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

Jinliang Li, Lifeng Wang, Haiyan Wang, Lianyang Bai, Zheming Yuan

Published Version Information

Citation: Li, J. L., Wang, L. F., Wang, H. Y., Bai, L. Y., & Yuan, Z. M. (2012). High-accuracy splice site prediction based on sequence component and position features. Genetics and Molecular Research, 11(3), 3432-3451.

Digital Object Identifier (DOI): doi:10.4238/2012.September.25.12

Publisher's Link: http://www.geneticsmr.com/articles/1890

High-Accuracy Splice Sites Prediction Based on Sequence Component and Position Features

（running title: High-Accuracy Splice Sites Prediction）

Jinliang Li[1, 2*], Lifeng Wang[1, 2*], Haiyan Wang[3], Lianyang Bai[2], Zheming Yuan[1, 2**]

([1] Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha 410128, China; [2] College of Bio-safety Science and Technology, Hunan Agricultural University, Changsha 410128, China; [3]Department of Statistics, Kansas State University, Manhattan, Kansas 66506, USA）

ABSTRACT Identification of splice sites plays a key role in annotation of genes and hence, the improvement of computational prediction of splice sites with high accuracy has great significance. In this article, we first quantitatively determined the length of window and the number and position of the consensus bases by a Chi-square test, and then extracted the sequence multi-scale component (MSC) features and the position (Pos) and adjacent positions relationship (APR) features of consensus sites. Then we constructed a novel classification model using SVM with above features and applied it to the HS^3D dataset. Compared with the results in current literatures, our method produces a great improvement in the 10-fold 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 test. The obtained results are superior to those in literature, which demonstrates the stability and superiority of this method. Satisfying results show that our method has high accuracy for prediction of splice sites.

Key words: Splice sites prediction，Multi-scale component features，Position features，Adjacent positions relationship features，Support vector machine

1. INTRODUCTION

Owing to the vigorous increase of extensive genomic sequence data, it becomes an urgent demand to improve the efficiency of computational algorithms for gene annotation (Sonnenburg et al., 2007). 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 of the eukaryotic protein coding genes are split genes, which are composed of exons and introns. Introns are the protein non-coding region and are removed by RNA splicing in transcription. The border between an exon and an intron is termed as the splice site. The splice sites consist of the donor site with 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 region. 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 region and the number of them accounts for approximately 1% in total (Burset et al., 2000). There exists a large number of dinucleotide GT and AG in eukaryotic genes, but only 0.1% of them are the real splice sites (Sonnenburg et al., 2007). How to identify whether a dinucleotide GT/AG is a real splice site or not is always one of the most important and challenging tasks in bioin-

[*] These authors contributed equally to this work.

[**] Corresponding author: Zheming Yuan, Fax: +86-731-84673775, Tel:+86-731-84613956 , E-mail: zhmyuan@sina.com

formatics (Sonnenburg et al., 2007; Baten et al., 2008). In this article, we refer real splice sites as positives and false ones as negatives.

In the literature, several statistical models have been constructed for splice sites 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 Bayesian network (BN) (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 (HMM) (Baten et al., 2007; Baten et al., 2008; Asa et al., 2008; Zhang et al., 2009; Zhang et al., 2010) and artificial neural network (ANN) etc. (Reese et al., 1997; Wang et al., 2009) were successively introduced. And a series of special prediction tools were improved for splice sites prediction, such as GeneSplicer (Pcrtea et al., 2001), DGSplicer (Chen et al., 2007), NNSplice (Reese et al., 1997), SpliceMachine (Kahn et al., 2007), etc. Those 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, and protein coding potentiality of exon, etc. The fusion of splicing signals and RNA secondary structure features (Mareshi et al., 2006) 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. Those elements are generally short sequence motifs composed of 6-10 bases, including the enhancer and silencer appeared 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 sites prediction (Sun et al., 2008).

The existing methods of splice site prediction have achieved acceptable level of accuracy. However, 1) It is of prime importance to increase prediction accuracy, especially since the amount of pseudo splice sites in 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） The present 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 those graphs is lack of quantitative criterion 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 potentiality of exon 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 centre is a problem that researchers must take into deep deliberation. 4) The protein coding potentiality of exon 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, this paper first quantitatively determine the length of window and the number and position of the consensus bases by a Chi-square test, then extract the sequence multi-scale component (MSC) features, the position (Pos) and adjacent positions relationship (APR) features of the consensus sites, and finally construct a SVM classifier. Satisfying results show that our method achieves high accuracy for prediction of splice sites.

