A Method for Functional Protein Classification Enhanced by Multiple Sequence Alignment

Pengda Zhang

***Abstract*—Functional protein identification is a key area of research in bioinformatics. Understanding the functions of unknown proteins is crucial, especially for newly discovered ones. This is often achieved by classifying them alongside proteins with known functions. The classification process typically involves three main steps: removing redundant amino acid sequences from protein datasets, extracting features that capture both statistical information and physicochemical properties, and fine-tuning classifier parameters for optimization. However, challenges remain in achieving accurate classifications for certain functional proteins, with ongoing debates regarding the relevance of specific amino acid sequence fragments to classification outcomes. To address these challenges, we propose a novel protein classification method that incorporates multiple sequence alignment as an intermediate step in the existing framework. Our tests, conducted on nine protein datasets using various feature extraction methods and classifiers, demonstrate improved classification results, suggesting that the inclusion of sequence alignment significantly enhances the effectiveness of protein classification.**

***Index Terms*—Feature extraction, classification, functional protein, multiple sequence alignment**

I. Introduction

Functional proteins perform physiological functions and various metabolic activities. Although effective and reliable, identifying them using biological methods inevitably faces challenges such as long experimental cycles, low efficiency, and high resource consumption. As a result, machine learning methods are commonly used for functional protein identification. As to newly discovered proteins, their functions can often be inferred by classifying them alongside proteins with known functions [1].

Typically, the classification of functional proteins involves three sequential steps: first, redundant amino acid sequences are eliminated from the original protein dataset; second, features that combine statistical information with physical and chemical characteristics are extracted from the protein sequences; and finally, classifiers are optimized by fine-tuning their parameters for effective functional protein classification. However, these steps do not consistently yield satisfactory results, as classification outcomes can sometimes be suboptimal [2]. This raises the need to explore factors that may influence these results, particularly the potential role of specific fragments within the original amino acid sequences in determining the classification of certain functional proteins.

In this article, a novel classification method for functional proteins is proposed, incorporating multiple sequence alignment as a preprocessing step before feature extraction. It is hypothesized that common subsequences may be shared between functional and non-functional proteins (i.e., positive and negative samples), which can adversely affect classification performance. By eliminating these common subsequences, the effectiveness of the classification process is aimed to be enhanced. Following the removal of these subsequences, standard feature extraction and classification methods are applied. Experimental results obtained from nine public protein datasets demonstrate that the performance of existing methods for classifying different functional proteins can be significantly improved by incorporating multiple sequence alignment.

II. Materials and Methods

A. Benchmark dataset

Nine datasets of functional proteins, all experimentally verified in biology, are considered benchmark datasets. These include Legionella pneumophila effector protein [3], apolipoprotein [4], acidase [5], immunoglobulin [6], bacterial type IV secretory effector protein [7], thermophilic protein [8], malaria parasite mitochondrial protein [9], malaria parasite secretory protein [10], and bacterial cell wall lyases [11].

B. Multiple sequence alignment

MAFFT [12] is utilized for performing multiple sequence alignments. Upon careful comparison of the results from multiple protein sequence alignments, it is observed that the

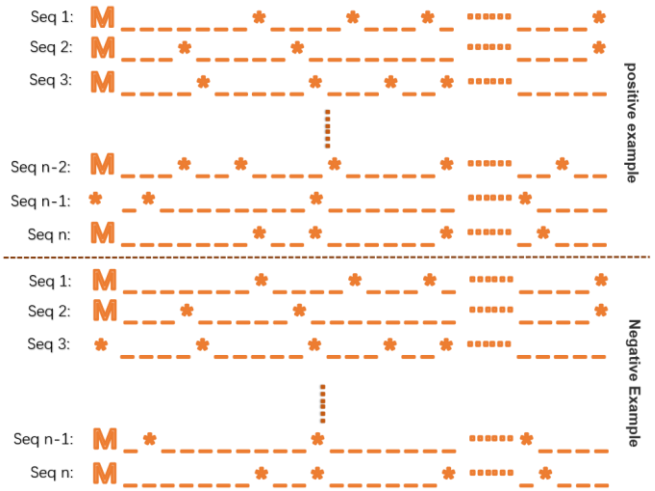

Fig. 1. Diagrammatic sketch of protein. 'M' represents amino acid M, '*' represents other amino acids except M, and the symbol '-' denotes a vacancy.

Manuscript received February 19, 2024; revised December 26, 2024.

Pengda Zhang is a research associate in Heilongjiang Institute of Atomic Energy, Heilongjiang Province 150081, China (corresponding author to provide e-mail: hljxlhj@163.com)

initial position of almost every amino acid sequence contains the amino acid 'M'. This finding is confirmed through the alignment and analysis of the sequences of the nine types of proteins, as depicted in Fig. 1. Therefore, the amino acid 'M' at the beginning of each protein sequence is removed from both functional and non-functional proteins.

