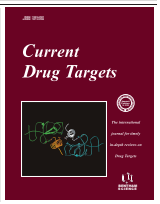
Current Drug Targets, 2019, *20*, 540-550

REVIEW ARTICLE



A Survey for Predicting Enzyme Family Classes Using Machine Learning Methods

Jiu-Xin Tan[1], Hao Lv[1], Fang Wang[1], Fu-Ying Dao[1], Wei Chen[1,2,3,*] and Hui Ding[1,*]

[1]*Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China;* [2]*Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China;* [3]*Gordon Life Science Institute, Boston, MA 02478, USA*

ARTICLE HISTORY

Received: July 21, 2018
Revised: August 17, 2018
Accepted: September 04, 2018

DOI:
10.2174/1389450119666181002143355

Abstract: Enzymes are proteins that act as biological catalysts to speed up cellular biochemical processes. According to their main Enzyme Commission (EC) numbers, enzymes are divided into six categories: EC-1: oxidoreductase; EC-2: transferase; EC-3: hydrolase; EC-4: lyase; EC-5: isomerase and EC-6: synthetase. Different enzymes have different biological functions and acting objects. Therefore, knowing which family an enzyme belongs to can help infer its catalytic mechanism and provide information about the relevant biological function. With the large amount of protein sequences influxing into databanks in the post-genomics age, the annotation of the family for an enzyme is very important. Since the experimental methods are cost ineffective, bioinformatics tool will be a great help for accurately classifying the family of the enzymes. In this review, we summarized the application of machine learning methods in the prediction of enzyme family from different aspects. We hope that this review will provide insights and inspirations for the researches on enzyme family classification.

Keywords: Enzyme, family, classification, machine learning methods.

1. INTRODUCTION

Enzymes, also known as biocatalysts, are a kind of proteins that have the high degree of specificity and catalytic efficiency in living cells. They are found in various cells, where they catalyze almost all chemical reactions of life processes ranging from cell growth to cell metabolism. The high efficiency, specificity, diversity and mild reaction conditions of enzymes can make the process of material metabolism in cells methodical. According to the EC number [1], enzymes can be generally classified into six families as shown in Fig. **1** : (I) oxidoreductase, (II) transferase, (III) hydrolase, (IV) lyase, (V) isomerase and (VI) synthetase.

Enzyme family classes can be determined through wet-experimental method, but the method is cost-ineffective and time-consuming. Therefore, based on their sequence information, a series of machine learning methods have been proposed for enzyme family classification. As early as 2002, Jensen *et al.* firstly predicted the enzyme families based on Artificial Neural Networks (ANN) through combining primary sequence information incorporating with sequence-related physicochemical features [2]. Later, Chou *et al.* published their technique called 'Go-PseAAC predictor', which used the nearest neighbor (NN) method to distinguish enzyme families with features vector consisting of gene product composition and pseudo amino acid composition (PseAAC) [3]. In the same year, Cai *et al.* proposed a support vector machine (SVM)-based technique for enzyme family classification [4]. Subsequently, they proposed a new method using protein functional domain composition (FunD) [5], FunD combined with PseAAC [6], and gene ontology (GO) combined with PseAAC to classify the enzyme families [7]. In 2007, by using FunD, Lu *et al.* developed an SVM-based method to investigate the enzyme family [8]. By fusing FunD with evolution information, Shen *et al.* developed a predictor named EzyPred [9]. Afterward, in 2009, Nasibov *et al.* used the frequency of the amino acid residue to represent the enzyme sequences to classify the enzyme families by adopting the KNN classifier with minimum distanced-based classifier [10]. In the same year, Concu *et al.* published two papers introducing the enzyme families classification based on 3D structure [11, 12]. In 2010, Qiu *et al.* conducted a new feature extraction method by using PseAAC and discrete wavelet transform information, and obtained good results based on SVM [13]. By using amino acid composition (AAC) and dipeptide composition (DC), low-frequency power spectral density and increment of diversity (ID) were calculated to represent the enzyme sequences for classifying

*Address correspondence to these authors at the Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China; Tel/Fax: +86-28-8320-8238; E-mail: hding@uestc.edu.cn and Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China; Tel/Fax: +86-315-3725715; E-mail: chenweiimu@gmail.com

1873-5592/19 $58.00+.00

enzyme families [14]. In addition, Volpato *et al.* adopted the N-to-1 Neural Networks method to predict enzyme family classes [15]. In 2015, Niu *et al.* introduced a new method based on the protein-protein network to predict enzyme family classes [16]. In 2016, Wu *et al.* adopted the SVM to identify human enzyme family classes by incorporating rigidity, flexibility and irreplaceability of amino acids into PseAAC [17]. All these studies did achieve encouraging results, which could provide important clues for enzyme function annotation and enzyme-related drug design.

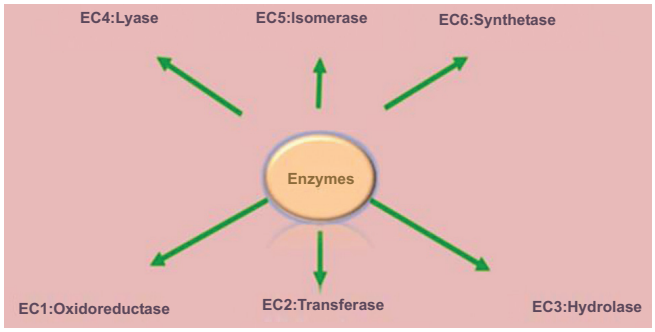


Fig. (1). Schematic drawing to show the 6 families of enzymes.

To provide scholars a whole background on enzyme family class prediction, we summarized recent development of machine learning methods on enzyme family classification in the following four aspects as shown in (Fig. **2**): *e.g.* (1) benchmark datasets; (2) feature expression; (3) classifier algorithms; (4) performance evaluation.

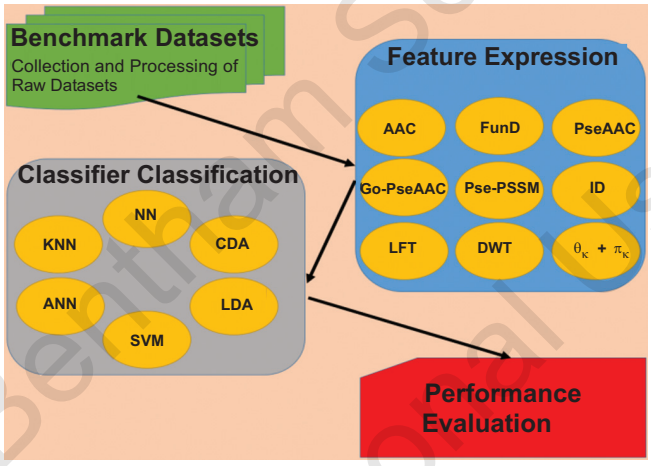


Fig. (2). The flow diagram for the enzyme classification.

