IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 1, 2019, accepted July 8, 2019, date of publication July 18, 2019, date of current version August 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929363

ProtDec-LTR3.0: Protein Remote Homology Detection by Incorporating Profile-Based Features into Learning to Rank

BIN LIU[1,2] AND YULIN ZHU[3]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 518055, China

Corresponding author: Bin Liu (bliu@bliulab.net)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672184, Grant 61732012, and Grant 61822306, in part by the Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China under Grant 161063, in part by the Shenzhen Overseas High Level Talents Innovation Foundation under Grant KQJSCX20170327161949608, in part by the Guangdong Natural Science Funds for Distinguished Young Scholars under Grant 2016A030306008, and in part by the Scientific Research Foundation in Shenzhen under Grant JCYJ20180306172207178.

ABSTRACT Protein remote homology detection is one of the most challenging problems in the field of protein sequence analysis, which is an important step for both theoretical research (such as the understanding of structures and functions of proteins) and drug design. Previous studies have shown that combining different ranking methods via learning to the rank algorithm is an effective strategy for remote protein homology detection, and the performance can be further improved by the protein similarity networks. In this paper, we improved the ProtDec-LTR1.0 and ProtDec-LTR2.0 predictors by incorporating three profile-based features (Top-1-gram, Top-2-gram, and ACC) into the framework of learning to rank via feature mapping strategies. The predictive performance was further refined by the pagerank (PR) algorithm and hyperlink-induced topic search (HITS) algorithm. Finally, a predictor called ProtDec-LTR3.0 was proposed. Rigorous tests on two widely used benchmark datasets showed that the ProtDec-LTR3.0 predictor outperformed both ProtDec-LTR1.0 and ProtDec-LTR2.0, and other nine existing state-of-the-art predictors, indicating that the ProtDec-LTR3.0 is an efficient method for protein remote homology detection, and will become a useful tool for protein sequence analysis. A user-friendly web server of the ProtDec-LTR3.0 predictor was established for the convenience of users, which can be accessed at http://bliulab.net/ProtDec-LTR3.0/.

INDEX TERMS Protein remote homology detection, profile-based features, feature mapping strategy, learning to rank, pagerank, hyperlink-induced topic search.

I. INTRODUCTION

The proteins belonging to the same superfamily but different families are remote homology proteins [1]. Homologous proteins refer to proteins belonging to the same family. Different homologous proteins may belong to the same superfamily. The sequence similarity between remote homologous proteins is usually less than 40%, while homologous proteins usually share less than 95% sequence similarity [1].

As one of the key tasks in the field of protein sequence analysis, protein remote homology detection is playing an important role in analyzing the structures and functions of proteins.

In order to efficiently detect the proteins sharing remote homology relationship, some computational methods have been proposed, which can be divided into two categories [1], [2]: discriminative methods and ranking methods.

Discriminative methods treat protein remote homology detection as a classification problem, where the proteins are represented as fixed length feature vectors, and then fed into classifiers to train the models. Finally, for the unknown samples, their homology relationship can be detected by these models. A key to improve their predictive performance is to find features to reflect the characteristics of proteins.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

In this regard, several powerful features have been proposed. Most of these features were generated only based on the sequence information, such as PseAAC [3], Kmer [4], physicochemical property [5], secondary structure [6], [7], hybrid features [8], etc. Among these features, the profile-based features showed highly discriminative power, because they take the evolutionary process of proteins into consideration, such as Top-n-gram [9], ACC [10], DT [11], SOFM [12], PDT [13], profile-based protein representation [14], etc. Recently, in order to explore more accurate profile-based features, the deep learning techniques have been employed to automatically generate the features from PSSMs [4]. Most of these features can be generated by the Pse-in-One [15].

Ranking methods treat protein remote homology detection as a ranking task or database retrieving task. Some early ranking methods were constructed based on the alignment methods, such as Smith–Waterman algorithm. [16], Local Alignment BLAST [17], PSI-BLAST [18], HAlign [19], [20] etc. Later, the ranking methods were improved by employing more accurate alignment algorithms based on multiple sequence alignments, such as Hmmer [21], Coma [22], HHblits [23], etc. Recently, some advanced techniques have been proposed to further facilitate the development of the ranking methods, such as semantic embedding [24], PageRank [25], Rank Aggregation [26], etc. These ranking methods have achieved the-state-of-the-art performance, and showed complementary predictive results. Therefore, the Learning to Ranking algorithm [27] have been employed to combine different ranking methods in a surprised manner, and a predictor called ProtDec-LTR [1] has been established. Later, the ProtDec-LTR2.0 [28] improved ProtDec-LTR by combining the profile-based protein representation and Learning to Ranking algorithm. Recently, HITS-PR-HHblits [29] performs PageRank algorithm (PR) [30] and Hyperlink-Induced Topic Search algorithm (HITS) [31] on the protein similarity network constructed by HHblits [23] to further improve the detection performance. For more information of these methods, please refer to a recent review paper [32].

All these computational methods have made great contributions to the development of this very important field. However, detection performance improvement is still desired for accurately investigating the structures and functions of proteins. As discussed above, the profile-based features showed the highest discriminative power, and ranking methods achieved the best performance. Can we incorporate the profile-based features into the ranking methods? In order to answer this question, we proposed the feature mapping method to incorporate the profile-based features into the Learning to Ranking algorithm and combine PageRank algorithm and HITS algorithm [29] to further improve the accuracy of protein remote homology detection results, and established a new predictor called ProtDec-LTR3.0, which is an important improved version of ProtDec-LTR1.0 [1] and ProtDec-LTR2.0 [28]. Experimental results showed that ProtDec-LTR3.0 outperformed other existing methods for protein remote homology detection. Finally, a web server of the proposed was established, which can be accessed at http://bliulab.net/ProtDec-LTR3.0/.

II. MATERIALS AND METHOD

A. BENCHMARK DATASET

In order to facilitate the comparison with other existing computational methods, and fairly evaluate the performance of the proposed method, two widely used benchmark datasets were used in this study [24], [26], [28], which were constructed based on the SCOP and SCOPe database [33]. The benchmark datasets can be found at the link http://SCOP.berkeley.edu/astral/.