C. Feature extraction

After the step of multiple sequence alignment and common subsequences removal, three feature extraction methods, i.e., adaptive-k-skip-2-gram-feature (400D) [13], AAC [14], and SVMProt-188D (188D) [15], are employed.

1) Adaptive-k-skip-2-gram-feature

The adaptive-k-skip-2-gram-feature (400D) is considered an enhanced n-gram feature extraction technique. K-skip is a procedure that allows skipping k elements during feature extraction; while, n-gram refers to a continuous sequence of n amino acids. Thus, k-skip-2-gram refers to allowing up to k-1 positions to be skipped between two amino acids. When applied to short protein sequences, the original method for n-gram feature extraction encounters the problem of sparse n-gram models. Thus, the adaptive-k-skip-2-gram-features (400D) method is considered. The distance between any two amino acids in a sequence can be computed as follows,

$$\mathrm{D}(A_i, A_j) = j - i - 1, \tag{1}$$

where A_i and A_j denote the amino acids at positions i and j. When A_i is adjacent to A_j, the distance is zero. Note that $\mathrm{D}(A_i, A_j)$ is abbreviated as D in the following part. The improved model takes into account not only adjacent residues in the conventional n-gram model, but also residues at distances ranging from 1 to k along the sequence. Therefore, the set of subsequences obtained by a-skip on the original sequence of length L can be expressed as,

$$Skip(D = a) = \{A_i A_{i+a} | 1 \le i \le L\text{-}k, 1 \le a \le k|\}. \tag{2}$$

Correspondingly, the union of all subsequences derived from 1-skip to k-skip can be expressed as,

$$T_{SkipGram} = \left\{\cup_{a=1}^{k} Skip(D = a)\right\}. \tag{3}$$

Using enumeration, the set $T_{SkipGram}$ can also be expressed as $T_{SkipGram} = \{m_j | 1 \le j \le p\}$. Thus, the k-skip-2-gram feature vector can be computed as,

$$FV_{SkipGram} = \left[\frac{N(m_j)}{N(T_{SkipGram})}\right], \tag{4}$$

where $1 \le j \le p$. Since the feature space dimension expands exponentially with n, n is limited to less than 3 to avoid overfitting. When n equals one, the k-skip-n-gram model simplifies to the traditional n-gram model. Hence, our focus remains on n equals two.

2) AAC

AAC is determined by computing the frequencies of the 20 natural amino acids within the sequence (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y in alphabetical order). It can be formally defined using the following formula,

$$(X_1, X_2, \cdots, X_{20}) = \frac{N_i}{L}, \tag{5}$$

where N_i denotes the count of amino acid i, and L represents the length of the amino acid sequence.

3) SVMProt-188D

SVMprot-188D (188D) is a feature extraction method derived from amino acid composition and physicochemical properties, resulting in a total of 188 feature dimensions. The initial 20 dimensions represent the occurrence frequency of the 20 amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y in alphabetical order) within the sequence. The calculation formula is equivalent to the upper formula.

The remaining 168 feature dimensions pertain to the physical and chemical properties of the amino acid sequences. Dimensions 21 to 41 represent hydrophobic characteristics, dimensions 42 to 62 correspond to van der Waals forces, dimensions 63 to 83 relate to polarity, dimensions 84 to 104 denote polarizability, dimensions 105 to 125 represent charge properties, dimensions 126 to 146 indicate surface tension, dimensions 147 to 167 represent secondary structure, and dimensions 168 to 188 correspond to solvent accessibility. More details can be seen in reference [15].

D. Classification

Three classifiers are utilized here for functional protein classification: multilayer perceptron, support vector machine, and random forest. The multilayer perceptron (MLP), also known as an artificial neural network (ANN), is a type of feedforward artificial neural network. It has strong learning capabilities, robustness, and the ability to include multiple hidden layers in addition to input and output layers.

The support vector machine (SVM) presents an alternative approach to the sequential minimal optimization (SMO), which addresses a quadratic programming problem rooted in the Karush-Kuhn-Tucker (KKT) conditions by introducing a novel coefficient, α. Renowned for its efficiency, quick computation of α, and high accuracy, the SMO algorithm particularly shines when dealing with large datasets. It excels in facilitating efficient SVM learning in such scenarios.

The random forest (RF) is an ensemble learning model where the outcomes of individual weak classifiers are learned and combined together to achieve superior learning outcomes compared to a single classifier. RF, a classical bagging model, employs a decision tree as its weak classifier. To ensure the model's generalization capacity, two fundamental principles, namely 'data randomness' and 'feature randomness', are maintained during each tree's construction. Data randomness involves the random extraction of data from the entire dataset to serve as training data for one of the decision tree models. Feature randomness involves the assumption that each sample has M dimensions, with a constant k<M specified. Here, k features are randomly selected from the pool of M features.