2. BENCHMARK DATASETS

The ENZYME database (https://enzyme.expasy.org/) [18], an information library for the enzyme naming database, provided the EC numbers of enzymes. The UniProt (http://www.uniprot.org/) [19] is a protein database annotated and maintained by the European Institute of bioinformatics (EBI). It collected protein sequences, reference information, taxonomy information, and annotation. Users can get EC numbers from the ENZYME database, and then obtain the corresponding protein sequences from the UniProt database according to the EC numbers. Based on the two databases, several benchmark datasets focusing on enzyme families can be constructed.

Establishing a set of high quality and reliable benchmark datasets is the first step for most of the computational studies [20-25]. The following steps have been widely used to build an objective benchmark dataset for enzyme classification.

(1) Removing the sequence containing irregular characters such as B, J, O, U, X, Z.

(2) Removing the sequence whose length is less than a certain number of residues.

(3) Removing redundant sequences by using PISCES [26, 27] or CD-HIT [28, 29].

After following these rigorous steps, 12 highly reliable benchmark datasets of the enzyme families have been constructed. Each of these benchmark datasets consists of six families: oxidoreductase (S1), transferase (S2), hydrolase (S3), lyase (S4), isomerase (S5) and synthetase (S6), respectively. Details for these datasets were summarized in Table **1**.

3. FEATURE EXPRESSION METHODS

After the benchmark datasets were built, the effective feature vectors should be extracted to formulate enzyme samples which can be handled by machine learning methods. Here, we summarized them as follows:

3.1. Amino Acid Composition (AAC)

The amino acid composition (AAC) [30-32] is the most widely used method to formulate a protein sample **P**, and can be formulated as:

$$\begin{cases} \mathbf{P} = [a_1, a_2, \ldots a_i, \ldots a_{20}]^T \\ a_i = \frac{n_i}{N} \ (1 \le i \le 20) \end{cases} \quad (1)$$

where n_i represents the number of amino acid residue *i*. *N* represents the total number of amino acid residues in an enzyme sequence, *T* denotes transport.

3.2. Function Domain Composition (FunD)

Proteins usually consist of one or more functional domains. Thus, the functional domains were used as features to predict enzyme families. Currently, several FunD databases have been established, such as SMART [33], KOG and COG [34], CDD [35]. By using program IPRSCAN [36] to retrieve functional domain information of enzymes in the InterPro database, a given enzyme **P** can be transferred to a 7785-D (dimensional) vector as follows:

$$\mathbf{P} = [p_1, p_2, \ldots p_i, \ldots p_{7785}]^T \quad (2)$$

where:

$$p_i = \begin{cases} 1, hit\ found\ in\ the\ IntroPro\ database \\ 0, otherwise \end{cases} \quad (3)$$

3.3. Pseudo Amino Acid Composition (PseAAC)

By adding these spatial structure and physicochemical properties into the amino acid frequency, the PseAAC was developed to formulate enzyme sequences [37]. The feature

Table 1. A list of benchmark datasets for 6 enzyme family classes.

Datasets	Published	Number of Each Enzyme Family Class						Total Number	Sequence	References
	Year	S1	S2	S3	S4	S5	S6		Identity	
E1	2004	560	980	945	285	149	176	3095	≤ 25%	[3]
E2	2005	153	290	385	82	33	57	1000	≤ 20%	[5]
E3	2005	1697	2582	2902	939	503	840	9463	≤ 40%	[6]
E4	2005	2314	3653	3246	1307	676	1324	12520	≤ 60%	[7]
E5	2007	436	832	741	170	114	150	2443	≤ 20%	[8]
E6	2007	1820	2847	3279	892	639	965	10442	≤ 40%	[9, 14]
E7	2009	200	200	200	200	200	200	1200	Not report	[10]
E8	2013	945	2110	2226	208	136	445	6081	≤ 30%	[15]
E9	2010	200	200	200	200	200	200	1200	≤ 40%	[13]
E10	2009	151	178	223	85	74	100	790	Not report	[11]
E11	2009	154	178	223	87	64	104	810	Not report	[12]
E12	2016	155	361	404	44	35	118	1117	≤ 30%	[17]

extraction strategy includes not only AAC in the sequence but also the correlations of physicochemical properties between two residues. It can balance between the sequence composition and physicochemical properties correlation as well, greatly enhancing the richness and expressiveness of the features. Hence, PseAAC has been widely used in many fields of computational proteomics [38-51].

Thus, a sample protein **P** can be described by a (20+γ)-dimensional vector as follows:

$$\mathbf{P} = [p_1, \cdots p_{20}, p_{20+1}, \cdots p_{20+\gamma}]^T \quad (4)$$

where:

$$p_\theta = \begin{cases} \frac{f_\theta}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\gamma} \tau_j} & (1 \le \theta \le 20) \\ \frac{\omega \tau_\theta}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\gamma} \tau_j} & (21 \le \theta \le 20+\gamma) \end{cases} \quad (5)$$

where f_θ represents the normalized frequency of 20 native amino acid residues. ω represents the weight factor and τ_j represents the *j*-tier sequence correlation factor and can be calculated by the following equations:

$$\tau_j = \frac{1}{L-j} \sum_{i=1}^{l-j} F(R_i, R_{i+1}), (j < L) \quad (6)$$

where τ_j represents the *j*-th tire correlation factor that reflects the sequence order correlation between all the *j*-th residues along a protein sequence. $F(R_i, R_{i+1})$ can be calculated as follows.

$$F(R_i, R_j) = \frac{1}{k}\left\{\left[H_1(R_j) - H_1(R_i)\right]^2 + \cdots + \left[H_k(R_j) - H_k(R_i)\right]^2\right\} \quad (7)$$

where *k* represents the number of factors; $H_1(R_j)$ represents the *l*-th physiochemical properties of the residue R_i, which can be calculated by:

$$H_1(R_j) = \frac{h_0^1(R_i) - \sum_{i=1}^{20}(h_0^1(R_i)/20)}{\sqrt{\frac{\sum_{i=1}^{20}\left[h_0^1(R_i) - \sum_{i=1}^{20}(h_0^1(R_i)/20)\right]^2}{20}}} \quad (8)$$

where $h_0^1(R_i)$ represents the original value of the *i*-th amino acid residue R_i. (*i*=1, 2, ⋯, 20) is the 20 native amino acid based on the alphabetical order [52, 53].

3.4. Gene Product Composition and Pseudo Amino Acid Composition (GO-PseAAC)

This feature expression method was established to represent a protein sequence by fusing the gene product composition and pseudo amino acid composition. And it can be obtained by the following steps:

(1) A list of data named "InterPro2Go" (ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go/) was obtained by mapping of InterPro (http://www.ebi.ac.uk/interpro) entries to the GO database. In "InterPro2Go", every entry of InterPro corresponds to one GO number.

(2) Because the GO numbers in InerProt2GO do not increase continuously and orderly, they need to be renumbered by adopting organization and compression procedure [54]. Then the GO_compress database was obtained from GO database, and the dimensions of the GO_compress database were reduced from 46416 to 1930.

(3) Based on each of the 1930 entries in the GO_compress database, a vector was established to represent a protein P by using program IPRSCAN [36] to retrieve each of the protein in the GO_compress database. Thus, a protein **P** can be formulated as:

$$\mathbf{P} = [a_1, a_2, \ldots a_i, \ldots a_{1930}]^T \quad (9)$$

where a_i= 1 if there is a hit for the *i*-th entry of the GO_compress database; otherwise, a_i = 0.

