Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 623149, 12 pages
http://dx.doi.org/10.1155/2014/623149

Research Article

iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition

Wei Chen,[1,2] Peng-Mian Feng,[3] Hao Lin,[2,4] and Kuo-Chen Chou[1,2,5]

[1] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*
[2] *Gordon Life Science Institute, Boston, MA 02478, USA*
[3] *School of Public Health, Hebei United University, Tangshan 063000, China*
[4] *Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[5] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Wei Chen; wchen@gordonlifescience.org and Hao Lin; hlin@uestc.edu.cn

Received 19 February 2014; Revised 22 April 2014; Accepted 23 April 2014; Published 21 May 2014

Academic Editor: Rita Casadio

In eukaryotic genes, exons are generally interrupted by introns. Accurately removing introns and joining exons together are essential processes in eukaryotic gene expression. With the avalanche of genome sequences generated in the postgenomic age, it is highly desired to develop automated methods for rapid and effective detection of splice sites that play important roles in gene structure annotation and even in RNA splicing. Although a series of computational methods were proposed for splice site identification, most of them neglected the intrinsic local structural properties. In the present study, a predictor called "iSS-PseDNC" was developed for identifying splice sites. In the new predictor, the sequences were formulated by a novel feature-vector called "pseudo dinucleotide composition" (PseDNC) into which six DNA local structural properties were incorporated. It was observed by the rigorous cross-validation tests on two benchmark datasets that the overall success rates achieved by iSS-PseDNC in identifying splice donor site and splice acceptor site were 85.45% and 87.73%, respectively. It is anticipated that iSS-PseDNC may become a useful tool for identifying splice sites and that the six DNA local structural properties described in this paper may provide novel insights for in-depth investigations into the mechanism of RNA splicing.

1. Introduction

In eukaryotic genomes, exons that code for proteins are typically interrupted by introns termed as protein noncoding regions. The borders between exons and introns are called splice sites (Figure 1). A splice site can be located at either the upstream or the downstream part of an intron. For the former, it is called the $5'$ splice site or donor site; for the latter, it is called the $3'$ splice site or acceptor site. The vast majority of the donor and acceptor sites are canonical or regular splice sites that are characterized by the presence of the GT and AG, respectively. During RNA splicing, both the donor and acceptor sites will be recognized by a large macromolecule called spliceosome that is comprised of more than 300 proteins and five small nuclear RNAs (snRNAs U1, U2, U4, U5, and U6) [1]. Once the splice sites are recognized, the spliceosome will remove introns through two sequential transesterification reactions (Figure 1). Removing introns from precursor messenger RNA (pre-mRNA) so that exons can be joined together to form mature mRNA is an essential step of gene expression. Therefore, to better understand the splicing process and mechanism, it is important to accurately detect the splice sites in the genome.

Although biochemical experimental approaches can provide some details about the splice sites, it is both time-consuming and expensive to rely on the biochemical experimental techniques alone. Hence, it is a big challenge and also highly desirable to develop computational methods for timely and effectively identifying the splice sites. In view of this, the present study was initiated in an attempt to develop a computational method for predicting splice sites.

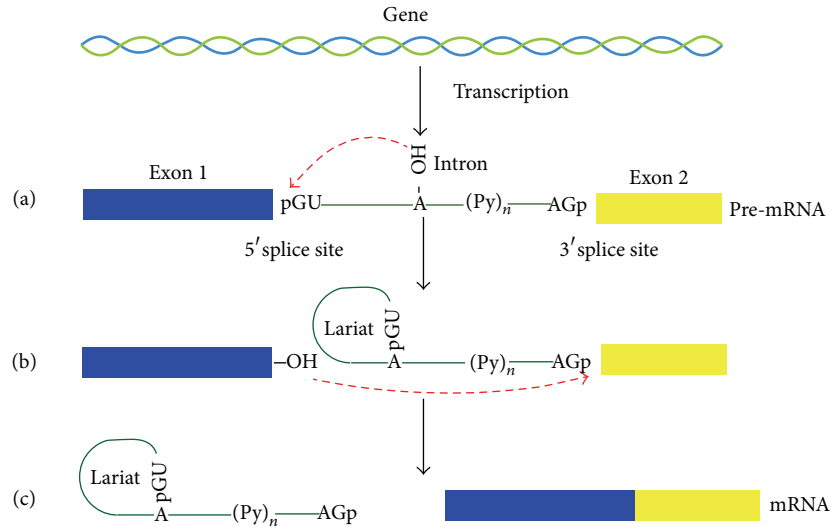

FIGURE 1: A schematic drawing to show the pathways of RNA splicing. (a) The 2′OH of the branchpoint nucleotide within the intron (solid line) carries out a nucleophilic attack at the first nucleotide of the intron at the 5′ splice site (GU) forming the lariat intermediate. (b) The 3′OH of the released 5′ exon then performs a nucleophilic attack at the last nucleotide of the intron at the 3′ splice site (AG). (c) Joining the exons and releasing the intron lariat.

According to a comprehensive review [2] and demonstrated by a series of recent publications [3–9], to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor. Below, let us describe how to deal with these procedures one by one.

2. Materials and Methods

2.1. Benchmark Dataset. The human splice site-containing sequences were obtained from the database HS^{3}D (http://www.sci.unisannio.it/docenti/rampone/), which contained the sequences of exons, introns, and splice regions extracted from GenBank Rel.123. All the splice site-containing sequences in HS^{3}D obey the GT-AG rule; that is, begin with the dinucleotides GT (GU in RNA) and end with the dinucleotides AG, and their lengths are of 140 nucleotides with the splice donor site GT (or acceptor site AG) in the middle positions.

At present, there are 2,796 (2,880) true splice donor (acceptor) site-containing sequences and 271,937 (329,374) false splice donor (acceptor) site-containing sequences in HS^{3}D. To balance the number of the true and false splice site-containing sequences and to avoid the overfitting problem in the model-training processes, we randomly selected out 2,800 false splice donor (acceptor) site-containing sequences from the 271,937 (329,374) false splice donor (acceptor) site-containing sequences.

As pointed out in a comprehensive review [10], there is no need to separate a benchmark dataset into a training dataset and a testing dataset for examining the performance of a prediction method if it is tested by the jackknife test or subsampling cross-validation test.

Finally, we obtained two benchmark datasets, one for the splice donor site-containing sequence, while the other for the splice acceptor, as can be formulated by

$$\begin{aligned} \mathbb{S}_1 &= \mathbb{S}_1^+ \cup \mathbb{S}_1^- \text{ for splice donor,} \\ \mathbb{S}_2 &= \mathbb{S}_2^+ \cup \mathbb{S}_2^- \text{ for splice acceptor,} \end{aligned} \tag{1}$$

where the positive dataset $\mathbb{S}_1^+$ contains 2,796 true splice donor site-containing sequences while the negative dataset $\mathbb{S}_1^-$ contains 2,800 false splice donor site-containing sequences; $\mathbb{S}_2^+$ contains 2,880 true splice acceptor site-containing sequences, while $\mathbb{S}_2^-$ contains 2,800 false splice acceptor site-containing sequences, and the symbol $\cup$ means the union in the set theory. The detailed sequences in the two benchmark datasets $\mathbb{S}_1$ and $\mathbb{S}_2$ are given in Supplementary Information S1 and Supplementary Information S2, respectively; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/623149.