2. MATERIAL AND METHOD

2.1 Dataset

To construct a reliable prediction model, we used the publicly available HS^3D (Pollastro and Rampone, 2002) splice site dataset (http://www.sci.unisannio.it/do-centi/rampone) as the model dataset, which was derived from human genes. The dataset contains 2796 confirmed real donor splice sites, 271937 pseudo donor sites, 2880 confirmed real acceptor sites, 329374 pseudo acceptor sites. The redundant information has already been removed. Each splice site sequence has the length of 140bp. For donor splice sites, the dinucleotide GT 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 real 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 real splice sites and that of pseudo splice sites is 1:1. We used this 1:1 data set to extract features for further modeling, and constructed another 1:10 (real sites: pseudo sites) data set to compare the performance of our model with that of Zhang *et al.*'s (2010).

In order 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 real donor splice sites, 4922 pseudo donor sites, 1324 confirmed real acceptor sites and 5553 pseudo acceptor sites. Each donor site sequence has the length of 15bp, and the dinucleotide GT is conserved in positions 8 and 9 of the sequences; each acceptor site sequence has the length of 90bp, and the AG is conserved in positions 69 and 70. For comparison of performance for donor sites, we selected 208 real samples and 782 pseudo samples as the test set and the rest 1116 real ones and 4140 pseudo ones as the training set. For acceptor sites, 208 real samples and 881 pseudo samples were selected as the test set and the rest as the training set. The selection consulted with references Baten et al. (2006), Sonnenburg et al. (2007), and Baten et al. (2008).

2.2 Feature extraction

2.2.1 Chi-square test

Consider donor sites as an example, for 2796 real donor sites sequences (positives) and 2796 pseudo donor sites sequences (negatives), we calculated the frequency of different bases (A, T, G, C) appeared 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 made accordingly a 2×4 contingency table (Table 1), and a Chi-square value could be calculated for each position by Equation 1. As the degree of freedom v=3, the critical value is 7.81 at the significant level of 0.05.

Table1. The 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

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

If the Chi-square test is significant for a certain position, it shows that the base distribution on this position has significant difference 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 could clearly determine the length of the left and right windows and the number and position of consensus bases.

2.2.2 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 bases string 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 bases (i.e. A/T/G/C). Then the probability of a bases string $\alpha_1\alpha_2\cdots\alpha_R$ appeared in this sequence is 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 multi-scale with value of *a*~*b*, there are $\sum_{k=a}^{b} 4^k$ multi-scale component (MSC) features. Because the features are selected separately for the left and right sequences around splice sites, there are 2×4^k SSC features and $2\times\sum_{k=a}^{b} 4^k$ MSC features to be extracted in total for each sequence. Due to the short length of sequences (less than 70bp), many features are 0 for large *k* and this is adverse for modeling. Hence the component scale *k* can not be enlarged unlimitedly.

2.2.3 Position (Pos) feature

The number and position of consensus bases have already been determined through aforementioned Chi-square test. Considering donor sites as an example, based on 2796 positives and 2796 negatives, we calculated the frequency of 4 bases α_i(A,T,C,G) for each conserved site respectively, 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 $4\times L$ probability distribution tables for positives and negatives, respectively. Then a $4\times L$ statistical difference table could be obtained by subtracting elementwise of those two probability distribution tables with value denoted as $\widehat{f}_{x(\alpha_i)} = f^{+}_{x(\alpha_i)} - f^{-}_{x(\alpha_i)}$. This statistical difference table could 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 could 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 se-

quence is T and it could be defined as (0, $\widehat{f}_{3(T)}$, 0, 0), and similarly for other sites. Suppose there are *L* consensus sites. Then $4 \times L$ features could be extracted for each sample.

2.2.4 Adjacent Positions Relationship (APR) feature

The Pos feature contains the information of a singe site while the APR feature takes the correlative information between two different sites into account. Consider a donor site GT (position 00) for an example, suppose that the position of the furthest conserved site upstream to the donor site is *-m*, and that in downstream is *n*. Then every two consecutive positions between *-m* and *n*, i.e., *(-m, -m+1), (-1, 1) … (n-1, n),* could constitute a 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 (α_i =AA, AT, AC, AG… GG) are calculated. Then two $16 \times (m+n-1)$ probability distribution tables of dinucleotides could be constructed for positives and negatives, respectively. By subtracting corresponding elements of those 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 $16 \times (m+n-1)$. This statistical difference table highlights the relevant differences between positives and negatives, and could 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 could 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.

2.3 Support vector machine (SVM)

Support vector machine (SVM) is one of the most important learning machines based on statistical learning theory, which contains Support vector classifier (SVC) and Support vector regression (SVR) (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, and local minimum point, etc. and also has the strong generalization ability (Vapnik, 1995). The software LIBSVM developed by Chang and Lin (2011) is the concrete realization of SVM. This paper adopted SVC (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.