(4) If there is no hit for all the 1930 entrances of the GO_compress database, the protein can be represented by PseAAC method (details in Section 3.3).

3.5. Pseudo Position-Specific Scoring Matrix (Pse-PSSM)

The PSSM was obtained from the evolution information of proteins. Using 0.001 as the E-value cutoff, the PSSM were calculated by using PSI-BLAST [55] to search the Swiss-Prot database *via* three iterations for multiple sequence alignment against the protein **P** sequence [9]. Shen *et al.* adopted a Pse-PSSM method by combining PseAAC with PSSM to represent a protein sequence in order to avoid the loss of the sequence-order information [9]. By using this approach, the Pse-PSSM can be described as:

$$\mathbf{P}_{PsePSSM}^{\mu} = \left[\delta_1^{\mu}, \delta_2^{\mu}, \ldots, \delta_i^{\mu}, \ldots, \delta_{20}^{\mu}\right] \quad (10)$$

where

$$\delta_i^{\mu} = \frac{\sum_{j=1}^{L-\mu}\left[V_{j\to i} - V_{(j+\mu)\to i}\right]^2}{L-\mu}, (i = 1, 2, \ldots, 20; \mu < L) \quad (11)$$

where δ_i^{μ} represents the correlation factor of amino acid residue *i*, and its continuous distance along the protein sequence is μ; $V_{j\to i}$ represents the normalized score obtained by PSI-BLAST; When μ =0, δ_i^0 represents the average score of the amino acid residues in protein **P**, which changes to amino acid residue *i* during the evolution process. By calculating the value μ of the best prediction accuracy when μ varied from 0 to *L*, a protein **P** can be described by using Pse-PSSM:

$$\mathbf{P}_{PsePSSM} = \left[\delta_1^0, \delta_2^0, \ldots, \delta_{20}^0, \delta_1^1, \delta_2^1, \ldots, \delta_{20}^1, \ldots, \delta_1^{\mu}, \delta_2^{\mu}, \ldots, \delta_{20}^{\mu}\right] \quad (12)$$

3.6. Increment of Diversity (ID)

Laxton gave a definition of diversity in 1978 [56], and the algorithm is defined in reference [56, 57]. The algorithm was firstly applied in protein structural class prediction by Li *et al.* [58], in which the ID was used as the sole parameter of protein structural class prediction, and achieved good results.

Given the state space of *t* dimension, m_i represents the number of the *i*-th state, and the diversity source *S:*($m_1, m_2 \ldots m_t$) is formulated as:

$$D_S = M\, log\, M - \sum_i m_i\, log\, m_i \quad (13)$$

Given two state space of *t* dimension, ID between the sources of diversity X: $(m_1, m_2 \ldots m_t)$ and Y: $(m_1, m_2 \ldots m_t)$ is formulated as:

$$ID(X,Y) = D(X+Y) - D(X) - D(Y) \quad (14)$$

It can also be rewritten as:

$$ID(X,Y) = D(M,N) - \sum_i D(m_i, n_i) \quad (15)$$

where

$$\begin{cases} D(M,N) = (M+N)\, log(M+N) - M\, log\, M - N\, log\, N \\ D(m_i,n_i) = (m_i+n_i)\, log(m_i+n_i) - m_i\, log\, m_i - n_i\, log\, n_i \end{cases} \quad (16)$$

where

$$\begin{cases} M = \sum_i m_i \\ N = \sum_i n_i \end{cases} \quad (17)$$

If m_i or n_i equals 0, $D(m_i, n_i)$=0. In the refrence [14], Shi converted the 400 dipeptides composition (DC) [59] to a 400 dimension space described as the follows:

$$S: (m_1, m_2, \ldots m_i, \ldots m_{400}) \quad (18)$$

where m_i represents the absoluted occurrence frequencies of the 400 dipeptides.

3.7. Low-frequency Fourier Transform (LFT)

Suppose a protein sequence **P** with *L* amino acids, it can be represented as follows:

$$\mathbf{P} = [R_1, R_2, \ldots R_i, \ldots R_L]^T \quad (19)$$

where R_i represents the amino acid residue at the polypeptide chain position *i* ($1 \leqslant i \leqslant L$). If using the hydrophilic value $h(R_i)$ [60] of the amino acid to describe the protein sequence, it can be formulated as follows：

$$\mathbf{P} = [h(R_1), h(R_2), \ldots h(R_i), \ldots h(R_L)]^T \quad (20)$$

Then it can be transformed into discrete Fourier sequence from the digital sequence of Eq.(21) as follows:

$$\mathbf{P} = [F_1, F_2, \ldots F_K, \ldots F_L]^T\ (K = 1,2, \ldots, L) \quad (21)$$

where

$$F_K = \sum_{i=1}^{L} h(R_i) e^{-j2\pi(n-1)(\frac{i-1}{L})} \quad (22)$$

Then the discrete Fourier sequence of Eq.(22) can be transformed into power spectral density as follows:

$$\mathbf{P} = [V_1, V_2, \ldots V_K, \ldots V_L]^T \quad (23)$$

where

$$V_K = lim_{C\to\infty} \frac{|F_K|^2}{C} \quad (24)$$

In the reference [14], by using discrete Fourier transform, Shi *et al.* obtained 512 frequency points from the digital sequence with different length, and then obtained the power spectral density. However, for all 512 power spectral density values, the low-frequency components are more important than high-frequency components with much more noisy [61-66]. Therefore, Shi *et al.* only needs to consider the λ low-frequency components and finally selected λ=16 as the best choice based on the statistical calculations. In addition, V_1 is excluded as it is a particular value that has nothing to do with the sequence information. Thus an enzyme sequence **P** can be described by 15 power spectral density values formulated as

$$\mathbf{P} = [V_2, V_3, V_4, \ldots V_{16}]^T \quad (25)$$

3.8. Discrete Wavelet Transform (DWT)

Nowadays, wavelet analysis has been widely used in the realm of bioinformatics, such as genome sequence analysis [67], gene expression data [68], protein structure prediction [69-71].

Based on DTW analysis, enzyme families can be predicted by breaking down the amino acid sequence into coefficients at different dilations, then removed the noisy component from profile. Therefore, DWT analysis can provide more effective and brief feature data that can reflect the

characteristics of the enzyme families. DWT was briefly described as:

(1) First, the amino acid residues were transformed into numerical numbers based on their hydrophobicity scales.

(2) Second, the hydrophobicity profile was broken down into wavelet coefficients based on DWT. More detail about this algorithm can be found in the reference [13].