2.2. DNA Sample Formulation. Given a DNA sample $\mathbf{D}$ with L nucleic acid residues, the most straightforward way to express the sample is to use the following sequential model:

$$\mathbf{D} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_L, \tag{2}$$

where R_1 represents the first nucleic acid residue at position 1, R_2 represents the second nucleic acid residue at position 2, and so forth. Although the sequential formulation of (2) contains the complete information of the DNA sample, it

is difficult to be handled for statistical prediction. This is because all the existing operation engines, such as optimization approach [11], covariance discriminant (CD) [12], neural network [13], support vector machine (SVM) [14–16], random forest [17, 18], conditional random field [8], nearest neighbor (NN) [19], K-nearest neighbor (KNN) [20], OET-KNN [21], fuzzy K-nearest neighbor [22–24], ML-KNN algorithm [25], and SLLE algorithm [26], can only handle vector but not sequence samples. Although some sequence-similarity-search-based tools, such as BLAST [27], can be used to directly search for those sequences with high similarity to the query sample, unfortunately, this kind of straightforward and intuitive approach failed to work when the query sample did not have significant similarity to any of the character-known sequences. Therefore, various nonsequential or discrete models to represent the DNA samples were proposed in hopes of establishing some sort of correlation or cluster manner through which the prediction could be more effectively carried out.

The simplest discrete model used to represent a DNA sample is its nucleic acid composition or NAC, as given below:

$$\mathbf{D} = \left[f(\mathrm{A}) \;\; f(\mathrm{C}) \;\; f(\mathrm{G}) \;\; f(\mathrm{T})\right]^{\mathbf{T}}, \tag{3}$$

where $f(\mathrm{A})$, $f(\mathrm{C})$, $f(\mathrm{G})$, and $f(\mathrm{T})$ are the normalized occurrence frequencies of adenine (A), cytosine (C), guanine (G), and thymine (T) in the DNA sequence, respectively; the symbol **T** is the transpose operator. However, as we can see from (3), all its sequence-order information is completely lost if using NAC to represent a DNA sample. Actually, one of the most important but also most difficult problems in computational biology is how to effectively formulate a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information.

One way to cope with such a problem is to represent the DNA segment with the k-tuple nucleotide composition, a vector with 4^k components; that is,

$$\mathbf{D} = \left[f_1^{K\text{-tuple}} \;\; f_2^{K\text{-tuple}} \;\; \cdots \;\; f_i^{K\text{-tuple}} \;\; \cdots \;\; f_{4^k}^{K\text{-tuple}}\right]^{\mathbf{T}}, \tag{4}$$

where $f_i^{K\text{-tuple}}$ is the normalized occurrence frequency of the ith k-tuple nucleotide in the DNA segment. As we can see from (4), the dimension of the vector is

$$4^k = \begin{cases} 64 & k = 3, \\ 256 & k = 4, \\ 1024 & k = 5, \\ 4096 & k = 6, \\ 16384 & k = 7, \\ \vdots & \vdots \end{cases} \tag{5}$$

indicating that by increasing the value of k, although the coverage scope of sequence order will be gradually increased, the dimension of the vector **D** will be rapidly increased as well. This will cause the high-dimension disaster [28] as reflected by the following disadvantages: (i) the overfitting problem that will make the predictor with a serious bias and extremely low capacity for generalization; (ii) the information redundancy or noise that will bring about the error of misrepresentation resulting in very poor prediction accuracy; and (iii) unnecessarily increasing the computational time.

To avoid the high-dimension disaster, here, the dinucleotide composition (DNC) was used to formulate the DNA sample, as given by

$$\begin{aligned} \mathbf{D} &= \left[f_1^{2\text{-tuple}} \;\; f_2^{2\text{-tuple}} \;\; \cdots \;\; f_i^{2\text{-tuple}} \;\; \cdots \;\; f_{16}^{2\text{-tuple}}\right]^{\mathbf{T}} \\ &= \left[f(\mathrm{AA}) \;\; f(\mathrm{AC}) \;\; f(\mathrm{AG}) \;\; f(\mathrm{AT}) \;\; \cdots \;\; f(\mathrm{TT})\right]^{\mathbf{T}}, \end{aligned} \tag{6}$$

where $f_1^{2\text{-tuple}} = f(\mathrm{AA})$ is the normalized occurrence frequency of AA in the DNA sequence, $f_2^{2\text{-tuple}} = f(\mathrm{AC})$ is that of AC, $f_3^{2\text{-tuple}} = f(\mathrm{AG})$ is that of AG, and so forth. By doing so, we can only incorporate the local sequence-order information between the most contiguous nucleotides, but none of the global or long-range sequence-order information can be reflected.

Actually, similar problem also occurred in computational proteomics, where, in order to incorporate the global or long-range sequence-order information for proteins, the pseudo amino acid composition [29] or Chou's PseAAC [30] was proposed. Since the concept of PseAAC was proposed in 2001 [29], it has been penetrating into almost all the fields of protein attribute predictions (see, e.g., [31–73]). Because it has been widely used, recently two types of open access software, called "PseAAC-Builder" [51] and "propy" [74], were established for generating various modes of PseAAC.

Encouraged by the successes of introducing the PseAAC approach into computational proteomics, Chen et al. [4] proposed the "pseudo dinucleotide composition" or PseDNC to identify recombination spots of DNA. The formulation of PseDNC is given by

$$\mathbf{D}_{\mathrm{PseDNC}} = \left[d_1 \;\; d_2 \;\; \cdots \;\; d_{16} \;\; d_{16+1} \;\; \cdots \;\; d_{16+\lambda}\right]^{\mathbf{T}}, \tag{7}$$

where

$$d_u = \begin{cases} \dfrac{f_u^{2\text{-tuple}}}{\sum_{i=1}^{16} f_i^{2\text{-tuple}} + w\sum_{j=1}^{\lambda}\theta_j}, & 1 \le u \le 16, \\ \dfrac{w\theta_u}{\sum_{i=1}^{16} f_i^{2\text{-tuple}} + w\sum_{j=1}^{\lambda}\theta_j}, & (16+1) \le u \le (16+\lambda), \end{cases} \tag{8}$$

where $f_i^{2\text{-tuple}}$ $(i = 1, 2, \ldots, 16)$ have the same meaning as those in (6), while θ_j is the jth tire correlation factor that reflects the sequence-order correlation between all the jth most contiguous dinucleotides along a DNA sequence (see Figure 2), as formulated by

$$\theta_j = \frac{1}{L-j-1}\sum_{i=1}^{L-j-1}\Theta\left(R_iR_{i+1}; R_{i+j}R_{i+1+j}\right) \quad \left(j = 1, 2, \ldots, \lambda < L\right). \tag{9}$$

