A Survey for Predicting Enzyme Family Classes Using Machine Learning Methods

Jiu-Xin Tan¹, Hao Lv¹, Fang Wang¹, Fu-Ying Dao¹, Wei Chen¹,²,³,* and Hui Ding¹,*

¹Key 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; ²Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China; ³Gordon Life Science Institute, Boston, MA 02478, USA

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 inflowing 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 distance-based classifier [10]. In the same year, Conc 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
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.](image1)

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.](image2)

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:

$$ P = [a_1, a_2, ..., a_i, ..., a_{20}]^T $$

where:

$$ a_i = \frac{n_i}{N} \quad (1 \leq i \leq 20) $$

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:

$$ P = [p_1, p_2, ..., p_i, ..., p_{7785}]^T $$

where:

$$ p_i = \begin{cases} 1, & \text{hit found in the InterPro database} \\ 0, & \text{otherwise} \end{cases} $$

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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Published</th>
<th>Number of Each Enzyme Family Class</th>
<th>Total Number</th>
<th>Sequence</th>
<th>Identity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year</td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
</tr>
<tr>
<td>E1</td>
<td>2004</td>
<td>560</td>
<td>980</td>
<td>945</td>
<td>285</td>
<td>149</td>
</tr>
<tr>
<td>E2</td>
<td>2005</td>
<td>153</td>
<td>290</td>
<td>385</td>
<td>82</td>
<td>33</td>
</tr>
<tr>
<td>E3</td>
<td>2005</td>
<td>1697</td>
<td>2582</td>
<td>2902</td>
<td>939</td>
<td>503</td>
</tr>
<tr>
<td>E4</td>
<td>2005</td>
<td>2314</td>
<td>3653</td>
<td>3246</td>
<td>1307</td>
<td>676</td>
</tr>
<tr>
<td>E5</td>
<td>2007</td>
<td>436</td>
<td>832</td>
<td>741</td>
<td>170</td>
<td>114</td>
</tr>
<tr>
<td>E6</td>
<td>2007</td>
<td>1820</td>
<td>2847</td>
<td>3279</td>
<td>892</td>
<td>639</td>
</tr>
<tr>
<td>E7</td>
<td>2009</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>E8</td>
<td>2013</td>
<td>945</td>
<td>2110</td>
<td>2226</td>
<td>208</td>
<td>136</td>
</tr>
<tr>
<td>E9</td>
<td>2010</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>E11</td>
<td>2009</td>
<td>154</td>
<td>178</td>
<td>223</td>
<td>87</td>
<td>64</td>
</tr>
<tr>
<td>E12</td>
<td>2016</td>
<td>155</td>
<td>361</td>
<td>404</td>
<td>44</td>
<td>35</td>
</tr>
</tbody>
</table>

The 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:

\[
P = [p_1, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\gamma}]^T
\]  

where:

\[
p_\theta = \begin{cases} 
\frac{\omega}{\gamma} & (1 \leq \theta \leq 20) \\
\frac{\omega}{\gamma} & (21 \leq \theta \leq 20 + \gamma) \end{cases}
\]

where \( \omega \) represents the normalized frequency of 20 native amino acid residues. \( \gamma \) represents the weight factor and \( \tau_j \) represents the \( j \)-th sequence correlation factor and can be calculated by the following equations:

\[
\tau_j = \frac{1}{L - j} \sum_{i=1}^{L-j} F(R_i, R_{i+j}), (j < L)
\]

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

\[
F(R_i, R_j) = \frac{1}{k} \{ [H_i(R_j) - H_i(R_i)]^2 + \cdots + [H_k(R_j) - H_k(R_i)]^2 \}
\]

where \( k \) represents the number of factors; \( H_i(R_j) \) represents the \( i \)-th physicochemical properties of the residue \( R_i \), which can be calculated by:

\[
H_i(R_j) = \frac{h^2_i(R_j) - \sum_{l=1}^{20} h^2_{l}(R_i)/20}{\sum_{l=1}^{20} h^2_{l}(R_i) - \sum_{l=1}^{20} h^2_{l}(R_i)/20}^2
\]

where \( h^2_{l}(R_i) \) represents the original value of the \( i \)-th amino acid residue \( R_i \), \( (i = 1, 2, \ldots, 20) \) 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 InterPro2Go 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:

\[
P = [a_1, a_2, \ldots, a_l, \ldots, a_{1930}]^T
\]

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:

\[
P_{\text{PsePSSM}}^\mu = [\delta_1^\mu, \delta_2^\mu, ..., \delta_{20}^\mu]
\]

where \( \delta_i^\mu \) represents the correlation factor of amino acid residue \( i \), and its continuous distance along the protein sequence is \( \mu \); \( \delta_i^0 \) represents the normalized score obtained by PSI-BLAST; When \( \mu = 0 \), \( \delta_i^0 \) 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 of \( \mu \) of the best prediction accuracy when \( \mu \) varied from 0 to L, a protein P can be described by using Pse-PSSM:

\[
P_{\text{PsePSSM}} = [\delta_1^0, \delta_2^0, ..., \delta_{20}^0, \delta_1^1, \delta_2^1, ..., \delta_{20}^1, ..., \delta_1^{20}, \delta_2^{20}, ..., \delta_{20}^{20}]
\]