3.9. Electrostatic (θ_κ) and HINT Potentials (π_κ)

This method is based on 3D-Potentials. Electrostatic θ_κ and HINT Potentials π_κ were used as inputs to construct the QSAR model. These features can be calculated by using the MARCH-INSIDE 2.0 method (Markov Chains Invariants for Network Simulation and Design) [72]. The calculated parameters describe the different types of amino acid-amino acid interactions that consist of all pairs of adjacent amino acids. Here, the protein was considered to be a structural network of amino acids, which was expressed as a node interconnected by the edges (interactions) which were represented by electrostatic (θ_κ), van der Waals (vdW), and hydrophobic + surface (HINT) fields. These values are useful for finding models that link protein structure with biological activity, also known as protein quantitative structure-property relationship (QSAR). And more detailed information and calculation can be found in references [72-74].

4. CLASSIFICATION METHODS

The third step in enzyme family classification is to choose a suitable classification algorithm. The following six algorithms have been applied in this field. Thus, we briefly introduced them as follows.

4.1. Nearest Neighbor (NN)

This algorithm [75] is especially suitable for the situation that the distribution of the samples is unknown. The basic theory of the algorithm can be described as follows.

Given N enzyme samples ($\mathbf{P_1}, \mathbf{P_2}, \dots, \mathbf{P_N}$), their vector can be represented as ($R_1, R_2, \dots, R_N$). The distance of an any query enzyme $\mathbf{P}$ can be formulated as :

$$\mathrm{D}(\mathbf{P}, \mathbf{P_i}) = 1 - \frac{R \cdot R_i}{|R| \cdot |R_i|} (i = 1,2, \dots, N) \quad (26)$$

where $\mathrm{D}(\mathbf{P}, \mathbf{P_i})$ represents the generalized distance between $\mathbf{P}$ and i-th sample $\mathbf{P_i}$. $R \cdot R_i$ represents the dot multiplication for vector R and vector R_i. $|R|$ and $|R_i|$ represent the modulus of R and R_i.

The value of $\mathrm{D}(\mathbf{P}, \mathbf{P_i})$ is between 0 and 1 ($0 \leq \mathrm{D}(\mathbf{P}, \mathbf{P_i}) \leq 1$). When $R=R_i$, the value of $\mathrm{D}(\mathbf{P}, \mathbf{P_i})$ is 0. If the value of $\mathrm{D}(\mathbf{P}, \mathbf{P_k})$ is the smallest between $\mathbf{P}$ and $\mathbf{P_k}$ (k = 1, 2, $\cdots$, N), then the query enzyme $\mathbf{P}$ belongs to the category that the $\mathbf{P_k}$ owns.

4.2. Covariant-discriminant Algorithm (CDA)

The CDA and quadratic discriminant (QD) function are widely used to deal with many bioinformatics problems [76, 77], such as protein structural class prediction [78], membrane protein prediction [79], conotoxin superfamily classification [80] and enzyme family classification [81, 82]. This method can be briefly described as:

Determine which category an enzyme sequence belongs to will be judged by:

$$Sgn(\propto) = Min[\Delta(Z, \bar{Z}^{\propto})] \ (\propto = 1,2, \dots, 6) \quad (27)$$

where $Sgn(\propto)$ is the argument of $\propto$ that minimized $\Delta(Z, \bar{Z}^{\propto})$, which is defined by:

$$\Delta(Z, \overline{Z^{\propto}}) = D_M^2(Z, \overline{Z^{\propto}}) + ln|S^{\propto}| \quad (28)$$

where

$$D_M^2(Z, \overline{Z^{\propto}}) = (Z - \overline{Z^{\propto}})^T S_{\propto}^{-1} (Z - \overline{Z^{\propto}}) \quad (29)$$

where $D_M^2(Z, \overline{Z^{\propto}})$ is the Mahalanobis distance [31, 83] between Z and $\overline{Z^{\propto}}$; T is the transposition operator; $S_{\propto}^{-1}$ is the inverse matrix of $S^{\propto}$; $|S^{\propto}|$ is the determinat of $S^{\propto}$; and $S^{\propto}$ is formulated as follows:

$$S^{\propto} = \begin{bmatrix} S_{1.1}^{\propto} & S_{1.2}^{\propto} & \cdots & S_{1.\theta}^{\propto} \\ S_{2.1}^{\propto} & S_{2.2}^{\propto} & \cdots & S_{2.\theta}^{\propto} \\ \vdots & \vdots & \ddots & \vdots \\ S_{\theta.1}^{\propto} & S_{\theta.2}^{\propto} & \cdots & S_{\theta.\theta}^{\propto} \end{bmatrix} \quad (30)$$

where

$$S_{i,j}^{\propto} = \frac{1}{N-1} \sum_{K=1}^{N} (z_{k,i}^{\propto} - \overline{z_i^{\propto}}) (z_{k,j}^{\propto} - \overline{z_j^{\propto}}) \ (i, j = 1,2, \dots, \theta) \quad (31)$$

4.3. K-Nearest Neighbor (KNN)

The basic principle of KNN algorithm is to assign the query data to majority of its KNN. This algorithm is based on the distance definition and can be briefly described as follows:

(1) Calculate the distance between the query enzyme sequence and each training enzyme sequence as follows:

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (32)$$

where D represents the Euclidean distance between testing data and training data; x_i represents the i-th value of the query enzyme sequence vector; y_i represents the i-th value of the training enzyme sequence vector.

(2) Draw the nearest k training enzymes as the neighbor for the query enzyme.

(3) The query enzyme belongs to the majority of these nearest neighbors' categories.

4.4. Artificial Neural Network (ANN)

The ANN is also referred as the neural network or the connection model. It is a mathematical model for simulating the behavior characteristics of the neural network of animals and the distributed parallel information processing. It is an adaptive nonlinear dynamic system composed of a large number of neurons. The structure and function of each neuron are relatively simple, but the systematic behavior produced by a large number of neurons is very complicated. This network relies on the complexity of the system, and by adjusting the interconnections among the large number of nodes inside, it achieves the purpose of processing information. Compared with digital computer, the structure principle and function

features of ANN are closer to the human brain. It's not a given program step by step in accordance with the operation, but can adapt to the environment itself, summary law, performs some operation, identification, or process control. In the last decade, ANN has demonstrated good intelligence in the prediction of enzyme family classes [2, 15].

4.5. Support Vector Machine (SVM)

In machine learning, SVM is a supervised learning model. The basic idea of SVM is transforming the input vector into a high-dimensional Hilbert space and finding a separating hyperplane in this space. It has the advantage that the parameter optimization process is relatively simple. Therefore, for a small amount of data, the binary classification problem can obtain higher prediction accuracy. Based on this fact, SVM was widely used in bioinformatics [8, 13, 14, 84-106]. The brief description of its basic ideas is as follows:

For a given benchmark dataset $x_i \in R^n$ $(i = 1, \ldots, N)$ with corresponding labels y_i (+1 or -1), SVM can provide a decisive function as:

$$f(x) = \sum_{i=1}^{N} y_i a_i K(x_i, x_j) + b \tag{33}$$

where K represents the kernel function (liner function, polynomial function and radial basis function are three common kernel functions); a_i represents the coefficient to be learned, which was trained by maximizing the Lagrangian expression as follow:

$$max_{a_i} = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \tag{34}$$

where $\sum_{j=1}^{N} a_i y_i = 0,\ 0 \le a_i \le \mathrm{C}$, C represents the penalty parameter.