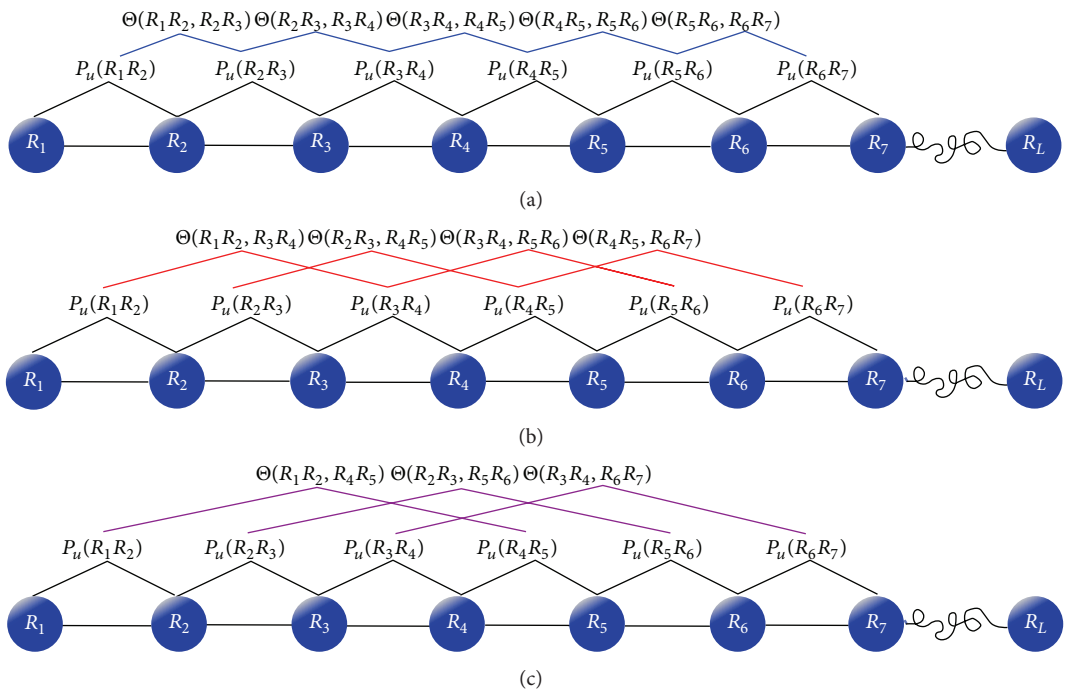

Figure 2: A schematic illustration to show the correlations of dinucleotides along a DNA sequence. (a) The first-tier correlation reflects the sequence-order mode between all the most contiguous dinucleotides. (b) The second-tier correlation reflects the sequence-order mode between all the second-most contiguous dinucleotides. (c) The third-tier correlation reflects the sequence-order mode between all the third-most contiguous dinucleotides.

In the above two equations, λ is the number of the total counted ranks or tiers of the correlations along a DNA sequence, and w is the weight factor. Their concrete values as well as the final value for k will be further discussed later. The correlation function $\Theta(R_iR_{i+1}; R_{i+j}R_{i+1+j})$ in (9) is defined by

$$\Theta\left(R_iR_{i+1}; R_{i+j}R_{i+1+j}\right) = \frac{1}{\mu}\sum_{\nu=1}^{\mu}\left[P_\nu(R_iR_{i+1}) - P_\nu(R_{i+j}R_{i+1+j})\right]^2, \tag{10}$$

where μ is the number of local DNA structural properties considered that is equal to 6 in the current study as will be explained below, $P_\nu(R_iR_{i+1})$ is the numerical value of the νth ($\nu = 1, 2, \ldots, \mu$) DNA local structural property for the dinucleotide R_iR_{i+1} at position i, and $P_\nu(R_{i+j}R_{i+1+j})$ is the corresponding value for the dinucleotide $R_{i+j}R_{i+1+j}$ at position $i + j$, as will be given below.

2.3. DNA Local Structural Property Parameters. A lot of evidences have shown that DNA local structural properties play important roles in many biological processes, such as protein-DNA interactions [75], formation of chromosomes [76], and meiotic recombination [4]. Generally speaking, the spatial arrangements of two successive base pairs can be characterized by six parameters, of which three are the local translational ones and the other three are the local angular ones (Figure 3), as formulated by

$$\text{translational} = \begin{cases} \text{slide}, \\ \text{shift}, \\ \text{rise}, \end{cases} \qquad \text{angular} = \begin{cases} \text{roll}, \\ \text{tilt}, \\ \text{twist}. \end{cases} \tag{11}$$

The six structural parameters of dinucleotides have been calculated by Goñi et al. [75] based on the long atomistic molecular dynamics (MD) simulations in water, and their concrete values are given in Table 1, which will be used to calculate the global or long-range sequence-order effects for the DNA sequences via (9) and (10).

Note that before substituting the values of physicochemical property into (10), they were all subjected to a standard conversion as described by the following equation:

$$P_\nu\left(R_iR_{i+1}\right) = \frac{P_\nu^0\left(R_iR_{i+1}\right) - \left\langle P_\nu^0\left(R_iR_{i+1}\right)\right\rangle}{\mathrm{SD}\left\langle P_\nu^0\left(R_iR_{i+1}\right)\right\rangle}, \tag{12}$$

where the symbols $\langle\rangle$ mean taking the average of the quantity therein over the 16 different combinations of A, C, G, T for R_iR_{i+1} and SD means the corresponding standard deviation [10]. The converted values obtained by (12) will have a zero mean value over the 16 different dinucleotides and will remain unchanged if going through the same conversion procedure again. Listed in Table 2 are the values of $P_\nu(R_iR_{i+1})$ ($\nu = 1, 2, \ldots, 6$) obtained via the standard conversion of (12) from those of Table 1.

