IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received February 15, 2020, accepted March 5, 2020, date of publication March 11, 2020, date of current version March 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2979888

INVITED PAPER

MF-EFP: Predicting Multi-Functional Enzymes Function Using Improved Hybrid Multi-Label Classifier

XUAN XIAO[1], LI-WEN DUAN[1], GUANG-FU XUE[1], GANG CHEN[1], PU WANG[2], AND WANG-REN QIU[1]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jingdezhen 333403, China
[2]Computer School, Hubei University of Arts and Science, Xiangyang 441053, China

Corresponding authors: Xuan Xiao (jdzxiaoxuan@163.com) and Wang-Ren Qiu (qiuone@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 31860312, Grant 31760315, and Grant 61841104, in part by the Natural Science Foundation of Jiangxi Province, China, under Grant 20171ACB20023, in part by the Department of Education of Jiangxi Province under Grant GJJ160866 and Grant GJJ180703, and in part by the China-Montenegro Intergovernmental S&T Cooperation under Grant 2018-3-3.

ABSTRACT Predicting enzymes function is an important and difficult problem, particularly when enzymes may have the multiplex character, i.e., some enzymes simultaneously have two or three function classes. Most of the existing enzyme function predictor can only be used to deal with the mono-functional enzymes. Actually, multi-functional enzymes should not be ignored because they usually possess diverse biological functions worthy of our special notice. By introducing the ''improved Hybrid Multi-label Classifier'' and ''neighbor score'', a new predictor, called **MF-EFP**, has been developed that can be used to deal with the systems containing both mono-functional and multi-functional enzymes. As demonstration, the jackknife cross-validation was performed with MF-EFP on a benchmark dataset of enzymes classified into the following 7 functional classes: (1) EC 1 Oxidoreductase, (2) EC 2 Transferase, (3) EC 3 Hydrolase, (4) EC 4 Lyase, (5) EC 5 Isomerase, (6) EC 6 Ligase, (7) EC7 Translocases, where none of enzymes included has $\geq$90% pairwise sequence identity to any other in a same subset. The subset accuracy and average precision thus obtained by MF-EFP was 85.62% and 94.16% respectively. Extensive experiments also show that MF-EFP can outperform the existing predictors that also have the capacity to deal with such a complicated and stringent system. As a user-friendly web-server, MF-EFP is freely accessible to the public at the web-site http:// www.jci-bioinfo.cn/MF-EFP.

INDEX TERMS Multi-functional enzyme, multi-label learning, neighbor score, hybrid method, function prediction.

I. INTRODUCTION

Enzymes play crucial roles in the catalysis of biological and chemical reactions and are considered to be one of the most important biocatalysts in all biological processes. As a biocatalyst, many chemical reactions can be accelerated after being catalyzed. Based on enzyme-catalyzed chemical reactions, different enzymes were numerically classified by enzymatic Commission number (EC). Enzyme function prediction is important as can be viewed from the following four aspects. (1) It is a significant step toward designing novel enzymes. (2) It can help our understanding of the intricate pathways that regulate biological processes at the cellular level. (3) It is very useful for diagnosing enzyme-related diseases. (4) It can speed up the process of prioritizing drug targets.

Protein sequences growth has spurted in the post-genomic age, it is highly desired to develop computational methods for timely and effectively identifying functions for newly found enzymes, due to both the high costs and time-consuming nature of wet-lab biochemical experiments. Actually, the automated prediction the enzymes functions has been an important topic in the field of bioinformatics. During the last decade, various predictors have been proposed for both mono-functional and multi-functional enzymes on

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

different datasets. These predictors each had their own advantages and played a role in stimulating the development of predicting enzyme function although they also each had their own limitations.