The best values for the regularization penalty parameter C and kernel parameter g were obtained based on the grid search strategy with cross-validation. For convenience, SVM software packages LibSVM can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm.

4.6. Linear Discriminant Analysis (LDA)

In LDA, also known as Fisher linear discriminant analysis, is based on the idea of finding a projection function and reducing high-dimensional problems to one-dimensional (vector w) problems. In addition, it requires the transformed one-dimensional data to have the following properties: the same category of samples gathers together as close as possible, and different category of samples are as far away as possible. First, the LDA determines the projection direction w and the threshold b (namely, calculates the linear discriminant function) through the given training data. Suppose the projection function for binary classification is:

$$y = w^T x + b \tag{35}$$

where

$$w = S_w^{-1}(\mu_1 - \mu_2) \tag{36}$$

where μ_1 and μ_2 represent the mean of two categories; S_w represents the within-class scatter matrix.

And a more detailed explanation of the principles and the derivation of the formulation can be found in reference [107]. Then, the test dataset's categories were obtained according to this linear discriminant function.

5. PERFORMANCE EVALUATION

5.1. Commonly-used Evaluation Method

In order to evaluate the performance of predictors, the following three cross-validation methods are often used to test a predictor for its effectiveness, independent database test, subsampling test and jackknife test [108-118]. Among the three methods, jackknife test is the most objective and most widely used one [38, 46, 93, 119-131]. It follows the principle of using n-1 data as the training set and the remaining one as the testing set, this ensures that only one result can be obtained. And the accuracy can be calculated as follow:

$$\begin{cases} \Lambda_i = \frac{m_i}{M_i} \\ \Lambda = \frac{\sum_{i=1}^{6} m_i}{\sum_{i=1}^{6} M_i} \end{cases} (i=1, 2, \ldots, 6) \tag{37}$$

where Λ_i represents the success rate for i-th enzyme family class; Λ represents the overall success rate for 6 enzyme families; m_i represents the number of the i-th enzyme family class that were correctly predicted; M_i represents the total number of the i-th enzyme family class.

5.2. Published Results

The detailed results about classifying 6 enzyme families obtained by these machine learning methods are listed in Table **2**.

Based on the benchmark dataset E1, Chou *et al.* used a descriptor named 'Go-PseAAC', which fused the gene ontology and the PseAAC, and then obtained an overall success rate of 89.0% for 6 enzyme family classes in the jackknife cross-validation test [3].

On the basis of benchmark dataset E2, Cai *et al.* further adopted the function domain composition method to construct the feature vector. By using the NN method to classify the enzyme family classes, they achieved an overall accuracy of 85.35% [5]. To further improve the accuracy, they changed the feature expression method to FunD-PseAAC which combined the function domain composition with the PseAAC. The accuracy was improved to 95.0% on the benchmark dataset E3 [6]. It can be seen from the results that the improvement of the success rate is obvious. Compared with the former's feature extraction method, the PseAAC and FunD are more accurate in expressing the sequence information of the enzyme. By adding the spatial structure and physicochemical properties into features vector extraction, the method greatly enhanced the richness and expressiveness of the features. Therefore, the purpose of improving the success rate is achieved.

Cai *et al.* also constructed the benchmark dataset E4 and developed the 'Go-PseAAC' method to classify the enzymes in E4. Finally, they obtained an overall accuracy of 98.48% for classifying the 6 enzyme family classes in the jackknife cross-validation test [7]. Based on the benchmark dataset E5, Lu *et al.* encoded the enzyme sequences with FunD. As a result, the overall success rate of 91.32% was achieved in the

Table 2. A list of published results for classifying 6 enzyme families.

Datasets	Methods	Accuracy of Each Enzyme Family Class (%)						Overall (%)	References
		S1	S2	S3	S4	S5	S6		
E1	'Go-PseAAC'	73.04	95.20	93.76	87.72	74.50	94.89	89.0	[3]
E2	FunD-NN	75.16	79.65	77.14	81.71	63.63	89.47	85.35	[5]
E3	'FunD-PseAAC'	85.56	97.38	95.62	96.70	94.63	97.86	95.0	[6]
E4	'Go-PseAAC'	98.98	99.13	98.87	97.18	96.77	96.75	98.48	[7]
E5	'ESC'	93.53	93.63	94.20	75.29	74.56	89.33	91.32	[8]
E6	'EzyPred'	86.7	95.8	95.9	94.4	93.3	98.3	93.7	[9]
E6	ACC,LFD,ID-SVM	88.1	98.4	99.3	94.3	94.5	94.0	95.5	[14]
E7	ACC-KNN	98	98	100	100	100	100	99	[10]
E9	PseAAC,DWT-SVM	89.3	92.2	87.5	93.3	95.2	94.2	91.9	[13]
E10	(3D)θ_κ+π_k-LDA	70.2	100	100	75.3	100	100	97.0	[11]
E11	(3D)θ_κ+π_k-LDA	88.3	71.3	91.0	78.2	45.3	69.2	78.4	[12]
E11	(3D)θ_κ+π_k-ANN	100	100	100	100	100	100	100	[12]

jackknife cross-validation test based on SVM algorithm [8]. Moreover, a web server Enzyme Classification System (ECS) was built and can be freely accessed at http://pcal.biosino.org/.

Two works were performed on the benchmark dataset E6. A novel method called 'EzyPred' was developed by using FunD and Pse-PSSM to formulate the enzyme samples and using the KNN method to classify the samples. This predictor could produce an overall accuracy of 93.7% [9]. The features used by 'EzyPred' were not only closely associated with the protein function, but also closely related to the evolution information. Based on the same data, Shi *et al.* used SVM combined with AAC, low-frequency Fourier transform (LFT) and increment of diversity (ID) calculated by occurrence frequencies of the 400 dipeptides to classify the enzyme family and obtained an overall accuracy of 95.9% [14]. As shown in Table **2**, the overall success rates for the classes of oxidoreductase, transferase, hydrolase and isomerase are higher than Shen's [9], but the overall success rates for the classes of lyase and ligase are lower than Shen's [9]. Compared with the method of Shen *et al.* [9], the method of Shi *et al.* has four advantages. At first, using ID to formulate the samples of enzymes can clearly reflect their sequence order and pattern information. Secondly, using LFT to formulate the samples of enzymes can clearly express the important information of sequence. Thirdly, SVM is a mature and successful machine learning technology which can always get satisfied results. Finally, the method can reduce the feature vectors dimension obviously, and can obviously simplify the formulation and reduce the operation time.

Based on the benchmark E7, Nasibov *et al.* tried three methods in this area, the best overall accuracy is 99% for classifying 6 enzyme family classes by adopting AAC-KNN method [10]. Based on the benchmark dataset E9, Qiu *et al.* used SVM to classify the samples based on the feature selection technique of PseAAC and DTW. The overall accuracy is 91.9% in the 10-fold cross-validation test [13]. DWT is a useful tool for analyzing time and frequency localization of protein sequences. It is similar to a mathematical microscope with the ability to amplify and translate the protein sequences. DWT analysis can decompose the hydrophobic value sequence into different expansion coefficients, and then remove the noise components from the hydrophobic section to give us a local sequence structure. Therefore, DWT was used as a novel feature extraction tool, which improved the prediction success rate of the enzyme families.