2.4 Model evaluation

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

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

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

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

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

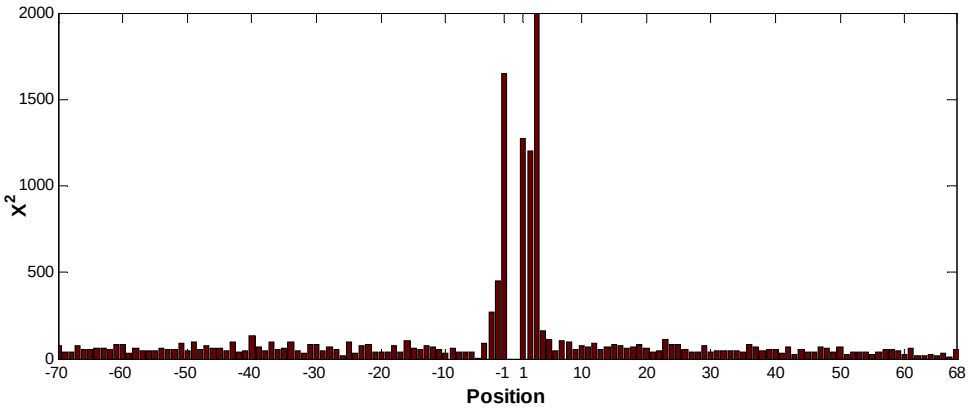
Plotting Sn against 1-Sp gives the Receiver Operator Characteristic Curve (ROC) (Fawcett, 2003). ROC analysis is an effective and widely used method to assess the performance of a classification method (Baten et al., 2001). Plotting the positive predictive value *PPV* = *TP*/ (*FP* + *TP*), i.e. the fraction of correct positive predictions among all positively predicted examples against the Sn; one obtains the Precision Recall Curve (PRC) (Davis and Goadrich, 2006). The area 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.

3. Results and analysis

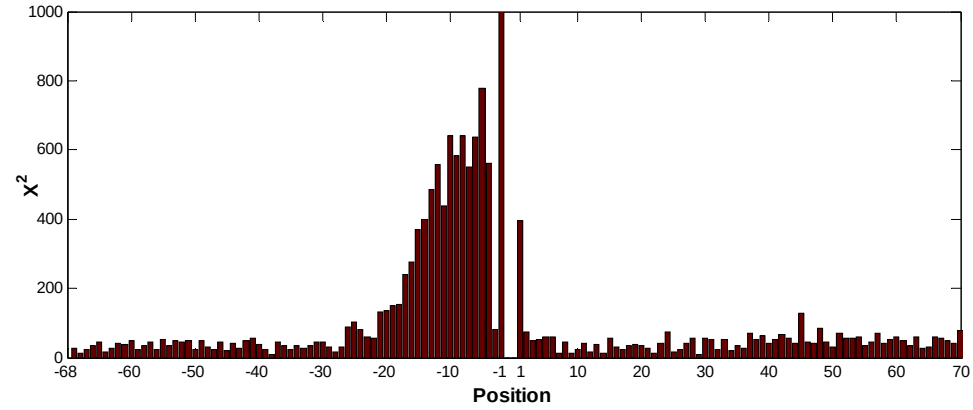
3.1 Chi-square independence test of sites

Based on the constructed 1:1 dataset (donor sites 2796/2796 and acceptor sites 2880/2880), the obtained Chi-square values of independence test for each position of positives and negatives are shown in Figure 1(a) and Figure 1(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 the position -5 of the donor site. This shows that the distribution of bases {A, T, G, C} between positives and negatives of all positions except for the position -5 of the donor site are significantly different, and 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_{left}=70，L_{right}=68 for donor sites and L_{left}=68，L_{right}=70 for acceptor sites).

Despite the fact that the Chi-square test is significant for almost all positions at individual significance level 0.05, the specific sites with conservatism should show the 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 reach 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 are above AVG_{donors}=106.31; for acceptors sites, the Chi-square values of positions -20~+1, 45 are above $AVG_{acceptors}$=107.20. However, the positions -39, 23 and 45 are isolated ones and relatively further away from the splice sites. We finally choose 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.



(a)



(b)

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

3.2 Parameter optimization based on SSC and MSC features

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

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

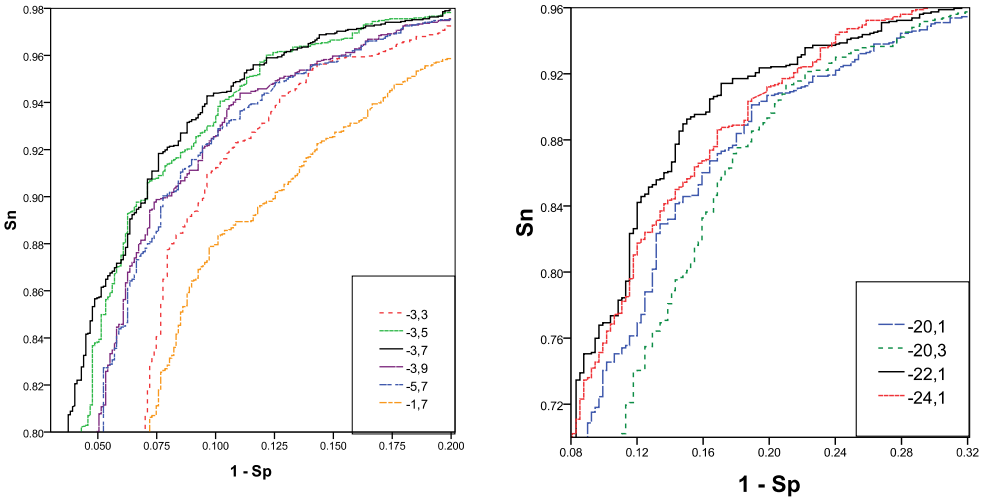
Table 2. Ten–fold cross validation based on different MSC features for HS^{3}D dataset