B. MAIN EXPERIMENTAL FLOW CHART OF PROTDEC-LTR3.0

Learning to Rank algorithm [34] is one of the most powerful machine learning techniques, which has been applied to the field of protein remote homology detection, and showed promising predictive performance [1], [28]. Previous study [29] shows that PageRank algorithm and HITS algorithm can improve the accuracy of query feedback results by constructing protein similarity networks.

In this study, we combined Learning to Rank model, PageRank and HITS to further improve the accuracy the feedback list results. The flow chart of ProtDec-LTR3.0 is shown in **Fig. 1**.

C. BASIC RANKING METHODS

In this study, three state-of-the-art ranking methods (PSI-BLAST [18], Hmmer [21] and HHblits [23]) are viewed as the basic ranking predictors. These methods are complementary, because they can produce different characteristics basing on different technologies, for example, PSI-BLAST [18] is a profile-sequence alignment method, which constructs profiles of query proteins and iteratively searches the database. Hmmer [21] is a HMM-sequence alignment method, which constructs HMM profiles of query proteins. The HMM-HMM alignment method HHblits [23] constructs HMM profiles for both query proteins and proteins in the database, and then iteratively searches the query HMM profile against the database of HMM profiles. The parameters of PSI-BLAST [18] were set as "-num_iterations = 3 and -outfmt = 6". The Jackhmmer was used as the implementation of Hmmer [21]. The parameters of Jackhammer and HHblits [23] were set as default.

For a given query protein q, a set of feedback protein sequences of q were obtained by three basic methods, including PSI-BLAST [18], Hmmer [21] and HHblits [23], which can be represented as:

$$\mathbb{C}(q) = \left\{ p_1 \quad p_2 \quad \cdots \quad p_l \right\} \tag{1}$$

D. BASIC RANKING METHODS

As demonstrated in previously studies [3], [4], [9], [10], profile-based features outperformed the other sequence-based features, because they consider the evolutionary

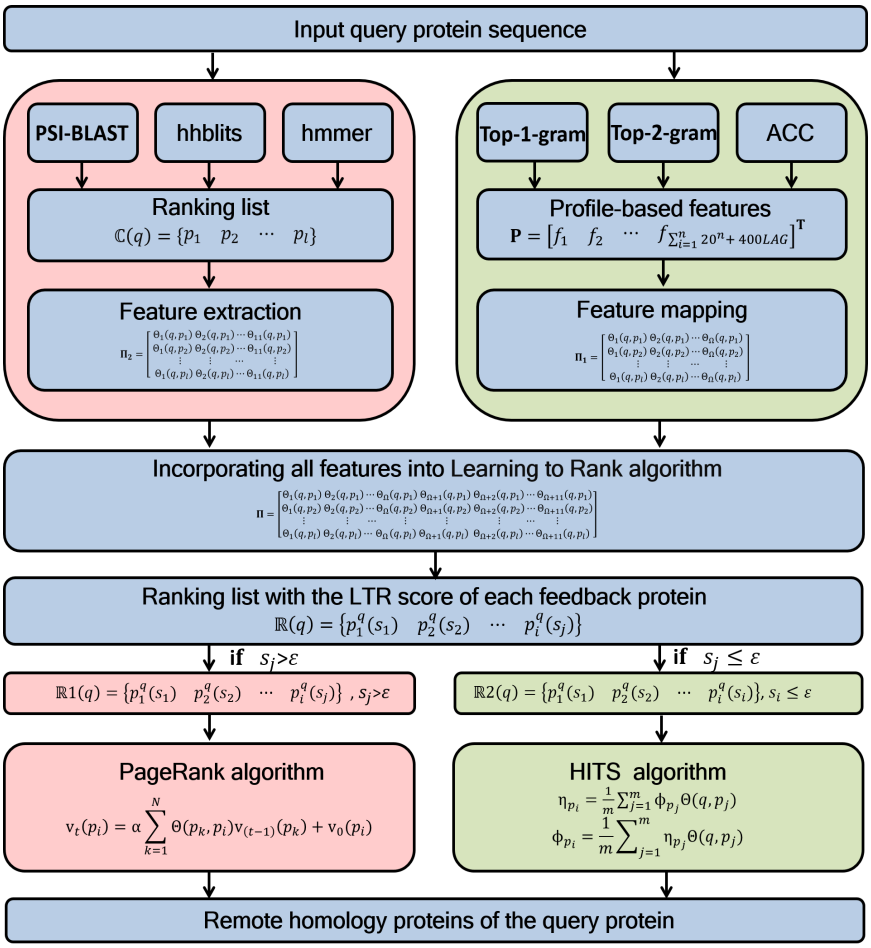

FIGURE 1. The flowchart to show how protdec-LTR3.0 works.

information extracted from multiple sequence alignments. Therefore, two profile-based features were employed in this study, including Top-n-gram [9] and ACC [10].

1) TOP-N-GRAM

Top-n-gram [9] is a novel profile-based building blocks of proteins based on the frequency profiles. In this study, the frequency profiles were generated by using PSI-BLAST [18], to search against the NCBI's nrdb90 database [35] with the parameters that set as default. A protein **P** can be represented by the normalized occurrence frequencies of Top-n-grams as [9]:

$$\mathbf{P} = \left[f_1^{\text{Top-n-gram}} \quad f_2^{\text{Top-n-gram}} \quad \cdots \quad f_{20^n}^{\text{Top-n-gram}} \right]^{\mathbf{T}} \quad (2)$$

where symbol 'T' means the transformation symbol in vector operations. n is the parameter of Top-n-gram [9], and $f_i^{\text{Top-n-gram}}$ represents the normalized frequency of the corresponding Top-n-gram occurring in **P**.