Based on the benchmark dataset E10, Concu *et al.* adopted LDA to classify the enzyme families based on the 3D structure (θ_κ+π_κ), the overall accuracy reached 97.0% in the jackknife cross-validation test [11]. Furthermore, they constructed the benchmark dataset E11 and used the same method to predict the enzyme classes. The overall accuracy of 78.4% was achieved. Moreover, Concu *et al.* even obtained an overall accuracy of 100% by adopting ANN classifier based on the same feature extraction method [12]. The excellent result shows the ANN-based model has a powerful performance. However, it should be noted that the ANN models are more complex than the LDA model. Moreover, it is more difficult to obtain the 3D entropy and spectral moment parameters than to obtain sequence information.

On the basis of benchmark dataset E12, Wu *et al.* applied the PseAAC to formulate enzyme sequences by combining with rigidity, flexibility and irreplaceability of amino acids. In addition, a feature selection method named F-score was adopted to optimize the best dataset. By adopting the SVM to classify the enzyme family classes with incremental feature selection (IFS) [132] process, an overall accuracy of 46.1% was obtained [17].

Besides the published papers on the prediction of the enzyme family classes as described above, the N-to-1 Neural

Networks method has also been introduced to predict enzyme family classes [15]. The overall accuracy was 96% on the benchmark dataset E8. In 2015, Niu *et al.* introduced a new method based on protein-protein network to predict enzyme family classes, and the accuracy is only 62.86% [16].

CONCLUSION

Different types of enzymes have different catalytic modes and different catalytic effects. Classification of enzyme families based on sequence information is a hot topic in bioinformatics, which could help us to understand the function of the enzyme. Accurately identifying the family of enzymes will also facilitate our understanding of the entire context of enzyme reactions.

How to effectively extract feature datasets from an enzyme sequence is a key step in the enzyme classification process. The ACC algorithm was first applied to extract protein sequence characteristics. The basic idea is to use the amino acid composition to characterize protein sequences. Although the principle is simple, the internal correlation information of the sequence is ignored. In response to this defect, the PseAAC method was proposed. It not only considers the general rule of the frequency of *k*-mer in the sequence, but also adds the calculation of the physical and chemical properties of several proteins, which can effectively enhance the expressiveness of the features. Inspired by the above algorithms, extracting sequence features after combining a plurality of feature extraction methods was widely used, such as FunD+PseAAC, Go+PseAAC, DTW+PseAAC and so on, to enhance flexibility in sequence feature extraction and improve prediction accuracy.

Choosing the best machine learning classification algorithm could guarantee high prediction accuracy in pattern recognition. Analyzing the classification results of each classifier, we find that the Nearest Neighbor (NN) discriminant, K-Nearest Neighbor (KNN) discriminant, Mahalanobis Distance discriminant can get good accuracies. However, due to its strong generalization ability and low computational complexity, SVM always displays better accuracy in comparison with other classifiers. With the deepening of research, other algorithms such as Random forest algorithm (RF) [133]; Naïve bayes classification algorithm [134], LibD3C algorithm [135], can also be applied in this field. However, we expect that deep learning [136-138] could dramatically improve the prediction performance for enzyme family classification.

Although the results of the current enzyme classification prediction were acceptable, they still can be improved. Since high dimensional features can lead to over-fitting of classification algorithms, it is necessary to screen the most useful or most representative features by using feature selection techniques such as F-score [139], mRMR [140, 141], which can also improve the performance of the predictor and make the predictor more explanatory. We wish more scholars will devote themselves to this field.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests.

This work was supported by the National Nature Scientific Foundation of China (31771471), the Fundamental Research Funds for the Central Universities of China (Nos. ZYGX2015Z006, ZYGX2016J125, ZYGX2016J118), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244), the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

ACKNOWLEDGEMENTS

Hui Ding and Wei Chen conceived and designed the paper. Jiu-Xin Tan, Hao Lv, Fang Wang and Fu-Ying Dao collected and analyzed the data. Jiu-Xin Tan, Hao Lv, and Fang Wang wrote the paper.

REFERENCES

[1] Webb EC. Enzyme nomenclature. Academic Press, SanDiego 1992.
[2] Jensen LJ, Skovgaard M, Brunak S. Prediction of novel archaeal enzymes from sequence-derived features. Protein Sci 2002; 11: 2894-98.
[3] Chou KC, Cai YD. Using GO-PseAA predictor to predict enzyme sub-class. Biochem Biophys Res Commun 2004; 325: 506-9.
[4] Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. Proteins 2004; 55: 66-76.
[5] Cai YD, Chou KC. Using functional domain composition to predict enzyme family classes. J Proteome Res 2005; 4: 109-111.
[6] Cai YD, Chou KC. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 2005; 4: 967-71.
[7] Cai YD, Zhou GP, Chou KC. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 2005; 234: 145-9.
[8] Lu L, Qian Z, Cai YD, Li Y. ECS: an automatic enzyme classifier based on functional domain composition. Comput Biol Chem 2007; 31: 226-32.
[9] Shen HB, Chou KC. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 2007; 364: 53-59.
[10] Nasibov E, Kandemir-Cavas C. Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction. Comput Biol Chem 2009; 33: 461-64.
[11] Concu R, Dea-Ayuela MA, Perez-Montoto LG, *et al.* Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of Leishmania proteins. J Proteome Res 2009; 8: 4372-82.
[12] Concu R, Dea-Ayuela MA, Perez-Montoto LG, *et al.* 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites. Biochim Biophys Acta 2009; 1794: 1784-94.
[13] Qiu JD, Huang JH, Shi SP, Liang RP. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 2010; 17: 715-22.
[14] Shi R, Hu X. Predicting enzyme subclasses by using support vector machine with composite vectors. Protein Pept Lett 2010; 17: 599-604.
[15] Volpato V, Adelfio A, Pollastri G. Accurate prediction of protein enzymatic class by N-to-1 Neural Networks. BMC Bioinformatics 2013; 14 Suppl 1: S11.
[16] Niu B, Lu Y, Lu J, *et al.* Prediction of enzyme's family based on protein-protein interaction network. Curr Bioinfor 2015; 10: 16-21.
[17] Wu Y, Tang H, Chen W, Lin, H. Predicting human enzyme family classes by using pseudo amino acid composition. Current Proteomics 2016; 13: 99-104.
[18] Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000; 28: 304-05.
[19] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data

