ISSN: 1813-4890

Prediction of Phage Enzymes and Hydrolases Based on Support Vector Machine

Mengying Deng, Fengmin Li *

Inner Mongolia Agricultural University College of Science, Hohhot 010018, China

Abstract

Hydrolases coded by phage are key enzymes secreted by phages during the process of infecting host bacteria. The presence of hydrolases not only enables phages to penetrate the polysaccharide capsule of host bacteria but also facilitates bacterial lysis through synergistic action with porin proteins. Predicting phage hydrolases is crucial for exploring the pathogenic mechanisms and therapeutic approaches of certain related diseases. In this study, based on the dataset constructed by Ding et al., we conducted predictive analysis of phage hydrolases. First, phage enzymes were identified from both phage enzymes and non-phage enzymes, and then phage hydrolases were further screened from the identified phage enzymes. Four feature parameters were extracted. Single-feature prediction and fused-feature prediction were performed on these four features. Preliminary dimensionality reduction was achieved using the Maximum Relevance-Minimum Redundancy (mRMR) method, followed by secondary dimensionality reduction using Analysis of Variance (ANOVA). Based on the Support Vector Machine (SVM) algorithm and under the Jackknife test, the highest prediction success rates for phage enzymes and phage hydrolases reached 85.88% and 95.16%, respectively.

Keywords

Hydrolases coded by phage, Phage enzymes, Feature information, Dimensionality reduction.

1. Introduction

Bacteriophages, a type of virus capable of infecting and replicating within bacteria, represent the most widely distributed group of viruses. They are commonly found in environments rich in bacterial communities, such as soil and the intestinal tracts of animals [1]. The interactions between bacteriophages and microbial communities play a significant role in influencing the Earth's chemical cycles [2]. Structurally, bacteriophages exhibit a complex symmetry. The head, which houses the genetic material, displays icosahedral symmetry, while the tail, characterized by helical symmetry, contains proteins that specifically bind to receptor proteins on the surface of host bacteria, facilitating the recognition and infection of target bacteria. Recent studies have highlighted the emergence of antibiotic resistance in certain bacteria during disease treatment, necessitating the exploration of novel methods to inhibit bacterial growth. Research has demonstrated that bacteriophages can lyse bacteria, offering a potential therapeutic approach for disease treatment [3]. Furthermore, in the realm of food safety testing, bacteriophages not only enable rapid and accurate detection of foodborne pathogens but also play a crucial role in microbial inactivation during raw material collection and processing [4].

Enzymes encoded by bacteriophages that disrupt the penetration layer of host cells are collectively referred to as hydrolases. hydrolases coded by phage are critical enzymes secreted during the infection of host bacteria, primarily responsible for degrading bacterial cell walls, capsules, or nucleic acids, thereby facilitating the processes of phage invasion, replication, and

ISSN: 1813-4890

progeny release [5]. The presence of hydrolases not only enables bacteriophages to penetrate the polysaccharide capsules of host bacteria but also collaborates with holing proteins to induce bacterial lysis [6]. Some hydrolases contain secretory signal peptide sequences, which can substitute for holing proteins and directly lyse bacteria [7]. Consequently, the accurate identification of bacteriophage-encoded hydrolases not only aids in elucidating the lysis mechanisms of the bacteriophage-bacteria system but also provides a foundational basis for the development of novel antimicrobial agents. In recent years, researchers have employed machine learning algorithms to predict hydrolases coded by phage.

In 2016, DING et al [8]. proposed a predictive model, termed PHYPred, specifically tailored for forecasting hydrolases coded by phage. This model employed g-gap dipeptide composition to delineate protein sequences and leveraged Analysis of Variance (ANOVA) and Incremental Feature Selection (IFS) methodologies to meticulously screen for optimal feature subsets. Classification and predictive tasks were then executed using the Support Vector Machine (SVM) algorithm, with model parameters being fine-tuned through grid search optimization. Moving forward to 2020, LI et al [9]. extracted four distinct feature parameters: G-gap Dipeptide Composition (GGDC), Pseudo Amino Acid Composition (PseAAC), Grouped Tripeptide Composition (GTPC), and the Composition, Transition, and Distribution (CTD) characteristics of amino acids. These parameters were seamlessly integrated through feature fusion, and the optimal features were rigorously selected using ANOVA. Ultimately, predictions were carried out using the SVM algorithm.

The identification process of hydrolases coded by phage involves two critical steps: first, distinguishing phage enzymes from non-phage enzymes, and subsequently differentiating hydrolases coded by phage from non-hydrolases within the identified phage enzymes. Utilizing the dataset constructed by Ding et al., four types of feature information were extracted: amino acid composition (AAC), dipeptide deviation from expected Mean (DDE), Composition/ Transition/ Distribution (CTD), natural vector method (NV).These feature parameters were then combined, and the resulting combinations were initially reduced in dimensionality using the minimum redundancy-maximum relevance (mRMR) algorithm. Subsequently, a secondary dimensionality reduction was performed through analysis of variance (ANOVA). The extracted feature information was input into a support vector machine (SVM) for prediction, yielding the final results.