3.6. Increment of Diversity (ID)

Laxton defined a measure 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, \( m_t \) represents the number of the \( t \)-th diversity, and the diversity source \( S = (m_1, m_2, ..., m_t) \) is formulated as:

\[
D = M \log M - \sum_i m_i \log m_i
\]

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

\[
ID(X, Y) = D(X + Y) - D(X) - D(Y)
\]

It can also be rewritten as:

\[
ID(X, Y) = D(M, N) - \sum_i D(m_i, n_i)
\]

where

\[
\begin{align*}
D(M, N) &= (M + N) \log (M + N) - M \log M - N \log N \\
D(m_i, n_i) &= (m_i + n_i) \log (m_i + n_i) - m_i \log m_i - n_i \log n_i
\end{align*}
\]

where

\[
\begin{align*}
M &= \sum_i m_i \\
N &= \sum_i n_i
\end{align*}
\]

If \( m_t \) or \( n_t \) equals 0, \( D(m_t, n_t) = 0 \). In the reference [14], Shi converted the 400 dipeptides composition (DC) [59] to a 400 dimension space described as the follows:

\[
S = (m_1, m_2, ..., m_t, m_{400}) \]

where \( m_t \) represents the absolute 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:

\[
P = [R_1, R_2, ..., R_t]_T
\]

where \( R_t \) represents the amino acid residue at the polypeptide chain position \( i \). If using the hydrophilic value \( h(R_t) \) of the amino acid to describe the protein sequence, it can be formulated as follows:

\[
P = [h(R_1), h(R_2), ..., h(R_t)]_T
\]

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

\[
P = [F_1, F_2, ..., F_k, ..., F_L]_T \quad (K = 1, 2, ..., L)
\]

where

\[
F_k = \sum_{i=1}^{L} h(R_i) e^{-j2\pi (n-1)(k-1)/L}
\]

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

\[
P = [V_1, V_2, ..., V_k, ..., V_L]_T
\]

where

\[
V_k = \lim_{c \to \infty} \frac{|F_k|^2}{c}
\]

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 \( \lambda \) low-frequency components and finally selected \( \lambda = 16 \) 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

\[
P = [V_2, V_3, V_4, ..., V_{16}]_T
\]

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 (θk) and HINT Potentials (πk)

This method is based on 3D-Potentials. Electrostatic θk and HINT Potentials πk 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 (θk), 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-activity, also known as protein quantitative structure-activity 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 (P1, P2, ..., PN), their vector can be represented as (R1, R2, ..., RN). The distance of any query enzyme P can be formulated as:

\[ D(P, P_i) = 1 - \frac{R - R_i}{R_i} \text{ (i = 1, 2, ..., N)} \]  

where \( D(P, P_i) \) represents the generalized distance between P and i-th sample \( P_i \), \( R \) 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 \( D(P, P_i) \) is between 0 and 1 (0 ≤ \( D(P, P_i) \) ≤ 1). When \( R = R_i \), the value of \( D(P, P_i) \) is 0. If the value of \( D(P, P_k) \) is the smallest between P and \( P_k \) \( (k = 1, 2, \ldots, N) \), then the query enzyme P belongs to the category that the \( P_k \) 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:

\[ Sgn(\alpha) = \min[\Delta(Z, Z^\alpha)] \text{ (}\alpha = 1, 2, ..., \theta) \]  

where \( Sgn(\alpha) \) is the argument of \( \alpha \) that minimized \( \Delta(Z, Z^\alpha) \), which is defined by:

\[ \Delta(Z, Z^\alpha) = D^2_M(Z, Z^\alpha) + |\ln S^\alpha| \]  

where \( D^2_M(Z, Z^\alpha) = (Z - Z^\alpha)^T S^{-1}\alpha (Z - Z^\alpha) \)

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:

\[ D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

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 \( x_i \in \mathbb{R}^n \) (\( i = 1, ..., N \)) with corresponding labels \( y_i (\pm 1 \text{ or } -1) \), SVM can provide a decisive function as:

\[
f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b
\]

where \( K \) represents the kernel function (linear 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:

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

where \( \sum_{i=1}^{N} a_i y_i = 0 \), \( 0 \leq a_i \leq C \), \( C \) 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:

\[
y = w^T x + b
\]

where

\[
w = S_w^{-1} (\mu_1 - \mu_2)
\]

where \( \mu_1 \) and \( \mu_2 \) 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:

\[
\begin{align*}
\Lambda_i &= \frac{m_i}{M_i} \\
\Lambda &= \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} M_i}
\end{align*}
\]

where \( \Lambda_i \) represents the success rate for \( i \)-th enzyme family class; \( \Lambda \) 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
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 (\( \theta_k \)). 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


PMID: 30277150