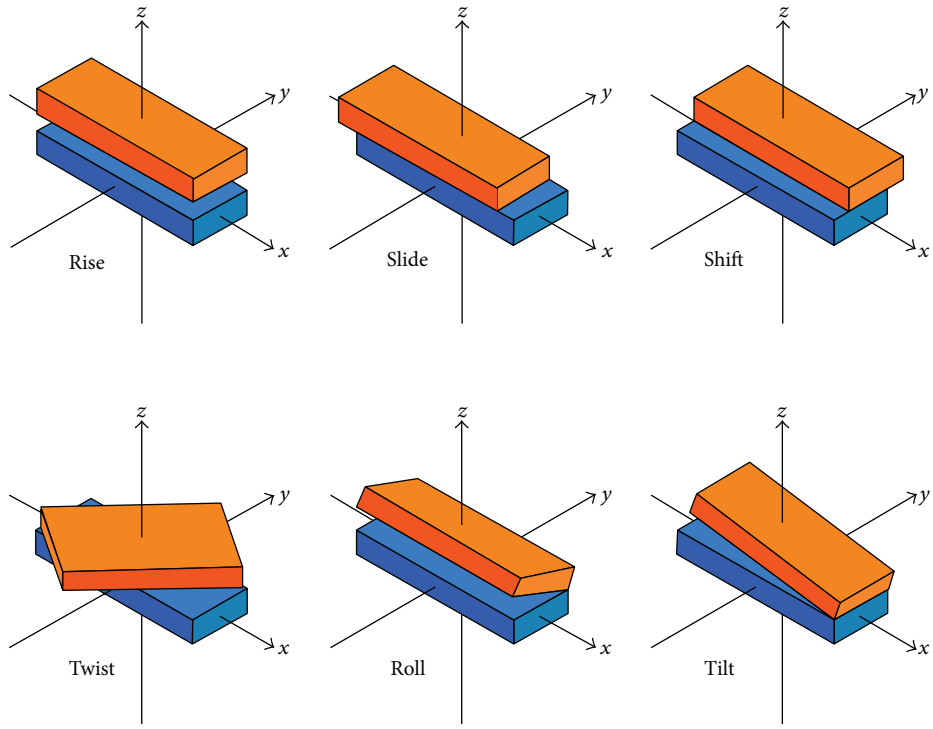

Figure 3: A schematic drawing to illustrate the six spatial arrangements between two neighboring base pairs in DNA. Of the six panels, three are for the local translational arrangements and the other three are for the local angular ones [6].

Table 1: The original values for the six DNA dinucleotide physical structures.

Dinucleotide	Physical structures[a]					
	$P_1(R_iR_{i+1})$	$P_2(R_iR_{i+1})$	$P_3(R_iR_{i+1})$	$P_4(R_iR_{i+1})$	$P_5(R_iR_{i+1})$	$P_6(R_iR_{i+1})$
AA	0.026	0.038	0.020	1.69	2.26	7.65
AC	0.036	0.038	0.023	1.32	3.03	8.93
AG	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA	0.016	0.025	0.017	1.07	1.78	6.38
CC	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2.00	6.23
CT	0.031	0.037	0.019	1.46	2.03	7.08
GA	0.025	0.038	0.020	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.20	2.61	9.53
GG	0.026	0.042	0.019	1.43	1.65	8.04
GT	0.036	0.038	0.023	1.32	3.03	8.93
TA	0.017	0.018	0.016	0.72	1.20	6.23
TC	0.025	0.038	0.020	1.32	1.93	8.56
TG	0.016	0.025	0.017	1.07	1.78	6.38
TT	0.026	0.038	0.020	1.69	2.26	7.65

[a]In this table, the following symbols were used to represent the six physical structures of dinucleotide: P_1 for "twist", P_2 for "tilt", P_3 for "roll", P_4 for "shift", P_5 for "slide", and P_6 for "rise". The data was obtained from [75].

2.4. Support Vector Machine (SVM). Support vector machine (SVM) is an effective method for supervised pattern recognition and has been widely used in the realm of bioinformatics [4, 14, 77, 78]. The basic idea of SVM is to transform the data into a high dimensional feature space and then determine the optimal separating hyperplane. A brief introduction about the formulation of SVM has been given in [14]. In this study, the SVM implementation was based on the freely available package LIBSVM 2.84 written by Chang and Lin [79], which can be downloaded

Table 2: The normalized values for the six DNA dinucleotide physical structures.

Dinucleotide	Physical structures[a]					
	$P_1(R_iR_{i+1})$	$P_2(R_iR_{i+1})$	$P_3(R_iR_{i+1})$	$P_4(R_iR_{i+1})$	$P_5(R_iR_{i+1})$	$P_6(R_iR_{i+1})$
AA	0.06	0.5	0.27	1.59	0.11	−0.11
AC	1.50	0.50	0.80	0.13	1.29	1.04
AG	0.78	0.36	0.09	0.68	−0.24	−0.62
AT	1.07	0.22	0.62	−1.02	2.51	1.17
CA	−1.38	−1.36	−0.27	−0.86	−0.62	−1.25
CC	0.06	1.08	0.09	0.56	−0.82	0.24
CG	−1.66	−1.22	−0.44	−0.82	−0.29	−1.39
CT	0.78	0.36	0.09	0.68	−0.24	−0.62
GA	−0.08	0.5	0.27	0.13	−0.39	0.71
GC	−0.08	0.22	1.33	−0.35	0.65	1.59
GG	0.06	1.08	0.09	0.56	−0.82	0.24
GT	1.50	0.50	0.80	0.13	1.29	1.04
TA	−1.23	−2.37	−0.44	−2.24	−1.51	−1.39
TC	−0.08	0.5	0.27	0.13	−0.39	0.71
TG	−1.38	−1.36	−0.27	−0.86	−0.62	−1.25
TT	0.06	0.5	0.27	1.59	0.11	−0.11

[a]See footnote a of Table 1 for further explanation.

from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Because of its effectiveness and speed in training process, the radial basis kernel function (RBF) was used to obtain the best classification hyperplane. The regularization parameter C and the kernel width parameter γ were tuned via the grid search method in the 10-fold cross-validation.

The predictor obtained via the above procedures is called iSS-PseDNC, where "i" stands for "identifying," "SS" for "splice site," "Pse" for "pseudo," "D" for "di," "N" for "nucleotide," and "C" for "composition."

2.5. Criteria for Performance Evaluation. To provide a more intuitive and easier-to-understand method to measure the prediction quality, the following set of four metrics based on the formulation used by Chou [80] in studying signal peptide prediction was adopted. According to Chou's formulation, the sensitivity (Sn), specificity (Sp), overall accuracy (Acc), and Matthew's correlation coefficient (MCC) can be expressed as follows [4, 7–9]:

$$\begin{aligned}
\mathrm{Sn} &= 1 - \frac{N_-^+}{N^+},\\
\mathrm{Sp} &= 1 - \frac{N_+^-}{N^-},\\
\mathrm{Acc} &= 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-},\\
\mathrm{MCC} &= \frac{1 - \left(\left(N_-^+/N^+\right) + \left(N_+^-/N^-\right)\right)}{\sqrt{\left(1 + \left(N_+^- - N_-^+\right)/N^+\right)\left(1 + \left(N_-^+ - N_+^-\right)/N^-\right)}},
\end{aligned} \tag{13}$$