2. Materials And Methods

2.1. Benchmark Dataset

Reliable and high-quality datasets are crucial for the construction of predictive models. In this work, samples were gained from Ding.et al. Consequently, the definitive benchmark dataset contains 255 proteins, of which 124 proteins belong to phage enzymes, and the remaining 131 are non-phage enzymes. Furthermore, 124 phage enzymes are divided into 69 hydrolases and 55 nonhydrolases, respectively. The following calculations are all based on these data.

2.2. Protein Feature Extraction

2.2.1. Amino Acid Composition(AAC)

In protein classification research, Amino Acid Composition (AAC) is one of the most widely used sequence feature descriptors [10]. It characterizes protein sequences based on the frequency of occurrence of the 20 different amino acids within the sequence. AAC forms a 20 denotes the dimension of the vector, which can be mathematically represented by the following formula:

$$P_{AAC} = [R_1, R_2, R_3, \cdots, R_i, \cdots, R_{20}] \quad (1)$$

ISSN: 1813-4890

$$R_i = \frac{m_i}{L}(i = 1, 2, \ldots, 20) \quad (2)$$

where L represents the length of the protein sequence, and m_i denotes the number of occurrences of the i-th amino acid in the protein.

2.2.2. Composition/Transition/Distribution (CTD)

The CTD feature was first introduced by Dubchak et al. in 1995 as a protein feature extraction method for protein folding class prediction [11]. It comprises three components: amino acid composition (CTDC), amino acid transition (CTDT), and amino acid distribution (CTDD). During the feature calculation process, 13 physicochemical properties were selected, and the 20 amino acids were categorized into three groups based on each physicochemical property. Each physicochemical property corresponds to a 21 denotes the dimension of the vector representing composition, transition, and distribution. Consequently, the CTD feature can be expressed as a 273 denotes the dimension of the vector (21 × 13). Table 1 provides the classification of amino acids based on their physicochemical properties.

CTDC refers to the proportion of individual amino acids with specific physicochemical properties within the entire protein sequence, which can be calculated using the following formula:

$$C_j^i = \frac{n_j^i}{L}(i = 1, \cdots, 13; j = 1, 2, 3) \quad (3)$$

Where L represents the length of the amino acid sequence, and n_j^i is the number of residues in the j-th group of the i-th physicochemical property.

CTDT refers to the transition probability between two adjacent amino acid residues belonging to two different groups, which can be calculated using the following formula:

$$T_{j,k}^i = \frac{m_{j,k}^i}{L-1}(i = 1, \cdots, 13; j = 1, 2, 3; j < k \le 3) \quad (4)$$

Where $m_{j,k}^i$ represents the number of dipeptides in the sequence where, according to the i-th physicochemical property, the first amino acid belongs to the j-th group and the second amino acid belongs to the k-th group, or the first amino acid belongs to the k-th group and the second amino acid belongs to the j-th group.

CTDD quantifies the distribution status of amino acids with specific residues in the sequence. It calculates the position of the first occurrence of each group of amino acids, as well as their distribution values at the 25%, 50%, 75%, and 100% positions within the sequence. The position of each residue is then normalized by dividing it by the total length of the sequence, as described by the following formula.

$$D_{j,q}^i = \frac{p_{j.q}^i}{L}(i = 1, \cdots, 13; j = 1, 2, 3; q = 1, 25, 50, 75, 100) \quad (5)$$

where $p_{j,q}^i$ represents the minimum sequence length that contains the first q% of the j-th group of amino acids classified according to the i-th physicochemical property.

Table 1 Classification of physical and chemical properties of amino acids

Physicochemical properties	Type 1	Type 2	Type 3
Hydrophobicity_PRAM900101	RKEDQN	GASTPHY	CLVIMFW
Hydrophobicity_ARGP820101	QSTNGDE	RAHCKMV	LYPFIW
Hydrophobicity_ZIMJ680101	QNGSWTDERA	HMCKV	LPFYI

ISSN: 1813-4890

Hydrophobicity_PONP930101	KPDESNQT	GRHA	YMFWLCVI
Hydrophobicity_CASG920101	KDEQPSRNTG	AHYMLV	FIWC
Hydrophobicity_ENGD860101	RDKENQHYP	SGTAW	CVLIMF
Hydrophobicity_FASG890101	KERSQD	NTPG	AYHWVMFLIC
Van der Waals Volume	GASTPDC	NVEQIL	MHKFRYW
polarity	LIFWCMVY	PATGS	HQRKNE
polarizability	GASDT	CPNVEQIL	KMHFRYW
charge	KR	ANCQGHILMFPSTWYV	DE
solvent accessibility	ALFCGIVW	RKQEND	MSPTHY
secondary structure	EALMQKRH	VIYCWFT	GNPSD