The development of predicting enzyme function has generally followed five main research directions. Firstly, based on the fact that structure similarity enzymes have similar functions, many researches, such as [1]–[3], focused on predicting the enzyme function by searching the similar structure enzymes that their function have been determined by experiments in the database or the library. However, enzyme structure prediction is necessary but as a matter of fact structure prediction is still relatively immature. The errors in the process of structure prediction and function prediction should have a negative effect on the final prediction result. Secondly, based on the fact that enzymes with similar sequence have similar functions, the most straightforward method is to use the sequence similarity search-based tools, for example, BLAST, to search enzyme database for those enzymes with high sequence similarity to the query enzyme. Subsequently, the function annotations of the enzymes thus found are used to deduce the function of query enzyme [4]–[6]. However, this kind of method failed to work when the enzymes have the remote homology enzyme means that the enzyme have same function but these sequences are not similar. Thirdly, extracting more useful information from enzyme sequences via different models and predicting the enzyme function using machine learning algorithms is the universal studied direction. Various discrete models to represent enzyme sequences were proposed in hopes to establish some sort of correlation through which the prediction could be more effectively carried out, such as from amino acid composition, to amino acid physico-chemical properties [7], to the various modes of pseudo amino acid composition [8], and to the higher-level forms of pseudo amino acid composition by conjoint triad feature and hierarchical context [9], sequential evolution information [10], InterPro signatures [11], and functional domain information. K-nearest neighbor method, Adaptive fuzzy K-nearest neighbor method [12], support vector machine [13], Neural network system [14], deep learning [15] have been proposed for enzyme classification. Fourthly, Enzyme Commission (EC) system specifies the function of an enzyme by four digits and has a tree structure, many researchers predict enzyme functional classes and subclasses from top to bottom method, Che *et al.* [10] predicted enzyme main six classes, corresponding to (1) Oxidoreductase, (2) Transferase, (3) Hydrolase, (4) Lyase, (5) Isomerase, (6) Ligase. EzyPrd is a three-layer predictor that can predict the enzyme main classes and subclasses [16]. Enzml and EFICAz are four-layer predictor that can predict four EC digit levels [17]. Fifthly, most of these existing methods were established based on the assumption that an enzyme can catalyze only one reaction specifically. Such an assumption is valid only for mono-functional enzymes but not for multi-functional enzymes that can catalyze two to six chemical reactions. Multi-functional enzymes actually constitute a relatively large part of all the enzymes. With regard to multifunctional enzyme prediction, Zou *et al.* [7] proposed two feature models to make predictions and obtained 99.54% and 98.73% accuracy by using 20-D and 188-D features, respectively; however, dataset redundancy was not mentioned in the paper. Subsequently, Zou and Xiao [18] and Che *et al.* [10] predict multifunctional enzyme in the case of taking redundancy into account. Zou used three feature extraction algorithms to compare results, the best one is SAAC with 90.57% accuracy. Che applied feature extraction from PSSM (position-specific scoring matrix), 91.25% accuracy is achieved in multi-functional enzyme prediction. Amidi *et al.* [3] combined structural and sequence information based on the ML-KNN and ML-SVM to predict enzyme function on a specific dataset Zou *et al.* [19] designed mIDEEPre to predict multi-functional enzyme function based on deep learning.

However, exiting predictors have following shortcomings. (1) Only 6 main functional classes can be predicted. In order to enhance the power of practical application, the coverage scope should be enlarged, such as from covering only 6 main functional classes to 7 main functional classes. (2) It was through an optimal threshold factor to control the prediction of multiple function, it would be more natural if we could find a more intuitive approach to deal with such a problem. The present study was initiated in an attempt to develop a new and more powerful predictor by addressing the above two problems.

II. MATERIALS AND METHODS

A. BENCHMARK DATASET

Enzyme sequences were taken from the Release 2019.4 of Enzyme nomenclature database at website http: //enzyme.expasy.org/, which allows one to select enzyme sequence entries according to their function classes. In order to collect multi-functional enzymes and meanwhile ensure a high-quality for the benchmark dataset, the following criteria should be strictly considered: (a) Enzyme sequences with keyword ''multi-functional'' were collected; (b) Sequences annotated with ''fragment'' were took out; also, sequences with < 50 amino acid residues were excluded because they might just be fragments; (c) To reduce the influence of redundancy and homology bias, the program CD-HIT was used to remove these enzymes that had > 90 % pairwise sequence identify to any other in a same subset.

Finally, we obtained 4479 different enzyme sequences which covers 7 different classes and can be formulated as follows:

$$\mathrm{S} = \mathrm{S}_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \cup S_7 \tag{1}$$

where $\cup$ stands for the symbol for ''union'' in the set theory, while S_1 represents the subset of oxidoreductases (EC 1), S_2 for transferases (EC 2), S_3 for hydrolases (EC 3), S_4 for lyase (EC 4), S_5 for isomerase (EC 5), S_6 for ligase (EC 6), S_7 for translocase (EC 7). The particular information about the above collected dataset is listed in Table 1.

