Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: www.elsevier.com/locate/biophyschem

Research Article

Identifying DNase I hypersensitive sites using multi-features fusion and F-score features selection via Chou's 5-steps rule

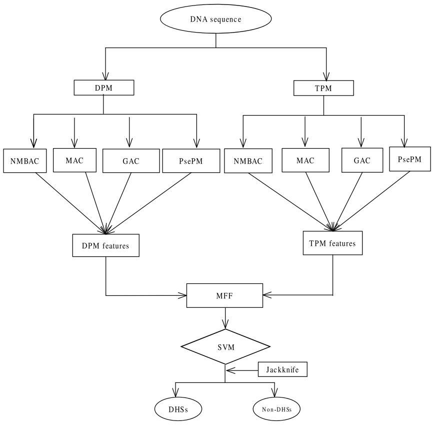
Yunyun Liang[a,*], Shengli Zhang[b]

[a] School of Science, Xi'an Polytechnic University, Xi'an 710048, PR China
[b] School of Mathematics and Statistics, Xidian University, Xi'an 710071, PR China

HIGHLIGHTS

- A novel identifying model named iDHSs-MFF is proposed based on di-nucleotide and trinucleotide property matrixs.
- F-score approach is performed for features selection.
- iDHSs-MFF model outperforms some highly related models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Hypersensitive sites
Multi-features fusion
Dinucleotide property matrix
Trinucleotide property matrix
F-score

ABSTRACT

DNase I hypersensitive sites (DHSs) are regarded as those regions of chromatin that are sensitive to cleavage by the DNase I enzyme. Identification of DNase I hypersensitive sites will provide useful insights for discovering DNA's functional elements from the non-coding sequences in the biomedical research. Because of the significance for DNase I hypersensitive sites, it is indispensable to develop an accurate, fast, robust, and high-throughput automated computational model. In this paper, we develop a model named iDHSs-MFF by combining multiple fusion features and F-score features selection approach. The multiple fusion features include three auto-correlation descriptors based on the dinucleotide property matrix and the trinucleotide property matrix (TPM), Pseudo-DPM and Pseudo-TPM. Evaluation by the jackknife cross-validation indicates that the selected features by F-score are effective in the identification of DNase I hypersensitive sites. Experimental results on two benchmark datasets demonstrate that the proposed model outperforms some highly related models. Systematic application of this computational approach will greatly facilitate the analysis of transcriptional regulatory elements. The datasets and Matlab source codes are freely available at: https://github.com/shengli0201/Datasets.

[*] Corresponding author.
E-mail address: liangyunyun@xpu.edu.cn (Y. Liang).

https://doi.org/10.1016/j.bpc.2019.106227
Received 19 April 2019; Received in revised form 4 July 2019; Accepted 10 July 2019
Available online 13 July 2019
0301-4622/

1. Introduction

In the human genome, DNA sequences containing active *cis*-regulatory DNA elements bounded to regulatory proteins undergo dynamic nucleosome displacements and have become hypersensitive to cleavage by the DNase I enzyme [1,2]. These specific genome regions are called DNase I hypersensitive sites (DHSs), which were recognized in 1980s [3,4]. The DHSs are reliable and generic markers of chromatin zones containing transcriptional regulatory elements, making them critical for discovering functional non-coding elements involved in gene regulation and understanding the regulatory mechanisms of gene expression. This has led to discovery of a wide variety of genomic regulatory elements including promoters, enhancers, insulators, silencers, and suppressors [4–6]. Accordingly, mapping DHSs is an extremely accurate method for detecting the positions of functional regulatory elements and has underpinned the discovery of most experimentally established distal *cis*-acting elements.

Southern blotting technique [7] for identifying DHSs is very traditional technique, however, obtaining information from Southern blot approach is a expensive, time-consuming, and inaccurate task. Recent advances in experimental methods based on high-throughput sequencing technology have been employed to identify DHSs genome widely in the human genome [8,9]. Unfortunately, performing whole-genome sequencing requires significant expenditure and skilled labor, therefore, it is urgent to develop automated computational means to efficiently and accurately identify DHSs. In fact, several computational methods have been proposed for addressing this issue based on sequence information, which include SVM-RevcKmer [10], SVM-PseDNC [11], iDHL-EL [12], DHSpred [13] and so on. Although these methods continuously improve the prediction accuracy, their provided predictive results are unsatisfactory, thus, a more effective computational model is needed to identify DHSs.

The information derived by the computational approaches are advantageous. The knowledge of protein 3D (three-dimensional) structures or their complexes with ligands is vitally important for rational drug design. Although X-ray crystallography is a powerful tool in determining these structures, it is time-consuming and expensive, and not all proteins can be successfully crystallized. Membrane proteins are difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein structures have been determined. NMR is indeed a very powerful tool in determining the 3D structures of membrane proteins [14–31], but it is also time-consuming and costly. To acquire the structural information in a timely manner, a series of 3D protein structures have been developed by means of structural bioinformatics tools [32–44]. Meanwhile, facing the explosive growth of biological sequences discovered in the post-genomic age, to timely use them for drug development, a lot of important sequence-based information, such as PTM (posttranslational modification) sites in proteins [45], protein-drug interaction in cellular networking [46], protein-protein interactions [47], DNA-methylation sites [48], recombination spots [49], and sigma-54 promoters [50], have been deducted by various sequential bioinformatics tools such as PseAAC approach [51] and PseKNC approach [52]. Actually, the rapid development in sequential bioinformatics and structural bioinformatics have driven the medicinal chemistry undergoing an unprecedented revolution [53], in which the computational biology has played increasingly important roles in stimulating the development of finding novel drugs. In view of this, the computational (or in silico) methods were also utilized in this study for Identifying DNase I hypersensitive sites.

In this paper, we focus on developing a identifying model named iDHSs-MFF by using Normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation descriptors based on dinucleotide and trinucleotide property matrixs, pseudo-dinucleotide property matrix and pseudo-trinucleotide property matrix. Then, a 1080-dimensional feature vector is obtained, which is too large to input into the SVM classifier. The large dimension will exist redundancy and increase computational complexity. F-score technique in this study is adopted due to its simplicity for a DHSs detection system with real applications. To objectively and rigorously evaluate our identifying model, the jackknife test is employed on two benchmark datasets, experimental results show that our model could achieve accuracies up to 86.63% and 86.94% for $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. As a result, our model provides a more powerful tool for identifying DHSs.

As demonstrated by a series of recent publications [46,48–50,54–69] and summarized in a comprehensive review [70], to develop a really useful predictor for a biological system, one needs to follow Chou's 5-step rule to go through the following five steps: (a) select or construct a valid benchmark dataset to train and test the predictor; (b) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm to conduct the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (e) establish a user-friendly web-server for the predictor that is accessible to the public. Papers presented for a developing a new sequence-analyzing method or statistical predictor by observing the guidelines of Chous 5-strp rules have the following notable merits: (1) crystal clear in logic development, (2) completely transparent in operation, (3) easily to repeat the reported results by other investigators, (4) with high potential in stimulating other sequence-analyzing methods, and (5) very convenient to be used by the majority of experimental scientists. Below, let us elaborate how to deal with these five steps.

2. Materials and methods

2.1. Datasets

To construct a promising computational model, there need valid benchmark dataset to train and test the model effectively. For this purpose, we have used dataset $\mathbb{S}_1$ in this paper, which have been taken from Noble et al. [10]. The benchmark dataset $\mathbb{S}_1$ contains 280 sequences for DNase I hypersensitive sites and 737 sequences for non-DNase I hypersensitive sites. Therefore, the dataset $\mathbb{S}_1$ for the current study can be expressed as

$$\mathbb{S}_1 = \mathbb{S}_1^+ \cup \mathbb{S}_1^- \tag{1}$$

where $\mathbb{S}_1^+$ is the subset for the DNase I hypersensitive sites and $\mathbb{S}_1^-$ is the subset of non-DNase I hypersensitive sites, and "∪" is a mathematical operator representing "union".

In order to avoid misleading results with an overestimated accuracy due to redundant sequences samples with high similarity, Feng et al. [11] deletes those DNA sequences for $\mathbb{S}_1$ that have $\geq 60\%$ pairwise sequence identity to each other. Finally, the benchmark dataset $\mathbb{S}_2$ is obtained, in other words, the sequence similarity between $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets is less than 60%, and the $\mathbb{S}_2$ contains 247 sequences for DNase I hypersensitive sites and 710 sequences for non-DNase I hypersensitive sites, which can be expressed as

$$\mathbb{S}_2 = \mathbb{S}_2^+ \cup \mathbb{S}_2^- \tag{2}$$

where $\mathbb{S}_2^+$ is the subset for the DNase I hypersensitive sites and $\mathbb{S}_2^-$ is the subset of non-DNase I hypersensitive sites.

