducibility and consistency of our method, we performed an additional evaluation on the NN269 dataset. As a benchmark dataset, the NN269 dataset is a compilation of human splice sites extracted from 269 genes (Reese et al., 1997). It contains 1324 confirmed true donor splice sites, 4922 pseudo-donor sites, 1324 confirmed true acceptor sites, and 5553 pseudo-acceptor sites. Each donor site sequence has the length of 15 bp, and the GT dinucleotide is conserved at positions 8 and 9 of the sequences; each acceptor site sequence has the length of 90 bp, and AG is conserved at positions 69 and 70. For comparison of performance for donor sites, we selected 208 true samples and 782 pseudo samples as the test set and the rest, 1116 true ones and 4140 pseudo-ones, as the training set. For acceptor sites, 208 true samples and 881 pseudo-samples were selected as the test set and the rest as the training set. The selection referred to the references Sonnenburg et al. (2007) and Baten et al. (2006, 2008).

Feature extraction

Chi-square test

Considering donor sites as an example, for 2796 true donor site sequences (positives) and 2796 pseudo-donor site sequences (negatives), we calculated the frequency of different bases (A, T, G, C) at each position (totally, 138 positions with donor site GT as the center, which was defined as the 00 position) in positives/negatives. We then make accordingly a 2 x 4-contingency table (Table 1), and a chi-square value can be calculated for each position by Equation 1. For degrees of freedom $v = 3$, the critical value is 7.81 at the significance level of 0.05.

$$\chi^2 = \frac{S^2}{R_1 \times R_2}[\sum_{i=1}^{4} \frac{a_i{}^2}{C_i} - \frac{R_1{}^2}{S}] \qquad \text{(Equation 1)}$$

If the chi-square test is significant for a certain position, it shows that the base distribution at this position is significantly different between positives and negatives. Making a graph with the position as the abscissa and the corresponding chi-square value as the ordinate, and then judging whether the chi-square value achieves the significance level of 0.05, we can clearly determine the length of the left and right windows and the number and position of consensus bases.

Table 1. Frequency distribution of bases between positives and negatives for a certain position.

Sample	Base				Total
	A	T	C	G	
True	a_1	a_2	a_3	a_4	R_1
False	b_1	b_2	b_3	b_4	R_2
Total	C_1	C_2	C_3	C_4	S

Component feature

As the length of the left and right windows is determined, alternative scale component features of each window are extracted, respectively. Let k be the component scale. For a sequence of length L, the overlap frequency of a string of bases with conjoined R bases $\alpha_1\alpha_2\cdots\alpha_R$ is represented by $f(\alpha_1\alpha_2\cdots\alpha_R)$, where each α_i is one kind of base (i.e., A/T/G/C). The probability of a string of bases $\alpha_1\alpha_2\cdots\alpha_R$ appearing in this sequence is then defined as follows:

$$P(\alpha_1\alpha_2\cdots\alpha_R) = \frac{f(\alpha_1\alpha_2\cdots\alpha_R)}{(L-R+1)} \qquad \text{(Equation 2)}$$

There are 4^k single-scale component (SSC) features to be extracted when the component scale k is set as a single scale. When component scale k is set as a multiscale with a value of $a\sim b$, there are $\sum_{k=a}^{b} 4^k$ MSC features. Because the features are separately selected for the left and right sequences around splice sites, there are 2 x 4^k SSC features and 2 x $\sum_{k=a}^{b} 4^k$ MSC features to be extracted in total for each sequence. Due to the short length of sequences (less than 70 bp), many features are 0 for large k and this is adverse for modeling. Hence, the component scale k cannot be enlarged indefinitely.

Pos feature

The number and position of consensus bases is already determined through the aforementioned chi-square test. Considering donor sites as an example, based on 2796 positives and 2796 negatives, we calculated the frequency of the 4 bases α_i(A,T,C,G) for each conserved site, which was defined as $f^{+}_{x(\alpha_i)}$ and $f^{-}_{x(\alpha_i)}$ ($x = 1,...,L$; L is the number of conserved sites). The frequencies from all conserved sites were made into two 4 x L-probability distribution tables for positives and negatives, respectively. A 4 x L-statistical difference table can be obtained by subtracting elementwise from these two probability distribution tables, with values denoted as $\widehat{f}_{x(\alpha_i)} = f^{+}_{x(\alpha_i)} - f^{-}_{x(\alpha_i)}$. This statistical difference table can reveal the difference between positives and negatives and be directly used for coding and evaluation for consensus sites of training and test samples as follows. By the coding method for a single base, a consensus base can be expressed as a four-dimensional vector according to the order of A, T, G, and C. For instance, the third conserved base site in a certain sequence is T and it can be defined as (0, $\widehat{f}_{3(T)}$, 0, 0), and similarly for other sites. Suppose there are L consensus sites, then 4 x L features can be extracted for each sample.

APR feature

The Pos feature contains the information of a single site, while the APR feature takes the correlative information between two different sites into account. Consider a donor site GT (position 00) as an example, and suppose that the position of the farthest conserved site upstream of the donor site is $-m$, and that downstream is n. Every two consecutive positions between $-m$ and n, i.e., ($-m$, $-m+1$), (-1, 1) ... ($n-1$, n), can then constitute an APR feature resulting in $m+n-1$ APR features. For each pair of positions, the frequencies $f^{+}_{x(\alpha_i)}$ and $f^{-}_{x(\alpha_i)}$ (for $x = 1,...,n$) for positives and negatives of 16 types of dinucleotides (α_i = AA, AT, AC, AG… GG) are calculated. Two 16 x ($m+n-1$)-probability distribution tables of dinucleotides can then be constructed for positives and negatives, respectively. By subtracting corresponding elements from these two distribution tables with the difference denoted as $\widehat{f}_{x(\alpha_i)} = f^{+}_{x(\alpha_i)} - f^{-}_{x(\alpha_i)}$, we finally obtain a statistical difference table for APR features with the size of 16 x ($m+n-1$). This statistical difference table highlights the relevant differences between positives and negatives, and can be directly used for coding and evaluation for consensus sites of training and test samples. For instance, if the $-i$ position of a certain sequence is base A, the $-i+1$ position is T, the difference can be expressed as $\widehat{f}_{i(AT)}$ and the rest are in the similar expressions. Based on the statistical difference table for adjacent bases, there are $m+n-1$ APR features to be extracted for each sample.