bank and its supplement TrEMBL. Nucleic Acids Res 1997; 25: 31-6.
[20] Cui T, Zhang L, Huang Y, *et al.* MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. Nucleic Acids Res 2018; 46: D371-D374.
[21] Zhang T, Tan P, Wang L, *et al.* RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Res 2017; 45: D135-D138.
[22] Yi Y, Zhao Y, Li C, *et al.* RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic Acids Res 2017; 45: D115-D118.
[23] Liang ZY, Lai HY, Yang H, *et al.* Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics 2017; 33: 467-9.
[24] Feng P, Ding H, Lin H, Chen W. AOD: the antioxidant protein database. Sci Rep 2017; 7: 7449.
[25] He B, Chai G, Duan Y, *et al.* BDB: biopanning data bank. Nucleic Acids Res 2016; 44: D1127-1132.
[26] Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003; 19: 1589-91.
[27] Zhu PP, Li WC, Zhong ZJ, *et al.* Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Mol Biosyst 2015; 11: 558-563.
[28] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22: 1658-59.
[29] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010; 26: 680-2.
[30] Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994; 269: 22014-20.
[31] Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995; 21: 319-44.
[32] Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. J Microbiol Methods 2011; 84: 67-70.
[33] Letunic I, Copley RR, Pils B, *et al.* SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 2006; 34: D257-260.
[34] Tatusov RL, Fedorova ND, Jackson JD, *et al.* The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003; 4: 41.
[35] Marchler-Bauer A, Anderson JB, Derbyshire MK, *et al.* CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 2007; 35: D237-240.
[36] Apweiler R, Attwood TK, Bairoch A, *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 2001; 29: 37-40.
[37] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001; 43: 246-55.
[38] Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem 2010; 34: 320-27.
[39] Nanni L, Lumini A, Gupta D, Garg A. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinform 2012; 9: 467-75.
[40] Nanni L, Lumini A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 2008; 34: 653-60.
[41] Qiu JD, Huang JH, Liang RP, Lu XQ. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal Biochem 2009; 390: 68-73.
[42] Mohabatkar H, Mohammad Beigi M, Esmaeili A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 2011; 281: 18-23.
[43] Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med Chem 2013; 9: 133-7.
[44] Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol 2014; 341: 34-0.
[45] Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett 2013; 20: 180-6.
[46] Esmaeili M, Mohabatkar H, Mohsenzadeh S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 2010; 263: 203-209.
[47] Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal Biochem 2013; 442:118-25.
[48] Feng PM, Ding H, Chen W, Lin H. Naive Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med 2013; 2013: 530696.
[49] Feng PM, Lin H, Chen W. Identification of antioxidants from sequence information using naive Bayes. Comput Math Methods Med 2013; 2013: 567529.
[50] Yang H, Tang H, Chen XX, *et al.* Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. Biomed Res Int 2016; 2016: 5413903.
[51] Chen XX, Tang H, Li WC, *et al.* Identification of bacterial cell wall lyases *via* pseudo amino acid composition. Biomed Res Int 2016; 2016:1654623.
[52] Tanford C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. Journal of the American Chemical Society 1962.
[53] Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981; 78: 3824-28.
[54] Chou KC, Cai YD. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 2003; 311: 743-47.
[55] Schaffer AA, Aravind L, Madden TL, *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001; 29: 2994-3005.
[56] Laxton RR. The measure of diversity. J Theor Biol 1978; 70: 51-67.
[57] Zhang L, Luo L. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Res 2003; 31: 6214-20.
[58] Li QZ, Lu ZQ. The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 2001; 213: 493-502.
[59] Liu W, Chou KC. Prediction of protein secondary structure content. Protein Eng 1999; 12: 1041-50.
[60] Weiss O, Herzel H. Correlations in protein sequences and property codes. J Theor Biol 1998; 190: 341-353.
[61] Liu H, Wang M, Chou KC. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 2005; 336: 737-9.
[62] Chou KC. The biological functions of low-frequency vibrations (phonons). VI. A possible dynamic mechanism of allosteric transition in antibody molecules. Biopolymers 1987; 26: 285-95.
[63] Chou KC. Biological functions of low-frequency vibrations (phonons). III. Helical structures and microenvironment. Biophys J 1984; 45: 881-9.
[64] Chou KC. Low-frequency motions in protein molecules. Beta-sheet and beta-barrel. Biophys J 1985; 48: 289-97.
[65] Chou KC. Low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 1988; 30: 3-48.
[66] Chou KC. Low-frequency resonance and cooperativity of hemoglobin. Trends Biochem Sci 1989; 14: 212-13.
[67] Haimovich AD, Byrne B, Ramaswamy R, Welsh WJ. Wavelet analysis of DNA walks. J Comput Biol 2006; 13: 1289-98.
[68] Turkheimer FE, Roncaroli F, Hennuy B, *et al.* Chromosomal patterns of gene expression from microarray data: methodology, validation and clinical relevance in gliomas. BMC Bioinformatics 2006; 7: 526.
[69] Mandell A, Selz K, Shlesinger M. Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families. Physical. Physical A Statistical Mechanics & Its Applications 1997; 244: 254-62.
[70] Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. Bioinformatics 2004; 20: 2572-78.
[71] Rezaei MA, Abdolmaleki P, Karami Z, *et al.* Prediction of mem-