TABLE 1. Distribution of multifunctional enzymes (release 2019.4) before and after CD-HIT (0.9).

Dataset	EC1	EC2	EC3	EC4	EC5	EC6	EC7	Total
Before redundancy	1840	4952	3488	1495	426	337	138	7718
After CD-HIT	1076	2814	1924	854	237	205	49	4479

In this study, the same benchmark dataset as investigated in Che *et al.* [10] was adopted for demonstration. It can be directly downloaded from the web-site at http://server.malab.cn/ MEC/download.jsp. The reasons we choose it as a compare dataset for the current study are that che's dataset contains the multifunctional enzymes have been well documented and reported in recent papers. Table 2 shows the distribution of multifunctional enzymes in the 6 functional classes. Of the 2181 different multifunctional enzymes sequences, 1983 to two classes, 189 to three classes, 9 to four classes.

B. REPRESENTATION OF ENZYME SAMPLE

Given a query enzyme sequence P as formulated by

$$P = Q_1 Q_2 Q_3 \cdots Q_L \tag{2}$$

where Q_1 represents the first residue in enzyme sequence, Q_2 represents the second residue, …, Q_L the L-th residue. How can we use its sequence information to predict which function(s) the enzyme P belongs to? In recent, many methods for predicting various protein attributes were based on the split Amino acid composition (SAAC) discrete model [18], [20] because the SAAC avoids completely losing the sequence-order information. Afridi and Lee have developed SAAC-based method and genetic ensemble classifier to predict mitochondrial achieved reasonable accuracy [21]. In our research, the SAAC and PseAAC (pseudo amino acid composition) hybrid model was proposed to represent the sample of an enzyme.

In SAAC model, the enzyme sequences are divided in parts and compositions of each part are calculated separately. In our SAAC model, each enzyme is divided into three parts: (i) 25 amino acids of N-termini, (ii) 25 amino acids of C-termini, and (iii) region between these two terminuses. As can be expressed by

$$P_1 = Q_1 Q_2 Q_3 \cdots Q_{25} \tag{3}$$

$$P_2 = Q_{26} Q_{27} Q_{28} \cdots Q_{L-25} \tag{4}$$

$$P_3 = Q_{L-24} Q_{L-23} Q_{L-22} \cdots Q_L \tag{5}$$

According to PseAAC, P_1, P_2, and P_3 enzyme sequence can be converted into a $20 + \Lambda$ dimension vector respectively, among the $20 + \Lambda$ elements, the first 20 represent the amino acid composition of the 20 native amino acids, while the latter Λ elements represent the sequence-order information. The sequence-order information can be indirectly represented by the following expression:

$$\delta_\eta = \frac{1}{L-\eta} \sum_{i=1}^{L-\eta} \Omega\left(R_i, R_{i+\eta}\right), \quad (\eta = 1, 2, \cdots, \Lambda) \tag{6}$$

In general, Λ should less than the length of the P_1, P_2, and P_3. The δ_η is the η-th tier correlation factor with that reflects the sequence-order information between all the η-th most contiguous residues separated by η. The correlation function $\Omega\left(R_i, R_j\right)$ can be defined as follows:

$$\Omega\left(R_i, R_j\right) = \frac{1}{3}\left\{\left[F\left(R_j\right) - F\left(R_i\right)\right]^2 + \left[G\left(R_j\right) - G\left(R_i\right)\right]^2 + \left[H\left(R_j\right) - H\left(R_i\right)\right]^2\right\} \tag{7}$$

where $F(R_i)$, $G(R_i)$ and $H(R_i)$ are the evaluated values of hydrophobicity, hydrophilicity, and mass, respectively. Before the three types of values were used, a standard conversion should be conducted using Eq. (7) of [22]

$$p_\varphi = \begin{cases} \dfrac{f_\varphi}{\sum_{i=1}^{20} f_i + \omega \sum_{\eta=1}^{\Lambda} \delta_n} & (1 \le \varphi \le 20) \\ \dfrac{\omega \delta_{\varphi-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{\eta=1}^{\Lambda} \delta_\eta} & (20+1 \le \varphi \le 20+\Lambda) \end{cases} \tag{8}$$

where ω is the weight factor, f_i $(i = 1, 2, 3, \cdots 20)$ represent the normalized occurrence frequencies of the 20 native amino acids, and δ_η is the η-th tier sequence-correlation factor, which can be computed by Eq. (6). According to Eqs. (3)–(8), the three parts of enzyme P can be formulated respectively, the enzyme P can been expressed by integrating three parts:

$$P = \left[p_1, p_2, \cdots, p_{60}, p_{60+1}, \cdots, p_{60+3\Lambda}\right]^T \tag{9}$$