SVM

SVM is one of the most important learning machines based on statistical learning theory, which contains support vector classifier and support vector regression (Muller et al., 2001). Based on structural risk minimization instead of empirical risk minimization, SVM can solve the problems of small-sample, non-linear, over-fit, dimension disaster, local minimum point, etc., and also has the strong generalization ability (Vapnik, 1995). The LIBSVM software developed by Chang and Lin (2011) is the concrete realization of SVM. This study adopted support vector classifier (subroutine of LIBSVM) to construct the classifier, where RBF kernel function is selected as kernel function and the grid.py of Python is adopted to optimize the lattice for parameter optimization.

Model evaluation

Sensitivity (Sn), specificity (Sp) and Matthew's correlation coefficients (Mcc) as common measures for determining the performance of a classification model are defined as follows:

$$Sn = \frac{TP}{TP+FN} \times 100 \qquad \text{(Equation 3)}$$

$$Sp = \frac{TN}{TN+FP} \times 100 \qquad \text{(Equation 4)}$$

$$Mcc = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}} \times 100 \qquad \text{(Equation 5)}$$

where *TP*, *FP*, *TN*, and *FN* represent the number of true positives, false positives, true negatives and false negatives, respectively.

Plotting Sn against 1-Sp gives the receiver operator characteristic (ROC) curve (Fawcett, 2003). ROC analysis is an effective and widely used method to assess the performance of a classification method (Baten et al., 2006). Plotting the positive predictive value *PPV* = *TP*/(*FP* + *TP*), i.e., the fraction of correct positive predictions among all positively predicted examples against Sn, one obtains the precision recall curve (PRC) (Davis and Goadrich, 2006). The areas under the ROC and PRC are denoted by AUC and auPRC, respectively. The larger the value of AUC and auPRC, the more accurate the model performance is.

RESULTS AND ANALYSIS

Chi-square independence test of sites

Based on the constructed 1:1 dataset (donor sites 2796/2796 and acceptor sites 2880/2880), the obtained values of the chi-square for independence test for each position of positives and negatives are shown in Figure 1A and B (where donor sites GT and acceptor sites AG are unified as position 00). The chi-square values of all positions exceed the critical value $\chi^2_{(0.05,\,3)} = 7.81$, except for that of position -5 of the donor site. This shows that the distributions of bases {A, T, G, C} between positives and negatives of all positions except for position -5 of the donor site are significantly different, and that the length of the left and right windows for splice sites should be extrapolated. Due to the limit of the length of sequence, we took the upper limit for the original sequence data to extract the component features ($L_{\text{left}} = 70$, $L_{\text{right}} = 68$ for donor sites and $L_{\text{left}} = 68$, $L_{\text{right}} = 70$ for acceptor sites).

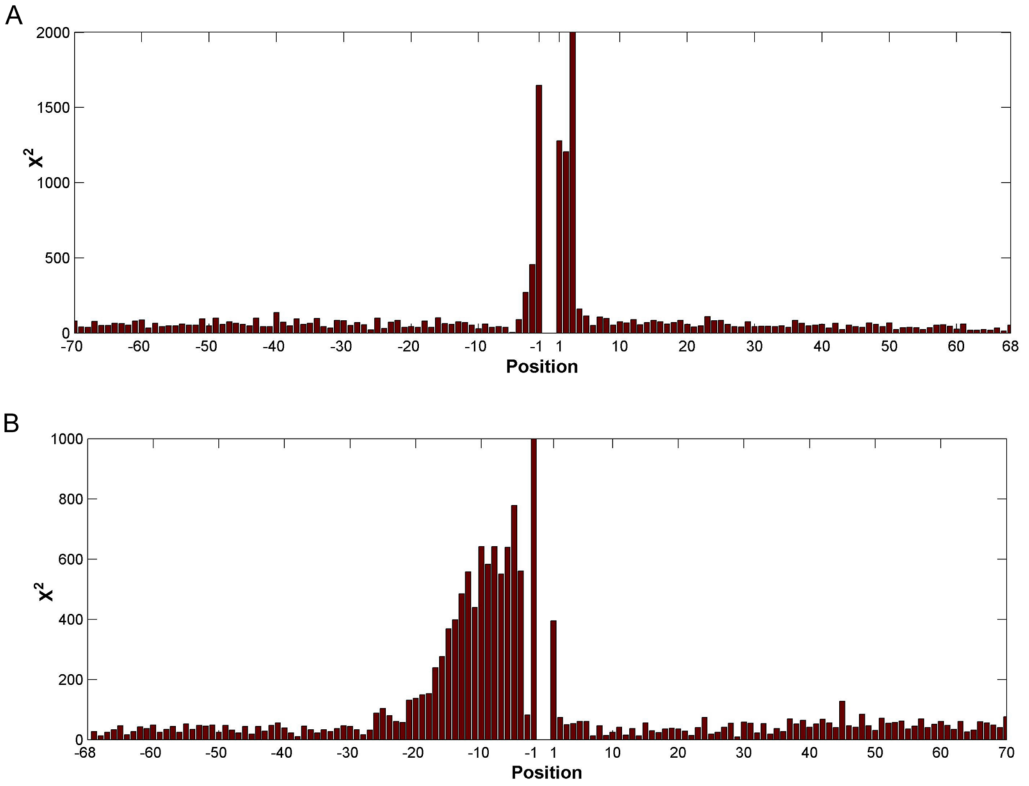

Figure 1. Chi-square test of each position for HS^{3}D dataset for: (**A**) donors and (**B**) acceptors.

Despite the fact that the chi-square test is significant for almost all positions at the individual significance level 0.05, the specific sites with conservatism should show an extremely significant difference at the distribution of bases between positives and negatives. We calculated the average value (AVG) of the chi-square values of all positions that reached the significance level and then took the AVG as the threshold to select the candidate positions to extract the Pos and APR features. For donor sites, the chi-square values of positions -39, -3~+5, 23 were above $AVG_{\text{donors}} = 106.31$; for acceptors sites, the chi-square values of positions -20~+1, 45 were above $AVG_{\text{acceptors}} = 107.20$. However, the positions -39, 23 and 45 are isolated ones and relatively further away from the splice sites. We finally chose the contiguous positions -3~+5 of donor sites and the positions -20~+1 of acceptor sites as the candidate positions to extract the Pos and APR features.