2.2.3. Dipeptide deviation from expected mean (DDE)

DDE is a feature descriptor related to dipeptide composition, which takes into account the degeneracy of codon encoding. Amino acids are determined by codons composed of three nucleotides. Since 61 codons encode the 20 amino acids, degeneracy exists, meaning that a single amino acid can be encoded by multiple codons [12]. Therefore, the theoretical frequency of dipeptide occurrence can be described by the degeneracy of codon encoding. For a protein sequence, the DDE feature parameter is obtained by directly calculating its dipeptide composition and then normalizing the results. This can be expressed using the following formula:

$$P_{DDE} = [f_1, f_2, \cdots f_i, \cdots, f_{400}] \tag{6}$$

$$f_i = \frac{DC_i - TM_i}{\sqrt{TV_i}} \tag{7}$$

$$DC_i = \frac{n_i}{L-1} (i = 1, 2, \cdots, 400) \tag{8}$$

$$TM_i = \frac{C_r}{C_N} \times \frac{C_s}{C_N} \tag{9}$$

$$TV_i = \frac{TM_i(1-TM_i)}{L-1} \tag{10}$$

In Equation (11), f_i represents the DDE value for each of the 400 possible dipeptides in the protein sequence, DC_i denotes the observed frequency of the i*i*-th dipeptide in the sequence, TM_i represents the theoretical mean, and TV_i represents the theoretical variance. In Equations (12), (13), and (14), n_i indicates the number of occurrences of the *i*-th dipeptide, L represents the length of the protein sequence. C_r and C_s represent the number of codons encoding the first and second amino acid residues in the dipeptide "*rs*" respectively, and C_N represents the total number of possible codons, excluding the three stop codons.

2.2.4. Natural Vector Method (NV)

The NV is a computational method designed to predict protein structure, function, interactions, and mutation effects [13]. It is capable of predicting secondary structures (such as α-helices and β-sheets), tertiary structures (spatial folding), and binding sites between proteins and other molecules, thereby enabling the classification of proteins. For each protein, three key pieces of information are extracted: the number of amino acids, the average position of amino

ISSN: 1813-4890

acids, and the second-order normalized central moment of amino acids. As a result, each protein sequence can be represented as a 60 denotes the dimension of the vector, as illustrated in Equation (11).

$$P_{NV} = [n_A, \mu_A, D_2^A, \cdots, n_k, \mu_k, D_2^k, \cdots, n_Y, \mu_Y, D_2^Y] \tag{11}$$

$$n_k = \sum_{i=1}^{L} w_k(R_i) \tag{12}$$

In Equation (12), n_k represents the number of occurrences of amino acid k in the protein sequence, R_i denotes the i-th amino acid in the protein sequence, and L represents the length of the protein sequence. For each amino acid k, it can be defined as:

$$w_k(\cdot) : \{A, C, D, \cdots, W, Y\} \to \{0,1\} \tag{13}$$

In Equation (13), $W_k(.)$ is a mapping function from the set {A,C,D,⋯,W,Y} to the set {0,1}. It takes any element from the set {A,C,D,...,W,Y} as input and outputs either 0 or 1. where $w_k(R_i)$=1, if R_i =k. Otherwise, $w_k(R_i)$=0.

$$\begin{cases} S_{(k)(i)} = i \times W_k(R_i) \\ T_k = \sum_{i=1}^{n_k} S_{(k)(i)} \\ \mu_k = \dfrac{T_k}{n_k} \end{cases} \tag{14}$$

$$D_2^k = \sum_{i=1}^{n_k} \frac{(S_{(k)(i)} - \mu_k)^2}{n_k \times L} \tag{15}$$

In Equation (14) and (15) Next, let $S_{(k)(i)}$ be the distance from the first amino acid (regarded as origin) to the i-th amino acid k in the protein sequence, T_k be the total distance of each set of the 20 amino acids, and μ_k be the mean position of the amino acid k.

2.3. Support Vector Machine

Support Vector Machine (SVM), first proposed by Vapnik and colleagues, constitute a learning methodology [14]. They have been successfully employed in bioinformatics research, encompassing protein subcellular localization and structural prediction of proteins. SVM are a prevalent supervised learning algorithm within the realm of machine learning, particularly for addressing binary classification tasks. The core concept of SVM involves mapping the originally extracted features of data into a multidimensional feature space, where an optimal decision hyperplane is sought by maximizing the margin between samples. This ensures that positive and negative samples in the training set are separated to the utmost degree within this feature space. A schematic illustration of this process is provided below:

As illustrated in Fig 1, squares and dots represent two distinct classes of samples to be classified, respectively. The line delineating these two classes is the classification line, with lines and representing the closest sample points to this line and being parallel to it. The distance between lines and is known as the classification margin. A wider classification margin corresponds to a smaller overall error for the classifier. The hyperplane that maximizes the separation between samples belonging to two different classes is referred to as the maximum margin hyperplane. For this paper, predictions were carried out using the LIBSVM SVM algorithm software package developed by Chang and Lin. In SVM, a kernel function is utilized to map data from a low-dimensional space to a high-dimensional space. Common kernel functions include the linear

ISSN: 1813-4890

kernel, Laplacian kernel, and radial basis function (RBF) kernel, among others. In this study, the RBF kernel function from LIBSVM was adopted as the kernel function [15].