2) AUTOCROSS-COVARIANCE (ACC)

Autocross-covariance [10] is a combination of auto covariance (AC) and cross covariance (CC) [10] based on PSSM profiles. AC is used to represent the correlation between the same amino acids and the distance of *lag*, which is described as follow [10]:

$$\mathrm{AC}(i, lag) = \sum_{j=1}^{L-lag} \frac{\left(m_{i,j} - \overline{m_i}\right)\left(m_{i,j+lag} - \overline{m_i}\right)}{(L - lag)} \quad (3)$$

where $m_{i,j}$ is the score of amino acid i in the j-th position in the PSSM, and $\overline{m_i}$ is the average score of amino acid i. PSSM profiles were generated by PSI-BLAST with the same parameters for generating Top-n-grams. L is the length of the protein sequence. Cross covariance CC represents the correlation between the different amino acids with a distance of *lag*, which is described as follow [10]:

$$\mathrm{CC}\,(i_1, i_2, lag) = \sum_{j=1}^{L-lag} \frac{\left(m_{i_1,j} - \overline{m_{i_1}}\right)\left(m_{i_2,j+lag} - \overline{m_{i_2}}\right)}{(L - lag)} \quad (4)$$

where i_1 and i_2 are different amino acids.

As the combination of AC and CC, the feature vector of ACC can be represented as [10]:

$$\mathbf{P} = \left[f_1^{\text{ACC}} \quad f_2^{\text{ACC}} \quad \cdots \quad f_{400LAG}^{\text{ACC}} \right]^{\mathbf{T}} \quad (5)$$

where *LAG* represents the maximum distance of amino acids, and the value of *LAG* is 2.

E. FEATURE MAPPING STRAEGY

As shown in many previous studies [3], [4], [9], [10], features based on profiles have strong discriminative power to distinguish the remote homology proteins. These features were widely used in many discriminative methods. The features extracted by ACC and top-n-gram cannot only reflect the evolutionary information of proteins, but also contain the characteristics of protein sequences. Can we combine these profile-based features and Learning to Rank algorithm to further improve the performance of the ProtDec-LTR [1] and ProtDec-LTR2.0 [28]? However, it is never an easy task, because all these features cannot directly reflect the relationship between the two proteins including query protein and feedback protein. In this regard, in this study, inspired by the idea of bitwise operation [36], we proposed the feature mapping strategy to measure this relationship. Here, we will introduce its steps.

Two profile-based features Top-n-gram [9] and ACC [10] are used to represent the proteins, and the resulting feature vector of a protein can be represented as:

$$\mathbf{P} = \left[f_1 \quad f_2 \quad \cdots \quad f_{\sum_{i=1}^{n} 20^n + 400LAG} \right]^{\mathbf{T}} \tag{6}$$

where the beginning $\sum_{i=1}^{n} 20^n$ dimensions represent the Top-n-gram features [9], and the last 400 *LAG* dimensions represent the ACC features [10].

The profile-based feature matrix $\mathbf{\Pi_1}$ is used to measure the relationship between the features of query protein and the corresponding features of all the feedback proteins in $\mathbb{C}(\mathbf{q})$ (cf. Eq. 1), which can be represented as:

$$\mathbf{\Pi}_1 = \begin{bmatrix} \Theta_1(q,p_1) & \Theta_2(q,p_1) & \cdots & \Theta_\Omega(q,p_1) \\ \Theta_1(q,p_2) & \Theta_2(q,p_2) & \cdots & \Theta_\Omega(q,p_2) \\ \vdots & \vdots & \cdots & \vdots \\ \Theta_1(q,p_l) & \Theta_2(q,p_l) & \cdots & \Theta_\Omega(q,p_l) \end{bmatrix} \tag{7}$$

where the value of Ω is $\sum_{i=1}^{n} 20^n + 400LAG$, and $\mathbf{\Theta}_u(q,p_i)$ can be calculated by

$$\mathbf{\Theta}_u(q,p_i) = \begin{cases} \left| f_u^{\text{Top-n-gram}}(q) - f_u^{\text{Top-n-gram}}(p_i) \right| & \left| 1 \le \boldsymbol{u} \le \sum_{k=1}^{n} 20^n \right. \\ \left| f_u^{\text{ACC}}(q) - f_u^{\text{ACC}}(p_i) \right| & \sum_{i=1}^{n} 20^n + 1 \le u \le \Omega \end{cases} \tag{8}$$

F. INCORPORATING ALL FEATURES INTO LEARNING TO RANK ALGORITHM

Following previous studies [1], [28], the alignment score feature matrix of the three ranking methods was constructed based on the alignment scores generated by three state-of-the-art ranking methods, including PSI-BLAST [18], Hmmer [21] and HHblits [23]. The alignment score feature matrix $\mathbf{\Pi_2}$ can be represented as :

$$\mathbf{\Pi_2} = \begin{bmatrix} \Theta_1(q,p_1) & \Theta_2(q,p_1) & \cdots & \Theta_{11}(q,p_1) \\ \Theta_1(q,p_2) & \Theta_2(q,p_2) & \cdots & \Theta_{11}(q,p_2) \\ \vdots & \vdots & \cdots & \vdots \\ \Theta_1(q,p_l) & \Theta_2(q,p_l) & \cdots & \Theta_{11}(q,p_l) \end{bmatrix} \tag{9}$$

where $\Theta_1(q,p_i)$, $\Theta_2(q,p_i)$, $\Theta_3(q,p_i)$ and $\Theta_4(q,p_i)$ repress-ent identity, E-value, bit score, and the reciprocal of the posi-tion generated by PSI-BLAST. [18]; $\Theta_5(q,p_i)$, $\Theta_6(q,p_i)$ and $\Theta_7(q,p_i)$ represent E-value, bit score, and the reciprocal of the position generated by Hmmer [21]; $\Theta_8(q,p_i)$, $\Theta_9(q,p_i)$, $\Theta_{10}(q,p_i)$, and $\Theta_{11}(q,p_i)$ represent prob, E-value, bit score, and the reciprocal of the position generated by HHblits [23].

Finally, matrix $\mathbf{\Pi_1}$ and matrix $\mathbf{\Pi_2}$ were combined to train the Learning to Rank model (10), as shown at the top of the next page.