Parameter optimization based on SSC and MSC features

The window size of component features has been determined in the chi-square test ($L_{\text{left}} = 70$, $L_{\text{right}} = 68$ for donor sites and $L_{\text{left}} = 68$, $L_{\text{right}} = 70$ for acceptor sites). Within the range of the windows, we extracted the SSC features and the MSC features for each sequence, and then carried out the 10-fold cross-validation. As can be inferred from Table 2, the best prediction based on SSC features for donor sites achieved an Mcc of 0.805 at $k = 4$, and the best prediction for acceptor sites achieved an Mcc of 0.753 at $k = 3$. In fact, Mcc first improves as *k* increases and then decreased as *k* gets too large. This illustrates that useless information increases as the value of *k* increases and correspondingly produces unfavorable effects for modeling with the SSC features.

Table 2. Ten-fold cross-validation based on different MSC features for HS^{3}D dataset.

k	Donor			Acceptor		
	Sn	Sp	Mcc	Sn	Sp	Mcc
1	78.69	69.53	0.484	82.36	71.39	0.541
2	84.51	80.54	0.651	87.01	83.16	0.702
3	88.98	88.84	0.778	88.96	86.32	0.753
4	90.77	89.70	0.805	88.13	86.22	0.744
5	82.90	81.97	0.649	82.40	85.80	0.682
1~2	88.77	85.23	0.727	90.17	84.06	0.744
2~3	93.46	90.67	0.842	91.53	87.12	0.787
3~4	93.78	92.71	0.865	88.13	87.08	0.752
4~5	85.09	83.91	0.670	83.51	86.01	0.696
1~3	93.67	91.35	0.850	91.88	87.08	0.790
2~4	**94.31**	**92.67**	**0.870**	90.04	87.74	0.778
3~5	86.09	84.51	0.706	83.72	86.53	0.703
1~4	94.06	92.60	0.868	**91.18**	**88.00**	**0.792**
2~5	85.51	85.48	0.710	84.13	86.81	0.710
1~5	86.37	85.27	0.716	84.37	86.91	0.713

Sn = sensitivity; Sp = specificity; Mcc = Matthew's correlation coefficients. Numbers in bold mean the best feature combinations.

The prediction results with features extracted on MSC with *k* values *a*~*b* are generally superior to those based on corresponding SSC with *k* equals *a* or *b*, where a and b assume values from 1 to 5. The best prediction for donor sites using MSC features has an Mcc of 0.870, which is achieved with *k* being 2~4 (there are 336 x 2 features for each sequence); the best prediction for acceptor sites achieved an Mcc of 0.792 with *k* being 1~4 (340 x 2 features for each sequence) (Table 2).

Parameter optimization based on the Pos feature

With AUC as a standard, we further searched the optimal window for Pos features around the consensus sites, which were preliminarily determined by the analysis of the chi-square test to be at positions -3~+5 for donor sites and positions -20~+1 for acceptor sites. This is done as follows. First, we extracted Pos features and constructed a model with the consensus sites as window, and obtained the corresponding prediction results. Second, we selected different sliding windows around the consensus site with two bases as the unit, and then extracted the Pos features and constructed different models to make the prediction. Finally, we selected the optimal model by comparing the performance of all the models. The results are shown in Figure 2. From Figure 2A, we can see that AUC was maximized when the window for donor sites was selected to be at positions of -3~+7, indicating that the positions -3~+7 were the optimal window for donor sites; for acceptor sites, the optimal window was at positions -22~+1 as shown in Figure 2B.

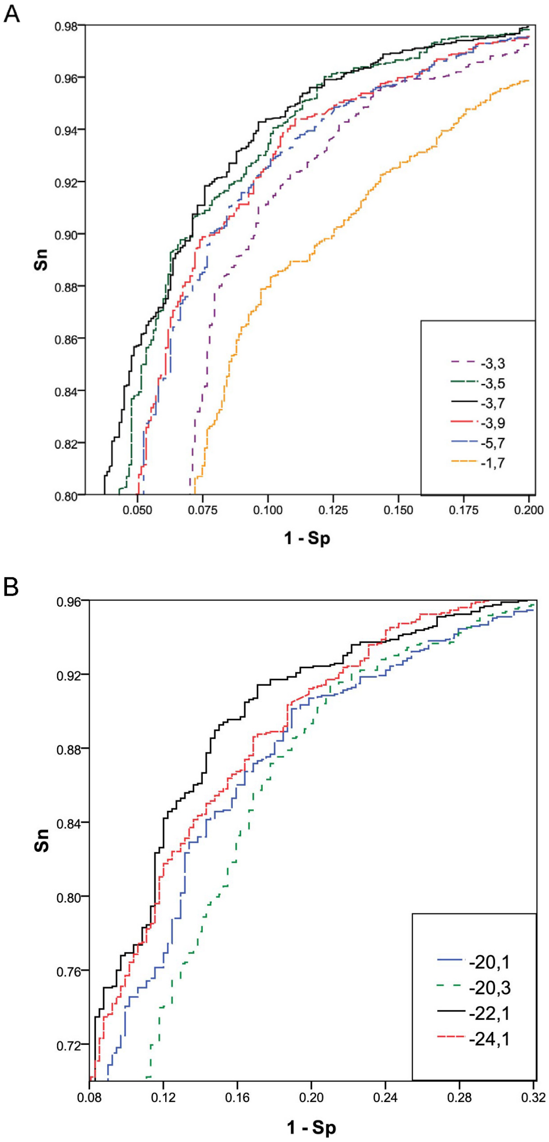

Figure 2. ROC curves of different Pos models for HS^{3}D dataset for: (**A**) donors and (**B**) acceptors. Sn = sensitivity; Sp = specificity.

Parameter optimization based on APR features

How to select the optimal parameters based on APR features was similar to that based on Pos features. With AUC as a standard, the optimal window sizes for APR features was further searched based on the consensus sites (the positions -3~+5 for donor sites and the positions -20~+1 for acceptor sites). With the consensus sites and groups of nearby different zones as the window, the APR features were extracted and the corresponding models were constructed to make prediction. The comparisons of different models are shown in Figure 3. As shown in Figure 3A, we can see that AUC was maximized when the window for donor sites was selected to be at positions of -3~+5, showing that the positions -3~+5 were the optimal window for APR features; for acceptor sites, the optimal window was at positions -22~+3 as shown in Figure 3B.