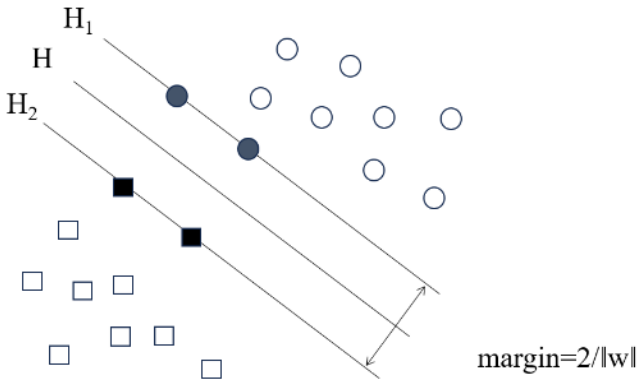

Fig. 1 Two or more references

2.4. Feature selection

2.4.1. Max-Relevance and Min-Redundancy (mRMR)

Feature selection is a critical step in the machine learning process, aimed at reducing data dimensionality and computational complexity, thereby enhancing the accuracy of predictive models. To date, numerous effective feature selection methods have been proposed, such as analysis of variance (ANOVA), maximum relevance-minimum redundancy (mRMR), and principal component analysis (PCA). In this study, mRMR is employed to reduce the dimensionality and redundancy of high-dimensional features. mRMR simultaneously considers the relevance between features and the target variable as well as the redundancy among features. This ensures that the selected features not only effectively explain the target variable but also avoid excessive redundancy, thereby improving the quality and stability of feature selection [16]. The method aims to maximize the relevance between features and the target class while minimizing the correlation among features. The approach is described as follows:

$$I(x,y)=\int\int P(x,y)\log\frac{P(x,y)}{P(x)P(y)}dxdy \tag{16}$$

In the above formula, $I(x;y)$ represents the mutual information between the random variables x and y, $P(x)$ and $P(y)$ denote the probability densities, respectively, and $P(x,y)$ represents the joint probability density of x and y.

The formulas for calculating the relevance between features and the target variable, as well as the correlation among features, are as follows:

$$\max D(S,c), D=\frac{1}{|S|}\sum_{x_i\in S} I(x_i;c) \tag{17}$$

$$\min R(S), R=\frac{1}{|S|^2}\sum_{x_i,x_j\in S} I(x_i;x_j) \tag{18}$$

Where S is the feature set, where n denotes the total number of features within this set. c represents the target class, and $I(x_i;c)$ signifies the mutual information between feature i and the target class c. Furthermore, $I(x_i;x_j)$ denotes the mutual information between feature i and feature j. D is defined as the mean of the mutual information values computed between each feature x_i in the feature set S and the class c. The redundancy among the features is quantified by calculating the mutual information between each pair of features within the feature set S, and this is denoted as R.

ISSN: 1813-4890

2.4.2. Analysis Of Variance（ANOVA）

Redundant or irrelevant features can reduce prediction accuracy and increase computational time. To eliminate such features, this study employs ANOVA as the second feature selection algorithm. The fundamental principle of ANOVA is that the value of a feature in a sample is primarily influenced by the differences between sample groups and the variations within individual samples. The F-value is calculated as the ratio of between-group differences to within-group differences. A higher F-value indicates that the feature exhibits more significant differences between sample groups compared to the variations within individual samples [17]. Consequently, ANOVA allows for the ranking of each feature based on its significance. The F-value for each feature is computed as follows.

$$F(i)=\frac{S_B{}^2(i)}{S_W{}^2(i)} \tag{19}$$

where is the score of the *i*-th feature, a high *F(i)*-value means a high ability to identify the sample; $S_W{}^2(i)$ is the variance within groups; $S_B{}^2(i)$ is the variance among groups; and they can be calculated as follows:

$$S_B{}^2(i)=\frac{SS_B(i)}{K-1} \tag{20}$$

$$S_W{}^2(i)=\frac{SS_W(i)}{N-K} \tag{21}$$

where $SS_B(i)$ is the sum of the squares between the groups; $SS_W(i)$ is the sum of squares within the groups; *K* is the total number of classes; *N* is the total number of samples. and they can be calculated as follows:

$$SS_B(i)=\sum_{j=1}^{2} m_j\left(\frac{\sum_{s=1}^{m_j} f_i(s,j)}{m_j}-\frac{\sum_{j=1}^{2}\sum_{s=1}^{m_j} f_i(s,j)}{\sum_{j=1}^{2} m_j}\right)^2 \tag{22}$$

$$SS_w(i)=\sum_{j=1}^{2}\sum_{s=1}^{m_j}\left(f_i(s,j)-\frac{\sum_{s=1}^{m_j} f_i(s,j)}{m_j}\right)^2 \tag{23}$$