G. COMBINING LEARNING OF RANK, PAGERANKE (PR) AND HYPERLINK-INDUCED TOPIC SEARCH ALGORITHM (HITS)

Previous study [29] indicated that using the similarity among proteins to construct protein similarity network can improve the accuracy of the existing ranking lists. In this study, we followed this study [29] to apply the PageRank (PR) and Hyperlink-Induced Topic Search algorithm (HITS) to further improve the performance of ProtDec-LTR3.0. Its detailed steps will show in the following sections.

The ranking results of query proteins obtained by Learning to Rank [34] model can be represented as:

$$\mathbb{R}(q) = \left\{ p_1^q(s_1) \quad p_2^q(s_2) \quad \cdots \quad p_i^q(s_j) \right\} \tag{11}$$

where p_i^q represents the i-th feedback in the ranking list of the query q, and s_j is the score of the j-th feedback protein p_j^q. The $\mathbb{R}(q)$ is sorted in descending order according to the corresponding feedback protein scores.

We assume proteins with higher scores in $\mathbb{R}(q)$ are more likely to be homologous proteins. According to the characteristics of homologous proteins, they should show a close relationship in the constructed protein similarity network with higher node weights. Therefore, homologous proteins should be detected by local network, while weak homologous proteins should be calculated by global network. Because HITS performs well in local network search, while PageRank performs better in global network search in information retrieval [29]. In this study, we divided the results ranking list of query feedback proteins into strong homology proteins and weak homology proteins according to the scores of feedback proteins, and employed different adjustment strategies for different parts. In this study, if the protein with score value higher than ε (the value of ε is 3), the HITS algorithm was performed on it, otherwise, the PageRank algorithm was employed.

$$\Pi = \Pi_1 + \Pi_2 = \begin{bmatrix} \Theta_1(q,p_1) & \Theta_2(q,p_1) & \cdots & \Theta_\Omega(q,p_1) \\ \Theta_1(q,p_2) & \Theta_2(q,p_2) & \cdots & \Theta_\Omega(q,p_2) \\ \vdots & \vdots & \cdots & \vdots \\ \Theta_1(q,p_l) & \Theta_2(q,p_l) & \cdots & \Theta_\Omega(q,p_l) \end{bmatrix} \begin{bmatrix} \Theta_{\Omega+1}(q,p_1) & \Theta_{\Omega+2}(q,p_1) & \cdots & \Theta_{\Omega+11}(q,p_1) \\ \Theta_{\Omega+1}(q,p_2) & \Theta_{\Omega+2}(q,p_2) & \cdots & \Theta_{\Omega+11}(q,p_2) \\ \vdots & \vdots & \cdots & \vdots \\ \Theta_{\Omega+1}(q,p_l) & \Theta_{\Omega+2}(q,p_l) & \cdots & \Theta_{\Omega+11}(q,p_l) \end{bmatrix} \tag{10}$$

The update rule of PageRank algorithm to update the value of each protein node in the network is calculated by [29]:

$$v_t(p_i) = \alpha \sum_{k=1}^{N} \Theta(p_k, p_i)\, v_{(t-1)}(p_k) + v_0(p_i) \tag{12}$$

where N is all the number of the proteins in the network. For the detailed information of other parameters for constructing protein similarity networks, please refer to [29].

Closed-loop links in the protein similarity network will result in Rank-Leak and Rank-sink [29], [30] during the iteration process. To solve this problem, in this study, we employed smoothing strategy to calculate node value of the protein similarity network. Because the score of feedback protein calculated by Learning to Rank has the same significance for all query-feedback proteins in the benchmark dataset, we set the initial weights of network nodes as their corresponding scores of feedback proteins, which can be represented as:

$$\Theta(p_i, p_j) = \begin{cases} 1, & if\ p_i, \quad p_j \in \mathbb{S} \wedge \mathbb{S}_F(p_i) = \mathbb{S}_F(p_j) \\ \log_{\boldsymbol{\beta}} \left| p_j^{score}(p_i) \right|, & if\ p_i = q, \quad p_i \in \mathbb{S} \wedge p_j \in \mathbb{R}(p_i) \\ \gamma, & otherwise \end{cases} \tag{13}$$

where p_i and p_j are two proteins in database $\mathbb{S}$. $\mathbb{S}_F(p_i) = \mathbb{S}_F(p_j)$ represents p_i and p_j are in the same superfamily. q is the query protein. $p_j^{\text{score}}(p_i)$ represents the score of p_jp$_j$ in the ranking list $\mathbb{R}(p_i)$. The logarithmic function is performed on the $|p_j^{\text{score}}(p_i)|$ to avoid the extreme values, which would have negative influence on the performance of PageRank and HITS. β is a regulator. The function of β is to enhance the score difference of feedback proteins, which is set as 8, which is the optimized value. The value of γ is set as 0.01, which is the default value in the field of information retrieval [37]. The edge weight value Θ' can be defined as:

$$\Theta'(p_i, p_j) = d * \frac{\Theta(p_i, p_j)}{\sum_{k=1}^{N} \Theta(p_i, p_k)} + \frac{1-d}{N} \tag{14}$$

where $\Theta'(p_i, p_j)$ represents the importance of protein pairs (p_i, p_j) in the whole network. The value of N is the number of all nodes. $(1-d)/N$ denotes the smooth value. The value of d is set as 0.99.

The update rule of Hyperlink-Induced Topic Search to calculate the hub value η_{p_i} and authority value ϕ_{p_i} of the protein p_i can be represented as [29]:

$$\eta_{p_i} = \frac{1}{m} \sum_{j=1}^{m} \phi_{p_j} \Theta(q, p_j) \tag{15}$$

$$\Phi_{p_i} = \frac{1}{m} \sum_{j=1}^{m} \eta_{p_j} \Theta(q, p_j) \tag{16}$$

where m is the number of nodes. $\Theta(q, p_j)$ represents the similarity between the feedback proteins p_j and the query protein q calculated by

$$\Theta(q, p_i) = \log_\beta \left| p_j^{\text{score}}(q) \right| \tag{17}$$

where $p_j^{\text{score}}(q)$ represents the score of feedback protein p_j in $\mathbb{R}(q)$. β is a regulator to regulate the score difference of feedback proteins, which was set as 8 in this study.