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

3.3 Parameter optimization based on Pos feature

Now with the AUC as a standard, we further search the optimal window for Pos features around the consensus sites, which were preliminarily determined by the former Chi-square test to be at positions -3~+5 for donor sites and positions -20~+1 for acceptor sites. This is done as follows. Firstly, we extract Pos features and construct a model with the consensus sites as window, and obtain the corresponding prediction results. Secondly, we select different sliding windows around the consensus site with two bases as the unit, and then extract the Pos features and construct different models to make prediction. Finally, we select the optimal model by comparing the performance of all the models. The results are shown in Figure 2. From Figure 2(a), we can see that the AUC is maximized when the window for donor sites is selected to be at positions of -3~+7, indicating that the positions -3~+7 are the optimal window for donor sites; for acceptor sites, the optimal window is at positions -22~+1 as Figure 2(b) shows.



(a) (b)

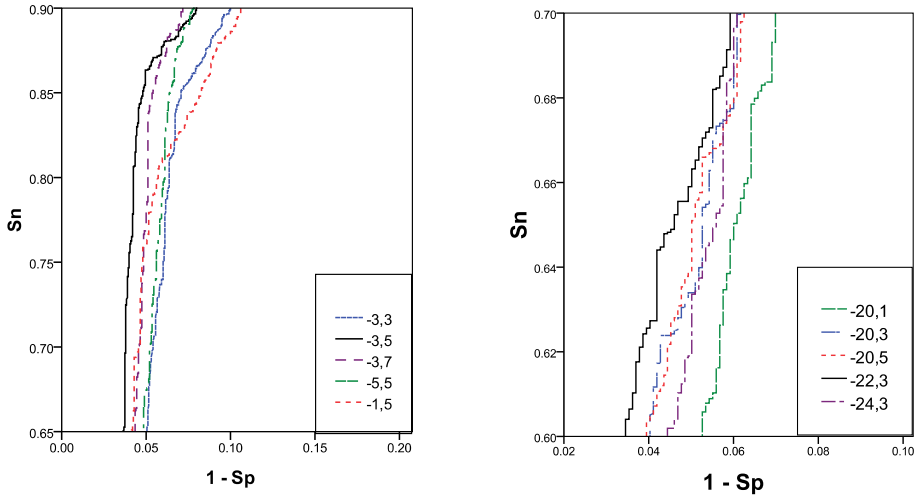
Figure 2. ROC curves of different Pos models for HS^{3}D dataset for: (a) Donors and (b) Acceptors

3.4 Parameter optimization based on APR features

How to select the optimal parameters based on APR features is similar to that based on Pos features. With the AUC as a standard, the optimal length of windows for APR features is further searched based on the consensus sites (the positions -3~+5 for donor sites and the po-

sitions -20~+1 for acceptor sites). With the consensus sites and groups of nearby different zones as the window, the APR features are extracted and the corresponding models are constructed to make prediction. The comparisons of different models are shown in Figure 3. As shown in Figure 3(a), we can see that the AUC is maximized when the window for donor sites is selected to be at positions of -3~+5, showing that the positions -3~+5 are the optimal window for APR features; for acceptor sites, the optimal window is at positions -22~+3 as Figure 3(b) shows.

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



(a) (b)

Figure 3. ROC curves of different APR models for HS^{3}D dataset for: (a) Donors and (b) Acceptors

3.5 Comparison of models with integrated multiple features for the 1:1 dataset

For the 1:1 dataset, we integrate 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 summary results are shown in Table 3. The Mcc's of the prediction results from the models with integrated MSC, Pos and APR (denote as MSC+Pos+APR) are 0.922 and 0.884 for donors and acceptors, respectively, which are superior to those of the three models with single feature. Moreover, the Mcc's of the models with integrated two features randomly selected from MSC, Pos and APR all exceed those of the corresponding two models with original single feature, which illustrates that the integrated features could improve the performance of the models. For donor sites, the optimal model is the one with integrated MSC, Pos and APR, and its Mcc is 0.922; but for acceptor sites, the optimal model is the one with integrated MSC and Pos, which has Mcc 0.887.