where N^+ is the total number of the true splice site-containing sequences investigated, while N_-^+ is the number of true splice site-containing sequences incorrectly predicted as the false splice site-containing sequences; N^- is the total number of the false splice site-containing sequences investigated, while N_+^- is the number of the false splice site-containing sequences incorrectly predicted as true splice site-containing sequences. From (13), we can easily see the following. When $N_-^+ = 0$ meaning that none of the true splice site-containing sequences was incorrectly predicted to be a false splice site-containing sequence, we have the sensitivity Sn = 1. When $N_-^+ = N^+$ meaning that all the true splice site-containing sequences were incorrectly predicted to be the false splice site-containing sequences, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning that none of the false splice site-containing sequences was incorrectly predicted to be a true splice site-containing sequence, we have the specificity Sp = 1, whereas when $N_+^- = N^-$ meaning that all the false splice site-containing sequences were incorrectly predicted to be the true splice site-containing sequences, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of the true splice site-containing sequences and none of the false splice site-containing sequences were incorrectly predicted, we have the overall accuracy Acc = 1 and Mathew's correlation coefficient MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the false splice site-containing sequences and all the true splice site-containing sequences were incorrectly predicted, we have Acc = 0 and MCC = −1, whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have Acc = 0.5 and MCC = 0 meaning no better than random prediction. As we can see from the above discussion based on (13), the meanings of the four metrics have become much more intuitive and easier to understand than the conventional

formulation often used in the literature, particularly for Mathew's correlation coefficient, which is usually used for measuring the quality of binary (two-class) classifications as in the case of the current study. However, it is instructive to point out that the set of the metrics in (13) is valid only for the single-label systems. For the multilabel systems whose existence has become more frequent in system biology [81–83] and system medicine [24, 84], a completely different set of metrics as defined in [25] is needed.

3. Results and Discussions

3.1. Graphic Profiles of True and False Splice Site-Containing Sequences. It has been reported that the DNA local structural properties, that is, angular parameters (twist, tilt, and roll) and translational parameters (shift, slide, and rise), play important roles in prokaryotic transcription initiation, protein-DNA interactions, and meiotic recombination [4, 75, 76, 85]. Accordingly, it is quite natural to ask whether these DNA structural properties may also play some role in regulating RNA splicing. Here, let us use the graphic approach to address this question. This is because using graphical approaches to study biological problems can provide an intuitive picture or useful insights for helping in analyzing complicated relations in these systems [30], as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions [86–89], inhibition of HIV-1 reverse transcriptase [90–93], inhibition kinetics of processive nucleic acid polymerases and nucleases [94], protein folding kinetics [95], drug metabolism systems [96], protein sequence evolutionary analysis [97], protein remote homology detection [5], and using Wenxiang diagram or graph [98] to study protein-protein interactions [99–102]. Shown in Figure 4 is a comparison of the graphic profiles between the true and false splice site-containing sequences. As we can see there, the divergence between the true and false splice site-containing sequence profiles is remarkable, clearly indicating that the six structural property parameters can indeed play important roles in RNA splicing. That was why we used them to calculate the global sequence-order effects as elaborated in Section 2.3.

3.2. Cross-Validation. How to properly evaluate the anticipated accuracy is an important step in developing a new predictor. Generally speaking, to avoid the "memory effect" [10] of the resubstitution test in which a same dataset was used to train and test a predictor, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling or K-fold (such as 5-fold, 7-fold, or 10-fold) test, and jackknife test. However, as elaborated by a penetrating analysis in [2], considerable arbitrariness exists in the independent dataset test. Also, as demonstrated by (28)–(30) in [2], the subsampling test (or K-fold cross-validation) cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., [42, 58, 59, 62, 64, 66, 67, 70, 103–107]). Therefore, in this study, the jackknife test was also used to examine the performance of the predictor. During the jackknife test, each sequence in the benchmark dataset $\mathbb{S}_1$ (or $\mathbb{S}_2$) was in turn singled out as an independent test sample and all the rule-parameters were derived based on the remaining data without including the one under the prediction.

Table 3: The prediction quality as measured by metrics of (13) by iSS-PseDNC in identifying the splice donor and acceptor sites, respectively.

Splice sites	Optimal parameters		Metrics			
	λ	w	Sn (%)	Sp (%)	Acc (%)	MCC
Donor[a]	4	0.3	86.66	84.25	85.45	0.71
Acceptor[b]	2	0.3	88.78	86.64	87.73	0.75

[a]See Supplementary Information S1 for benchmark dataset of donor.
[b]See Supplementary Information S2 for benchmark dataset of acceptor.

Table 4: The prediction quality as measured by metrics of (13) by using BLAST [109] and sequence similarity principle in identifying splice acceptor and donor sites, respectively.

Splice sites	Metrics			
	Sn (%)	Sp (%)	Acc (%)	MCC
Acceptor[a]	39.09	40.20	39.62	−0.21
Donor[b]	42.75	37.63	40.23	0.20

[a]See footnote a of Table 3 for further explanation.
[b]See footnote b of Table 3 for further explanation.

3.3. Parameter Optimization. As we can see from (8), the predictive accuracy of the present model depends on the two parameters w and λ, where w is the weight factor which was usually within the range from 0 to 1 and λ is the number of the correlation tiers to be counted for the global sequence-order information. Generally speaking, the greater the λ is, the more global sequence-order information the model will contain. However, if λ is too large, it would reduce the cluster-tolerant capacity [108] so as to lower down the cross-validation accuracy due to overfitting or "high dimension disaster" [28] problem. Therefore, our searching for the optimal values of the two parameters was confined in the range

$$0 \le w \le 1, \qquad 1 \le \lambda \le 10. \tag{14}$$

Furthermore, to reduce the computational time during the search process, the 10-fold cross-validation approach was adopted. Once the optimal values thus obtained for the two parameters were determined, the rigorous jackknife test was utilized to evaluate the anticipated accuracy of the predictor.

Listed in Table 3 are the jackknife test results of the iSS-PseDNC predictor in identifying the splice donor site-containing sequences and the splice acceptor site-containing sequences on the benchmark datasets $\mathbb{S}_1$ and $\mathbb{S}_2$, respectively, where the optimal values for w and λ are also explicitly given.

To further show the power of the iSS-PseDNC predictor, we also did some comparison calculations as described below.