H. EVALUATION METHODOLOGY

5-fold cross-validation was used to evaluate the performance of different methods [38], [39]. The average ROC1 and ROC50 scores [40] were employed to evaluate the performance of each method. The higher ROC1 and ROC50 are, the better the methods perform.

III. RESULT AND DISCUSSION

A. PROTDEC-LTR3.0 OUTPERFORMS OTHER LTR-BASED METHODS

In order to prove whether incorporating profile-based features, HITS and PR can improve the performance of Learning to Rank or not, the ProtDec-LTR3.0 predictors with different feature combinations were constructed and compared, and the results are shown in **TABLE 1** and **Fig. 2**. The statistical significance between ProtDec-LTR3.0 (AT, PH) and other LTR-based methods was estimated by using Wilcoxon signed rank test [41], [42], and the results are shown in **TABLE 2**. We can see the followings: 1) All the two ProtDec-LTR3.0 predictors outperform the ProtDec-LTR1.0 and ProtDec-LTR2.0, indicating that combining the profile-based features into the framework of Learning to Ranking is an efficient way to improve the predictive performance; 2) ProtDec-LTR3.0 (AT, PH) significantly outperforms all the other three predictors,

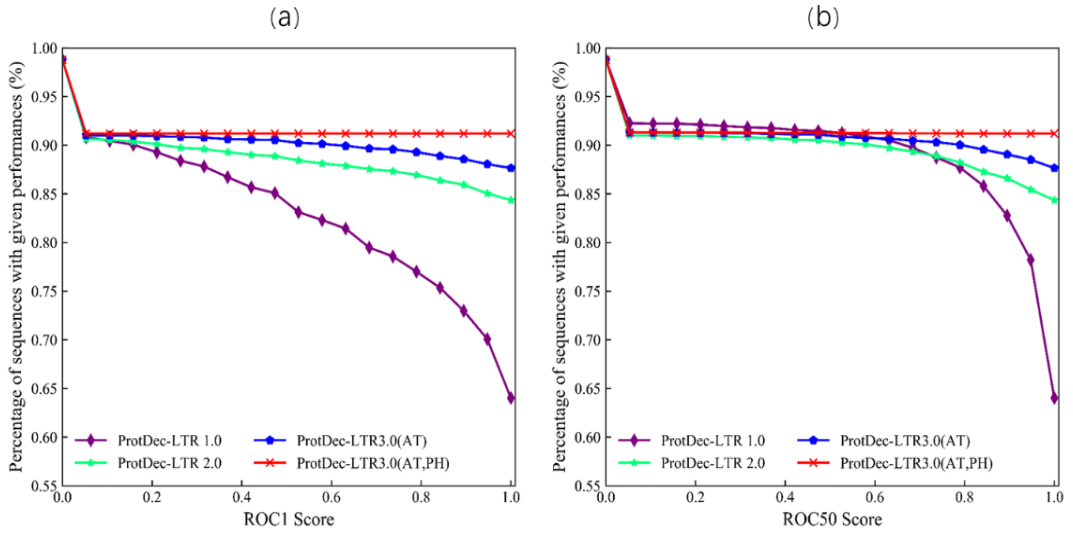

FIGURE 2. Performance comparison of protdec-LTR1.0, protdec-LTR2.0, and two protdec-LTR3.0 predictors based on different features in terms of ROC1 scores (a), and ROC50 scores (b) on the SCOP benchmark dataset. The higher the curve, the better performance of the corresponding predictor is.

TABLE 1. Performance of protdec-LTR3.0 predictors with different features and the reordering strategies on SCOP benchmark dataset via 5-fold cross-validation.

Methods	ROC1	ROC50
ProtDec-LTR1.0 [a]	0.8278±0.0075	0.8944±0.0067
ProtDec-LTR2.0 [b]	0.8839±0.0037	0.8954±0.0043
ProtDec-LTR3.0(AT) [c]	0.8999±0.0048	0.9057±0.0053
ProtDecLTR3.0(AT,PH) [d]	0.9117±0.0061	0.9121±0.0064

[a] represents the ProtDec-LTR1.0 predictor;
[b] represents the ProtDec-LTR2.0 predictor using pseudo protein sequences;
[c] represents the ProtDec-LTR3.0 predictor using three basic ranking methods, and profile-based features (ACC, Top-1-gram and Top-2-gram);
[d] represents the ProtDec-LTR3.0 predictor using all the features and combining Learning to Rank algorithm with reordering strategy including PR and HITS.

TABLE 2. Statistical significance of differences between protdec-LTR3.0(AT,PH) and other three LTR-based predictors on SCOP benchmark dataset [a].

Methods	P-value of ROC1	P-value of ROC50
ProtDec-LTR1.0	1.233e-21	1.111e-18
ProtDec-LTR2.0	2.818e-05	1.815e-04
ProtDecLTR3.0(AT)	3.906e-04	7.040e-04

[a] For the explanations of AT and PH, see the footnote of TABLE I.

indicating that the feature mapping strategy and the reordering strategies (PR and HITS) are able to significantly improve the predictive performance, and it is useful for protein remote homology detection, which is full consistent with previous studies [14], [29].

B. PERFORMANCE COMPARISON WITH HIGHLY RELATED METHODS

The performance of the proposed ProtDec-LTR3.0 predictor is compared with other 11 state-of-the-art predictors, including Hmmer [21], Coma [22], HHblits [23], PSI-BLAST [18], PsePro-Coma [26], PsePro-HHblits [26], PsePro-Hmmer [26], PsePro-PSI-BLAST [26], ProtDec-LTR 1.0 [1], ProtDec-LTR2.0 [28], and HITS-PR-HHblits [29].

TABLE 3. Performance comparison of the protdec-LTR3.0 with 11 state-of-the-art methods on SCOP benchmark dataset via 5-fold cross-validation.