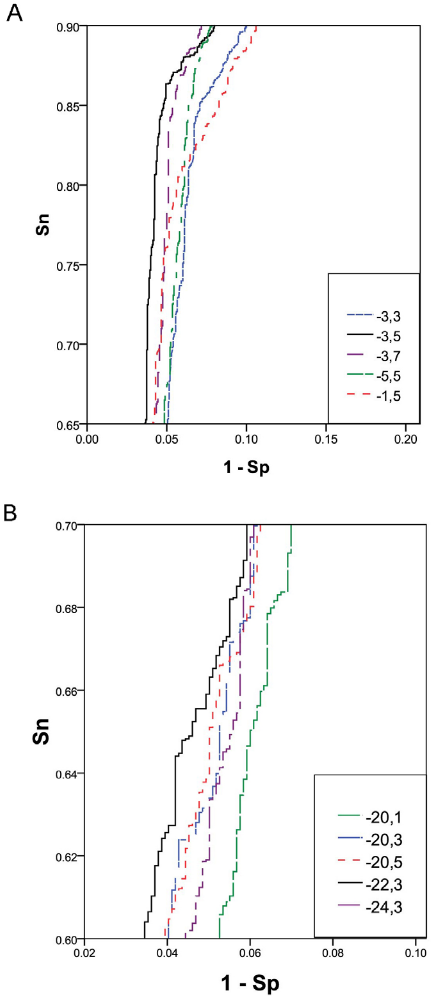

Figure 3. ROC curves of different APR models for HS^{3}D dataset for: (**A**) donor and (**B**) acceptor comparisons of models with integrated multiple features for the 1:1 dataset. Sn = sensitivity; Sp = specificity.

The parameter optimization based on Pos and APR features suggested that the optimal windows determined by the precise search were similar to the conserved region determined by the chi-square test, which indicated that the chi-square independence test could ensure the reliability of the consensus sites.

For the 1:1 dataset, we integrated the aforementioned optimal MSC features (*k* = 2~4 for donors and *k* = 1~4 for acceptors), Pos features (positions -3~+7 for donors, positions -22~+1 for acceptors) and APR features (positions -3~+5 for donors, positions -22~+3 for acceptors) to construct models for predictions of splice sites. The summarized results are shown in Table 3. The Mccs of the prediction results from the models with integrated MSC, Pos and APR (denoted as MSC+Pos+APR) were 0.922 and 0.884 for donors and acceptors, respectively, which were superior to those of the three models with single feature. Moreover, the Mccs of the models with two integrated features randomly selected from MSC, Pos and APR all exceeded those of the corresponding two models with original single feature, illustrating that the integrated features could improve the performance of the models. For donor sites, the optimal model was the one with integrated MSC, Pos and APR, and its Mcc was 0.922, but for acceptor sites, the optimal model was the one with integrated MSC and Pos, which had an Mcc of 0.887.

Table 3. Comparison of the models under 1:1 HS^{3}D dataset.

Methods	Donor			Acceptor		
	Sn	Sp	Mcc	Sn	Sp	Mcc
MSC	94.31	92.67	0.870	91.18	88.00	0.792
Pos	95.60	90.56	0.852	91.53	87.36	0.790
APR	93.02	89.31	0.825	90.94	86.39	0.774
MSC+Pos	96.42	93.85	0.903	95.38	93.26	0.887
MSC+APR	95.92	93.88	0.898	94.41	92.54	0.870
Pos+APR	94.78	90.67	0.855	91.01	88.06	0.791
MSC+Pos+APR	97.21	94.99	0.922	95.17	93.23	0.884
SVM+B	94.31	90.99	0.854	90.90	88.16	0.791
MM1-SVM	93.06	91.31	0.844	90.24	87.57	0.779
MDD/WWAM	93.60	93.60	0.840	93.30	87.70	0.791

SVM+B denotes the prediction method using SVM with a Bayes kernel; MM1-SVM is a prediction method that used probabilistic parameters and SVM classifier (Zhang et al., 2010), and MDD/WWAM denotes the method using maximum dependence decomposition and windowed weight array model (Tavares et al., 2009). MSC = multi-scale component; Pos = position; APR = adjacent position relationship; SVM = support vector machine; Sn = sensitivity; Sp = specificity; Mcc = Matthew's correlation coefficients.

Compared to SVM+B and MM1-SVM from Zhang et al. (2010) and MDD/WWAM from Tavares et al. (2009), our method gave a better performance. For donor sites, our MSC+Pos+APR model gave the best prediction with an Mcc of 0.922 m, which was 0.068 higher than that of SVM+B and 0.082 higher than that of MDD/WWAM. For acceptor sites, our MSC+Pos model gave the best prediction with an Mcc of 0.887, which was 0.096 higher than that of SVM+B and MDD/WWAM and 0.106 higher than that of MM1-SVM (Table 3).

Prediction results for 1:10 data set

Considering the fact that there are many more pseudo-splice sites than true ones in real genome sequences, we constructed the 1:10 (positives:negatives) dataset to verify the practical applicability of the models obtained. Based on the optimal features found in the

1:1 dataset, we extracted the following features for the 1:10 dataset and made the following prediction: MSC (*k* = 2~4), Pos (-3~+7), APR (-3~+5), and MSC (*k* = 1~4), Pos (-22~+1) for donors and acceptors, respectively. The comparison of prediction results between the 1:10 and 1:1 datasets is shown in Figure 4. As shown in Figure 4A, the AUC for donors of the 1:10 dataset was 99.03% while that of the 1:1 dataset was 98.84%, which indicates that the model for donors showed comparable or even better performance in the 1:10 dataset than in the 1:1 dataset. For the acceptor model, the AUCs were 96.43 and 98.32% for the 1:10 and 1:1 datasets (Figure 4B), respectively, indicating that model accuracy decreased marginally for the 1:10 dataset but was still at an excellent level. In summary, our novel models constructed with the integrated features could give a favorable performance in both the 1:10 and 1:1 datasets. This suggests that our method for prediction of splice sites can be widely applied in practice.