2.2. Feature extraction

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms (such as "Optimization" algorithm [71], "Covariance Discriminant" or "CD" algorithm [72,73], "Nearest

Neighbor" or "NN" algorithm [74], and "Support Vector Machine" or "SVM" algorithm [74,75]) can only handle vectors as elaborated in a comprehensive review [53]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [51] or PseAAC [76] was proposed. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all the areas of computational proteomics [77–88]. Because it has been widely and increasingly used, four powerful open access soft-wares, called 'PseAAC' [89], 'PseAAC-Builder' [90], 'propy' [91], and 'PseAAC-General' [92], were established: the former three are for generating various modes of Chou's special PseAAC [93]; while the 4th one for those of Chou's general PseAAC [70], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see eqs. 9–10 of [70]), "Gene Ontology" mode (see eqs. 11–12 of [70]), and "Sequential Evolution" or "PSSM" mode (see eqs. 13–14 of [70]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, the concept of PseKNC (Pseudo *K*-tuple Nucleotide Composition) [52] was developed for generating various feature vectors for DNA/RNA sequences [94–96] that have proved very useful as well. Particularly, recently a very powerful web-server called 'Pse-in-One' [97] and its updated version 'Pse-in-One 2.0' [98] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

2.2.1. Dinucleotide and trinucleotide property matrixs

Suppose a DNA sequence D with L nucleotides, i.e., having length L is represented with the following expression:

$$D = R_1R_2R_3R_4R_5R_6\cdots R_i\cdots R_{L-1}R_L, \tag{3}$$

where $R_i \in \{A(adenine), C(cytosine), G(guanine), T(thymine)\}$ denotes the nucleic acid residue at the sequence position i $(i = 1, 2, \cdots, L)$. Arbitrary two nucleotides paired is called dinucleotide, as a result, there are totally 4*4 = 16 basic dinucleotides. Analogously, arbitrary permutation and combination of three nucleotides is called trinucleotide, as a result, there are totally 4*4*4 = 64 basic trinucleotides.

DNA physicochemical property is evolutionarily more constrained than the underlying actual sequence, and the topography-informed constrained regions usually correlate with functional noncoding elements such as enhancers [99]. Hence, it is reasonable to use the physicochemical properties of nucleotides to exact features based on the DNA sequences. The 15 DNA dinucleotide properties and 12 DNA trinucleotide properties are listed in Tables 1 and 2 [100], We further normalize each DNA property as follows:

$$\frac{P - P_{min}}{P_{max} - P_{min}}, \tag{4}$$

where P is the original property value, P_{min} and P_{max} are the minimum and the maximum property values, respectively. Then, according to the operation in [101], each dinucleotide is replaced by a value corresponding to a physicochemical property in a DNA sequence, with the result that each DNA sequence in $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets can be converted into a matrix $P^{di} = (p_{i,\ j}{}^{di})_{(L-1)\times 15}$, where 15 represents the number of dinucleotide property, and $p_{i,\ j}{}^{di}$ represents the value of the ith dinucleotide pair, in other words, two adjacent nucleotides corresponding to the jth property in the DNA sequence, then P^{di} is named the dinucleotide property matrix (DPM). Analogously, each DNA sequence in $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets can be converted into a matrix $P^{tri} = (p_{i,\ j}{}^{tri})_{(L-2)\times 12}$, where 12 represents the number of trinucleotide property, and $p_{i,\ j}{}^{tri}$ represents the value of the ith trinucleotide combination, in other words, three adjacent nucleotides corresponding to the jth property in the DNA sequence, then P^{tri} is named the trinucleotide property matrix (TPM).

Table 1
List of 15 DNA physicochemical properties of dinucleotides.

Number	Property
1	F-roll
2	F-tilt
3	F-twist
4	F-slide
5	F-shift
6	F-rise
7	Roll
8	Tilt
9	Twist
10	Slide
11	Shift
12	Rise
13	Energy
14	Enthalpy
15	Entropy

Table 2
List of 12 DNA physicochemical properties of trinucleotides.

Number	Property
1	Bendability (DNase)
2	Bendability (consensus)
3	Trinucleotide GC content
4	Nucleosome positioning
5	Consensus_roll
6	Consensus-Rigid
7	DNase I
8	DNase I-Rigid
9	MW-Daltons
10	MW-kg
11	Nucleosome
12	Nucleosome-Rigid

2.2.2. Three auto-correlation descriptors based on DPM and TPM

A DNA sequence can be viewed as a time sequence of the corresponding the physicochemical properties of nucleotides. In this paper, dinucleotide and trinucleotide properties represented in the form of the dinucleotide property matrix (DPM) and the trinucleotide property matrix (TPM) are adopted as the considered properties. Here, each column is taken as one property, so the DPM contains 15 different properties, and the TPM contains 12 different properties, of which can be considered as the time sequences.

To transform the DPM and TPM of different lengths into equal length vector, and avoid the loss of the sequence-order information, three different autocorrelation descriptors based on DPM and TPM are adopted, which include normalized Moreau-Broto autocorrelation [102], Moran autocorrelation [103] and Geary autocorrelation [104]. Autocorrelation descriptor is a powerful statistical tool and defined based on the distribution of nucleic acid residue properties along the DNA sequence, which measures the correlation between two dinucleotide or trinucleotide separated by a distance of λ in DPM and TPM, respectively, and they are defined as:

(1) Normalized Moreau-Broto autocorrelation descriptor

$$N_j^{di,\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} P_{i,j}^{di} \times P_{i+\lambda,j}^{di},\ (j = 1, 2, \cdots, 15; 0 < \lambda < L), \tag{5}$$

$$N_j^{tri,\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} P_{i,j}^{tri} \times P_{i+\lambda,j}^{tri},\ (j = 1, 2, \cdots, 12; 0 < \lambda < L), \tag{6}$$

where $N_j^{di,\ \lambda}$ and $N_j^{tri,\ \lambda}$ are the Moreau-Broto correlation factor of the jth dinucleotide and trinucleotide property, respectively. λ is the lag of the autocorrelation along the column in DPM and TPM, $P_{i,\ j}{}^{di}$ and $P_{i+\lambda,}$

$_{j}^{di}$ represent the values in the ith and $(i + \lambda)$th position of the jth column in DPM, $P_{i,\ j}^{tri}$ and $P_{i+\lambda,\ j}^{tri}$ represent the values in the ith and $(i + \lambda)$th position of the jth column in TPM. Here, the value of λ varies from 1 to 10. Then, a 150-dimensional feature vector is obtained by Normalized Moreau-Broto autocorrelation and DPM (NMBAC-DPM), and a 120-dimensional feature vector is obtained by Normalized Moreau-Broto autocorrelation and TPM (NMBAC-TPM).

(2) Moran autocorrelation descriptor

$$M_j^{di,\lambda} = \frac{\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}(P_{i,j}^{di} - \overline{P}_j^{di})(P_{i+\lambda,j}^{di} - \overline{P}_j^{di})}{\frac{1}{L}\sum_{i=1}^{L}(P_{i,j}^{di} - \overline{P}_j^{di})^2}, (j = 1, 2, \cdots, 15; 0 < \lambda < L), \tag{7}$$

$$M_j^{tri,\lambda} = \frac{\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}(P_{i,j}^{tri} - \overline{P}_j^{tri})(P_{i+\lambda,j}^{tri} - \overline{P}_j^{tri})}{\frac{1}{L}\sum_{i=1}^{L}(P_{i,j}^{tri} - \overline{P}_j^{tri})^2}, (j = 1, 2, \cdots, 12; 0 < \lambda < L), \tag{8}$$

where $M_j^{di,\ \lambda}$ and $M_j^{tri,\ \lambda}$ are the Moran correlation factor of the jth dinucleotide and trinucleotide property, respectively. λ, $P_{i,\ j}^{di}$, $P_{i+\lambda,\ j}^{di}$, $P_{i,\ j}^{tri}$ and $P_{i+\lambda,\ j}^{tri}$ are the same as the above, $\overline{P}_j^{di}$ and $\overline{P}_j^{tri}$ represent the average value of the jth in DPM and TPM, respectively. Here, the value of λ varies from 1 to 10. Then, a 150-dimensional feature vector is obtained by Moran autocorrelation and DPM (MAC-DPM), and a 120-dimensional feature vector is obtained by Moran autocorrelation and TPM (MAC-TPM).