E. Measurements

To evaluate the performance, four indicators are employed, i.e., accuracy (ACC), sensitivity (SN), specificity (SP), and Matthews's correlation coefficient (MCC). These metrics are defined as follows,

$$\mathrm{SN} = \frac{TP}{TP + FN},$$

$$\mathrm{SP} = \frac{TN}{TN + FP}, \tag{6}$$

$$\mathrm{ACC} = \frac{TN + TP}{TN + FP + FN + FP} \times 100\%,$$

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

where TP, FP, TN and FN represent true positives, false positives, true negatives and false negatives, respectively. SN and SP illustrate the model's predictive capability in positive and negative samples, respectively. Both ACC and MCC assess the overall model performance. Higher scores on all four indicators indicate better performance.

III. Experimental Results

The primary goal of the experiments is to assess whether incorporating multiple sequence alignment into the protein classification process improves classification performance. Therefore, the experimental design focuses on the following key aspects.

A. Feature extraction with multiple sequence alignment

First, three feature extraction methods (i.e., 188D, 400D, and AAC) are applied to both the original protein sequences (denoted as 'O') and the sequences with the starting 'M' amino acid removed (denoted as 'M'). The sequences are then classified using multiple classifiers (MLP, SVM, and RF), and performance is evaluated through five-fold cross-validation. This approach allows for a comparison of the impact of removing the common amino acid subsequence and identification of the best-performing feature extraction method. The experimental results for the benchmark datasets are presented in Figs. 2 to 4.

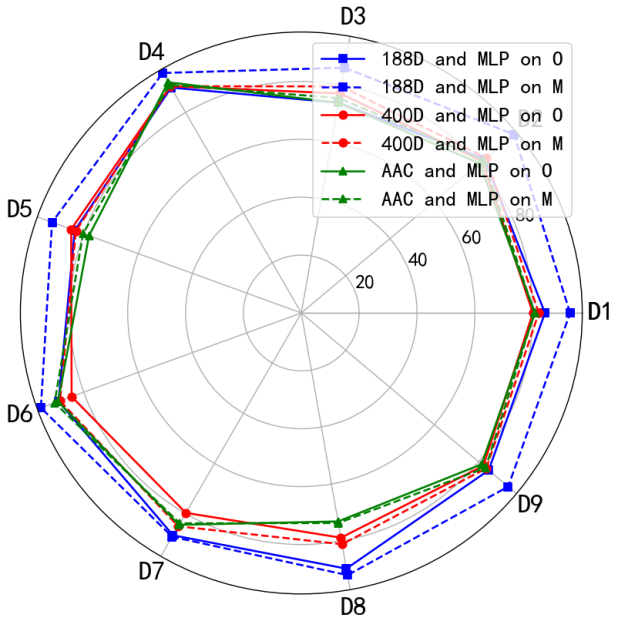

Fig. 2. Radar chart of average accuracy (ACC) values using different feature extraction methods (188D, 400D, AAC) with MLP classifier. Directions represent nine benchmark datasets. 'M' and 'O' denote sequences with and without the initial 'M' amino acid removed, respectively.

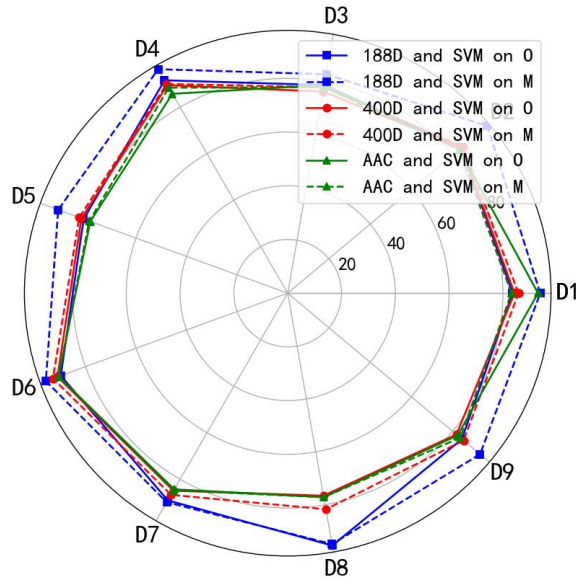

Fig. 3. Radar chart of average accuracy (ACC) values derived from different feature extraction methods (i.e., 188D, 400D, AAC) with SVM as the classification algorithm. Each of the nine directions corresponds to one of the nine benchmark datasets, with 'M' and 'O' indicating amino acid sequences with and without the initial 'M' amino acid, respectively.