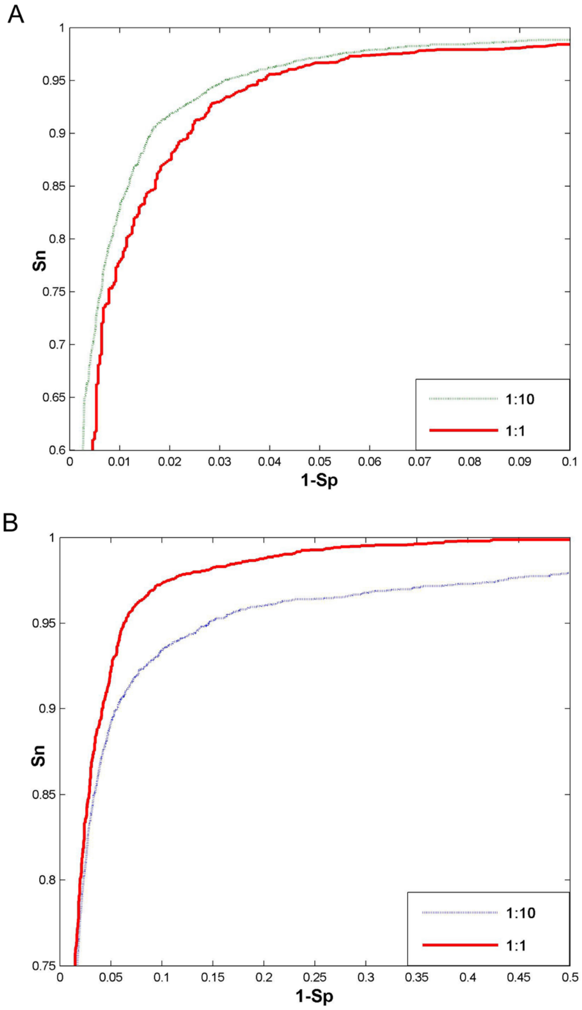

Figure 4. Comparison of results between 1:1 and 1:10 HS^{3}D dataset for: (**A**) donors and (**B**) acceptors. Sn = sensitivity; Sp = specificity.

Zhang et al. (2010) also adopted the methods LVMM2, LVWMM2, OLVWMM2, SVM+B, and MM1-SVM to make predictions for the 1:10 dataset. Among these methods, the OLVWMM2 gives an optimal performance for donors with Sn of 94.17% and Sp of 92.91%, and the LVMM2 shows the best performance for acceptors with corresponding Sn of 91.22% and Sp of 89.70%. In comparison, our MSC+Pos+APR model has Sn of 98.28% and Sp of 92.91% for donor sites. The 4.11% increase in Sn for our model indicates that our MSC+Pos+APR model is significantly better than the OLVWMM2 model. For acceptor sites, the Sn and Sp of our MSC+Pos model were 93.54 and 89.70%, a 2.32% increase for Sn in our model compared to LVMM2.

It can be concluded through the comparisons that the performance of our novel model with integrated MSC features and Pos features is significantly superior to that of available methods in both the 1:1 dataset and 1:10 dataset.

Evaluation on NN269

Here, we applied our method to the evaluation of dataset NN269 in the following 5 steps using the training set. Step 1 = Using the chi-square independence test, the consensus sites were determined to be at positions -3~+4 and -16~+1 for the donor and acceptor sites, respectively (Figure S1). Step 2 = Through contrast screening, the optimal MSC features with *k* = 1~3 and *k* = 1~2 were selected for the donor and acceptor sites, respectively (Figure S2). Step 3 = For extraction of the Pos features, the optimal windows were fixed at positions -3~+4 and -16~+3 for the donors and acceptors, respectively (Figure S3). Step 4 = For extraction of the APR features, the optimal windows were fixed at positions -3~+4 and -16~+1 for the donors and acceptors, respectively (Figure S4). Step 5 = The model with integrated MSC+Pos+APR gave the best performance for the prediction of both donor sites (AUC of 98.58%) and acceptor sites (AUC of 98.40%), as shown in Figure S5.

The optimal models for donors and acceptors were then used for prediction in the test set. Because AUC and auPRC were adopted as the evaluation indices in related published studies (Sonnenburg et al., 2007; Baten et al., 2006, 2008), our results were also translated into those indices for convenience of comparison. Table 4 summarizes the predictive accuracy of our models and other models in terms of the AUC and auPRC for the NN269 dataset. From Table 4, for donor sites, the predictive accuracy AUC and auPRC of our model were as high as 98.93 and 95.11%, higher than that of the available optimal model by 0.43 and 2.25%, respectively; for acceptor sites, the AUC and auPRC of our model were as high as 98.81 and 95.57%, higher than that of the available optimal model by 0.16 and 1.21%, respectively. Hence, our method gave the best predictive performance in the dataset NN269.

Table 4. Comparison of different models on NN269 dataset.

Methods	MC	LIK	WD	WDS	MC-SVM	MM1-SVM	IC-S-SVM	Ours
Donor								
AUC	98.18	98.04	98.50	98.13	97.64	97.90	96.66	98.93
auPRC	92.42	92.65	92.86	92.47	89.57	-	-	95.11
Acceptor								
AUC	96.78	98.19	98.16	98.65	96.74	97.41	96.28	98.81
auPRC	88.41	92.48	92.53	94.36	88.33	-	-	95.57

MC = Markov chain (Durbin et al., 1998); LIK = support vector machine (SVM) using the locality improved kernel (Zien et al., 2000); WD = weighted degree kernel (Rätsch et al., 2004); WDS = weighted degree kernel with shifts (Rätsch et al., 2005); MC-SVM = Markov chain-SVM (Baten et al., 2006); MM1-SVM = first-order Markov model-SVM (Baten et al., 2008); IC-S-SVM = IC Shapiro SVM (Baten et al., 2008); AUC = area under the ROC curve; auPCR = area under the precision recall curve.

DISCUSSION AND CONCLUSIONS

In this paper, we present a method that first determines the window size and the number and position of consensus sites by the chi-square independence test, then integrates the MSC features and the Pos features of consensus sites, and finally applies the SVM classifier to predict the splice sites. This method gave a much better performance than currently available methods reported in the literature in the results of the 10-fold cross-validation for the 1:1 and 1:10 training sets. We also applied this method to the NN269 dataset for further evaluation as a test for independence. The results obtained were also superior to those of the available methods. This demonstrated the stability and superiority of our method. Satisfactory results showed that our method has a high predictive accuracy for splice stes.

For the identification of splice sites and other "signals", we suggest that the "content" features of the left and right sequences in a certain length around the "signal" be extracted first. Earlier studies usually adopted the trial-and-error method to optimize the window sizes. In this paper, we found that the chi-square independence test integrating the sites of the positives and negatives provides a quantitative standard to precisely determine the window size. As for the selection of consensus sites, predecessors have mostly made the information content graphs for the positives and negatives based on Weblogo, which takes the "signal" as center. However, only the unbalanced distribution of bases {A, T, G, C} of a certain site in positives is not enough to determine whether this site is a consensus one or not. This is because the base distribution of this site may also be similarly unbalanced for negatives, such that this site contributes very little in differentiating the positives and negatives. In this study, we developed a chi-square independence test that integrates the sites of the positives and negatives, through which the determination of consensus sites is obviously more reasonable. Furthermore, our method highlights the differences in base distribution for consensus sites between positives and negatives through the statistical difference table. The protein coding potential of an exon is usually evaluated by the statistical frequency of nucleotide triplets ($k = 3$). For the investigation of an object, multiscale is more reasonable than single scale, in theory. The results of this study confirm that MSC features ($1{\sim}k$) are superior to SSC features (k). However, the values of many extracted features are 0 as k becomes relatively large due to the insufficient length of the sequence. This will lead to a decline in model accuracy. The regulatory element motifs generally need to be considered as comprising 6 nucleotides ($k = 6$), and if a mismatch is allowed, then $k = 5{\sim}6$. We postulate that $k = 4$ already satisfies the need of the scale for the regulatory element motifs to a greater degree. The results in the literature also confirm this standpoint.