Compared with SVM+B and MM1-SVM from Zhang et al. (2010) and MDD/WWAM from Tavares (2009), our method produces a better performance. For donor sites, our MSC+Pos+APR model gives the best prediction with Mcc of 0.922 that is 0.068 higher than that of SVM+B and 0.082 higher than that of MDD/WWAM. For acceptor sites, our

MSC+Pos model gives the best prediction with Mcc of 0.887 that is 0.096 higher than that of SVM+B and MDD/WWAM and 0.106 higher than that of MM1-SVM (Table 3).

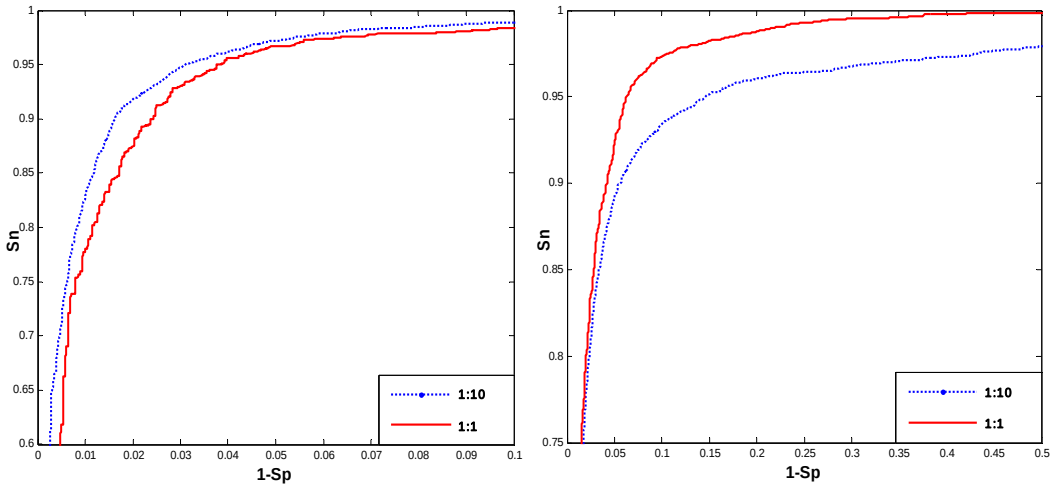
Table 3. Comparison of the models under 1:1 HS^{3}D dataset

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).

3.6 Prediction results for 1:10 data set

Considering the fact that there are many more pseudo splice sites than true ones in real genome sequence, we construct the 1:10 (positives: negatives) dataset to verify the practical applicability of the obtained models. Based on the optimal features found in the 1:1 dataset, we extracted the following features for the 1:10 dataset and make the 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 data sets are shown in Figure 4. As shown in Figure 4 (a), the AUC for donors of the 1:10 dataset is 99.03% while that of the 1:1 dataset is 98.84%, which indicates that the model for donors produces comparable or even better performance in the 1:10 dataset than that in the 1:1 dataset. For the acceptors model, the AUCs are 96.43% and 98.32% for the 1:10 and 1:1 dataset (Figure 4 (b)), respectively, which indicates that the model accuracy has a marginal decrease for the 1:10 dataset but is still at an excellent level. In summary, our novel models constructed with the integrated features could produce favorable performances in both the 1:10 and 1:1 datasets. This suggests that our method for prediction of splice sites can be applied widely in practice.



(a) (b)

Figure 4. Comparison of results between 1:1 and 1:10 HS^{3}D dataset for: (a) donors and (b) acceptors

Zhang *et al*. (2010) also adopted the methods LVMM2, LVWMM2, OLVWMM2, SVM+B and MM1-SVM to make prediction for the 1:10 dataset. Among these methods, the OLVWMM2 has the optimal performance for donors with Sn of 94.17% and Sp of 92.91%, and the LVMM2 produces 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 are 93.54% and 89.70%, a 2.32% increase for Sn of our model compared to the LVMM2.

It can be concluded through the comparisons that the performance of our novel model with integrated multi-scale component features and position features is significantly superior to those of available methods in both the 1:1 dataset and 1:10 dataset.

3.7 Evaluation on NN269

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

The optimal models for donors and acceptors are then used for the prediction for the test set. Because AUC and auPRC were adopted as the evaluation indices in referring literatures (Baten et al., 2006; Sonnenburg et al., 2007; Baten et al., 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 reach 98.93% and 95.11%, higher than that of available optimal model by 0.43% and 2.25%, respectively; for acceptor sites, the AUC and auPRC of our model reach 98.81% and 95.57%, higher than that of available optimal model by 0.16% and 1.21%, respectively. Hence our method produces the best predictive performance in 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: SVMs 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, 2006); 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).

4. DISCUSSIONS AND CONCLUSIONS

In this paper, we presented a method that first determines the length of window and the number and position of consensus sites by the Chi-square independence test, then integrates the MSC features and the position features of consensus sites, and finally applies the SVM classifier to perform prediction of splice sites. This method produces a much better performance than present literatures 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 an independence test. The obtained results are also superior to those of the available methods. This demonstrates the stability and superiority of our method. Satisfying results show that our method has high prediction accuracy for splice sites.