Here, $f_i(s,j)$ represents the F-value of the *k*-th feature in the *j*-th sample of the *i*-th category.

2.5. Performance Evaluation Metrics

Accurately and comprehensively evaluating the predictive performance of models is a crucial task in machine learning-based protein function prediction. In this study, we employed the Jackknife test method and the following performance evaluation metrics: Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and Matthews Correlation Coefficient (MCC). Sn represents the proportion of positive samples correctly identified; Sp indicates the proportion of negative samples correctly identified; Acc denotes the proportion of all samples correctly identified. MCC measures the correlation between the classifier's predictions and the actual classifications, serving as a comprehensive evaluation metric. It takes into account True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), making it suitable for evaluating binary classification problems even when there is a significant imbalance between positive and negative samples. The specific formulas for these four evaluation metrics are as follows:

$$Sn=\frac{TP}{TP+FN} \tag{24}$$

$$Sp=\frac{TN}{TN+FP} \tag{25}$$

$$Acc=\frac{TP+TN}{TP+FN+TN+FP} \tag{26}$$

ISSN: 1813-4890

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \quad (27)$$

In the aforementioned formulas, TP refers to the number of bacteriophage viral proteins correctly predicted; TN represents the number of bacteriophage non-viral proteins correctly predicted; FP indicates the number of bacteriophage non-viral proteins incorrectly predicted; and FN denotes the number of bacteriophage viral proteins incorrectly predicted. These terms are essential for evaluating the predictive performance of the model in distinguishing between bacteriophage viral and non-viral proteins.

3. Results and discussion

3.1. Bacteriophage Enzyme Prediction Results

3.1.1. Prediction Results of Single Feature Parameters

In this study, based on the SVM algorithm, we extracted features including AAC, DDE, CTD, and NV to predict phage enzymes under the Jackknife test. The detailed results are presented in Table 2.

Table 2 Prediction results of single characteristic parameters

Features	Sn/%	Sp/%	MCC	Acc/%
AAC	79.84	67.94	0.48	73.73
DDE	70.16	75.57	0.46	72.94
CTD	80.65	67.18	0.48	73.73
NV	76.61	73.28	0.50	74.90

As shown in Table 2, the highest MCC and Acc values reached 0.50 and 74.90%, respectively, both achieved by the NV. In contrast, the prediction results of DDE were relatively poor, which may be attributed to its insufficient capture of the unique structural characteristics and functional site information of phage enzymes, or the presence of redundant features.

3.1.2. Prediction Results of Combined Features

Compared to single feature encoding methods, integrating feature information can more comprehensively and accurately represent the complex characteristics of protein sequences. Therefore, to enhance model performance, we adopted a feature fusion strategy. The feature parameters AAC, DDE, CTD, and NV were fused for prediction, as shown in Table 3.

Table 3 The prediction results of fusion feature information

Features	Sn/%	Sp/%	MCC	Acc/%
AAC+CTD	79.03	67.94	0.47	73.33
AAC+DDE	70.16	75.57	0.46	72.94
AAC+NV	74.19	71.76	0.46	72.94
CTD+DDE	72.58	74.81	0.47	73.73
CTD+NV	62.90	80.15	0.44	71.76
DDE+NV	63.71	80.15	0.45	72.16
AAC+CTD+DDE	74.19	70.99	0.45	72.55
AAC+CTD+NV	72.58	73.28	0.46	72.94
AAC+DDE+NV	64.52	78.63	0.44	71.76
CTD+DDE+NV	74.19	74.05	0.48	74.12
AAC+CTD+DDE+NV	76.61	74.81	0.51	75.69

As can be seen from Table 3, the fusion of feature parameters AAC, DDE, CTD, and NV achieved the highest accuracy (Acc) of 75.69%, with an MCC of 0.51, indicating that multi-feature fusion

ISSN: 1813-4890

can integrate the advantages of different features and effectively enhance the predictive capability of the model. In contrast, among the two-feature parameter combinations, the combination of Dipeptide Deviation from Expected Mean (DDE) and Natural Vector (NV) performed relatively poorly, with a prediction accuracy (Acc) of 71.76%. This may be due to the limited overlap in feature information between these two parameters, which hindered their complementary effects. Although the prediction accuracy (Acc) of some feature combinations improved compared to single features, the extent of improvement was modest, likely due to information redundancy among the features. Therefore, the next step involves dimensionality reduction on the fused features to further enhance prediction accuracy.