C. PREDICTION ENGINE

A novel multi-label classification algorithm hML-KNN is proposed by integrating the feature score, neighbor score, and a self-adapted label assigning threshold, described in the following sections. hML-KNN is based on hMuLab multi-label algorithm. The performance of hMuLab is accurate and stable in biomedical multi-label classification. The main reason is that there are two complementary scoring methods for comprehensive modeling. The measurement feature score $f_1(e, y_j)$ is calculated to evaluate whether the enzyme e has the functional class y_j based on a regression model. In the feature scoring method, the linear decision function is calculated according to the global information in the training data. The neighbor score $f_2(e, y_j)$ is used to evaluate how significantly the neighbors of an enzyme e tend to have the functional label y_j. However, neighborhood scoring method uses the non-linear aggregation of label information to incorporate label correlation into the model. The functional class y_j whether belongs to enzyme e is quantified as a weighted sum of feature score $f_1(e, y_j)$ and neighbor score $f_2(e, y_j)$ as

$$f\left(e, y_j\right) = a f_1\left(e, y_j\right) + (1-a) f_2(e, y_j) \tag{10}$$

where $1 \le \mathrm{j} \le 7$, a is the weight factor and $0 \le a \le 1$. The functional classes belong to enzyme e are defined as

$$\mathrm{h}\,(e) = \left\{y_j \,|f(e, y_j) \ge 0, 1 \le j \le 7\right\} \tag{11}$$

Therefore, the hybrid method hMuLab provides a valid way to integrate different information sources for predicting multi-label enzyme functional classification problems, which may be a complementary solution to existing algorithms. A detailed description about how the classifier works is clearly described in [23]. The predictor established in this work has the ability in predicting the functional classes of both singleplex and multiplex functional enzymes.

In hMuLab, an optimal threshold factor to control the prediction of multifunctional enzymes is to set a constant as described in Eq.11. However, the constant threshold does not consider the difference between different enzymes. It would be more natural if we could find a more intuitive approach to deal with such a problem. Now, for a query enzyme e, its functional classes will be predicted according to the following steps.

Step 1: The number of how many different functional classes it belongs to will be determined by its nearest neighbor enzyme in train dataset S. For example, suppose e^* is the nearest enzyme to e in S. if e^* has only one functional class, then e will also have only one function; if e^* has two functional classes, then e will also have two functional classes; and so forth.

Step 2: However, the concrete functional class(es) to which e belongs will not be the same as e^* does, but determined by the element(s) in Eq.11 that has(have) the highest score(s). for example, if e is found belonging to only one function in Step 1, and the highest score in Eq.11 is $f(e, y_7)$, then enzyme e will be predicted that it has the translocase function. If e is found belonging to three functional classes in Step 1, and the first highest scores in Eq. 11 are $f(e, y_1)$, $f(e, y_3)$, and $f(e, y_6)$, then e will be predicted that it have function oxidoreductases, hydrolases, and ligase simultaneously.

The entire classifier thus established is called MF-EFP, which can be used to predict the functional class of both mono-functional and multi-functional enzymes.

III. MEASUREMENT

In single-label learning, the metrics such as accuracy, recall, precision, F-measure are frequently used. But in multi-label learning, each sample is a label set, which makes the evaluation index much more complex.