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 at first. Earlier researches usually adopt the trial-and-error method to optimize the length of windows. 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 length of windows. As for the selection of consensus sites, predecessors mostly make the information content graphs for the positives and negatives based on Weblogo which takes the "signal" as center. However, only the imbalance 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 bases distribution of this site may be also similarly imbalanced for negatives such that this site makes very little contribution to differentiate the positives and negatives. This paper 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 of bases distribution for consensus sites between positives and negatives through the statistical difference table. The protein coding potentiality of exon is usually evaluated by the statistical frequency of nucleotide triplets (k=3). For the investigation of an object, multi scale is more reasonable than single scale in theory. The results in this paper confirm that the MSC features (1~k) are superior to SSC features (k). However, the values of many extracted features are 0 as k gets relatively large due to the insufficient length of the sequence. This will lead to a decline of the model accuracy. The regulatory element motifs generally need to be considered to contain 6 nucleotides (k=6) if a mismatch is allowed and then k=5~6. We postulate that k=4 already satisfies the need of the scale for the regulatory element motifs in a greater degree. The results in literature also confirmed this standpoint.

There is still some possibility for the results of our methods to be further improved. Firstly, the number of the features generated with MSC features alone is too large. Hence effective screening method should be improved hereafter to prune the useless or inhibiting number of features to improve the accuracy of the models and reduce the time cost for prediction. Secondly, the splice sites prediction conducted in this paper may be validated by more completely independent test set and by more datasets derived from other species. Particularly, 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

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. This work was supported by a grant from the Science Foundation for Distinguished Young Scholars of Hunan Province, China (No. 10JJ1005), the Research Fund for the Doctoral Program of Higher Education of China (No. 200805370002).

REFERENCES

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

Baten AKMA, Chang, BCH, Halgamuge S K and Li J (2006). Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics*. 7(Suppl 5).

Baten AKMA, Halgamuge SK, Chang B, et al (2007). Biological sequence data preprocessing for classification: A case study in splice site identification. *Advances in Neural Networks*. 4492:1221-1230.

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

Burset M, Seledtsov IA and Solovyev VV (2000). Analysis of canonical and non-canonical splices sites in mammalian genomes. *Nucleic Acids Research*. 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. *Transactions on Intelligent Systems and Technology*. 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 G E, Hon G, Chandonia JM and Brenner SE (2004). Weblogo: A sequence logo generator. *Genome Research*. 14:1188-1190.

Davis J and Goadrich M (2006). The relationship between Precision-Recall and ROC curves. *ICML*. 233-240.

Durbin R, Eddy S, Krogh A and Mitchison G (1998). Biological sequence analysis probabilistic models of proteins and nucleic acids Cambridge, UK, Cambridge University Press;.

Fawcett T (2003). ROC graphs: Notes and practical considerations for data mining researchers. In Technical report hpl-2003-4 HP Laboratories, Palo Alto, CA, USA.

Kahn AB, Ryan CM, Liu HF, 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, Pezechk H, et al (2008). Impact of RNA structure on the prediction of donor and acceptor splice sites. *BMC Bioinformatics*. 7:297.

Muller KR, Mika S, Ratsch G, et al (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*. 12:181-201.

Pcrtea M, Lin XY 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). HS3D, a dataset of Homo sapiens splice regions, and its extraction procedure from a major public database. *International Journal of Modern Physics C*.13:1105-1117.

Rätsch G, Sonnenburg S, Srinivasan J, et al (2007). Improving the caenorhabditis elegans genome annotation using machine learning. *PLoS Computational Biology*. 3:313-322.

Rätsch G and Sonnenburg S (2004). Accurate splice site detection for caenorhabditis elegans. In Kernel Methods in Computational Biology Edited by: B Schölkopf KT, Vert JP. MIT Press.

Rätsch G, Sonnenburg S and Schölkopf B (2005). RASE: Recognition of alternatively spliced exons in C. elegans. *Bioinformatics*. 21:369-377.

Reese MG, Eeckman F, Kupl D and Haussler D (1997). Improved splice site detection in Genie. *Journal of Computational Biology*. 4:311-324.

Schneider TD and Stephens RM (1990). Sequence logos: a new way 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).

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

Sun ZX, Sang LJ, Ju LN, et al (2008). Splice site prediction based on splicing information and motif sequences character. *Chinese Science Bulletin*. 53: 2298-2306.