(3) Geary autocorrelation descriptor

$$G_j^{di,\lambda} = \frac{\frac{1}{2(L-\lambda)}\sum_{i=1}^{L-\lambda}(P_{i,j}^{di} - P_{i+\lambda,j}^{di})^2}{\frac{1}{L-1}\sum_{i=1}^{L}(P_{i,j}^{di} - \overline{P}_j^{di})^2}, (j = 1, 2, \cdots, 15; 0 < \lambda < L), \tag{9}$$

$$G_j^{tri,\lambda} = \frac{\frac{1}{2(L-\lambda)}\sum_{i=1}^{L-\lambda}(P_{i,j}^{tri} - P_{i+\lambda,j}^{tri})^2}{\frac{1}{L-1}\sum_{i=1}^{L}(P_{i,j}^{tri} - \overline{P}_j^{tri})^2}, (j = 1, 2, \cdots, 12; 0 < \lambda < L), \tag{10}$$

where $G_j^{di,\ \lambda}$ and $G_j^{tri,\ \lambda}$ are the Geary correlation factor of the jth dinucleotide and trinucleotide property, respectively. λ, $P_{i,\ j}^{di}$, $P_{i+\lambda,\ j}^{di}$, $P_{i,\ j}^{tri}$, $P_{i+\lambda,\ j}^{tri}$, $\overline{P}_j^{di}$ and $\overline{P}_j^{tri}$ are the same as the above. Here, the value of λ varies from 1 to 10. Then, a 150-dimensional feature vector is obtained by Geary autocorrelation and DPM (GAC-DPM), and a 120-dimensional feature vector is obtained by Geary autocorrelation and TPM (GAC-TPM).

2.2.3. Pseudo-property matrix

Encouraged by the success of using the pseudo-position specific scoring matrix (PsePSSM) idea [105] to deal with the protein sequences, PseDPM and PseTPM are proposed based on dinucleotide property matrix (DPM) and the trinucleotide property matrix to deal with the DNA sequences, respectively, which are expressed as

$$\theta_j^{di,\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}(P_{i,j}^{di} - P_{i+\lambda,j}^{di})^2, (j = 1, 2, \cdots, 15; 0 < \lambda < L), \tag{11}$$

$$\theta_j^{tri,\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}(P_{i,j}^{tri} - P_{i+\lambda,j}^{tri})^2, (j = 1, 2, \cdots, 12; 0 < \lambda < L), \tag{12}$$

where $\theta_j^{di,\ \lambda}$ and $\theta_j^{tri,\ \lambda}$ are the correlation factor of the jth dinucleotide and trinucleotide property, whose contiguous distance is λ along the column in the DPM and TPM, respectively. $P_{i,\ j}^{di}$ and $P_{i+\lambda,\ j}^{di}$ represent the values in the ith and $(i + \lambda)$th position of the jth column in DPM, $P_{i,\ j}^{tri}$ and $P_{i+\lambda,\ j}^{tri}$ represent the values in the ith and $(i + \lambda)$th position of the jth column in TPM. Here, the value of λ varies from 1 to 10. Then, a 150-dimensional feature vector is obtained by Pseudo-DPM (PseDPM), and a 120-dimensional feature vector is obtained by Pseudo-TPM (PseTPM). PseDPM and PseTPM are collectively called Pseudo-property matrix (PsePM).

2.3. F-score algorithm

The features selection can help the original classification system achieve a better predictive performance and a lower computational cost by removing any redundant features. F-score is a simple and effective algorithm including variable ranking as a principal selection mechanism. Given the ith feature vector $\{x_{i1}, x_{i2}, \cdots x_{iN}\}$ with the N instances, the total number of positive and negative instances are n^+ and n^-, respectively, then the F-score of the ith feature is defined as

$$F(i) = \frac{(\overline{x}_i^{(+)} - \overline{x}_i)^2 + (\overline{x}_i^{(-)} - \overline{x}_i)^2}{\frac{1}{n^+ - 1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \overline{x}_i^{(+)})^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \overline{x}_i^{(-)})^2}, \tag{13}$$

where $\overline{x}_i^{(+)}$ stands for the mean value of the ith feature of entire positive instances, $\overline{x}_i^{(-)}$ for the mean value of ith feature of entire negative instances, $\overline{x}_i$ for the mean value of the ith feature of the total instances. $x_{k,\ i}^{(+)}$ for the value of the ith feature of the kth instance in the positive dataset, and $x_{k,\ i}^{(-)}$ for the value of the ith feature of the kth instance in the negative dataset. The numerator indicates the discrimination between the positive and the negative sets, and the denominator is the sum of the deviation within each feature set.

The larger the F-score is, the more discriminative the feature is [106]. The F-score algorithm is widely used in the field of computational biology and bioinformatics, such as identification of anticancer peptides [107], prediction of DNA N4-methylcytosine sites [108] and prediction of anti-hypertensive peptides [109]. We adopt the F-score approach in this paper due to its simplicity of its use in a identifying system with real applications.

2.4. Support vector machine

The concept of support vector machine (SVM) was firstly introduced by Vapnik and his coworkers [110]. SVM is a supervised machine learning approach based on statistical learning theory and has been extensively used in many kinds of pattern recognition problems, such as protein structural classes prediction [111,112], protein subcellular localization prediction [113], proline cis/trans isomerization prediction [114], taxonomy-based protein fold recognition [115], identifying bacterial secreted proteins [116], predicting protein-ATP binding sites [117], prediction of phosphorylation sites [118], predicting protein oxidation sites [119]. The basic idea of SVM is to project instances with low-dimensional feature into a high-dimension Hilbert space, it searches and constructs a separating hyperplane which could classify positive and negative instances with the maximal margin in the space by using the decision function:

$$f(\vec{x}) = sgn\left[\sum_{i=1}^{N} y_i\alpha_i \cdot K(\vec{x}, \vec{x}_i) + b\right] \tag{14}$$

where $\vec{x}$ is the ith training vector, y_i denotes the type of the ith training vector. $K(\vec{x}, \vec{x}_i)$ is called a kernel function which defines an inner product in a high dimensional feature space. In this paper, the LIBSVM [120] is used to perform the prediction, and the radial basis function (RBF) $K(\vec{x}, \vec{x}_i) = \exp(-\gamma\|\vec{x}_i - \vec{x}_j\|)$ is selected as the kernel function, since it is suitable for nonlinear classification. In order to construct the optimal model, the regularization parameter C and the kernel width parameter γ are optimized via an optimization procedure using a grid search approach, of which the search spaces for C and γ are $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^5]$.

2.5. Cross-validation

The cross-validation methods are often used to examine the quality of a predictor and its effectiveness in the pattern recognition problem. The independent dataset test, subsampling or K-fold cross validation test and jackknife test or leave-one-out (LOO) test are the most cross-

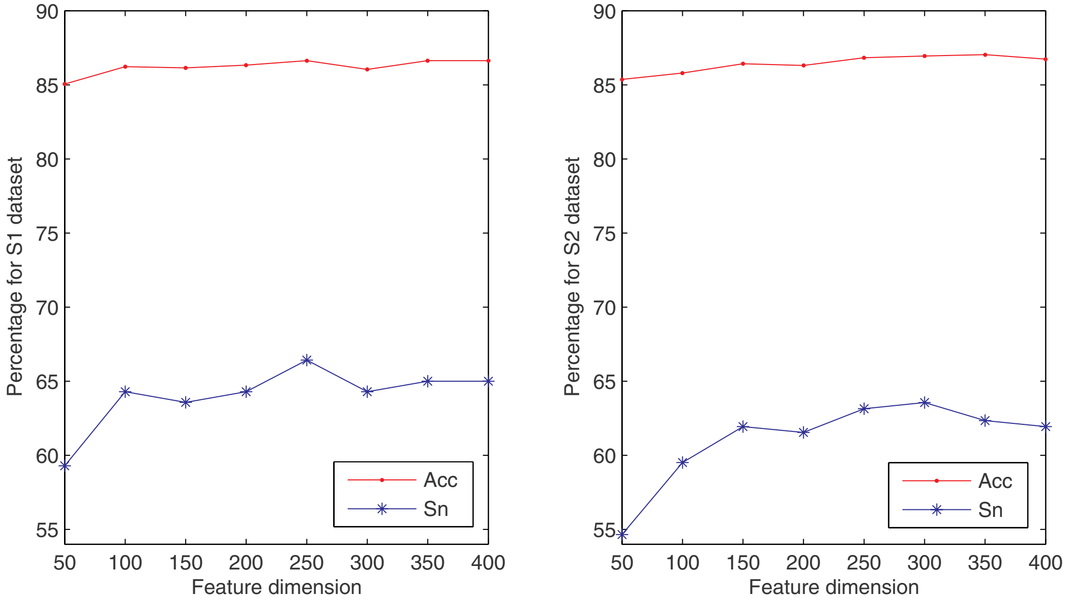


Fig. 1. The selection of the optimal features by F-score algorithm.

validations. The jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset [121,122]. Therefore, it has been widely recognized and increasingly utilized to examine the quality of various predictors [115,123–126]. For the jackknife cross-validation, each instance in the dataset is in turn singled out as an independent test instance and all the rule parameters are calculated based on the remaining instances without including the one being identified. Therefore, we also use the jackknife cross-validation to examine the proposed model.