3.1.3. Prediction Results After Dimensionality Reduction of Combined Features

Feature selection is a crucial step in machine learning and data analysis, primarily aimed at identifying the most useful features from the original feature set while eliminating irrelevant or redundant ones. Due to the significant increase in dimensionality after feature parameter fusion, this chapter employs the mRMR (minimum Redundancy Maximum Relevance) algorithm for preliminary feature screening, followed by the ANOVA (Analysis of Variance) algorithm for secondary dimensionality reduction (results are detailed in Table 4), ultimately obtaining the optimal feature subset.

Table 4 The prediction results after performing variance analysis for dimensionality reduction

Features	Sn/%	Sp/%	MCC	Acc/%
AAC+CTD	79.03	81.68	0.61	80.39
AAC+DDE	86.29	81.68	0.68	83.92
AAC+NV	79.84	83.87	0.64	81.96
CTD+DDE	87.90	81.68	0.70	84.71
CTD+NV	82.26	84.73	0.67	83.53
DDE+NV	81.68	78.63	0.60	80.15
AAC+CTD+DDE	85.48	83.21	0.69	84.31
AAC+CTD+NV	81.68	79.39	0.61	80.53
AAC+DDE+NV	79.84	85.50	0.65	82.75
CTD+DDE+NV	87.90	83.97	0.72	85.88
AAC+CTD+DDE+NV	87.90	82.44	0.70	85.10

As shown in Table 4, after initial dimensionality reduction using mRMR, further application of ANOVA for secondary dimensionality reduction resulted in improved prediction accuracy and a further reduction in feature dimensionality. Specifically, after two rounds of dimensionality reduction, the CTD+DDE+NV feature combination continued to perform the most prominently, achieving a prediction Acc, Sn, Sp, and MCC of 85.88%, 87.9%, 83.97%, and 0.72, respectively. These values represent improvements of 2.35%, 3.32%, 1.53%, and 0.05 compared to the results after the initial mRMR dimensionality reduction, fully demonstrating the effectiveness of the secondary dimensionality reduction.

3.2. Hydrolases coded by phage Prediction Results

3.2.1. Prediction Results of Single Feature Parameters

In this section, three feature parameters, including AAC, DDE, CTD, and NV, were extracted. Based on the support vector machine (SVM) algorithm, the prediction of phage lytic enzymes was conducted under the Jackknife test. The specific results are presented in Table 5.

Table 5 Prediction results of single characteristic parameters

Features	Sn/%	Sp/%	MCC	Acc/%
AAC	73.91	50.91	0.26	63.71

ISSN: 1813-4890

DDE	76.81	58.18	0.36	68.55
CTD	69.57	69.09	0.38	69.35
NV	79.71	54.55	0.36	68.55

As shown in Table 5, the prediction results of the five feature parameters are relatively low, with Acc all below 70.00%. Among them, the feature with the highest Sp, MCC, and Acc is DDE, achieving 69.09%, 0.38, and 69.35%, respectively. This indicates that the DDE feature has a strong capability in distinguishing non-phage lytic enzymes. The feature with the highest Sn is NV, at 79.71%, suggesting that the NV feature has an advantage in capturing key information about phage lytic enzymes and can more accurately identify positive samples. The feature with the poorest prediction performance is AAC, with an Acc of 63.71% and the lowest MCC of 0.26.

3.2.2. Prediction Results of Combined Features

Building on the single-feature predictions, the feature parameters AAC, DDE, CTD, and NV were fused, and the SVM algorithm was employed to predict under the Jackknife test. The results of different combination methods are presented in Table 6.

Table 6 The prediction results of fusion feature information

Features	Sn/%	Sp/%	MCC	Acc/%
AAC+CTD	66.67	67.27	0.34	66.94
AAC+DDE	75.36	65.45	0.41	70.97
AAC+NV	75.36	61.82	0.38	69.35
CTD+DDE	82.61	69.09	0.52	76.61
CTD+NV	79.71	65.45	0.46	73.39
DDE+NV	76.81	69.09	0.46	73.39
AAC+CTD+DDE	75.36	67.27	0.43	71.77
AAC+CTD+NV	75.36	65.45	0.41	70.97
AAC+DDE+NV	72.46	70.91	0.43	71.77
CTD+DDE+NV	78.26	67.27	0.46	73.39
AAC+CTD+DDE+NV	78.26	69.09	0.48	74.19

As shown in Table 6, the accuracy (Acc) of most feature combinations improved after feature fusion. Among the two-feature parameter combinations, only the AAC+CTD combination showed a decrease in prediction accuracy (Acc) compared to the single CTD feature, with a reduction of 2.41%. Among them, the CTD+DDE combination achieved the highest Acc of 76.61%, which is the best prediction result among all combinations. Its sensitivity (Sn), specificity (Sp), and Matthews correlation coefficient (MCC) were 82.61%, 69.09%, and 0.52, respectively. Compared to the single CTD feature, the prediction accuracy increased by 7.26%, and compared to the single DDE feature, the Acc improved by 8.12%. Among the three-feature combinations, the CTD+DDE+NV combination achieved the highest Acc of 73.39%. For the fusion of all four features, the AAC+CTD+DDE+NV combination achieved an Acc of 74.19%.