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

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

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

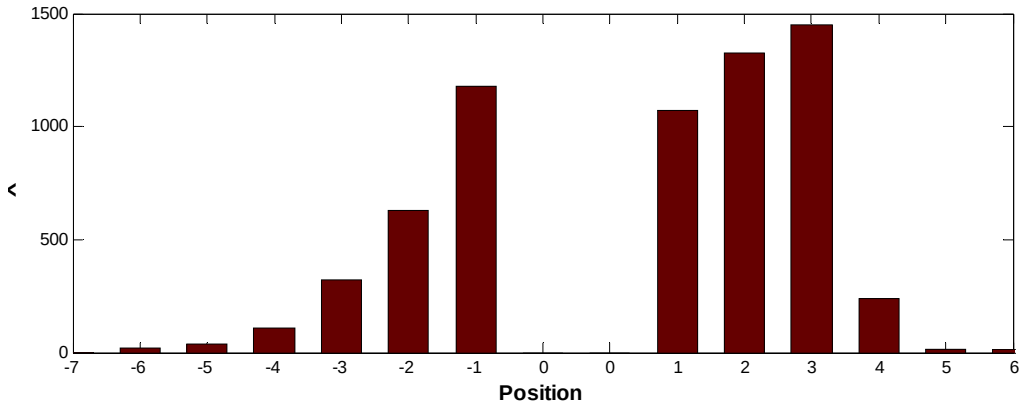
Zhang Y, Chu CH, Chen YX, et al (2006). Splice site prediction using support vector machines with a Beyes kernel. *Expert System with Application*. 30:73-81.

Zhang QW, Peng QK and Xu T (2009). DNA splice site sequences clustering method for conservativeness analysis. *Progress in Natural Science*. 19:511-516.

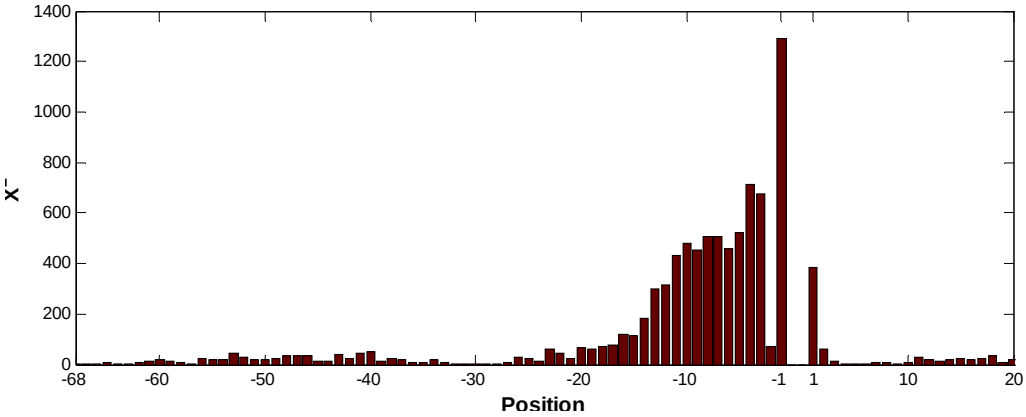
Zhang QW, Peng QK, Zhang Q, et al(2010). Splice sites prediction of human genome using length-variable Markov model and feature selection. *Expert Systems with Applications*. 37:2771-2782.

Zien A, Rätsch G, Mika S, et al (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*. 16:799-807.

APPENDIX

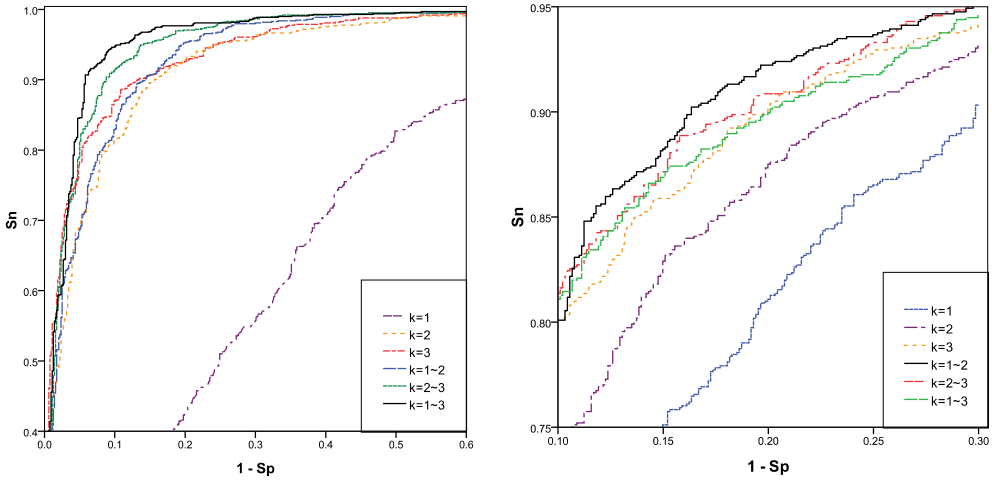


(a)