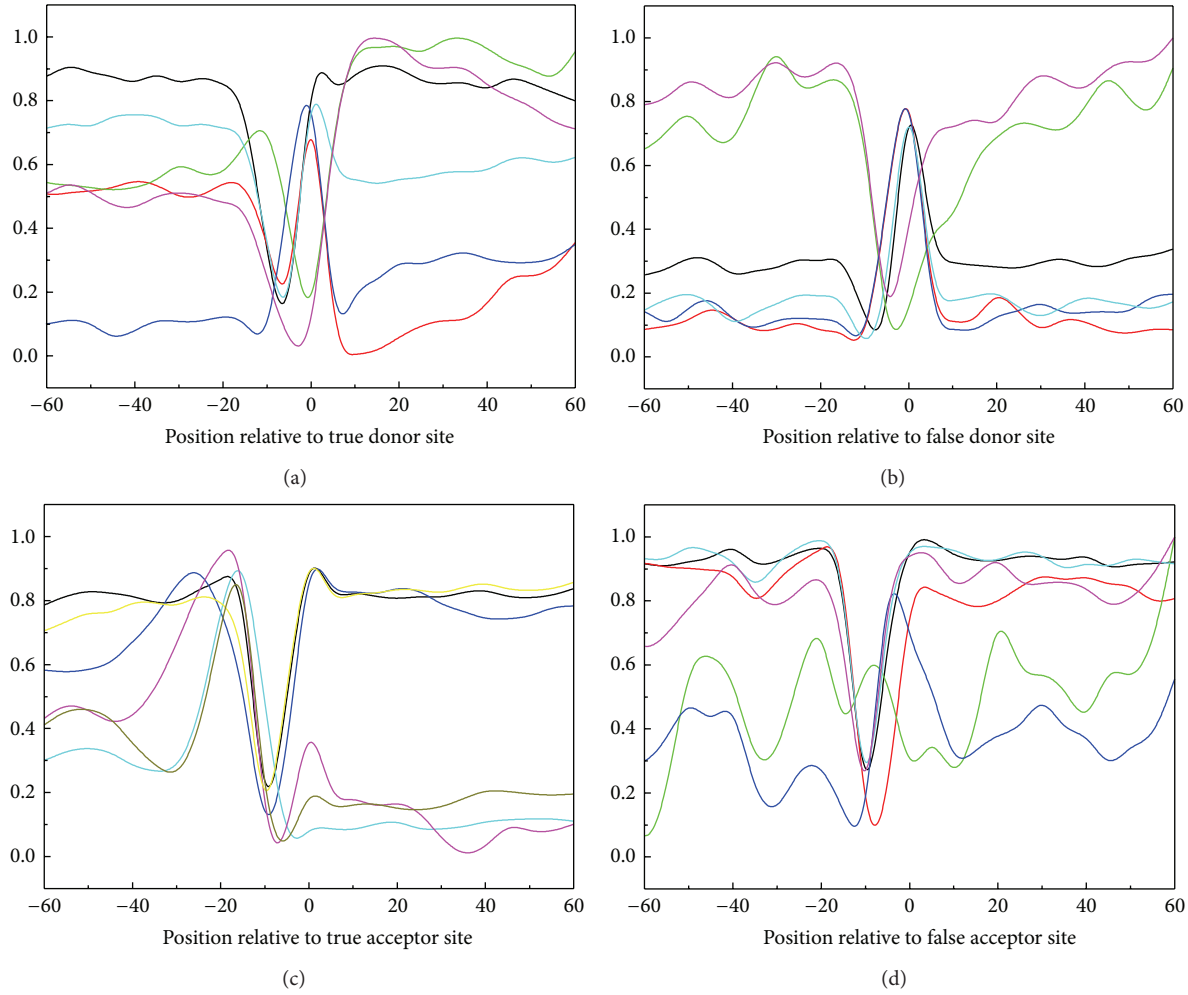

Figure 4: Graphic profiles to show the difference between the true and false splice site-containing sequences. The profiles of six DNA structural properties (i.e., rise (black), slide (red), shift (blue), twist (orange), roll (green) and tilt (purple)) for (a) true splice donor site-containing sequences, (b) false splice donor site- containing sequences, (c) true splice acceptor site-containing sequences, and (d) false acceptor donor site-containing sequences. The profiles are plotted with a window size of 10 bp and a step size of 5 bp.

First, based on the sequence similarity principle, we used BLAST [109] to conduct the jackknife test on the same benchmark dataset as used by the iSS-PseDNC predictor. The results thus obtained are given in Table 4, from which we can see that the percentage rates for Sn, Sp, and Acc by BLAST are about 40% lower than those by iSS-PseDNC and that the rates of MCC by BLAST are about 0.5 lower than those by iSS-PseDNA, for the cases of both donor and acceptor.

Second, rather than pseudo dinucleotide composition (7), we used the dinucleotide compositions (6) to represent the DNA samples for prediction. The corresponding results thus obtained are given in Table 5, from which we can see that the rates for Sn, Sp, Acc, and MCC are all lower than those reported in Table 3, clearly implying that the additional components in the pseudo nucleotide composition did play a role in enhancing the prediction quality.

All these results indicate that the iSS-PseDNC model as proposed in this paper is quite promising and may become a useful tool in identifying splice sites.

Table 5: The prediction quality as measured by metrics of (13) by using the dinucleotide composition (6) to formulate the DNA samples in identifying the splice donor and acceptor sites, respectively.

Splice sites	Metrics			
	Sn (%)	Sp (%)	Acc (%)	MCC
Donor[a]	81.23	84.42	82.58	0.67
Acceptor[b]	83.39	85.60	83.78	0.68

[a]See footnote a of Table 3 for further explanation.
[b]See footnote b of Table 3 for further explanation.

4. Conclusions

RNA splicing is a complicated biological process that involves interactions among DNA, RNA, and proteins. Hence, it is reasonable to analyze the structural properties that can be used to describe these interactions. In view of this, we firstly plotted the profiles of the six DNA structural properties

(twist, tilt, roll, shift, slide, and rise) for splice site-containing sequences and found the differences between true and false splice site-containing sequences. The structural divergences surrounding splice sites may facilitate the removal of the introns by spliceosome.

By defining PseDNC using the above six DNA structural properties, we proposed a model, namely, iSS-PseDNC, for identifying splice sites. The predictive performance demonstrated that our model is helpful for splice site recognitions. Since user-friendly and publicly accessible web-servers represent the direction of developing practically more useful models [110], simulated methods, or predictors, we will make efforts in our future work to provide a web-server for the approach presented in this paper.

It has not escaped our notice that the web-server PseKNC (pseudo *K*-tuple nucleotide composition) developed very recently [111] will be very useful for further improving the prediction quality in identifying the splicing sites.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

The authors wish to thank the editor for taking time to edit this paper and the anonymous reviewers for the constructive comments, which were very helpful for strengthening the presentation of this paper. This work was supported by the National Nature Scientific Foundation of China (nos. 61100092 and 61202256) and the Nature Scientific Foundation of Hebei Province (no. C2013209105).

References

[1] A. A. Hoskins and M. J. Moore, "The spliceosome: a flexible, reversible macromolecular machine," *Trends in Biochemical Sciences*, vol. 37, no. 5, pp. 179–188, 2012.

[2] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[3] X. Xiao, P. Wang, W. Z. Lin, and J. H. Jia, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[4] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, p. e69, 2013.

[5] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[6] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, and W. Chen, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, 2014.

[7] W. R. Qiu and X. Xiao, "iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 1746–1766, 2014.

[8] Y. Xu, J. Ding, and L. Y. Wu, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[9] Y. Xu, X. J. Shao, L. Y. Wu, and N. Y. Deng, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, 2013.