Although the fusion of three-feature and four-feature information introduces more feature information, their prediction accuracy (Acc) does not surpass that of the CTD+DDE combination. The reason may be that the three-feature and four-feature fusions were not subjected to dimensionality reduction, leading to redundant information in the feature space, which obscured the critical discriminative information in the sequences and ultimately affected the final prediction results. Additionally, while appropriate feature combinations can improve prediction outcomes, not all combinations have a positive impact. Therefore, dimensionality reduction is necessary to filter out the most discriminative feature combinations, thereby achieving higher prediction accuracy.

ISSN: 1813-4890

3.2.3. Prediction Results After Dimensionality Reduction of Combined Features

Due to the significant increase in dimensionality after feature parameter fusion, this chapter employs the mRMR (minimum Redundancy Maximum Relevance) algorithm for preliminary feature screening. Subsequently, ANOVA (Analysis of Variance) is used for secondary dimensionality reduction (results are detailed in Table 7), ultimately obtaining the optimal feature combination.

Table 7 The prediction results after performing variance analysis for dimensionality reduction

Features	Sn/%	Sp/%	MCC	Acc/%
AAC+CTD	92.75	87.27	0.80	90.32
AAC+DDE	89.86	83.64	0.74	87.10
AAC+NV	91.30	83.64	0.75	87.90
CTD+DDE	92.75	85.45	0.79	89.52
CTD+NV	89.86	89.09	0.79	89.52
DDE+NV	95.65	89.09	0.85	92.74
AAC+CTD+DDE	94.20	96.36	0.90	95.16
AAC+CTD+NV	91.30	86.44	0.78	89.06
AAC+DDE+NV	95.65	89.09	0.85	92.74
CTD+DDE+NV	94.20	90.90	0.85	92.74
AAC+CTD+DDE+NV	95.65	87.27	0.84	91.94

As shown in Table 7, after initial dimensionality reduction using mRMR, further application of ANOVA for secondary dimensionality reduction resulted in improved accuracy (Acc) in most cases, along with a further reduction in feature dimensionality. The Acc after two rounds of dimensionality reduction exceeded 87.00%. Among them, the feature combination with the best prediction performance was AAC+CTD+DDE, which, after dimensionality reduction, was reduced to 129 dimensions. Its sensitivity (Sn), specificity (Sp), Matthews correlation coefficient (MCC), and accuracy (Acc) were 94.20%, 96.36%, 0.90, and 95.16%, respectively.

3.3. Comparison with Existing Methods

To further validate the effectiveness of the proposed prediction model, a comparison was made with the results obtained from previous studies. The specific prediction results are presented in Table 8 and Table 9.

Table 8 Comparison of predicted results of Phage enzymes

Method	Sn/%	Sp/%	MCC	Acc/%
Ding et al	87.10	87.10	-	84.30
Our prediction model	87.90	83.97	0.72	85.88

As can be seen from Table 8, the prediction accuracy (Acc) of our model reaches 85.88%, which is 1.58% higher than that of Ding et al.'s method. Therefore, the prediction results for phage enzymes in our model outperform those of other methods.

Table 9 Comparison of predicted results of hydrolases coded by phage

Method	Sn/%	Sp/%	MCC	Acc/%
Ding et al	94.50	92.80	-	93.50
Our prediction model	94.20	96.36	0.90	95.16

As shown in Table 9, the prediction results of the model proposed in this paper outperform those of Ding et al., with the prediction accuracy (Acc) being 1.66% higher than that of Ding et

al.'s method. Therefore, the prediction results of the model in this paper for phage lytic enzymes are superior to those of other methods.