There is still some possibility for the performance of our methods to be further improved. First, the number of the features generated with MSC features alone is too large. Hence, an effective screening method should be implemented hereafter to prune the useless or inhibiting number of features to improve the accuracy of the models and reduce the time cost for prediction. Second, the splice site prediction conducted in this paper may be validated by a more completely independent test set and by more datasets derived from other species. In particular, we expect that our method could be applied to a whole genome to identify the potential unknown splice sites. Finally, this paper does not involve the prediction of alternative splice sites, which is a more complicated problem.

ACKNOWLEDGMENTS

Research supported by the Science Foundation for Distinguished Young Scholars of Hunan Province, China (#10JJ1005) and the Research Fund for the Doctoral Program of Higher Education of China (#200805370002). We thank Yuan Chen, Zhijun Dai and Wei Zhou (Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization) for their assistance and useful comments.

REFERENCES

Asa BH, Cheng SO and Sonnenburg S (2008). Support vector machines and kernels for computational biology. *PLoS* 4: 1-10.

Baten AK, Chang BC, Halgamuge SK and Li J (2006). Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics* 7 (Suppl 5): S15.

Baten AK, Halgamuge SK, Chang B and Wickramarachchi N (2007). Biological sequence data preprocessing for classification: A case study in splice site identification. *Adv. Neural Netw.* 4492: 1221-1230.

Baten AK, Halgamuge SK and Chang BC (2008). Fast splice site detection using information content and feature reduction. *BMC Bioinformatics* 9 (Suppl 12): S8.

Burset M, Seledtsov IA and Solovyev VV (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28: 4364-4375.

Cai D, Delcher A, Kao B and Kasif S (2000). Modeling splice sites with Bayes networks. *Bioinformatics* 16: 152-158.

Chang CC and Lin CJ (2011). LIBSVM: a library for support vector machines. *Trans. Intell. Syst. Technol.* 2: 278-289.

Chen TM, Lu CC and Li WH (2005). Prediction of splice sites with dependency graphs and their expanded Bayesian networks. *Bioinformatics* 21: 471-482.

Crooks GE, Hon G, Chandonia JM and Brenner SE (2004). WebLogo: a sequence logo generator. *Genome Res.* 14: 1188-1190.

Davis J and Goadrich M (2006). The Relationship Between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML), New York, 233-240.

Durbin R, Eddy S, Krogh A and Mitchison G (1998). Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids Cambridge. Cambridge University Press, Cambridge.

Fawcett T (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Laboratories, Palo Alto.

Kahn AB, Ryan MC, Liu H, Zeeberg BR, et al. (2007). SpliceMiner: a high-throughput database implementation of the NCBI evidence viewer for microarray splice variant analysis. *BMC Bioinformatics* 8: 75.

Mareshi SA, Eslahchi C and Pezechk H (2008). Impact of RNA structure on the prediction of donor and acceptor splice sites. *BMC Bioinformatics* 7: 297.

Muller KR, Mika S and Ratsch G (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* 12: 181-201.

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

Pollastro P and Rampone S (2002). HS^3D, a dataset of *Homo sapiens* splice regions, and its extraction procedure from a major public database. *Int. J. Mod. Phys. C* 13: 1105-1117.

Rätsch G and Sonnenburg S (2004). Accurate Splice Site Detection for Caenorhabditis Elegans. In: Kernel Methods in Computational Biology (Schölkopf KT and Vert JP, eds.). MIT Press, Cambridge.

Rätsch G, Sonnenburg S and Schölkopf B (2005). RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics* 21: i369-i377.

Rätsch G, Sonnenburg S, Srinivasan J, Witte H, et al. (2007). Improving the *Caenorhabditis elegans* genome annotation using machine learning. *PLoS Comput. Biol.* 3: e20.

Reese MG, Eeckman F, Kupl D and Haussler D (1997). Improved splice site detection in Genie. *J. Comp. Biol.* 4: 311-324.

Schneider TD and Stephens RM (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097-6100.

Sonnenburg S, Schweikert G, Philips P, Behr J, 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.

Sun ZX, Sang LJ and Ju LN (2008). Splice site prediction based on splicing information and motif sequences character. *Chin. Sci. Bull.* 53: 2298-2306.

Tavares LG, Lopes HS and Lima CRE (2009). Evaluation of weight matrix models in the splice junction recognition problem. *Bioinform. Biomed. Workshop* 1: 14-19.

Vapnik VN (1995). The Nature of Statistical Learning Theory. Springer Verlag, New York.

Wang K, Ussery DW and Brunak S (2009). Analysis and prediction of gene splice sites in four *Aspergillus* genomes. *Fungal Genet. Biol.* 4: 14-18.

Zhang QW, Peng QK and Xu T (2009). DNA splice site sequences clustering method for conservativeness analysis. *Prog. Nat. Sci.* 19: 511-516.

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

Zhang Y, Chu CH and Chen YX (2006). Splice site prediction using support vector machines with a Beyes kernel. *Expert Syst. Appl.* 30: 73-81.

Zien A, Rätsch G and Mika S (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 16: 799-19.