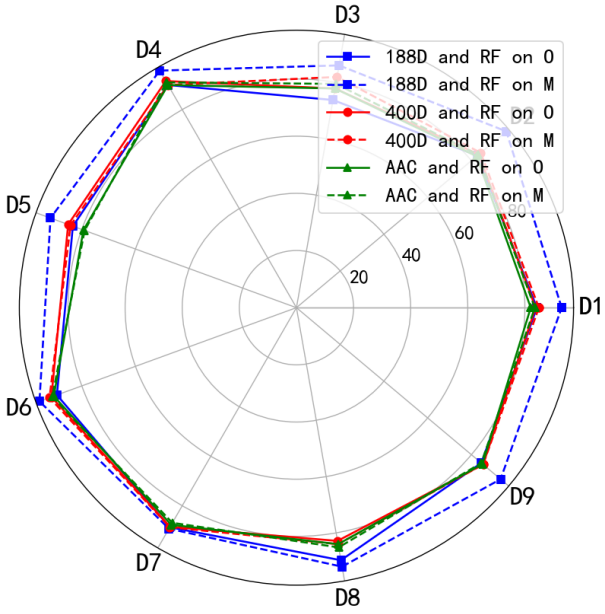

Fig. 4. Radar chart based on average accuracy (ACC) values derived from multiple feature extraction methods (188D, 400D, AAC) with RF as the classifier. The nine directions correspond to nine benchmark datasets, where 'M' and 'O' denote amino acid sequences with and without the initial 'M' amino acid, respectively.

The impact of different feature extraction methods on protein classification was evaluated using the following experimental procedures and analysis methods. First, three feature extraction methods-188D, 400D, and AAC-were applied to nine benchmark datasets labeled D1 through D9. Each method was used on both the original protein sequences (denoted as 'O') and the sequences with the starting 'M' amino acid removed (denoted as 'M'). During this process, a multilayer perceptron (MLP) was selected as the classifier, and the average accuracy (ACC) values, defined by Eq. (6), were obtained through five-fold cross-validation. For each dataset, the five-fold average ACC values were calculated for both scenarios (with and without the starting 'M' amino acid). These results are presented as radar charts in Fig. 2, where each direction represents a dataset.

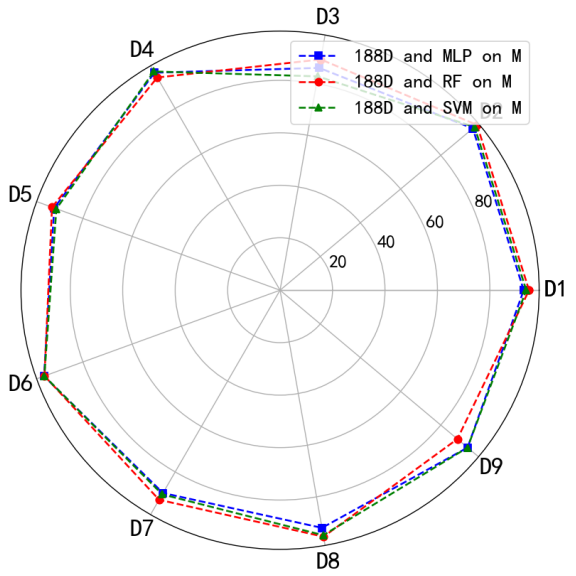

Fig. 5. Radar chart of average accuracy (ACC) values using 188D for feature extraction and different classifiers (MLP, SVM, RF). The nine directions correspond to nine benchmark datasets, where 'M' indicates the amino acid sequence with the starting 'M' amino acid removed.

In Fig. 2, the average ACC values obtained using the same feature extraction method (188D, 400D, or AAC) with MLP as the classifier are connected across the datasets by lines with different colors and styles. The legend labels 'M' and 'O' correspond to sequences with and without the starting 'M' amino acid removed, respectively. The figure clearly shows that when the starting 'M' amino acid is removed, 188D consistently achieves the highest average ACC values across the nine benchmark datasets, indicating that 188D is the most effective feature extraction method. This trend is similarly observed when support vector machines (SVM) and random forests (RF) are used as classifiers, as shown in Figs. 3 and 4.

Additionally, Figs. 2, 3, and 4 reveal that the dashed lines, representing the average ACC values for sequences with the starting 'M' amino acid removed, consistently outperform the solid lines, which represent the average ACC values for the original sequences. This indicates that removing the common amino acid subsequence at the start of protein sequences (i.e., the initial 'M') after multiple sequence alignment improves classification accuracy, further confirming the positive impact of this modification on classification performance.

B. Classifier performance after removing the starting 'M'

Second, it is vital to evaluate the performance of different classifiers on the nine benchmark datasets after confirming that removing the starting 'M' amino acid improves classification results. The goal is to compare average ACC values and other metrics to determine the relative impact of the classifier choice on classification outcomes, and identify the most effective classifier once the feature extraction method is established.

Instead of detailed comparisons across Figs. 2, 3, and 4, Fig. 5 provides a more intuitive summary of the relevant results. It illustrates the average ACC values for different classifiers (i.e., MLP, SVM, and RF) using 188D as the feature extraction method. The differently colored dashed lines represent the average classification accuracy achieved by various classifiers across different datasets, with the results connected for clarity, while 'M' in the legend indicates sequences with the starting 'M' amino acid removed. The results reveal that none of the classifiers consistently achieve the highest ACC value across all nine datasets after the removal of the starting 'M'. This suggests that the improvement in classification performance is primarily attributable to the 188D feature extraction method, rather than the choice of classifier (i.e., MLP, SVM, or RF).