A test set consisting of m multiple label samples expressed by $S = \{(x_i, y_i) \mid i = 1, 2, 3 \cdots m\}$, where $x_i \in \chi$ is a feature vector in sample space and $y_i \in \Gamma$ is the label set to which x_i belongs. Here, we employed both example-based and label-based methods.

$$TP_j = |\{x_i \mid l_i \in y_i \wedge l_i \in h(x_i), 1 \le i \le m\}| \quad (12)$$

$$FP_j = |\{x_i \mid l_i \notin y_i \wedge l_i \in h(x_i), 1 \le i \le m\}| \quad (13)$$

$$TN_j = |\{x_i \mid l_i \notin y_i \wedge l_i \notin h(x_i), 1 \le i \le m\}| \quad (14)$$

$$FN_j = |\{x_i \mid l_i \in y_i \wedge l_i \notin h(x_i), 1 \le i \le m\}| \quad (15)$$

where $h(x_i)$ indicates the set of predicted the labels of the enzyme x_i, and $1 \le j \le 7$. $B(TP_j, FP_j, TN_j, FN_j)$ denotes the use of a binary evaluation index $B \in \{\text{Precison, Recall, F} - \text{score}\}$ for the $j-$ th label, while the label-based multi-label evaluation index has the following two modes:

Macro-averaging:

$$B_{macro} = \frac{1}{7}\sum_{j=1}^{7} B(TP_j, FP_j, TN_j, FN_j) \quad (16)$$

Micro-averaging:

$$B_{micro} = B\left(\sum_{j=1}^{7} TP_j, \sum_{j=1}^{7} FP_j, \sum_{j=1}^{7} TN_j, \sum_{j=1}^{7} FN_j\right) \quad (17)$$

The performance of the learner on each sample is examined by the evaluation index of the sample, and then the average results of all samples are taken.

Hamming Loss:

$$\text{HammingLoss} = \frac{1}{m}\sum_{i=1}^{m} \frac{|h(x_i)\, \Delta y_i|}{7} \quad (18)$$

where Δ denotes the symmetric difference between two sets, $|\cdot|$ is used to find the cardinals of the set (the number of elements),

Subset Accuracy:

$$\text{SubsetAccuracy} = \frac{1}{m}\sum_{i=1}^{m} I(h(x_i) = y_i) \quad (19)$$

where I is the Kronecker delta:

$$\begin{cases} I(h(x_i) = y_i) = 1, & \textit{if and only all the labels in } h(x_i) \\ & \textit{are equal to those in } y_i \\ I(h(x_i) = y_i) = 0, & \textit{otherwise} \end{cases} \quad (20)$$

One Error:

$$\text{OneError} = \frac{1}{m}\sum_{i=1}^{m} I\left(\left[argmax_{l \in Y} f(x_i, l)\right] \notin y_i\right) \quad (21)$$

where l is a known label of x_i. The index of label sorting is defined according to the sorting relation of the output value of tag. If the multi-label learner has a real-valued output function $f(\cdot, \cdot)$, all the labels can be sorted in order of their output values from large to small, and the ranking value of the label l can be represented by $r(x, l)$.

Coverage:

$$\text{Coverage} = \frac{1}{m}\sum_{i=1}^{m} max_{l \in y_i} r(x_i, l) - 1 \quad (22)$$

Ranking Loss:

$$\text{RankingLoss} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|y_i|\,|\bar{y}_i|} \left|\left\{(l, l') \mid f(x_i, l) \le f(x_i, l'), (l, l') \in y_i \times \bar{y}_i\right\}\right| \quad (23)$$

where y_i is the subset of class labels associated with sample x_i, $\bar{y}_i$ is the complementary subset of y_i.

Average precision:

$$\text{AveragePrecision} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{|y_i|}\sum_{l \in y_i} \frac{\left|\left\{l' \mid r(x_i, l') \le r(x_i, l), l' \in y_i\right\}\right|}{r(x_i, l)} \quad (24)$$

TABLE 2. Distribution of multifunctional enzymes (Che 2016) before and after CD-HIT (0.9).

Dataset	EC1	EC2	EC3	EC4	EC5	EC6	Total
Before redundancy	1534	1924	2657	1698	616	179	4076
After CD-HIT	859	984	1419	926	290	91	2181

Of the 2181 different enzymes sequences, 1983 to two classes, 189 to three classes, 9 to four classes.

TABLE 3. The jackknife test result of multifunctional enzymes on dataset I showed in Table 1 based on different representation of enzyme sample.

	MF-EFP	PseAAC	SAAC	GM2
Hamming Loss	0.0407	0.0693	0.0588	0.0730
Ranking Loss	0.0388	0.0568	0.0531	0.0595
One Error	0.0853	0.1369	0.1210	0.1407
Coverage	0.8464	0.9661	0.9420	0.9866
Average Precision	0.9416	0.9082	0.9165	0.9042
Subset Accuracy	0.8562	0.7180	0.7450	0.6979

TABLE 4. The 5-fold cross-validation results of Multi-label classification on dataset II showed in Table 2.