SUPPLEMENTARY MATERIAL

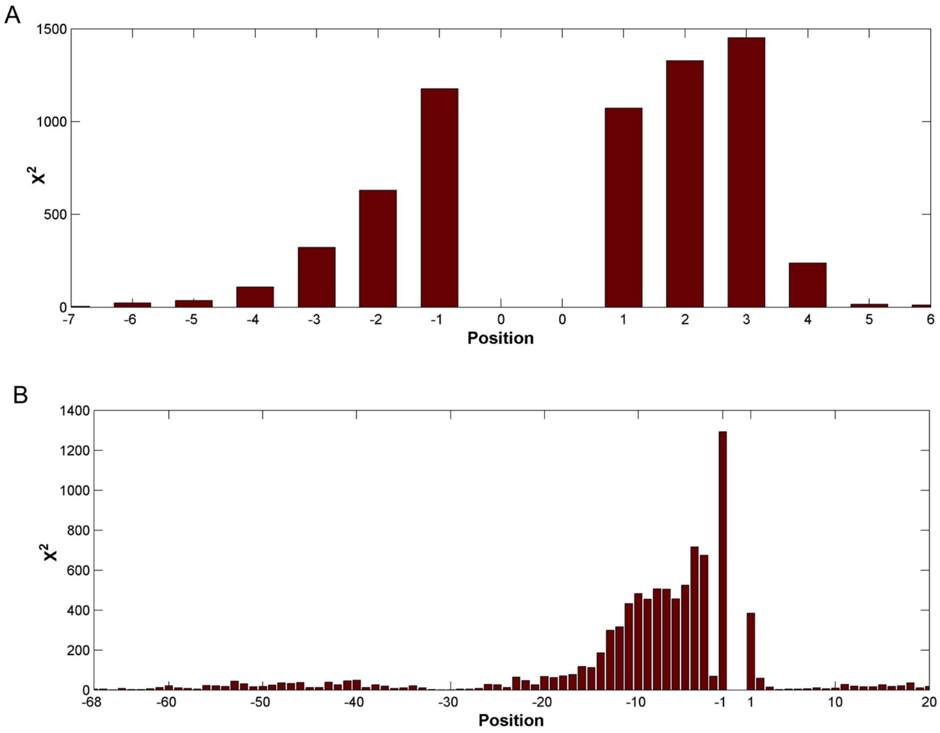

Figure S1. Chi-square test of each position for NN269 dataset for: (**A**) donors and (**B**) acceptors.

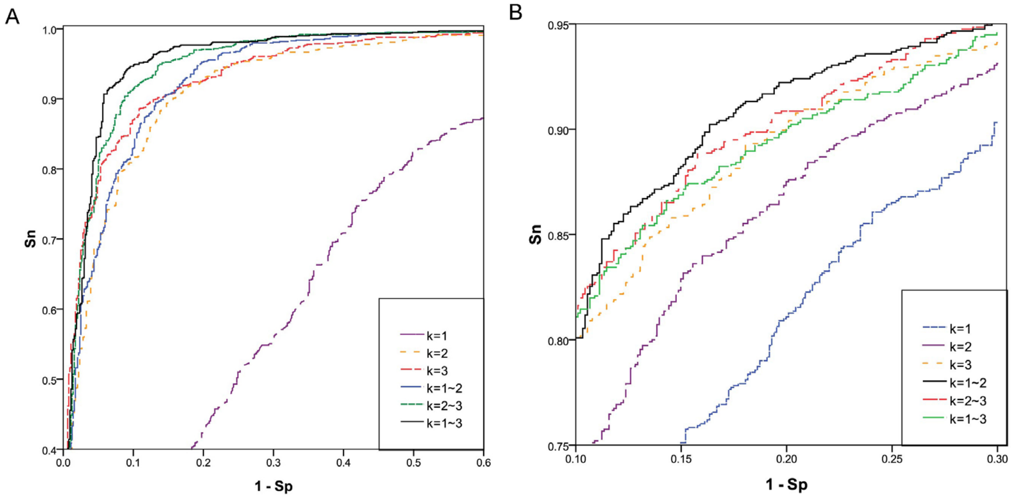

Figure S2. ROC curves of different MSC models for NN269 dataset for: (**A**) donors and (**B**) acceptors. Sn = sensitivity; Sp = specificity.

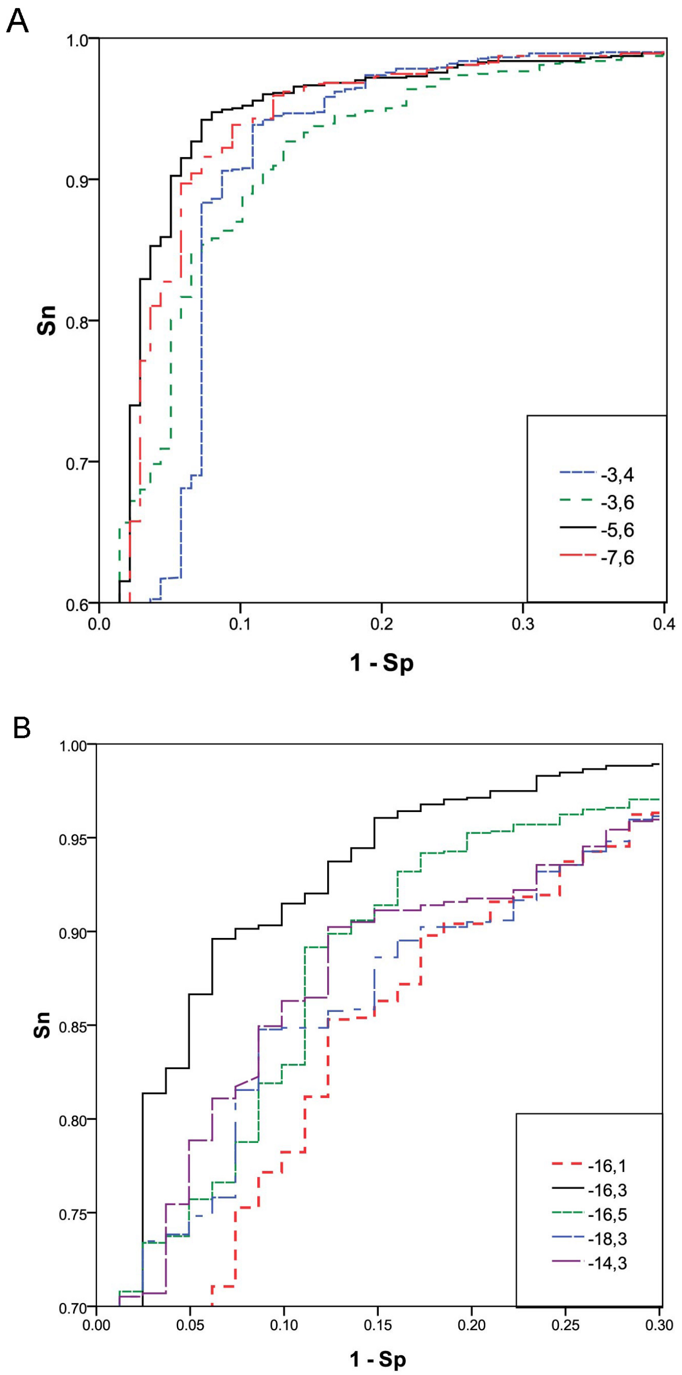

Figure S3. ROC curves of different Pos models for NN269 dataset for: (**A**) donors and (**B**) acceptors. Sn = sensitivity; Sp = specificity.