Further evidence is provided by the quantitative results shown in Tables I, II, and III, which are calculated using the metrics defined in Eq. (6). In these tables, 'D*_O' represents the original protein sequence, 'D*_M' denotes the sequence with the starting 'M' amino acid removed, and 'D1_*' to 'D9_*' correspond to the nine benchmark datasets. Note that here '*' serves as a placeholder for letters or numbers. The best classification results are highlighted in bold.

From Tables I to III, it is evident that the average values of the performance indicators obtained through five-fold cross-validation are generally higher when multiple sequence alignment and the removal of the common amino acid subsequence are applied. Furthermore, none of the three classifiers (i.e., MLP, SVM, and RF) consistently achieves the best classification results across the nine benchmark datasets, even after the initial 'M' amino acid has been removed from each sequence.

TABLE I
CLASSIFICATION RESULTS USING 188D AND MLP

Data	ACC	SN	SP	MCC
D1_O	84.20	0.843	0.841	0.660
D1_M	92.90	0.908	0.941	0.849
D2_O	82.73	0.745	0.866	0.610
D2_M	95.95	0.949	0.964	**0.992**
D3_O	73.77	0.689	0.781	0.473
D3_M	86.06	0.862	0.859	**0.965**
D4_O	89.94	0.933	0.814	0.752
D4_M	95.76	**0.970**	0.924	**0.976**
D5_O	83.33	0.858	0.808	0.668
D5_M	91.66	**0.938**	0.894	0.834
D6_O	89.28	0.913	0.871	0.786
D6_M	**95.78**	**0.966**	0.947	**0.988**
D7_O	88.69	0.888	0.885	0.774
D7_M	89.28	0.899	**0.886**	0.786
D8_O	89.71	0.761	0.939	0.714
D8_M	92.00	0.772	0.954	0.777
D9_O	84.53	0.569	0.910	0.491
D9_M	**93.33**	0.830	**0.954**	**0.772**

TABLE II
CLASSIFICATION RESULTS USING 188D AND SVM

Data	ACC	SN	SP	MCC
D1_O	83.58	0.805	0.852	0.648
D1_M	93.87	**0.943**	0.936	0.870
D2_O	84.26	0.786	0.867	0.640
D2_M	96.83	0.969	0.967	0.928
D3_O	78.68	0.741	0.828	0.573
D3_M	82.78	0.811	0.840	0.651
D4_O	91.53	0.934	0.862	0.788
D4_M	**96.29**	**0.970**	0.942	0.908
D5_O	80.70	0.826	0.787	0.614
D5_M	90.78	0.929	0.885	**0.909**
D6_O	89.75	0.903	0.890	0.794
D6_M	95.66	0.965	0.948	0.913
D7_O	89.08	0.926	0.860	0.784
D7_M	89.88	0.935	0.868	0.800
D8_O	95.42	0.921	0.963	0.869
D8_M	94.85	**0.918**	0.956	0.851
D9_O	84.53	0.638	0.867	0.387
D9_M	**93.33**	0.905	0.937	0.763

TABLE III
CLASSIFICATION RESULTS USING 188D AND RF

Data	ACC	SN	SP	MCC
D1_O	81.81	0.831	0.812	0.608
D1_M	**94.94**	0.942	**0.953**	**0.892**
D2_O	82.40	0.824	0.823	0.590
D2_M	**98.03**	**0.979**	**0.980**	0.956
D3_O	78.68	0.804	0.776	0.567
D3_M	**89.34**	**0.886**	**0.898**	0.784
D4_O	91.00	0.894	0.973	0.774
D4_M	93.65	0.918	**1.000**	0.843
D5_O	82.01	0.819	0.820	0.639
D5_M	**92.54**	0.918	**0.933**	0.851
D6_O	88.87	0.896	0.880	0.776
D6_M	95.60	0.943	**0.972**	0.912
D7_O	90.27	0.976	0.848	0.815
D7_M	**92.26**	**0.986**	0.873	**0.853**
D8_O	95.42	0.900	0.970	0.870
D8_M	**95.42**	0.900	**0.970**	**0.870**
D9_O	84.26	0.909	0.840	0.328
D9_M	88.26	**1.000**	0.847	0.556

To sum up, on the basis of the determined 188D feature extraction method, it cannot be simply assumed that a certain classifier has an absolute advantage. Instead, it is necessary to comprehensively consider the performance of different classifiers according to specific situations to select a more suitable classifier and achieve better protein classification results.