[10] K.-C. Chou and H.-B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[11] C.-T. Zhang and K.-C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Science*, vol. 1, no. 3, pp. 401–408, 1992.

[12] W. Chen, H. Lin, P. M. Feng, C. Ding, and Y. C. Zuo, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[13] T. B. Thompson, K.-C. Chou, and C. Zheng, "Neural network prediction of the HIV-1 protease cleavage sites," *Journal of Theoretical Biology*, vol. 177, no. 4, pp. 369–379, 1995.

[14] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003.

[15] P. M. Feng, W. Chen, H. Lin, and K.-C. Chou, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[16] X. Xiao, P. Wang, and K.-C. Chou, "iNR-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix," *PLoS ONE*, vol. 7, no. 2, Article ID e30869, 2012.

[17] K. K. Kandaswamy, K.-C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[18] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "iDNA-prot: identification of DNA binding proteins using random forest with grey model," *PLoS ONE*, vol. 6, no. 9, Article ID e24756, 2011.

[19] Y.-D. Cai and K.-C. Chou, "Predicting subcellular localization of proteins in a hybridization space," *Bioinformatics*, vol. 20, no. 7, pp. 1151–1156, 2004.

[20] T. Denoeux, "κ-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[21] K.-C. Chou and H.-B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1728–1734, 2007.

[22] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein & Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[23] X. Xiao, J. L. Min, and P. Wang, "iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints," *Journal of Theoretical Biology*, vol. 337, pp. 71–79, 2013.

[24] X. Xiao, P. Wang, W. Z. Lin, and J. H. Jia, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and

their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[25] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[26] M. Wang, J. Yang, Z.-J. Xu, and K.-C. Chou, "SLLE for predicting membrane protein types," *Journal of Theoretical Biology*, vol. 232, no. 1, pp. 7–15, 2005.

[27] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers and Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[28] T. Wang, J. Yang, H.-B. Shen, and K.-C. Chou, "Predicting membrane protein types by the LLDA algorithm," *Protein & Peptide Letters*, vol. 15, no. 9, pp. 915–921, 2008.

[29] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *PROTEINS: Structure, Function, and Genetics*, vol. 43, pp. 246–255, 2001, (Erratum: ibid., 2001, Vol. 44, 60).

[30] S. X. Lin and J. Lapointe, "Theoretical and experimental biology in one," *Journal of Biomedical Science and Engineering*, vol. 6, pp. 435–442, 2013.

[31] X.-B. Zhou, C. Chen, Z.-C. Li, and X.-Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," *Journal of Theoretical Biology*, vol. 248, no. 3, pp. 546–551, 2007.

[32] T.-L. Zhang, Y.-S. Ding, and K.-C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," *Journal of Theoretical Biology*, vol. 250, no. 1, pp. 186–193, 2008.

[33] X. Jiang, R. Wei, Y. Zhao, and T. Zhang, "Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location," *Amino Acids*, vol. 34, no. 4, pp. 669–675, 2008.

[34] H. Lin, "The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008.

[35] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[36] G.-Y. Zhang and B.-S. Fang, "Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 253, no. 2, pp. 310–315, 2008.

[37] S.-W. Zhang, W. Chen, F. Yang, and Q. Pan, "Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach," *Amino Acids*, vol. 35, no. 3, pp. 591–598, 2008.

[38] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[39] Z.-C. Li, X.-B. Zhou, Z. Dai, and X.-Y. Zou, "Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis," *Amino Acids*, vol. 37, no. 2, pp. 415–425, 2009.

[40] H. Lin, H. Wang, H. Ding, Y.-L. Chen, and Q.-Z. Li, "Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition," *Acta Biotheoretica*, vol. 57, no. 3, pp. 321–330, 2009.

[41] J.-D. Qiu, J.-H. Huang, R.-P. Liang, and X.-Q. Lu, "Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform," *Analytical Biochemistry*, vol. 390, no. 1, pp. 68–73, 2009.

[42] Y.-H. Zeng, Y.-Z. Guo, R.-Q. Xiao, L. Yang, L.-Z. Yu, and M.-L. Li, "Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach," *Journal of Theoretical Biology*, vol. 259, no. 2, pp. 366–372, 2009.

[43] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[44] H. Mohabatkar, "Prediction of cyclin proteins using Chou's pseudo amino acid composition," *Protein & Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[45] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[46] L. Yu, Y. Guo, Y. Li et al., "SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 267, no. 1, pp. 1–6, 2010.

[47] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[48] M. M. Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[49] J.-D. Qiu, S.-B. Suo, X.-Y. Sun, S.-P. Shi, and R.-P. Liang, "OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition," *Journal of Molecular Graphics and Modelling*, vol. 30, pp. 129–134, 2011.

[50] D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[51] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[52] G.-L. Fan and Q.-Z. Li, "Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 304, pp. 88–95, 2012.

[53] S. Mei, "Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization," *Journal of Theoretical Biology*, vol. 293, pp. 121–130, 2012.

[54] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *Journal of Theoretical Biology*, vol. 310, pp. 80–87, 2012.

[55] L. Nanni, S. Brahnam, and A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, no. 2, pp. 657–665, 2012.

[56] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on vari-

ants of Chou's Pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[57] X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang, and R. P. Liang, "Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform," *Molecular BioSystems*, vol. 8, no. 12, pp. 3178–3184, 2012.

[58] T. H. Chang, L. C. Wu, T. Y. Lee, S. P. Chen, H. D. Huang, and J. T. Horng, "EuLoc: a web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC," *Journal of Computer-Aided Molecular Design*, vol. 27, no. 1, pp. 91–103, 2013.

[59] Y. K. Chen and K. B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 318, pp. 1–12, 2013.

[60] G. L. Fan and Q. Z. Li, "Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 334, pp. 45–51, 2013.

[61] M. K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition," *SAR and QSAR in Environmental Research*, vol. 24, no. 7, pp. 597–609, 2013.

[62] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *Journal of Theoretical Biology*, vol. 341, pp. 34–40, 2014.

[63] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, no. 1, pp. 50–57, 2013.

[64] C. Huang and J. Q. Yuan, "A multilabel model based on chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types," *The Journal of Membrane Biology*, vol. 246, no. 4, pp. 327–334, 2013.

[65] C. Huang and J. Q. Yuan, "Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions," *Journal of Theoretical Biology*, vol. 335, pp. 205–212, 2013.

[66] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein & Peptide Letters*, vol. 20, no. 2, pp. 180–186, 2013.

[67] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[68] Y. F. Qin, L. Zheng, and J. Huang, "Locating apoptosis proteins by incorporating the signal peptide cleavage sites into the general form of Chou's Pseudo amino acid composition," *International Journal of Quantum Chemistry*, vol. 113, no. 11, pp. 1660–1667, 2013.

[69] A. N. Sarangi, M. Lohani, and R. Aggarwal, "Prediction of essential proteins in prokaryotes by incorporating various physico-chemical features into the general form of Chou's pseudo amino acid composition," *Protein & Peptide Letters*, vol. 20, no. 7, pp. 781–795, 2013.