Methods	ROC1	ROC50
ProtDec-LTR3.0[a]	0.9117±0.0061	0.9121±0.0064
HITS-PR-HHblits	0.8852±0.0028	0.8860±0.0029
ProtDec-LTR2.0	0.8839±0.0037	0.8954±0.0043
ProtDec-LTR1.0	0.8510±0.0075	0.8969±0.0067
PsePro-PSI-BLAST	0.7851±0.0102	0.8363±0.0076
PsePro-HHblits	0.8295±0.0056	0.8804±0.0068
PsePro-Hmmer	0.8137±0.0093	0.8302±0.0089
PsePro-Coma	0.7293±0.0105	0.8119±0.0083
PSI-BLAST	0.7499±0.0046	0.8005±0.0066
HHblits	0.8427±0.0077	0.8834±0.0086
Hmmer	0.7894±0.0063	0.7915±0.0061
Coma	0.6972±0.0102	0.7774±0.0081

[a] represents ProtDec-LTR3.0 predictor using all the features (AT) and combining Learning to Rank algorithm with reordering strategies including PR and HITS.

TABLE 3 and **Fig.3** show the predictive results of various methods. **TABLE 4** is the statistical significance of differences between ProtDec-LTR3.0 and other state-of-the-art methods on SCOP benchmark dataset by using Wilcoxon signed rank test [41], [42]. The following conclusions can be reached that the ProtDec-LTR3.0 predictor significantly outperforms other competing methods. These results further confirm that the ProtDec-LTR3.0 predictor is an accurate approach for protein remote homology detection, and will facilitate the development of protein sequence analysis.

C. PERFORMANCE OF PROTDEC-LTR3.0 PREDICTORS ON THE SCOPE BENCHMARK DATASET

For the SCOP benchmark dataset, there are 359 protein superfamilies with only one protein. Therefore, when these proteins are used as the queries, their homology proteins cannot be detected without the template in the dataset. In other

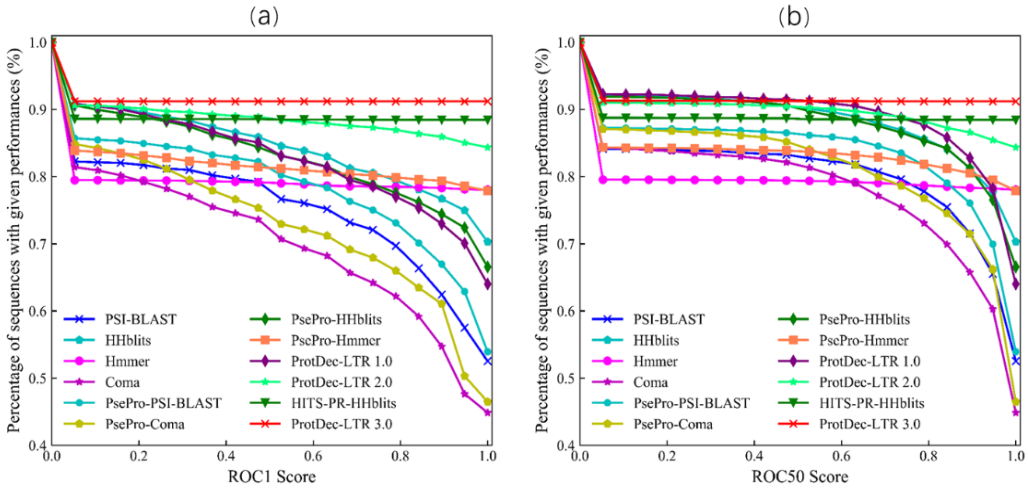

FIGURE 3. Performance comparison of various methods in terms of ROC1 scores (a), and ROC50 scores (b). The higher the curve is, the better performance of the corresponding predictor is.

TABLE 4. Statistical significance of differences between protdec-LTR3.0 with others state-of-the-art methods on SCOP benchmark dataset.

Methods	P-value of ROC1	P-value of ROC50
HITS-PR-HHblits	4.680e-07	2.357e-07
ProtDec-LTR2.0	2.812e-05	1.815e-04
ProtDec-LTR1.0	1.233e-21	1.111e-18
PsePro-PSI-BLAST	2.287e-39	1.747e-35
PsePro-HHblits	1.591e-16	4.212e-13
PsePro-Hmmer	5.426e-13	6.214e-12
PsePro-Coma	9.335e-54	9.454e-48
PSI-BLAST	5.810e-63	2.230e-56
HHblits	1.819e-12	3.460e-10
Hmmer	2.333e-19	6.811e-18
Coma	5.811e-63	2.230e-56

words, when using such dataset to evaluate the ProtDec-LTR3.0 predictor, its performance will be underestimated. This is the main reason for the abrupt decline of ROC1 and ROC50 scores starting from 0 to the next immediate point for the **Figs. 2**,**3**. Therefore, we used a more updated and comprehensive dataset SCOPe benchmark dataset to further evaluate the ProtDec-LTR3.0 predictors. In order to reduce the influence of the query proteins without templates, we construct another SCOPe benchmark dataset SCOPe-R via removing the superfamilies with only one protein sequence. The results are shown in **TABLE 5**, from which we can see the followings: 1) As we expected, all the four ProtDec-LTR3.0 predictors trained with SCOPe database (see **TABLE 5**) outperform the ProtDec-LTR3.0 predictors trained with SCOP database (see **TABLE 1**). The reason is that the SCOPe database is more comprehensive than the SCOP database, leading to a more accurate Learning to Rank model and a more comprehensive network, based on which the remote homology relationship of proteins can be accurately detected with the help of PR and HITS. 2) For both the two ProtDec-LTR3.0 predictors, their predictive results on SCOPe-R are

TABLE 5. Performance of protdec-LTR3.0 predictors with different features on SCOPe and SCOPe-R benchmark datasets via 5-fold cross-validation [a].

Methods	Datasets	ROC1	ROC50
ProtDec-LTR1.0 [a]	SCOPe	0.9596±0.0026	0.9617±0.0026
ProtDec-LTR2.0 [b]	SCOPe-R	0.9835±0.0024	0.9842±0.0025
ProtDec-LTR3.0 (AT) [c]	SCOPe	0.9717±0.0024	0.9789±0.0024
ProtDec-LTR3.0 (AT,PH) [d]	SCOPe-R	0.9912±0.0019	0.9916±0.0018

[a] For the explanations of AT and PH, see the footnote of **TABLE I**.

significantly higher than on SCOPe, indicating that the proteins without templates do have impact on the detection performance, because they cannot be detected.