C. *Results of removing specific amino acids*

Third, the impact of removing the starting 'M' amino acid is further investigated by comparing it with the removal of the second amino acid ('S') and the last amino acid ('L'). The 188D feature extraction method and MLP classifier, which showed good performance in previous experiments, are used to re-evaluate classification performance across the nine datasets. Accordingly, five-fold cross-validation is applied

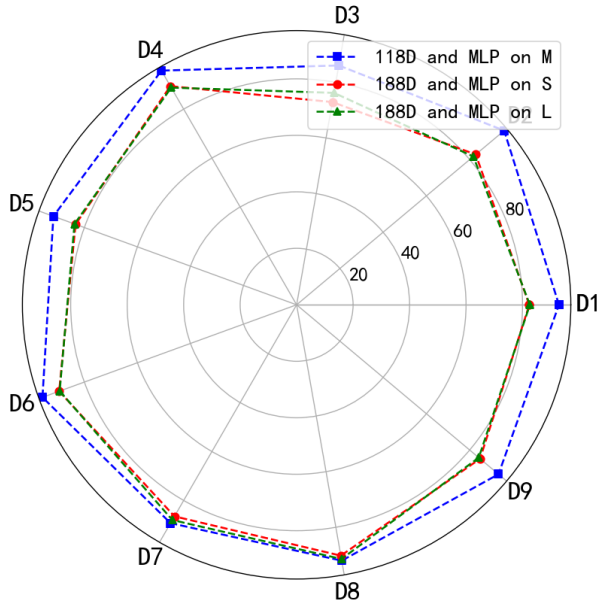

Fig. 6. Radar chart displaying the ACC values using 188D as the feature extraction method and MLP as the classifier. Nine directions correspond to nine benchmark datasets. 'M' denotes the sequence with the initial 'M' amino acid removed, 'S' represents the sequence with the second amino acid removed, and 'L' indicates the sequence with the last amino acid removed.

and performance metrics such as SN, SP, ACC, and MCC are analyzed. This comparison aims to assess whether removing the initial 'M' amino acid after sequence alignment enhances classification performance and how it compares to the removal of other amino acid positions. In this context, 'M' refers to sequences with the initial 'M' amino acid removed, 'S' represents sequences with the second amino acid removed, and 'L' refers to sequences with the last amino acid removed. It is important to note that 'S' and 'L' represent

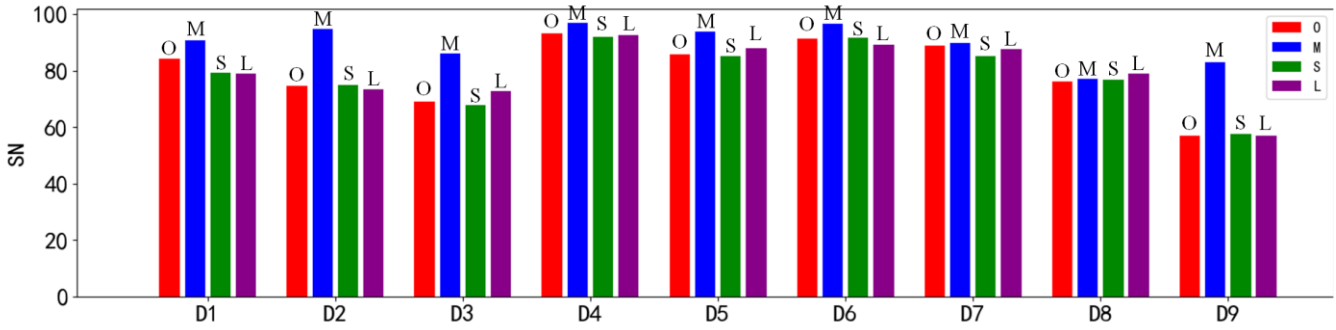

Fig. 7. Bar chart of sensitivity (SN) values calculated on nine public datasets using 188D as the feature extraction method and MLP as the classifier. 'O' represents the original amino acid sequence, 'M' indicates the sequence with the initial 'M' amino acid removed, 'S' refers to the sequence with the second amino acid removed, and 'L' denotes the sequence with the last amino acid removed.

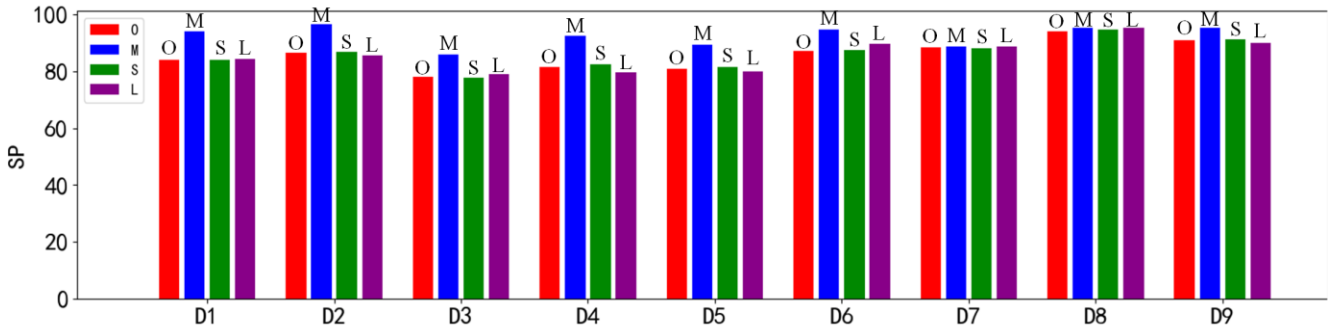

Fig. 8. Bar chart of specificity (SP) values calculated on nine public datasets using 188D as the feature extraction method and MLP as the classifier. 'O' represents the original amino acid sequence, 'M' indicates the sequence with the initial 'M' amino acid removed, 'S' refers to the sequence with the second amino acid removed, and 'L' denotes the sequence with the last amino acid removed.