4. Conclusion

This paper extracted amino acid monopeptide component information (AAC), amino acid dipeptide deviation from the expected average value information (DDE), amino acid composition, transition and distribution information (CTD), and amino acid natural vector information (NV) based on phage enzymes and hydrolases coded by phage. The prediction results of single feature parameters and fused feature parameters were discussed in detail. Then, the fused features were initially reduced in dimension using the minimum redundancy-maximum relevance (mRMR) method, and then further reduced in dimension using analysis of variance (ANOVA) to identify the best fused features. The support vector machine algorithm was used, and the prediction results were discussed in detail under the Jackknife test. After identifying the best fused features, the prediction results of phage enzymes and hydrolases coded by phage were compared with those of different algorithms. Ultimately, the highest prediction success rate of single feature parameters before dimension reduction for phage enzymes was 74.90%, and the highest prediction success rate of fused features was 76.61%. After two rounds of dimension reduction using the mRMR method and ANOVA, the highest prediction success rate of fused features reached 85.88%. The highest prediction success rate of single feature parameters before dimension reduction for hydrolases coded by phage was 69.35%, and the highest prediction success rate of fused features was 75.69% to 76.61%. After two rounds of dimension reduction using the mRMR method and ANOVA, the highest prediction success rate of fused features reached 95.16%. The prediction results of the support vector machine algorithm were higher than those of random forest and logistic regression. The prediction results of phage enzymes and phage hydrolases were better than those of previous studies, indicating that this model is a reliable tool for predicting hydrolases coded by phage.

References

[1] R.J. Clark, B.J. March: Bacteriophages and Biotechnology: Vaccines, Gene Therapy and Antibacterials, Trends in Biotechnology, 24 (2006) No.5, p.212-218.

[2] N. Auslander, A.B. Gussow, S. Benler, et al. Seeker: Alignment-Free Identification of Bacteriophage Genomes by Deep Learning, Nucleic Acids Research, 48 (2020) No.21, p.e121.

[3] Y.Q. Zhang, Z.Y. Li: RF Phage Virion: Classification of Phage Virion Proteins with a Random Forest Model, Frontiers in Genetics, 13 (2023), p.1103783.

[4] A. Vikram, J. Woolston, A. Sulakvelidze: Phage Biocontrol Applications in Food Production and Processing, Current Issues in Molecular Biology, 40 (2020) No.1, p.267-302.

[5] M. Rashel,J. Uchiyama, I. Takemura, et al. Tail-Associated Structural Protein Gp61 of Staphylococcus Aureus Phage Phi MR11 Has Bifunctional Lytic Activity, FEMS Microbiology Letters, 284 (2008) No.1, p.9-16.

[6] H. Nishikawa, M. Yasuda,J. Uchiyama J, et al. T-Even-Related Bacteriophages as Candidates for Treatment of Escherichia Coli Urinary Tract Infections, Archives of Virology, 153 (2008) No.3, p.507-515.

[7] D. Grandgirard, J.M. Loeffler, V.A. Fischetti, S.L. Leib: Phage Lytic Enzyme Cpl-1 for Antibacterial Therapy in Experimental Pneumococcal Meningitis, The Journal of Infectious Diseases, 197 (2008) No.11, p.1519-1522.

[8] H. Ding, W. Yang, H. Tang, P.PM Feng, J. Huang, W. Chen, et al: PHYPred: A Tool for Identifying Bacteriophage Enzymes and Hydrolases, Virologica Sinica, 31 (2016) No.4, p.350-352.

[9] H.F. Li, X.F. Wang, H. Tang: Predicting Bacteriophage Enzymes and Hydrolases by Using Combined Features, Frontiers in Bioengineering and Biotechnology, 8 (2020), p.183.

[10] Z.P. Feng: Prediction of the Subcellular Location of Prokaryotic Proteins Based on a New Representation of the Amino Acid Composition, Biopolymers, 58 (2001) No.5, p.491-499.

[11] Huang Y A, You Z H, et al: Improved Protein-Protein Interactions Prediction via Weighted Sparse Representation Model Combining Continuous Wavelet Descriptor and PseAAcomposition, BMC Systems Biology, 10 (2016) No.Supplement 4, p.485-494.

[12] S. Vijayakumar, G. Namasivayam: Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor, OMICS: A Journal of Integrative Biology, 19 (2015) No.10, p.648-658.

[13] J.T. Xin, S.L. Hao, Z.Z. Mei, et al: Identification of Hormone Binding Proteins Based on Machine Learning Methods, Mathematical Biosciences and Engineering: MBE, 16 (2019) No.4, p.2466-2480.

[14] M. Deng, C.L. Yu, Q. Liang, et al: A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications, PLOS ONE, 6 (2011) No.3, p.e17293.

[15] H.L. Zou: iAHTP-LH: Integrating Low-Order and High-Order Correlation Information for Identifying Antihypertensive Peptides, International Journal of Peptide Research and Therapeutics, 28 (2021) No.4, p.2651-2659.

[16] C. Ding, H.C. Peng: Minimum Redundancy Feature Selection from Microarray Gene Expression Data, Journal of Bioinformatics and Computational Biology, 3 (2005) No.2, p.185-205.

[17] S.S.Yuan, D. Gao, et al: IBPred: A Sequence-Based Predictor for Identifying Ion Binding Protein in Phage, Computational and Structural Biotechnology Journal, 20 (2022) No.7, p.4942-4951.