Please note that although ProtDec-LTR3.0 employs three profile-based features, including PSI-BLAST, ACC and Top-n-grams. These features extract the evolutionary information from profiles in different approaches, and previous study [4] showed that these features are complementary, for examples PSI-BLAST calculates the alignment scores between two proteins. ACC calculates the correlation between amino acids at different positions in the PSSMs. Top-n-gram removes the noise information in the PSFMs by the occurrences of amino acids with high frequencies. Therefore, the performance of ProtDec-LTR3.0 can be improved by combining these features.

D. WEB SERVER AND USER GUIDE

Construction of publicly accessible web servers is a key step for developing useful bioinformatics tools. In this regard, the corresponding web server of ProtDec-LTR3.0 has been established, by which users only need to submit the query protein sequences in FASTA format, and their homology proteins will be automatically detected, and shown. Its detailed steps are as follows.

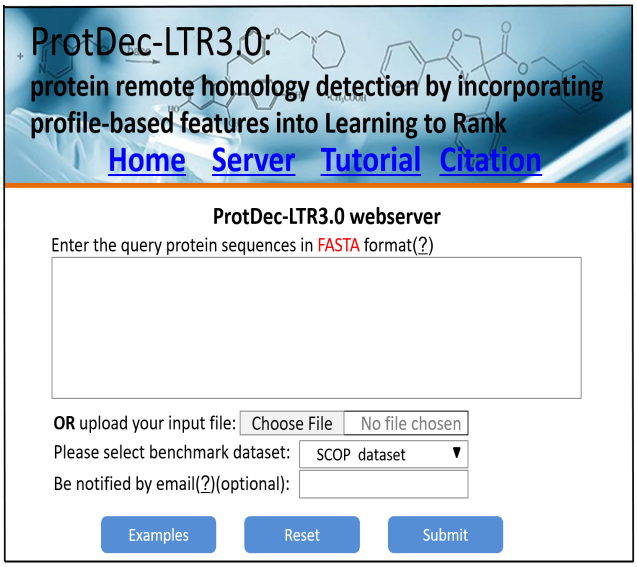

FIGURE 4. **The screenshot of the protdec-LTR3.0 web server, which can be accessed from http://bliulab.net/ProtDec-LTR3.0/.**

Step 1. Clicking the link http://bliulab.net/ProtDec-LTR3.0/, you will see the page of ProtDec-LTR3.0 web server as shown in **Fig. 4**.

Step 2. Click on the sever button, and the system will jump to the server interface. In the input box of the server interface, you can type, copy and paste FASTA formatted proteins, or click the Browse button to upload directly.

Step 3. In order to satisfy different precision requirements of users, we provide two models for users. Users can select different models through different models through dataset selection buttons. Then click the submit button, and the predictive results will be displayed on the screen. For example, when you select the default model and use two query sequences in the Example window as input, and then click the Submit button, you will see for the first query sequence, its 18 homology proteins are detected, and for the second query sequence, its 4 homology proteins are detected, which are completely consistent with experimental observations. The 3D structure information of the homology proteins can also be visualized.

ACKNOWLEDGMENT

The authors are very much indebted to the two anonymous reviewers, whose constructive comments are very helpful in strengthening the presentation of this article. They also would like to thank Xiaopeng Jin for his helpful discussion.

REFERENCES

[1] B. Liu, J. Chen, and X. Wang, "Application of learning to rank to protein remote homology detection," (in English), *Bioinformatics*, vol. 31, no. 21, pp. 3492–3498, Nov. 2015.doi: 10.1093/bioinformatics/btv413.

[2] L. Wei and Q. Zou, "Recent progress in machine learning-based methods for protein fold recognition," *Int. J. Mol. Sci.*, vol. 17, no. 12, p. 2118, 2016.

[3] B. Liu, J. Chen, and X. Wang, "Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis," *Mol. Genet. Genomics*, vol. 290, no. 5, pp. 1919–1931, 2015.

[4] B. Liu, "BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Brief Bioinf.*, Dec. 2017. doi: 10.1093/bib/bbx165.

[5] C. Lin, "Hierarchical classification of protein folds using a novel ensemble classifier," (in English), *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e56499. doi: 10.1371/journal.pone.0056499.

[6] L. Wei, M. Liao, X. Gao, and Q. Zou, "An improved protein structural classes prediction method by incorporating both sequence and structure information," *IEEE Trans. Nanobiosci.*, vol. 14, no. 4, pp. 339–349, Jun. 2015. doi: 10.1109/tnb.2014.2352454.

[7] L. Wei, M. Liao, X. Gao, and Q. Zou, "Enhanced protein fold prediction method through a novel feature extraction technique," (in English), *IEEE Trans. Nanobiosci.*, vol. 14, no. 6, pp. 649–659, Sep. 2015. doi: 10.1109/Tnb.2015.2450233.

[8] X. Zhao, Q. Zou, B. Liu, and X. Liu, "Exploratory predicting protein folding model with random forest and hybrid features," (in English), *Current Proteomics*, vol. 11, no. 4, pp. 289–299, 2014.

[9] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-*n*-grams and latent semantic analysis," (in English), *BMC Bioinf.*, vol. 9, no. 1, p. 510, 2008. doi: 10.1186/1471-2105-9-510.

[10] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," (in English), *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, Oct. 2009. doi: 10.1093/bioinformatics/btp500.

[11] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between top-n-gram and residue pairs for protein remote homology detection," (in English), *BMC Bioinf.*, vol. 15, p. S3, Jan. 2014. doi: 10.1186/1471-2105-15-S2-S3.

[12] K. Yan, X. Fang, Y. Xu, and B. Liu, "Protein fold recognition based on multi-view modeling," *Bioinformatics*, Jan. 2019. doi: 10.1093/bioinformatics/btz040.

[13] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," (in English), *PLoS ONE*, vol. 7, no. 9, Sep. 2012, Art. no. e46633. doi: 10.1371/journal.pone.0046633.