2.6. Performance measurement

We used four different measures that are commonly used in binary classification tasks to evaluate the performances of the models [127–129], the following measures are used such as sensitivity (Sn), specificity (Sp), accuracy (Acc) and Matthews correlation coefficient (MCC). Sensitivity is the percentage of correct predictions from positive case (DNase I hypersensitive sites), while specificity represents for that of negative case (non-DNase I hypersensitive sites). Accuracy reflects the overall proportion of correctively predicted DNase I hypersensitive sites (DHSs) and non-DNase I hypersensitive sites (non-DHSs). The MCC is considered as a more reliable measure of the quality of binary classification. The values of MCC range from −1 to +1, +1 represents a perfect prediction, while −1 and 0 are revealed for opposite and random predictions, respectively. The four statistical measures mentioned above can be could be defined by the following set of equations:

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+} \qquad 0 \le Sn \le 1, Sp = 1 - \frac{N_+^-}{N^-} \qquad 0 \le Sp \le 1, Acc = \Lambda \end{cases}$$

$$= 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \quad 0 \le Acc \le 1, MCC$$

$$= \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \qquad -1 \le MCC \le 1 \tag{15}$$

where the total numbers of DHSs instance and non-DHSs are denoted by N^+ and N^-, respectively. The number of DHSs instances incorrectly predicted to be of non-DHSs is denoted by $N_-{}^+$, while the number of non-DHSs instances incorrectly predicted to be of DHSs is by $N_+{}^-$.

As demonstrated by many credible publications [47,49,50,58,59,63,130–142], the meanings of Sens, Spec, G-mean, MCC and OA have become a more intuitive and easier-to-understand method to measure the prediction quality when using Eq. (15).

Either the set of traditional metrics copied from math books or the intuitive metrics derived from the Chou's symbols [143–145] are valid only for the single-label systems (where each sample only belongs to one class). For the multi-label systems (where a sample may simultaneously belong to several classes), whose existence has become more frequent in system biology [146–152], system medicine [153,154] and biomedicine [155], a completely different set of metrics as defined in [156] is absolutely needed.

3. Results and discussion

3.1. The selection of the optimal features

For multi-fusion features, the selection of optimal features, that is dimensionality reduction, is a key problem. In this study, we select the optimal features based on the ranked features obtained by F-score algorithm, and we choose top 50, 100, 150, 200, 250, 300, 350, 400 features from the ranked features to calculate the Acc and Sn for $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. From Fig. 1, we can see that the ACC and Sn obtain the highest values with 86.63% and 66.43%, respectively, when we choose top 250 features for $\mathbb{S}_1$ dataset. For $\mathbb{S}_2$ dataset, the ACC and Sn obtain 87.04% and 62.35%, respectively, when we choose top 350 features, and the ACC and Sn obtain 86.94% and 63.56%, respectively, when we choose top 300 features. Comprehensively, we should choose top 300 features to be more reasonable.

Table 3
Performance of the iDHSs-MFF on the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets by jackknife test.

Dataset	Acc (%)	MCC	Sn (%)	Sp (%)
$\mathbb{S}_1$	86.63	0.65	66.43	94.30
$\mathbb{S}_2$	86.94	0.64	63.56	95.07

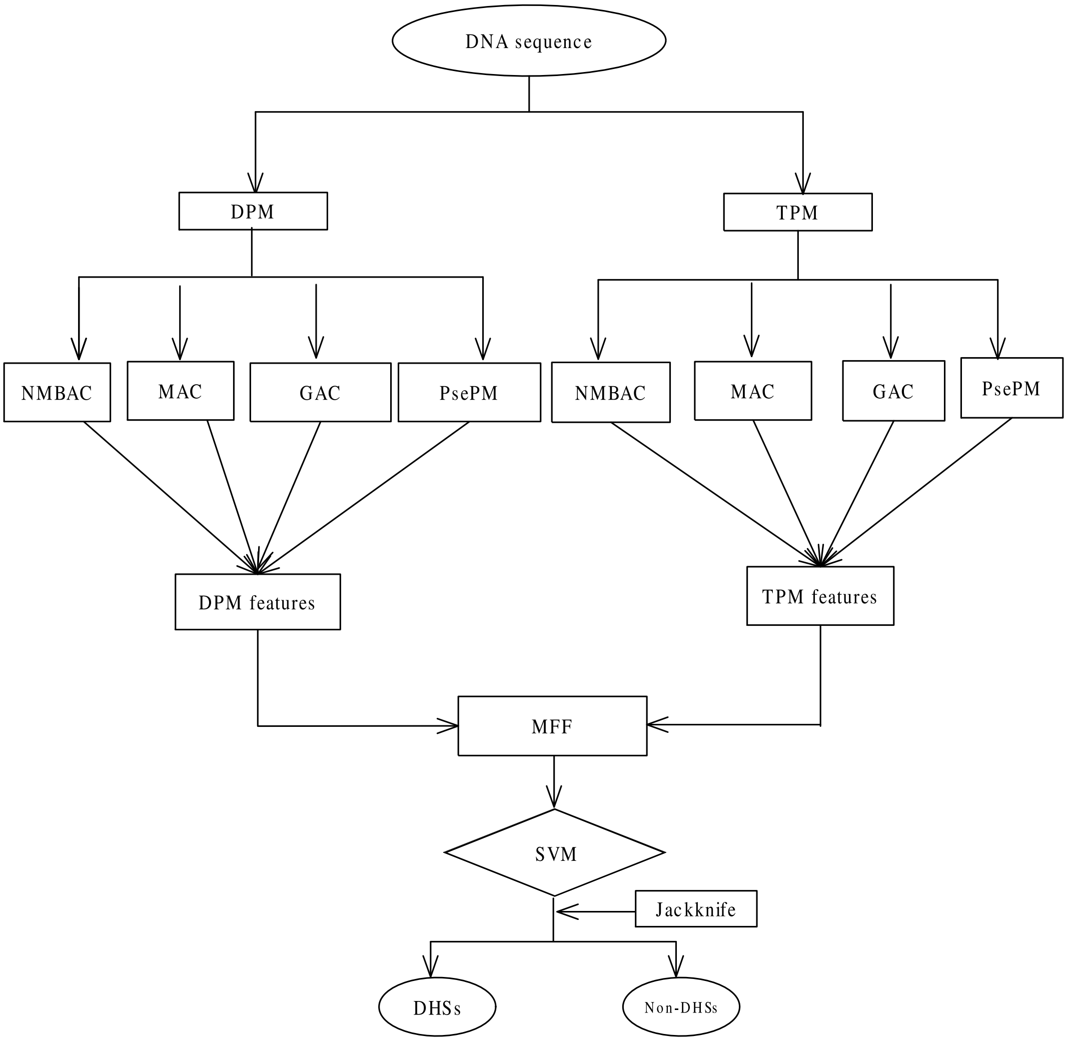


Fig. 2. The flowchart of our iDHSs-MFF model.

3.2. Prediction performance of our model

In this study, a novel predictor iDHSs-MMF is developed by aggregating multi-features and F-score. The multi-features include NMBAC-DPM, NMBAC-TPM, MAC-DPM, MAC-TPM, GAC-DPM, GAC-TPM, PseDPM and PseTPM, and a 1080-dimensional feature vector is obtained by means of these eight submodels. To avoid “the curse of the dimension”, informative features are selected by F-score approach to input SVM classifier with the RBF kernel function. The simple grid-search approach is used on the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets for finding the best parameters of C and γ. Finally, the C = 2048 and γ = 0.0039063 for $\mathbb{S}_1$, C = 5792.6188 and γ = 0.00069053 for $\mathbb{S}_2$ are found. The results obtained by the new predictor iDHSs-MMF on the two benchmark datasets via the bias-free jackknife test are given in Table 3.

Using graphic approaches to study biological and medical systems can provide an intuitive vision and useful insights for helping analyze complicated relations therein, as indicated by many previous studies on a series of important biological topics [157–170], particularly what happened is for the topics of enzyme kinetics, protein folding rates [165,171–173], and low-frequency internal motion [174,175]. The flowchart of iDHSs-MMF predictor is shown in Fig. 2.

As listed in Table 3, the accuracy reaches 86.63% and 86.94% for the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. Meanwhile, the values of Sn, Sp, and MCC reach 66.43%, 94.30%, 0.65 and 63.56%, 95.07%, 0.64 for $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. From the results of prediction performance, we can see that the ACC of $\mathbb{S}_2$ is higher than that of $\mathbb{S}_1$, however, the stability of $\mathbb{S}_1$ is better than that of $\mathbb{S}_2$. The numerical experiment results indicate that iDHSs-MMF model achieves excellent performance by fusing multiple statistical features and F-score technique.