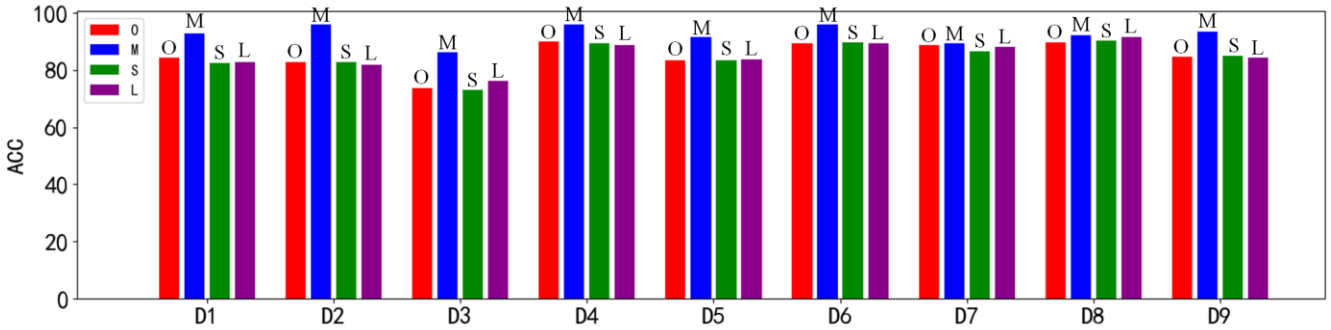

Fig. 9. Bar chart of accuracy (ACC) values calculated on nine public datasets using 188D as the feature extraction method and MLP as the classifier. 'O' represents the original amino acid sequence, 'M' indicates the sequence with the initial 'M' amino acid removed, 'S' refers to the sequence with the second amino acid removed, and 'L' denotes the sequence with the last amino acid removed.

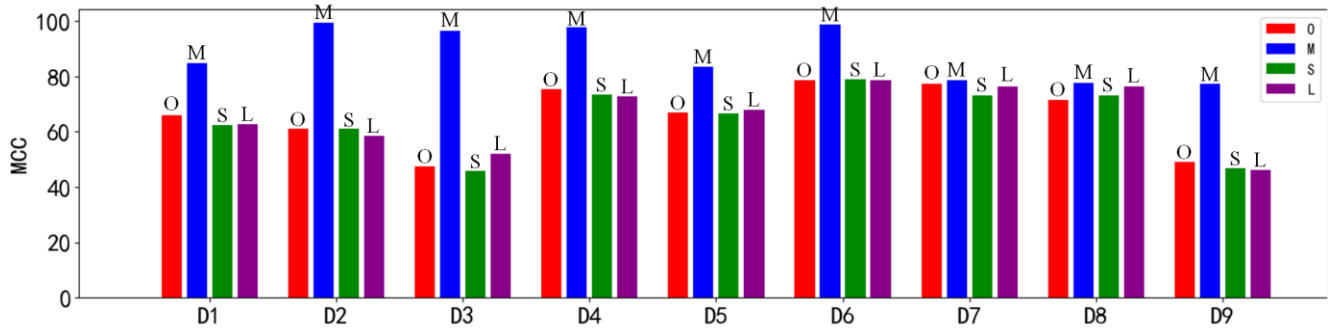

Fig. 10. Bar chart of Matthews correlation coefficient (MCC) values calculated on nine public datasets using 188D as the feature extraction method and MLP as the classifier. 'O' represents the original amino acid sequence, 'M' denotes the sequence with the initial 'M' amino acid removed, 'S' refers to the sequence with the second amino acid removed, and 'L' indicates the sequence with the last amino acid removed.

specific positions, not the results of multiple sequence alignment.

Since no classifier consistently outperformed others in Fig. 5 and Tables I to III, MLP is chosen for further analysis. Additionally, 188D is selected as the feature extraction method due to its superior performance in Figs. 2, 3, and 4. Five-fold cross-validation ensures robust results, with experimental outcomes presented in Fig. 6.

In Fig. 6, the average ACC values for different datasets, using 188D and MLP, are connected by dashed lines in varying colors. The results show that removing the common amino acid subsequence (i.e., the initial 'M' amino acid) after sequence alignment significantly improves classification performance. In contrast, removing the second ('S') or last ('L') amino acids does not yield comparable benefits.