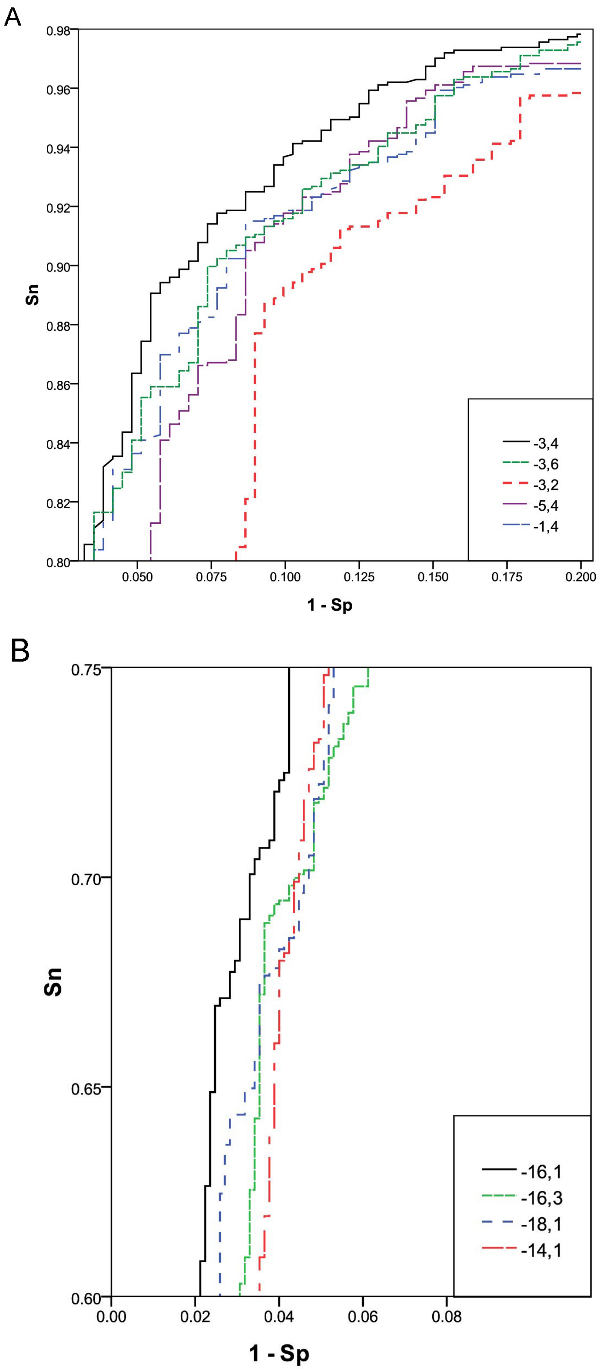

Figure S4. ROC curves of different APR models for NN269 dataset for: (**A**) donors and (**B**) acceptors. Sn = sensitivity; Sp = specificity.

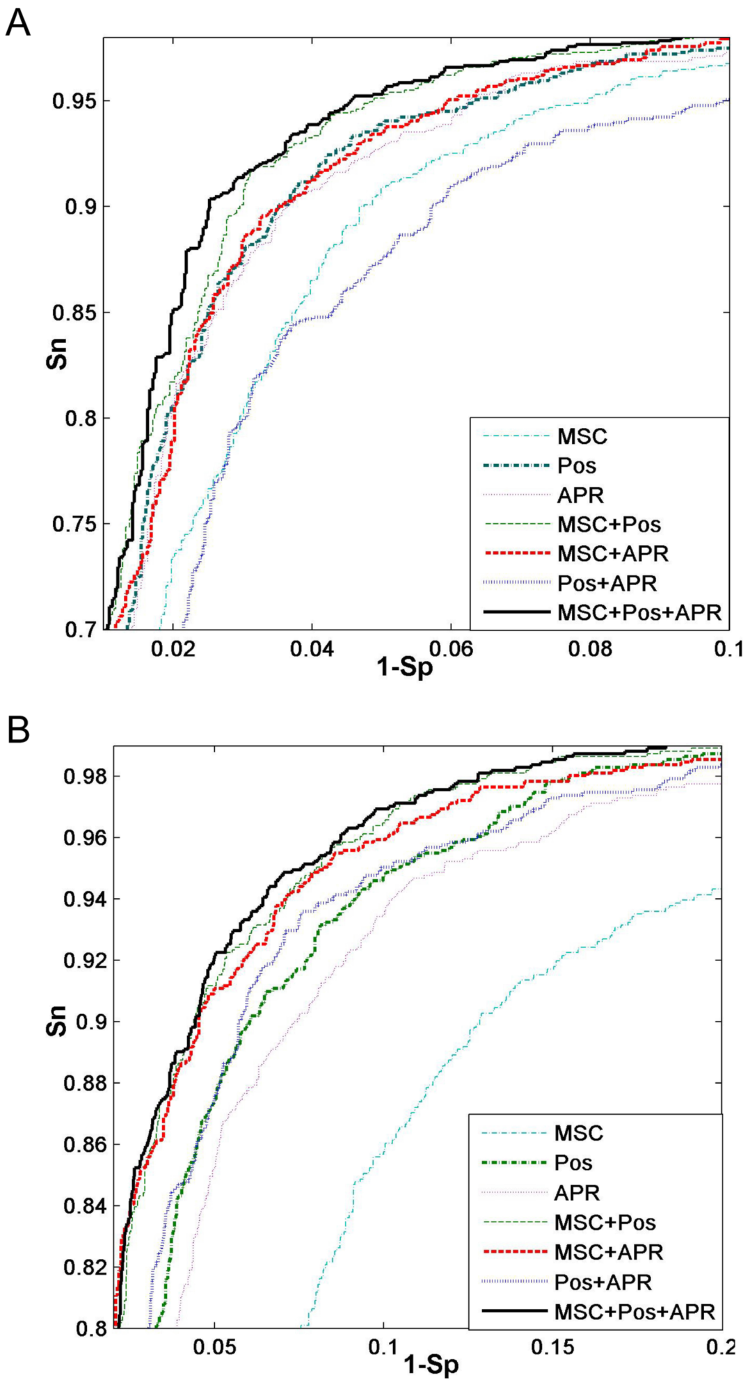

Figure S5. ROC curves of different hybird models for NN269 dataset for: (**A**) donors and (**B**) acceptors. For abbreviations, see legend to Table 3.