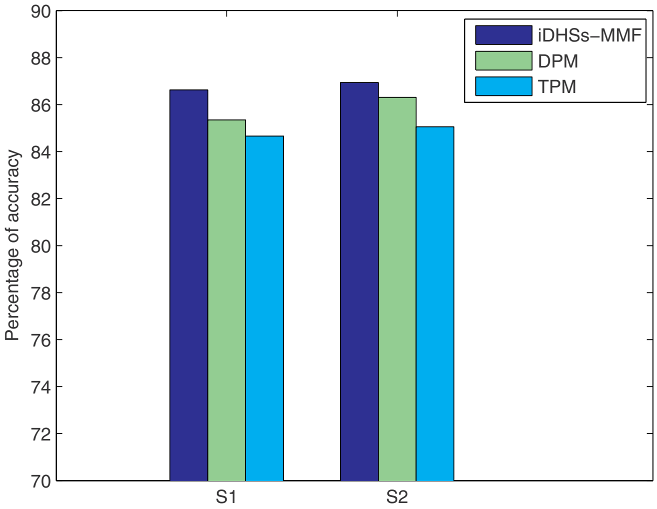


Fig. 3. Performance comparison of the features submodels based on property matrix.

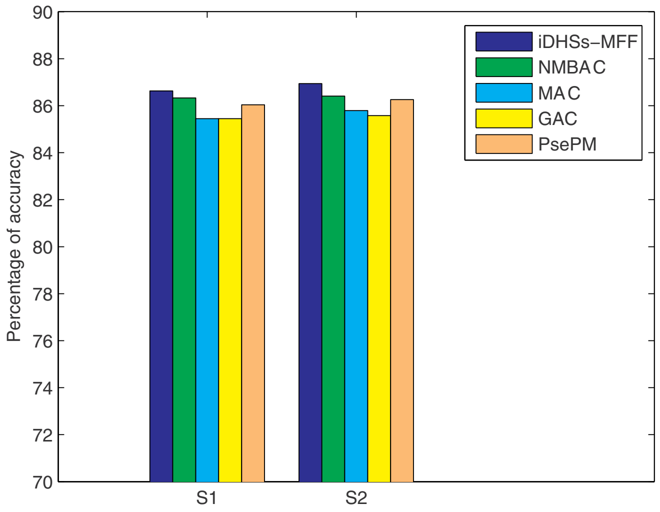


Fig. 4. Performance comparison of the features submodels based on descriptor.

Table 4
Performance comparison of different models on the $\mathbb{S}_1$ dataset.

Predictor	Acc (%)	MCC	Sn (%)	Sp (%)
SVM-RevcKmer [10]	85.25	0.62	65.36	92.81
SVM-PseDNC [11]	83.68	0.57	61.07	92.26
iDHS-EL [12]	86.14	0.64	64.64	94.30
DHSpred [13]	87.1	0.66	65.5	95.2
iDHSs-MFF	**86.63**	**0.65**	**66.43**	**94.30**

The bold value represents the value of our model iDHSs-MFF.

Table 5
Performance comparison of different models on the $\mathbb{S}_2$ dataset.

Predictor	Acc (%)	MCC	Sn (%)	Sp (%)
SVM-RevcKmer [10]	80.12	0.52	70.43	84.23
SVM-PseDNC [11]	83.00	0.57	72.12	86.78
iDHS-EL [12]	86.14	0.66	64.64	94.30
iDHSs-MFF	**86.94**	**0.64**	**63.56**	**95.07**

The bold value represents the value of our model iDHSs-MFF.

3.3. Feature submodels analysis

In this paper, eight feature submodels are proposed based on the dinucleotide property matrix (DPM) and the trinucleotide property matrix (TPM), NMBAC-DPM, MAC-DPM, GAC-DPM and PseDPM belong to DPM-based features, while NMBAC-TPM, MAC-TPM, GAC-TPM and PseTPM belong to TPM-based features. From the perspective of property matrix, the DPM-based model is obviously better than the TPM-based model for the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively, as shown in Fig. 3. That is to say, the relationship between DNase I hypersensitivity sites and dinucleotide properties is closer than that of trinucleotide. The DPM-based model and the TPM-based model complement each other for the final accuracy. From the perspective of the descriptor, NMBAC-DPM and NMBAC-TPM belong to NMBAC-based model, MAC-DPM and MAC-TPM belong to MAC-based model, GAC-DPM and GAC-TPM belong to GAC-based model, PseDPM and PseTPM belong PsePM-based model. As shown in Fig. 4, NMBAC-based model is superior to PsePM-based model, PsePM-based model is superior to MAC-based model, and also superior to GAC-based model for the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. The effect of MAC-based model is almost the same as that of GAC-based model for the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. But integrating these features, the Acc reaches 86.63% and 86.94% for the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. In words, these four descriptors also complement each other for the final accuracy.

3.4. Performance comparison with other models

In order to more conveniently illustrate the superiority of our model, we list the measures values of Acc, Sn, Sp and Mcc for our iDHSs-MFF model, SVM-RevcKmer model [10], SVM-PseDNC model [11], iDHS-EL model [12], DHSpred [13] in Tables 4 and 5, respectively, where SVM-RevcKmer is proposed by combining the DNA nucleotide composition with the SVM; SVM-PseDNC is developed by incorporating the pseudo dinucleotide composition into the DNA structural properties; iDHS-EL is formed by fusing three individual random forest (RF) classifiers into an ensemble predictor; DHSpred model is developed by selecting the optimal feature candidates using RF from a large set of features, which included nucleotide composition and di- and trinucleotide physicochemical properties. The accuracy is 1.38%, 2.95% and 1.38% higher than that obtained by SVM-RevcKmer, SVM-PseDNC and iDHS-EL models for the $\mathbb{S}_1$ dataset, respectively, and is 6.28%, 3.94% and 0.8% higher than that obtained by SVM-RevcKmer, SVM-PseDNC and iDHS-EL models for the $\mathbb{S}_2$ dataset, respectively. For DHSpred model, although ACC is 0.47% higher than that of our model, Sn is 0.93% lower than that of our model. For the $\mathbb{S}_1$ dataset, our model iDHSs-MFF is most stable compared with the other three models. For the $\mathbb{S}_2$ dataset, although our model iDHSs-MFF is slightly less stable than the iDHS-EL model, we still obtain satisfactory results. This is because we adopt the most rigorous leave-one-out test. In future research, we will further focus on improving accuracy while enhancing the stability of model.

4. Conclusions

In this paper, a feature extraction model named iDHSs-MFF is constructed by using multi-features fusion and F-score approach. These multi-features are generated via four descriptors, which include Normalized Moreau-Broto autocorrelation, Moran autocorrelation and Geary autocorrelation descriptors based on DPM and TPM, pseudo-DPM and pseudo-TPM. By combining these statistical descriptors, a 1080-dimensional feature vector is obtained for each instance on the $\mathbb{S}_1$ and $\mathbb{S}_2$ datasets, respectively. And then, F-score is carried out to select important features. The SVM with RBF kernal and the objective jackknife test are used to predict and evaluate the results, and our model provides a more accurate automated calculation method for identification of DNase I hypersensitive sites.

It is our desire to build an open platform which could provide more useful guidance for experimental workers of identification of DNase I hypersensitive sites. As pointed out in [176], user-friendly and publicly accessible web-servers represent the future direction for reporting various important computational analyses and findings [61,64,130,141,142,146–153,177–185]. Actually, they have significantly enhance the impacts of computational biology on medical science [53], driving medical science into an unprecedented revolution [88]. In our future work we shall strive to establish a web-server for the findings presented in this paper.

Acknowledgements

The authors would like to thank the anonymous reviewers and editor for their helpful comments on our manuscript. This work was supported by the National Natural Science Foundation of China (No. 11601407), the Natural Science Basic Research Plan in Shaanxi Province of China (Nos. 2018JM1037, 2019JQ-279) and the Doctoral Scientific Research Foundation of Xi'an Polytechnic University (No. BS1710).

References

[1] S. Henikoff, J.G. Henikoff, A. Sakai, G.B. Loeb, K. Ahmad, Genome-wide profiling of salt fractions maps physical properties of chromatin, Genome Res. 19 (2009) 460–469.

[2] C.Y. Jin, C.Z. Zang, G. Wei, K.R. Cui, W.Q. Peng, K.J. Zhao, F. Gary, H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions, Nat. Genet. 41 (2009) 941–945.

[3] C. Wu, P. M. Bingham, K.J. Livak, R. Holmgren, S.C.R. Elgin, The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence, Cell 16 (1979) 797–806.