The effectiveness of removing the initial 'M' amino acid is further illustrated in Figs. 7 to 10, which display the values of SN, SP, ACC, and MCC for the nine datasets. Different colors-red, blue, green, and purple-represent sequences labeled 'O' (original sequence), 'M' (initial amino acid 'M' removed), 'S' (second amino acid removed), and 'L' (last amino acid removed). As shown in Figs. 7 to 10, the removal of the initial 'M' amino acid, considered as part of multiple sequence alignment, leads to the best classification results, highlighting the effectiveness of incorporating sequence alignment as a preprocessing step before feature extraction.

IV. Discussion

There are three key points that can be concluded from the experimental results. Firstly, 188D is considered to be a more effective feature extraction method for the classification of various functional proteins. The experimental results in Fig. 2, Fig. 3, and Fig. 4 can support the above viewpoint. Secondly, selecting the classifiers such as MLP, SVM, and RF does not improve the results of functional protein classification once 188D is designated as the feature selection method (see Fig. 5). Thirdly, it is the removal of the common amino acid subsequence (i.e., the amino acid 'M' at the beginning of a protein sequence) that significantly improves the functional protein classification results, especially after 188D has been chosen as the feature extraction method.

All these points suggest that the initial 'M' amino acid is likely inversely correlated with the physical and chemical properties of the protein sequence being classified. In fact, it is 188D, but not other feature extraction methods, that retains these physical and chemical characteristics. This may explain why 188D performs exceptionally well in classifying the nine benchmark datasets, where the initial M amino acids have been removed from the protein sequences. Further discussion is needed to provide a corresponding biological explanation.

V. Conclusion

A novel approach for protein classification is proposed by integrating multiple sequence alignment into an existing procedure consisting of three steps: sequence elimination, feature extraction, and protein classification. By removing common subsequences (i.e., the initial 'M' amino acid) from both functional and non-functional proteins, improved classification results are achieved using 188D, which captures the physical and chemical properties of amino acid sequences on benchmark datasets. This improvement is observed regardless of whether the proteins originate from animals, plants, or microbes.

REFERENCES

[1] Q.W. Zhang, and X.X. Guo, "Dual-population firefly algorithm based on gender differences for detecting protein complexes," *Engineering Letters*, vol. 32, no. 5, pp. 1062-1072, 2024.

[2] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.C. Chou, "iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC," *Oncotarget*, vol. 7, no. 23, pp. 34558-34570, 2016.

[3] Z.E. Ashari, K.A. Brayton, and S.L. Broschat, "Using an optimal set of features with a machine learning-based approach to predict effector proteins for legionella pneumophila," *PloS One*, vol. 14, no. 1, pp. e0202312, 2016.

[4] T. Hua, P. Zou, C. Zhang, R. Chen, and H. Lin, "Identification of apolipoprotein using feature selection technique," *Scientific Reports*, vol. 6, pp. 30441, 2016.

[5] L. Hao, C. Wei, D. Hui, and V.N. Uversky, "AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes," *PloS One*, vol. 8, no. 10, pp. e75726, 2013.

[6] T. Hua, C. Wei, and L. Hao, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269-1275, 2016.

[7] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.Q. Wei, "PredT4SE-stack: prediction of bacterial Type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers in Microbiology*, vol. 9, pp. 2571, 2018.

[8] L. Hao, and C. Wei, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67-70, 2011.

[9] H. Ding, and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329-333, 2015.

[10] H. Thang, C.M. Zhang, R. Chen, P. Huang, C.G. Duan, and P. Zou, "Identification of secretory proteins of malaria parasite by feature selection technique," *Letters in Organic Chemistry*, vol. 14, no. 9, pp. 621-624, 2017.

[11] S. Jiao, L. Xu, and Y. Ju, "CWLy-RF: A novel approach for identifying cell wall lyases based on random forest classifier," *Genomics*, vol. 113, no. 5, pp. 2919-2924, 2021.

[12] K.M. Wong, M.A. Suchard, and J.P. Huelsenbeck, "Alignment uncertainty and genomic analysis," *Science*, vol. 319, no. 5862, pp. 473-476, 2008.

[13] L. Wei L, J. Tang, and Q. Zou, "SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC Genomics*, vol. 18, no. 7, pp. 1-11, 2017.

[14] M. Bhasin, and G. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp 23262-23266, 2004.

[15] Q. Zou, Z. Wang, X. Guan, B. Liu, and Y. Wu, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Research International*, vol. 2013, no. 4, pp. 686090-686090, 2013.

Pengda Zhang received the B.S. degree in computer science and technology from Northeast Forestry University and the M.S. degree in intelligent instrument from Harbin Institute of Technology, Harbin, China, in 2003 and 2010, respectively. Currently, he is a research associate in Heilongjiang Institute of Atomic Energy, Harbin, China. His research interest includes radiation detection and protection, image processing, machine learning and bioinformatics.