brane protein types by means of wavelet analysis and cascaded neural networks. J Theor Biol 2008; 254: 817-20.
[72] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics 2008; 8: 750-78.
[73] Concu R, Podda G, Uriarte E, Gonzalez-Diaz H. Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials. J Comput Chem 2009; 30: 1510-20.
[74] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 2008; 8: 1676-90.
[75] Li BQ, Zhang YH, Jin ML, Huang T, Cai YD. Prediction of Protein-Peptide Interactions with a Nearest Neighbor Algorithm. Current Bioinformatics 2018; 13: 14-24.
[76] Zhao W, Feng YE. Identify Protein 8-class secondary structure with quadratic discriminant algorithm based on the feature combination. Letters In Organic Chemistry 2017; 14: 625-31.
[77] Yuan LZ, Yong EF, Wei Z, Shan KG. Using quadratic discriminant analysis to predict protein secondary structure based on chemical shifts. Current Bioinformatics 2017; 12: 52-6.
[78] Lin H, Li QZ. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 2007; 28: 1463-66.
[79] Lin H. The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition. J Theor Biol 2008; 252: 350-56.
[80] Lin H, Li QZ. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 2007; 354: 548-51.
[81] Chou KC, Elrod DW. Prediction of enzyme family classes. J Proteome Res 2003; 2: 183-90.
[82] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005; 21: 10-9.
[83] Mahalanobis PC. On the generalised distance in statistic. Proc. Natl. Sci. India 1936; 2: 49-35.
[84] Zhou XB, Chen C, Li ZC, Zou XY. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 2007; 248: 546-551.
[85] Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. J Mol Biol 2005; 345: 187-99.
[86] Gaonkar B, Davatzikos C. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 2013; 78: 270-83.
[87] Cuingnet R, Rosso C, Chupin M, *et al.* Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Med Image Anal 2011; 15: 729-37.
[88] Su ZD, Huang Y, Zhang ZY, *et al.* iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics 2018, DOI: 10.1093/bioinformatics/bty508.
[89] Feng P, Yang H, Ding H, Lin H, Chen W, Chou KC. iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 2018, DOI: 10.1016/j.ygeno.2018.01.005.
[90] Lin H, Liang ZY, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM Trans Comput Biol Bioinform 2017, DOI: 10.1109/TCBB.2017.2666141.
[91] Zhang J, Feng P, Lin H, Chen W. Identifying RNA N(6)-methyladenosine sites in escherichia coli genome. Front Microbiol 2018; 9: 955.
[92] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017; 33: 3518-23.
[93] Yang H, Qiu WR, Liu G, *et al.* iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. Int J Biol Sci 2018; 14: 883-91.
[94] Tang H, Zhao YW, Zou P, *et al.* HBPred: a tool to identify growth hormone-binding proteins. International Journal Of Biological Sciences 2018; 14: 957-64.
[95] Qiu WR, Sun BQ, Tang H, Huang J, Lin H. Identify and analysis crotonylation sites in histone by using support vector machines. Artif Intell Med 2017; 83: 75-81.
[96] Zhao YW, Su ZD, Yang W, *et al.* Ionchanpred 2.0: a tool to predict ion channels and their types. Int J Mol Sci 2017; 18: 1838.
[97] Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-Based prediction of phage virion proteins using a support vector machine. Front Microbiol 2018; 9: 476.
[98] Manavalan B, Lee J. SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 2017; 33: 2496-503.
[99] Ye J, Chen W, Jin DC. Predicting the types of plant heat shock proteins. Letters In Organic Chemistry 2017; 14: 684-89.
[100] Tang H, Zhang CM, Chen R, *et al.* Identification of secretory proteins of malaria parasite by feature selection technique. Lett Org Chem 2017; 14: 621-4.
[101] Lei GC, Tang JJ, Du PF. Predicting s-sulfenylation sites using physicochemical properties differences. Lett Org Chem 2017; 14: 665-72.
[102] Jiang LM, Liao ZJ, Su R, Wei LY. Improved identification of cytokines using feature selection techniques. Lett Org Chem 2017; 14: 632-41.
[103] Loh SK, Low ST, Chai LE, *et al.* A Review of computational approaches to predict gene functions. Curr Bioinformatics 2018; 13: 373-86.
[104] Yang H, Lv H, Ding H, Chen W, Lin H. iRNA-2OM: A sequence-based predictor for identifying 2'-O-methylation sites in Homo sapiens. J Computational Biol 2018, DOI: 10.1089/cmb.2018.0004..
[105] Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 2018, DOI: doi.org/10.1093/bioinformatics/bty451.
[106] Li DP, Ju Y, Zou Q. Protein folds prediction with hierarchical structured svm. Current Proteomics 2016; 13: 79-85.
[107] Bishop C. Pattern recognition and machine learning. Springer; 2006.
[108] Dao FY, Yang H, Su ZD, *et al.* Recent advances in conotoxin classification by using machine learning methods. Mol 2017; 22: 1057.
[109] Song J, Wang Y, Li F, *et al.* iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief Bioinform 2018, DOI: doi.org/10.1093/bib/bby028.
[110] Song J, Li F, Leier A, *et al.* PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. Bioinformatics 2018; 34: 684-7.
[111] Li F, Li C, Marquez-Lago TT, *et al.* Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics 2018, DOI: doi.org/10.1093/bioinformatics/ bty522.
[112] Bao Y, Marini S, Tamura T, *et al.* Toward more accurate prediction of caspase cleavage sites: a comprehensive review of current methods, tools and features. Brief Bioinform 2018, DOI: doi.org/10.1093/bib/bby041.
[113] He WY, Jia CZ, Duan YC, Zou Q. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. Bmc Systems Biology 2018; 12: 44.
[114] Zou Q, Wan SX, Ju Y, Tang JJ, Zeng XX. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. Bmc Systems Biology 2016; 10: 114.
[115] Cao RZ, Adhikari B, Bhattacharya D, *et al.* QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 2017; 33: 586-8.
[116] Cao R, Freitas C, Chan L, *et al.* ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network. Mol 2017; 22: E1732.
[117] Cao RZ, Bhattacharya D, Hou J, Cheng JL. DeepQA: improving the estimation of single protein model quality with deep belief networks. Bmc Bioinformatics 2016; 17: 495.
[118] Tang H, Cao RZ, Wang W, *et al.* A two-step discriminated method to identify thermophilic proteins. Int J Biomath 2017; 10: 1750050.
[119] Mohabatkar H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 2010; 17: 1207-14.
[120] Chou KC, Wu ZC, Xiao X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol Biosyst 2012; 8: 629-41.
[121] Qin YF, Wang CH, Yu XQ, *et al.* Predicting protein structural class by incorporating patterns of over-represented k-mers into the general form of Chou's PseAAC. Protein Pept Lett 2012; 19: 388-97.
[122] Chou KC, Wu ZC, Xiao X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex

eukaryotic proteins. PLoS One 2011; 6: e18258.
[123] Zhao XW, Ma ZQ, Yin MH. Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition. Protein Pept Lett 2012; 19: 492-500.
[124] Tang H, Chen W, Lin H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Mol Biosyst 2016; 12: 1269-75.
[125] Li WC, Deng EZ, Ding H, Chen W, Lin H. iORI-PseKNC: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. Chemometrics Intelligent Laborat Sys 2015; 141: 100-106.
[126] Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 2014; 42: 12961-72.
[127] Ding H, Deng EZ, Yuan LF, *et al.* iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed Res Int 2014; 2014: 286419.
[128] Manavalan B, Shin TH, Lee G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 2018; 9: 1944-56.
[129] Manavalan B, Basith S, Shin TH, *et al.* MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 2017; 8: 77121-36.
[130] Lin YQ, Min XP, Li LL, *et al.* Using a machine-learning approach to predict discontinuous antibody-specific b-cell epitopes. Curr Bioinform 2017; 12: 406-15.
[131] Lai HY, Chen XX, Chen W, Tang H, Lin H. Sequence-based predictive modeling to identify cancerlectins. Oncotarget 2017; 8: 28169-28175.
[132] Li BQ, Hu LL, Niu S, Cai YD, Chou KC. Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. J Proteomics 2012; 75: 1654-65.
[133] Ho TK. The random subspace method for constructing decision forests. IEEE Transactoins on Pattrern Analysis & Machine Intelligence 1998.
[134] Voelz VA, Shell MS, Dill KA. Predicting peptide structures in native proteins from physical simulations of fragments. PLoS Comput Biol 2009; 5: e1000281.
[135] Lin C, Chen W, Qiu C, *et al.* LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 2014; 123: 424-35.
[136] Peng L, Peng MM, Liao B, *et al.* The advances and challenges of deep learning application in biological big data processing. Curr Bioinformatics 2018; 13: 352-59.
[137] Patel S, Tripathi R, Kumari V, Varadwaj P. DeepInteract: deep neural network based protein-protein interaction prediction tool. Curr Bioinform 2017; 12: 551-7.
[138] Long HX, Wang M, Fu HY. Deep convolutional neural networks for predicting hydroxyproline in proteins. Curr Bioinform 2017; 12: 233-8.
[139] Chen W, Lin H, Feng PM, *et al.* iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 2012; 7: e47843.
[140] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005; 3: 185-205.
[141] Naseem I, Khan S, Togneri R, Bennamoun M. ECMSRC: A sparse learning approach for the prediction of extracellular matrix proteins. Curr Bioinform 2017; 12: 361-8.

PMID: 30277150