[14] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, K.-C. Chou, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," (in English), *Bioinformatics*, vol. 30, no. 4, pp. 472–479, Feb. 2014. doi: 10.1093/bioinformatics/btt709.

[15] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," (in English), *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015. doi: 10.1093/nar/gkv458.

[16] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, 1981.

[17] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," (in English), *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990. doi: 10.1016/S0022-2836(05)80360-2.

[18] S. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," (in English), *Faseb J.*, vol. 12, no. 8, p. A1326, Apr. 1998.

[19] Q. Zou, Q. Hu, M. Guo, and G. Wang, "HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy," (in English), *Bioinformatics*, vol. 31, no. 15, pp. 2475–2481, Aug. 2015. doi: 10.1093/bioinformatics/btv177.

[20] S. Wan and Q. Zou, "HAlign-II: Efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing," (in English), *Algorithms Mol. Biol.*, vol. 12, p. 25, Sep. 2017. doi: 10.1186/s13015-017-0116-x.

[21] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER Web server: Interactive sequence similarity searching," (in English), *Nucleic Acids Res.*, vol. 39, pp. W29–W37, May 2011. doi: 10.1093/nar/gkr367.

[22] M. Margelevičius , M. Laganeckas, and C. Venclovas, "COMA server for protein distant homology search," (in English), *Bioinformatics*, vol. 26, no. 15, pp. 1905–1906, Aug. 2010. doi: 10.1093/bioinformatics/btq306.

[23] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," (in English), *Nature Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012. doi: 10.1038/Nmeth.1818.

[24] I. Melvin, J. Weston, W. S. Noble, and C. Leslie, "Detecting remote evolutionary relationships among proteins by large-scale semantic embedding," (in English), *PLoS Comput. Biol.*, vol. 7, no. 1, Jan. 2011, Art. no. e1001047. doi: 10.1371/journal.pcbi.1001047.
[25] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble, "Protein ranking: From local to global structure in the protein similarity network," (in English), *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 17, pp. 6559–6563, Apr. 2004. doi: 10.1073/pnas.0308067101.
[26] J. Chen, R. Long, X.-L. Wang, B. Liu, and K.-C. Chou, "dRHP-PseRA: Detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation," (in English), *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 32333. doi: 10.1038/srep32333.
[27] T. Qin, T.-Y. Liu, J. Xu, and H. Li, "LETOR: A benchmark collection for research on learning to rank for information retrieval," (in English), *Inf. Retr.*, vol. 13, no. 4, pp. 346–374, Aug. 2010. doi: 10.1007/s10791-009-9123-y.
[28] J. Chen, M. Guo, S. Li, and B. Liu, "ProtDec-LTR2.0: An improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank," (in English), *Bioinformatics*, vol. 33, no. 21, pp. 3473–3476, Nov. 2017. doi: 10.1093/bioinformatics/btx429.
[29] B. Liu, S. Jiang, and Q. Zou, "HITS-PR-HHblits: Protein remote homology detection by combining PageRank and hyperlink-induced topic search," *Briefings Bioinf.*, Nov. 2018. doi: 10.1093/bib/bby104.
[30] M. Franceschet, "PageRank: Standing on the shoulders of giants," *Commun. ACM*, vol. 54, no. 6, pp. 92–101, Jun. 2011. doi: 10.1145/1953122.1953146.
[31] Y. Du, X. Tian, W. Liu, M. Wang, W. Song, Y. Fan, and X. Wang, "A novel page ranking algorithm based on triadic closure and hyperlink-induced topic search," (in English), *Intell. Data Anal.*, vol. 19, no. 5, pp. 1131–1149, 2015. doi: 10.3233/Ida-150762.
[32] J. Chen, M. Guo, X. Wang, and B. Liu, "A comprehensive review and comparison of different computational methods for protein remote homology detection," (in English), *Briefings Bioinf.*, vol. 19, no. 2, pp. 231–244, Mar. 2018. doi: 10.1093/bib/bbw108.
[33] J.-M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL compendium in 2004," *Nucleic Acids Res.*, vol. 32, pp. D189–D192, Jan. 2004.
[34] A. Trotman, "Learning to rank," *Inf. Retr.*, vol. 8, no. 3, pp. 359–381, Jan. 2005. doi: 10.1007/s10791-005-6991-7.
[35] L. Holm and C. Sander, "Removing near-neighbour redundancy from large protein sequence collections," (in English), *Bioinformatics*, vol. 14, no. 5, pp. 423–429, Jun. 1998. doi: 10.1093/bioinformatics/14.5.423.
[36] M. S. Mulekar and C. S. Brown, *Distance and Similarity Measures*. New York, NY, USA: Springer, 2014.
[37] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, vol. 15. Upper Saddle River, NJ, USA: Prentice-Hall, 2004, no. 5, pp. 1211–1214.
[38] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, Sep. 2004.
[39] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," (in English), *J. Econ.*, vol. 187, no. 1, pp. 95–112, Jul. 2015. doi: 10.1016/j.jeconom.2015.02.006.
[40] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[41] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945. doi: 10.2307/3001968.
[42] S. Wan, M.-W. Mak, and S.-Y. Kung, "mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction," (in English), *Anal. Biochem.*, vol. 473, pp. 14–27, Mar. 2015. doi: 10.1016/j.ab.2014.10.014.

BIN LIU received the Ph.D. degree from the Harbin Institute of Technology, China, in 2010. From 2010 to 2012, he was a Postdoctoral Researcher with The Ohio State University, USA. He was a Professor with the Harbin Institute of Technology, Shenzhen, from 2012 to 2019. He is currently a Full Professor with the Beijing Institute of Technology. He is also putting the focus on exploring the language models of biological sequences and proposing computational predictors for some important tasks in bioinformatics based on natural language processing techniques. His research interests include bioinformatics, machine learning, and natural language processing.

YULIN ZHU is currently pursuing the master's degree with the Computer Science and Technology, Harbin Institute of Technology, Shenzhen. His research interests include bioinformatics and machine learning.