[4] D.S. Gross, W.T. Garrard, Nuclease hypersensitive sites in chromatin, Annu. Rev. Biochem. 57 (1988) 159–197.

[5] G. Felsenfeld, Chromatin as an essential part of the transcriptional mechanism, Nature 355 (1992) 219–224.

[6] G. Felsenfeld, M. Groudine, Controlling the double helix, Nature 421 (2003) 448–453.

[7] G.E. Crawford, I.E. Holt, J. Whittle, B.D. Webb, D. Tai, S. Davis, E.H. Margulies, Y. Chen, J.A. Bernat, D. Ginsburg, Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS), Genome Res. 16 (2006) 123–131.

[8] S.E. Celniker, L.A.L. Dillon, M.B. Gerstein, et al., Unlocking the secrets of the genome, Nature 459 (2009) 927–930.

[9] L. Song, G.E. Crawford, DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells, Cold Spring Harb. Protocol. 2010 (2010) (pdb. prot5384).

[10] W.S. Noble, S. Kuehn, R. Thurman, M. Yu, J. Stamatoyannopoulos, Predicting the in vivo signature of human gene regulatory sequences, Bioinformatics 21 (2005) i338–i343.

[11] P.M. Feng, N. Jiang, N. Liu, Prediction of DNase I hypersensitive sites by using pseudo nucleotide compositions, Sci. World J. 2014 (2014) 740506.

[12] B. Liu, R. Long, K.C. Chou, iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework, Bioinformatics 32 (2016) 2411–2418.

[13] B. Manavalan, T.H. Shin, G. Lee, DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest, Oncotarget 9 (2017) (1944C1956).

[14] J.J. Chou, H. Matsuo, H. Duan, G. Wagner, Solution structure of the RAIDD CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment, Cell 94 (1998) 171–180.

[15] K. Oxenoid, Y.S. Dong, C. Cao, T. Cui, Y. Sancak, A.L. Markhard, Z. Grabarek, L. Kong, Z. Liu, B. Ouyang, Y. Cong, V.K. Mootha, J.J. Chou, Architecture of the mitochondrial calcium uniporter, Nature 533 (2016) 269–273.

[16] J. Dev, D. Park, Q. Fu, J. Chen, H.J. Ha, F. Ghantous, T. Herrmann, W. Chang, Z. Liu, G. Frey, M.S. Seaman, B. Chen, J.J. Chou, Structural basis for membrane anchoring of HIV-1 envelope spike, Science 353 (2016) 172–175.

[17] J.R. Schnell, J.J. Chou, Structure and mechanism of the M2 proton channel of influenza A virus, Nature 451 (2008) 591–595.

[18] M.J. Berardi, W.M. Shih, S.C. Harrison, J.J. Chou, Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching, Nature 476 (2011) 109–113.

[19] B. OuYang, S. Xie, M.J. Berardi, X.M. Zhao, J. Dev, W. Yu, B. Sun, J.J. Chou, Unusual architecture of the p7 channel from hepatitis C virus, Nature 498 (2013) 521–525.

[20] J. Wang, R.M. Pielak, M.A. McClintock, J.J. Chou, Solution structure and functional analysis of the influenza B proton channel, Nat. Struct. Mol. Biol. 16 (2009) 1267–1271.

[21] Q. Fu, T.M. Fu, A.C. Cruz, P. Sengupta, S.K. Thomas, S. Wang, R.M. Siegel, H. Wu, J.J. Chou, Structural basis and functional role of intramembrane trimerization of the Fas/CD95 death receptor, Mol. Cell 61 (2016) 602–613.

[22] J.J. Chou, H. Li, G.S. Salvessen, J. Yuan, G. Wagner, Solution structure of BID, an intracellular amplifier of apoptotic signalling, Cell 96 (1999) 615–624.

[23] J.J. Chou, S. Li, C.B. Klee, A. Bax, Solution structure of Ca2+ −calmodulin reveals flexible hand-like properties of its domains, Nat. Struct. Biol. 8 (2001) 990–997.

[24] K. Oxenoid, J.J. Chou, The structure of phospholamban pentamer reveals a channel-like architecture in membranes, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 10870–10875.

[25] M.E. Call, J.R. Schnell, C. Xu, R.A. Lutz, J.J. Chou, K.W. Wucherpfennig, The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor, Cell 127 (2006) 355–368.

[26] M.E. Call, K.W. Wucherpfennig, J.J. Chou, The structural basis for intramembrane assembly of an activating immunoreceptor complex, Nat. Immunol. 11 (2010) 1023–1029.

[27] E. Gagnon, C. Xu, W. Yang, H.H. Chu, M.E. Call, J.J. Chou, K.W. Wucherpfennig, Response multilayered control of T cell receptor phosphorylation, Cell 142 (2010) 669–671.

[28] S. Bruschweiler, Q. Yang, C. Run, J.J. Chou, Substrate-modulated ADP/ATP-transporter dynamics revealed by NMR relaxation dispersion, Nat. Struct. Mol. Biol. 22 (2015) 636–641.

[29] C. Cao, S. Wang, T. Cui, X.C. Su, J.J. Chou, Ion and inhibitor binding of the double-ring ion selectivity filter of the mitochondrial calcium uniporter, Proc. Natl Acad. Sci. 114 (2017) E2846–E2851 current issue.

[30] A. Piai, J. Dev, Q. Fu, J.J. Chou, Stability and water accessibility of the trimeric membrane anchors of the HIV-1 envelope spikes, J. Am. Chem. Soc. 139 (2017) 18432–18435.

[31] L. Pan, T.M. Fu, W. Zhao, L. Zhao, W. Chen, C. Qiu, W. Liu, Z. Liu, A. Piai, Q. Fu, S. Chen, H. Wu, J.J. Chou, Higher-order clustering of the transmembrane anchor of DR5 drives signaling, Cell 176 (2019) 1477-1489(e14).

[32] K.C. Chou, A.G. Tomasselli, R.L. Heinrikson, Prediction of the tertiary structure of a caspase-9/inhibitor complex, FEBS Lett. 470 (2000) 249–256.

[33] K.C. Chou, D. Jones, R.L. Heinrikson, Prediction of the tertiary structure and substrate binding site of caspase-8, FEBS Lett. 419 (1997) 49–54.

[34] K.C. Chou, Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor, Biochem. Biophys. Res. Commun. 319 (2004) 433–438.

[35] K.C. Chou, Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein, J. Proteome Res. 4 (2005) 1681–1686.

[36] K.C. Chou, W.J. Howe, Prediction of the tertiary structure of the beta-secretase zymogen, Biochem. Biophys. Res. Commun. 292 (2002) 702–708.

[37] K.C. Chou, Insights from modelling the tertiary structure of BACE2, J. Proteome Res. 3 (2004) 1069–1072.

[38] K.C. Chou, Insights from modelling three-dimensional structures of the human potassium and sodium channels, J. Proteome Res. 3 (2004) 856–861.

[39] K.C. Chou, Modeling the tertiary structure of human cathepsin-E, Biochem. Biophys. Res. Commun. 331 (2005) 56–60.

[40] K.C. Chou, Insights from modeling the 3D structure of DNA-CBF3b complex, J. Proteome Res. 4 (2005) 1657–1660.

[41] S.Q. Wang, Q.S. Du, Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases, Biochem. Biophys. Res. Commun. 354 (2007) 634–640.

[42] S.Q. Wang, Q.S. Du, R.B. Huang, D.W. Zhang, Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus, Biochem. Biophys. Res. Commun. 386 (2009) 432–436.

[43] X.B. Li, S.Q. Wang, W.R. Xu, R.L. Wang, Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method, PLoS One 6 (2011) e28111.

[44] Y. Ma, S.Q. Wang, W.R. Xu, R.L. Wang, Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach, PLoS One 7 (2012) e38546.

[45] Y. Xu, J. Ding, 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 8 (2013) e55844.

[46] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, X. Cheng, iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach, J. Biomol. Struct. Dyn. 33 (2015) 2221–2233.

[47] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, J. Theor. Biol. 377 (2015) 47–56.

[48] Z. Liu, X. Xiao, W.R. Qiu, iDNA-methyl: identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77.

[49] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e68.

[50] H. Lin, E.Z. Deng, H. Ding, W. Chen, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.

[51] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins Struct. Funct. Genet. 43 (2001) 246–255.

[52] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.

[53] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.

[54] P.M. Feng, W. Chen, H. Lin, iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, Anal. Biochem. 442 (2013) 118–125.

[55] W. Chen, P.M. Feng, E.Z. Deng, H. Lin, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.

[56] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, H. Lin, W. Chen, iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels, Biomed. Res. Int. 2014 (2014) 286419.

[57] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy, J. Theor. Biol. 385 (2015) 153–159.