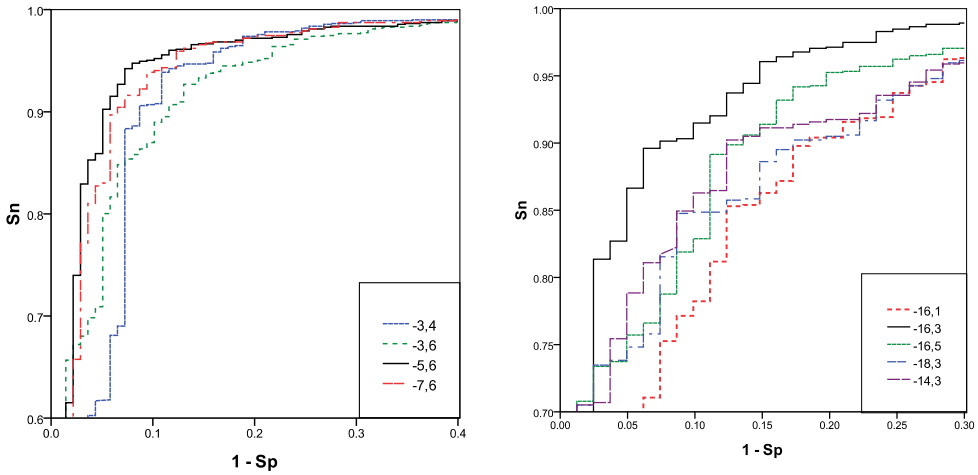
(b)

Figure 1S. Chi-square test of each position for NN269 dataset for: (a) donors and (b) acceptors



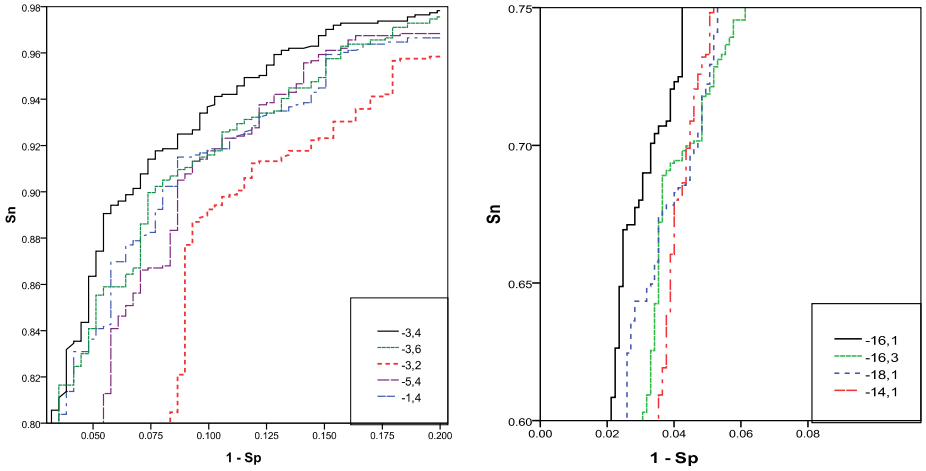
(a) (b)

Figure 2S. ROC curves of different MSC models for NN269 dataset for: (a) Donors and (b) Acceptors



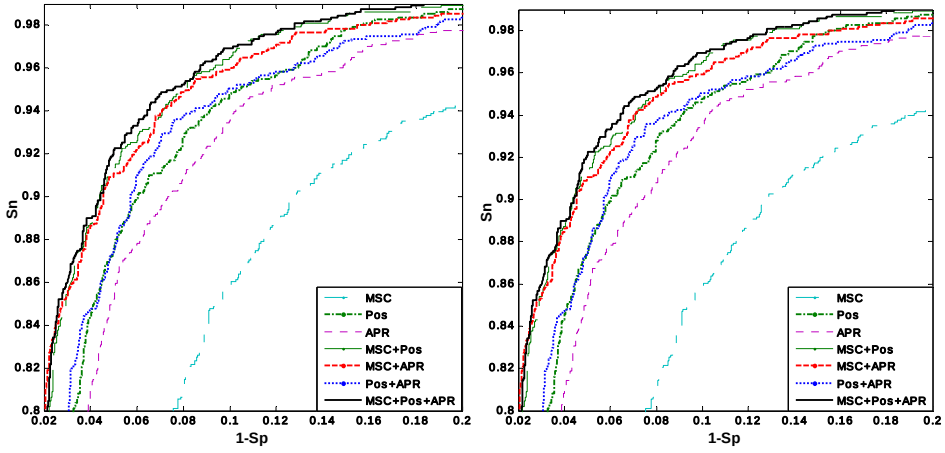
(a) (b)

Figure 3S. ROC curves of different Pos models for NN269 dataset for: (a) Donors and (b) Acceptors



(a) (b)

Figure 4S. ROC curves of different APR models for NN269 dataset for: (a) Donors and (b) Acceptors



(a) (b)

Figure 5S. ROC curves of different hybird models for NN269 dataset for: (a) Donors and (b) Acceptors