[70] S. Wan, M. W. Mak, and S. Y. Kung, "GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[71] X. Wang, G. Z. Li, and W. C. Lu, "Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition," *Protein & Peptide Letters*, vol. 20, no. 3, pp. 309–317, 2013.

[72] N. Xiaohui, L. Nana, X. Jingbo et al., "Using the concept of Chou's pseudo amino acid composition to predict protein solubility: an approach with entropies in information theory," *Journal of Theoretical Biology*, vol. 332, pp. 211–217, 2013.

[73] H. L. Xie, L. Fu, and X. D. Nie, "Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC," *Protein Engineering, Design and Selection*, vol. 26, no. 11, pp. 735–742, 2013.

[74] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[75] J. R. Goñi, A. Pérez, D. Torrents, and M. Orozco, "Determining promoter location based on DNA structure first-principles calculations," *Genome Biology*, vol. 8, no. 12, article R263, 2007.

[76] J. R. Goñi, C. Fenollosa, A. Pérez, D. Torrents, and M. Orozco, "DNAlive: a tool for the physical analysis of DNA at the genomic scale," *Bioinformatics*, vol. 24, no. 15, pp. 1731–1732, 2008.

[77] P. M. Feng, H. Ding, W. Chen, and H. Lin, "Naive Bayes classifier with feature selection to identify phage virion proteins," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 530696, 6 pages, 2013.

[78] W. Chen, P. Feng, and H. Lin, "Prediction of replication origins by calculating DNA structural properties," *FEBS Letters*, vol. 586, no. 6, pp. 934–938, 2012.

[79] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines. pp.Software," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[80] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[81] X. Xiao, Z.-C. Wu, and K.-C. Chou, "A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites," *PLoS ONE*, vol. 6, no. 6, Article ID e20592, 2011.

[82] X. Xiao, Z.-C. Wu, and K.-C. Chou, "iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 42–51, 2011.

[83] Z.-C. Wu, X. Xiao, and K.-C. Chou, "ILoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites," *Molecular BioSystems*, vol. 7, no. 12, pp. 3287–3297, 2011.

[84] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, "Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities," *PLoS ONE*, vol. 7, no. 4, Article ID e35254, 2012.

[85] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. van de Peer, "Generic eukaryotic core promoter prediction using structural features of DNA," *Genome Research*, vol. 18, no. 2, pp. 310–323, 2008.

[86] K. C. Chou and S. Forsén, "Graphical rules for enzyme-catalysed rate laws," *Biochemical Journal*, vol. 187, no. 3, pp. 829–835, 1980.

[87] G. P. Zhou and M. H. Deng, "An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways," *Biochemical Journal*, vol. 222, no. 1, pp. 169–176, 1984.

[88] K. C. Chou, "Graphic rules in steady and non-steady state enzyme kinetics," *Journal of Biological Chemistry*, vol. 264, no. 20, pp. 12074–12079, 1989.

[89] J. Andraos, "Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws—new methods based on directed graphs," *Canadian Journal of Chemistry*, vol. 86, no. 4, pp. 342–357, 2008.

[90] I. W. Althaus, K. M. Franks, M. R. Diebel et al., "The benzylthiopyrididine U-31, 355 is a potent inhibitor of HIV-1 reverse transcriptase," *Biochemical Pharmacology*, vol. 51, pp. 743–750, 1996.

[91] I. W. Althaus, J. J. Chou, A. J. Gonzales et al., "Kinetic studies with the non-nucleoside human immunodeficiency virus type-1 reverse transcriptase inhibitor U-90152E," *Biochemical Pharmacology*, vol. 47, no. 11, pp. 2017–2028, 1994.

[92] I. W. Althaus, J. J. Chou, A. J. Gonzales et al., "Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E," *Journal of Biological Chemistry*, vol. 268, no. 9, pp. 6119–6124, 1993.

[93] I. W. Althaus, J. J. Chou, A. J. Gonzales et al., "Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E," *Biochemistry*, vol. 32, no. 26, pp. 6548–6554, 1993.

[94] K.-C. Chou, F. J. Kezdy, and F. Reusser, "Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases," *Analytical Biochemistry*, vol. 221, no. 2, pp. 217–230, 1994.

[95] K.-C. Chou, "Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems," *Biophysical Chemistry*, vol. 35, no. 1, pp. 1–24, 1990.

[96] K.-C. Chou, "Graphic rule for drug metabolism systems," *Current Drug Metabolism*, vol. 11, no. 4, pp. 369–378, 2010.

[97] Z.-C. Wu, X. Xiao, and K.-C. Chou, "2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids," *Journal of Theoretical Biology*, vol. 267, no. 1, pp. 29–34, 2010.

[98] K. C. Chou, W. Z. Lin, and X. Xiao, "Wenxiang: a web-server for drawing wenxiang diagrams," *Natural Science*, vol. 3, no. 10, pp. 862–865, 2011.

[99] G.-P. Zhou, "The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 142–148, 2011.

[100] N. Kurochkina and T. Choekyi, "Helix-helix interfaces and ligand binding," *Journal of Theoretical Biology*, vol. 283, no. 1, pp. 92–102, 2011.

[101] G.-P. Zhou, "The structural determinations of the leucine zipper coiled-coil domains of the cGMP-dependent protein kinase Iα and its interaction with the myosin binding subunit of the myosin light chains phosphase," *Protein & Peptide Letters*, vol. 18, no. 10, pp. 966–978, 2011.

[102] G. P. Zhou and R. B. Huang, "The pH-triggered conversion of the PrP(c) to PrP(sc)," *Current Topics in Medicinal Chemistry*, vol. 13, no. 10, pp. 1152–1163, 2013.

[103] S.-W. Zhang, Y.-L. Zhang, H.-F. Yang, C.-H. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.

[104] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," *Protein Journal*, vol. 17, no. 8, pp. 729–738, 1998.

[105] S. Ding, Y. Li, X. Yang, and T. Wang, "A simple k-word interval method for phylogenetic analysis of DNA sequences," *Journal of Theoretical Biology*, vol. 317, pp. 192–199, 2013.

[106] X. Jingbo, Z. Silan, S. Feng et al., "Using the concept of pseudo amino acid composition to predict resistance gene against Xanthomonas oryzae pv. oryzae in rice: an approach from chaos games representation," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 16–23, 2011.

[107] M. Hayat and A. Khan, "Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 10–17, 2011.

[108] K.-C. Chou, "A key driving force in determination of protein structural classes," *Biochemical and Biophysical Research Communications*, vol. 264, no. 1, pp. 216–224, 1999.

[109] A. A. Schäffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[110] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 2, pp. 63–92, 2009.

[111] W. Chen, T. Y. Lei, D. C. Jin, and H. Lin, "PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, 2014.