[58] J. Jia, Z. Liu, X. Xiao, B. Liu, iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset, Anal. Biochem. 497 (2016) 48–56.

[59] J. Jia, L. Zhang, Z. Liu, X. Xiao, pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC, Bioinformatics 32 (2016) 3133–3141.

[60] B. Liu, L. Fang, R. Long, X. Lan, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.

[61] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences, Oncotarget 8 (2017) 4208–4217.

[62] W. Chen, H. Ding, X. Zhou, H. Lin, iRNA(m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition, Anal. Biochem. 561 (2018) 59–65.

[63] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites, Mol. Ther. Nucleic Acid 11 (2018) 468–474.

[64] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, J.H. Jia, iKcr-PseEns: identify lysine

crotonylation sites in histone proteins with pseudo components and ensemble classifier, Genomics 110 (2018) 239–246.

[65] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, iDNA6mA-PseKNC: identifying DNA N (6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC, Genomics 111 (2019) 96–102.

[66] W. Hussain, S.D. Khan, N. Rasool, S.A. Khan, SPalmitoylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins, Anal. Biochem. 568 (2019) 14–23.

[67] W. Hussain, Y.D. Khan, N. Rasool, S.A. Khan, SPrenylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins, J. Theor. Biol. 468 (2019) 1–11.

[68] J. Jia, X. Li, W. Qiu, X. Xiao, iPPI-PseAAC(CGR): identify protein-protein interactions by incorporating chaos game representation into PseAAC, J. Theor. Biol. 460 (2019) 195–203.

[69] Y.D. Khan, M. Jamil, W. Hussain, N. Rasool, S.A. Khan, pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments, J. Theor. Biol. 463 (2019) 47–55.

[70] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), J. Theor. Biol. 273 (2011) 236–247.

[71] C.T. Zhang, An optimization approach to predicting protein structural class from amino acid composition, Protein Sci. 1 (1992) 401–408.

[72] K.C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors, J. Proteome Res. 1 (2002) 429–433.

[73] K.C. Chou, Y.D. Cai, Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition, J. Cell. Biochem. 90 (2003) 1250–1260.

[74] L. Hu, T. Huang, X. Shi, W.C. Lu, Y.D. Cai, Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties, PLoS One 6 (2011) e14556.

[75] Y.D. Cai, K.Y. Feng, W.C. Lu, Using LogitBoost classifier to predict protein structural classes, J. Theor. Biol. 238 (2006) 172–176.

[76] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

[77] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.

[78] M. Behbahani, H. Mohabatkar, M. Nosrati, Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition, J. Theor. Biol. 411 (2016) 1–5.

[79] M. Kabir, M. Hayat, iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples, Mol. Gen. Genomics. 291 (2016) 285–296.

[80] P.K. Meher, T.K. Sahu, V. Saini, A.R. Rao, Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC, Sci. Rep. 7 (2017) 42362.

[81] Z. Ju, J.J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC, J. Mol. Graph. Model. 76 (2017) 356–363.

[82] B. Yu, S. Li, W.Y. Qiu, C. Chen, R.X. Chen, L. Wang, M.H. Wang, Y. Zhang, Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising, Oncotarget 8 (2017) 107640–107665.

[83] J. Ahmad, M. Hayat, MFSC: multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components, J. Theor. Biol. 463 (2018) 99–109.

[84] S. Akbar, M. Hayat, iMethyl-STTNC: identification of N(6)-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences, J. Theor. Biol. 455 (2018) 205–211.

[85] E. Contreras-Torres, Predicting structural classes of proteins by incorporating their global and local physicochemical and conformational properties into general Chou's PseAAC, J. Theor. Biol. 454 (2018) 139–145.

[86] S.L. Zhang, Y.Y. Liang, Predicting apoptosis protein subcellular localization by integrating auto-cross correlation and PSSM into Chou's PseAAC, J. Theor. Biol. 457 (2018) 163–169.

[87] M. Tahir, M. Hayat, S.A. Khan, iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition, Mol. Gen. Genomics. 294 (2019) 199–210.

[88] K.C. Chou, An unprecedented revolution in medicinal chemistry driven by the progress of biological science, Curr. Top. Med. Chem. 17 (2017) 2337–2358.

[89] H.B. Shen, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.

[90] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo amino acid compositions, Anal. Biochem. 425 (2012) 117–119.

[91] D.S. Cao, Q.S. Xu, Y.Z. Liang, Propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.

[92] P. Du, S. Gu, Y. Jiao, PseAAC-general: fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (2014) 3495–3506.

[93] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteomics 6 (2009) 262–274.

[94] W. Chen, H. Lin, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, Mol. BioSyst. 11 (2015) 2620–2634.

[95] B. Liu, F. Yang, D.S. Huang, iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC, Bioinformatics 34 (2018) 33–40.

[96] M. Tahir, H. Tayara, K.T. Chong, iRNA-PseKNC(2methyl): identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components, J. Theor. Biol. 465 (2019) 1–6.

[97] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.

[98] B. Liu, H. Wu, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (2017) 67–91.

[99] S.C. Parker, L. Hansen, H.O. Abaan, T.D. Tullius, E.H. Margulies, Local DNA topography correlates with functional noncoding regions of the human genome, Science 324 (2009) 389–392.

[100] S.X. Zhang, J.H. Li, L. Su, Z.P. Zhou, pDHS-DSET: Prediction of DNase I hypersensitive sites in plant genome using DS evidence theory, Anal. Biochem. 564 (2019) 54–63.

[101] L.C. Zhang, L. Kong, iRSpot-ADPM: identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components, J. Theor. Biol. 441 (2018) 1–8.

[102] G. Moreau, P. Broto, Autocorrelation of molecular structures, application to SAR studies, New J. Chem. 4 (1980) 757–764.

[103] P.A. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1950) 17–23.

[104] R.C. Geary, The contiguity ratio and statistical mapping, Inc. Stat. 5 (1954) 115–145.

[105] H.B. Shen, K.C. Chou, NUC-PLOC: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM, Protein Eng. Des. Sel. 20 (2007) 561–567.

[106] M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Syst. Appl. 36 (2009) 3240–3247.

[107] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, D.C. Yang, mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides, Int. J. Mol. Sci. (8) (2019) 1964.

[108] L. Wei, S. Luan, L.A.E. Nagai, R. Su, Q. Zou, Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species, Bioinformatics 35 (2019) 1326–1333.

[109] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/bty1047.

[110] V. Vapnik, Statistical Learning Theory, Wiley, NewYork, 1998.

[111] T.G. Liu, X.Q. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, Biochimie 92 (2010) 1330–1334.

[112] J.R. Wang, C. Wang, J.J. Cao, X.Q. Liu, Y.H. Yao, Q. Dai, Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features, Gene 554 (2015) 241–248.

[113] T.G. Liu, X.Q. Zheng, C.H. Wang, J. Wang, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation, Protein Pept. Lett. 17 (2010) 1263–1269.

[114] K.P. Exarchos, C. Papaloukas, T.P. Exarchos, A.N. Troganis, D.I. Fotiadis, Prediction of cis/trans isomerization using feature selection and support vector machines, J. Biomed. Inform. 42 (2009) 140–149.

[115] J.Y. Yang, X. Chen, Improving taxonomy-based protein fold recognition by using global and local features, Proteins 79 (2011) 2053–2064.

[116] L.Z. Yu, Y.Z. Guo, Y.Z. Li, G.B. Li, M.L. Li, J.S. Luo, W.J. Xiong, W.L. Qin, SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition, J. Theor. Biol. 267 (2010) 1–6.

[117] Y.N. Zhang, D.J. Yu, S.S. Li, Y.X. Fan, Y. Huang, H.B. Shen, Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features, BMC Bioinforma. 13 (2012) 1–11.

[118] Y.C. Dou, B. Yao, C. Zhang, PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine, Amino Acids 46 (2014) 1459–1469.

[119] S. Niu, L.L. Hu, L.L. Zheng, Predicting protein oxidation sites with feature selection and analysis approach, J. Biomol. Struct. Dyn. 29 (2012) 650–658.

[120] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, (2001).

[121] K.C. Chou, C.T. Zhang, Review: prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[122] K.C. Chou, H.B. Shen, Review: recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[123] G.L. Fan, 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, J. Theor. Biol. 304 (2012) 88–95.

[124] S.Y. Ding, S.L. Zhang, A gram-negative bacterial secreted protein types prediction method based on PSI-BLAST profile, Biomed. Res. Int. 3206741 (2016) 1–5.

[125] M. Kabir, D.J. Yu, Predicting DNase I hypersensitive sites via un-biased pseudo trinucleotide composition, Chemom. Intell. Lab. Syst. 167 (2017) 78–84.

[126] B.Q. Liu, Y.M. Liu, X.P. Jin, X.L. Wang, B. Liu, iRSpot-DACC: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance, Sci. Rep. 6 (2016) 33483.

[127] S. Basith, B. Manavalan, T.H. Shin, G. Lee, iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree, Comput. Struct. Biotechnol. J. 16 (2018) 412–420.

[128] R. Su, J. Hu, Q. Zou, B. Manavalan, L. Wei, Empirical comparison and analysis of

web-based cell-penetrating peptide prediction tools, Brief. Bioinform. (2019), https://doi.org/10.1093/bib/bby124.

[129] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783.

[130] B. Liu, S. Wang, R. Long, iRSpot-EL: identify recombination spots with an ensemble learning approach, Bioinformatics 33 (2017) 35–41.

[131] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.

[132] H. Yang, W.R. Qiu, G. Liu, F.B. Guo, W. Chen, H. Lin, iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC, Int. J. Biol. Sci. 14 (2018) 883–891.

[133] J. Jia, Z. Liu, X. Xiao, B. Liu, Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC), J. Biomol. Struct. Dyn. 34 (2016) 1946–1961.

[134] J. Jia, Z. Liu, X. Xiao, B. Liu, pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach, J. Theor. Biol. 394 (2016) 223–230.

[135] J. Jia, Z. Liu, X. Xiao, B. Liu, iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC, Oncotarget 7 (2016) 34558–34570.

[136] J. Jia, Z. Liu, X. Xiao, B. Liu, iPPBS-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets, Molecules 21 (2016) E95.

[137] Z. Liu, X. Xiao, D.J. Yu, J. Jia, W.R. Qiu, pRNAm-PC: predicting N-methyladenosine sites in RNA sequences via physical-chemical properties, Anal. Biochem. 497 (2016) 60–67.

[138] X. Xiao, H.X. Ye, Z. Liu, J.H. Jia, iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition, Oncotarget 7 (2016) 34180–34189.

[139] J. Song, Y. Wang, F. Li, T. Akutsu, N.D. Rawlings, G.I. Webb, iProt-sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites, Brief. Bioinform. (2018), https://doi.org/10.1093/bib/bby028.

[140] W. Chen, H. Tang, J. Ye, H. Lin, iRNA-PseU: identifying RNA pseudouridine sites, Mol. Ther. Nucleic Acids 5 (2016) e332.

[141] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, iRNA-PseColl: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC, Mol. Ther. Nucleic Acids 7 (2017) 155–163.

[142] B. Liu, F. Yang, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function, Mol. Ther. Nucleic Acids 7 (2017) 267–277.

[143] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins Struct. Funct. Genet. 42 (2001) 136–139.

[144] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.

[145] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.

[146] X. Cheng, X. Xiao, pLoc-mPlant: predict subcellular localization of multi-location plant proteins via incorporating the optimal GO information into general PseAAC, Mol. BioSyst. 13 (2017) 1722–1727.

[147] X. Cheng, X. Xiao, pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC, Gene 628 (2017) 315–321.

[148] X. Xiao, X. Cheng, G. Chen, Q. Mao, Ploc–Bal-Mgpos: predict subcellular localization of gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC, Genomics (2018), https://doi.org/10.1016/J.Ygeno.2018.05.017.

[149] X. Cheng, S.G. Zhao, W.Z. Lin, X. Xiao, pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites, Bioinformatics 33 (2017) 3524–3531.

[150] X. Xiao, X. Cheng, S. Su, Q. Nao, pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins, Nat. Sci. 9 (2017) 331–349.

[151] X. Cheng, X. Xiao, pLoc-mGneg: predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general PseAAC, Genomics 110 (2018) 231–239.

[152] X. Cheng, X. Xiao, pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC, Genomics 110 (2018) 50–58.

[153] X. Cheng, S.G. Zhao, X. Xiao, iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals, Bioinformatics 33 (2017) 341–346.

[154] X. Cheng, S.G. Zhao, X. Xiao, iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals, Oncotarget 8 (2017) 58494–58503.

[155] W.R. Qiu, B.Q. Sun, X. Xiao, Z.C. Xu, iPTM-mLys: identifying multiple lysine PTM sites and their different types, Bioinformatics 32 (2016) 3116–3123.

[156] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. BioSyst. 9 (2013) 1092–1100.

[157] K.C. Chou, S.P. Jiang, W.M. Liu, C.H. Fee, Graph theory of enzyme kinetics: 1. Steady-state reaction system, Sci. Sinica 22 (1979) 341–358.

[158] K.C. Chou, S. Forsen, Graphical rules for enzyme-catalyzed rate laws, Biochem. J. 187 (1980) 829–835.

[159] K.C. Chou, S. Forsen, G.Q. Zhou, Three schematic rules for deriving apparent rate constants, Chem. Scr. 16 (1980) 109–113.

[160] K.C. Chou, R.E. Carter, S. Forsen, A new graphical method for deriving rate equations for complicated mechanisms, Chem. Scr. 18 (1981) 82–86.

[161] K.C. Chou, S. Forsen, Graphical rules of steady-state reaction systems, Can. J. Chem. 59 (1981) 737–755.

[162] G.P. Zhou, M.H. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways, Biochem. J. 222 (1984) 169–176.

[163] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics, J. Biol. Chem. 264 (1989) 12074–12079.

[164] I.W. Althaus, J.J. Chou, A.J. Gonzales, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E, J. Biol. Chem. 268 (1993) 6119–6124.

[165] K.C. Chou, Review: applications of graph theory to enzyme kinetics and protein folding kinetics, steady and non-steady state systems, Biophys. Chem. 35 (1990) 1–24.

[166] I.W. Althaus, A.J. Gonzales, J.J. Chou, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase, J. Biol. Chem. 268 (1993) 14875–14880.

[167] K.C. Chou, Graphic rule for drug metabolism systems, Curr. Drug Metab. 11 (2010) 369–378.

[168] G.P. Zhou, The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism, J. Theor. Biol. 284 (2011) 142–148.

[169] I.W. Althaus, J.J. Chou, A.J. Gonzales, M.R. Diebel, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E, Biochemistry 32 (1993) 6548–6554.

[170] K.C. Chou, W.Z. Lin, X. Xiao, Wenxiang: a web-server for drawing wenxiang diagrams, Nat. Sci. 3 (2011) 862–865.

[171] K.C. Chou, S. Forsen, Diffusion-controlled effects in reversible enzymatic fast reaction system: critical spherical shell and proximity rate constants, Biophys. Chem. 12 (1980) 255–263.

[172] K.C. Chou, T.T. Li, S. Forsen, The critical spherical shell in enzymatic fast reaction systems, Biophys. Chem. 12 (1980) 265–269.

[173] H.B. Shen, J.N. Song, Prediction of protein folding rates from primary sequence by fusing multiple sequential features, J. Biomed. Sci. Eng. 2 (2009) 136–143.

[174] K.C. Chou, N.Y. Chen, S. Forsen, The biological functions of low-frequency phonons: 2. Cooperative effects, Chem. Scr. 18 (1981) 126–132.

[175] K.C. Chou, Review: low-frequency collective motion in biomacromolecules and its biological functions, Biophys. Chem. 30 (1988) 3–48.

[176] K.C. Chou, H.B. Shen, Recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (2009) 63–92.

[177] X. Cheng, X. Xiao, pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information, Bioinformatics 34 (2018) 1448–1456.

[178] W.R. Qiu, S.Y. Jiang, Z.C. Xu, X. Xiao, iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition, Oncotarget 8 (2017) 41178–41188.

[179] W.R. Qiu, B.Q. Sun, X. Xiao, D. Xu, iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory, Mol. Inform. 36 (2017) (UNSP 1600010).

[180] X. Cheng, X. Xiao, pLoc–bal-mGneg: predict subcellular localization of gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC, J. Theor. Biol. 458 (2018) 92–102.

[181] X. Cheng, X. Xiao, pLoc–bal-mPlant: predict subcellular localization of plant proteins by general PseAAC and balancing training dataset, Curr. Pharm. Des. 24 (2018) 4013–4022.

[182] K.C. Chou, X. Cheng, X. Xiao, pLoc–bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset, Genomics (2018), https://doi.org/10.1016/j.ygeno.2018.08.007.

[183] X. Xiao, X. Cheng, G. Chen, Q. Mao, pLoc–bal-mVirus: predict subcellular localization of multi-label virus proteins by PseAAC and IHTS treatment to balance training dataset, Med. Chem. 15 (2018) 1–14.

[184] X. Cheng, W.Z. Lin, X. Xiao, pLoc–bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC, Bioinformatics 35 (2019) 398–406.

[185] K.C. Chou, X. Cheng, X. Xiao, pLoc–bal-mEuk: predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset, Med. Chem. 15 (2019) 1–14.