	MF-EFP	IBLR-ML	BRkNN	RAkEL	HOMER
Micro-Averaged Precision	0.9325	0.8406	0.8279	0.8090	0.7519
Micro-Averaged Recall	0.9407	0.8178	0.8285	0.8126	0.8233
Micro-Averaged F-Measure	0.9366	0.8290	0.8282	0.8108	0.7859
Macro-Averaged Precision	0.9203	0.6792	0.7341	0.7364	0.6056
Macro-Averaged Recall	0.8851	0.6705	0.7379	0.6917	0.6619
Macro-Averaged F-Measure	0.9003	0.6737	0.7347	0.7004	0.6305
Average Precision	0.9601	0.8940	0.8583	0.8910	0.8407

IV. RESULTS AND DISCUSSION

In statistical prediction, independent dataset test, sub-sampling test (such as five- or ten-fold cross-validation), and jackknife test have often been used to evaluate the performance of the prediction. Among the three test methods, the jackknife test was considered as the least arbitrary that can always yield a unique result for a given benchmark dataset. However, the more numbers of subsets (functional classes) a benchmark covers, the more difficult to achieve a high overall success rate in using the jackknife method for cross-validation.

The newest multifunctional enzymes (release 2019.4) in Table 2 as benchmark dataset S which contains 4479 proteins covers 7 functional classes. It is worthy pointing out that the data are imbalanced, with 2021 belong to one class, 2248 to two classes, 198 to three classes, 12 to four classes. Using the criteria for multi-label classification algorithm that have been discussed in section Measurement, we evaluated MF-EFP and compared it with other feature extraction methods, obtaining the performance results shown in Table 3. Among average precision, subset accuracy, one error, coverage, hamming loss, and ranking loss, the higher subset accuracy and average precision are, the better the multi-label classification model performance are; and vice versa for the other four measurements. In general, the subset accuracy and average precise are deemed to be more strict measurements. It can be seen from the Table 3. that the average precision is about 94.16% in identifying multi-functional enzymes among their seven main functional classes, indicating that, even for the stringent benchmark dataset in which covers seven functional classes versus many existed predictors only can predicted six functional classes, MF-EFP predictor can yield quite reliable results. Meanwhile, we have also noticed that, the subset accuracy of MF-EFP is 85.62% shows that the vast majority of enzymes whose predicted label set are identical with the real label set. The advantage of the algorithm is obvious.

We also compared our representation of enzyme sample with other popular protein prediction methods, such as PseAAC (which is the most commonly used to prediction diverse protein attributes), SAAC (which has been used in predicting enzyme functional class) [24], and GM2 proposed by Xiao *et al.* (which can catch the essence of a protein sequence and better reflect its overall pattern by grey dynamic

TABLE 5. Comparative results of MLKNN, hMuLab and hML-KNN in dataset I and dataset II.

Measurement	Dataset Ⅰ			Dataset Ⅱ		
	MLKNN	hMuLab	hML-KNN	MLKNN	hMuLab	hML-KNN
Hamming Loss	0.0629	0.0491	0.0407	0.0772	0.0514	0.0445
Subset Accuracy	0.7372	0.7852	0.8562	0.7772	0.8391	0.8973
Average Precision	0.9114	0.9325	0.9416	0.9321	0.9564	0.9601
Coverage	0.9612	0.8759	0.8464	1.4631	1.3384	1.2999
One Error	0.1297	0.1016	0.0853	0.0853	0.0518	0.0536
Ranking Loss	0.0563	0.0427	0.0388	0.0640	0.0404	0.0373
Macro-Averaged Precision	0.8733	0.9596	0.8802	0.8643	0.9398	0.9203
Macro-Averaged Recall	0.6858	0.6917	0.8432	0.7670	0.7970	0.8851
Macro-Averaged F-Measure	0.7548	0.7621	0.8592	0.8055	0.8390	0.9003
Micro-Averaged Precision	0.9025	0.9241	0.9058	0.9051	0.9370	0.9325
Micro-Averaged Recall	0.8123	0.8551	0.9173	0.8702	0.9142	0.9407
Micro-Averaged F-Measure	0.8550	0.8883	0.9115	0.8873	0.9254	0.9366

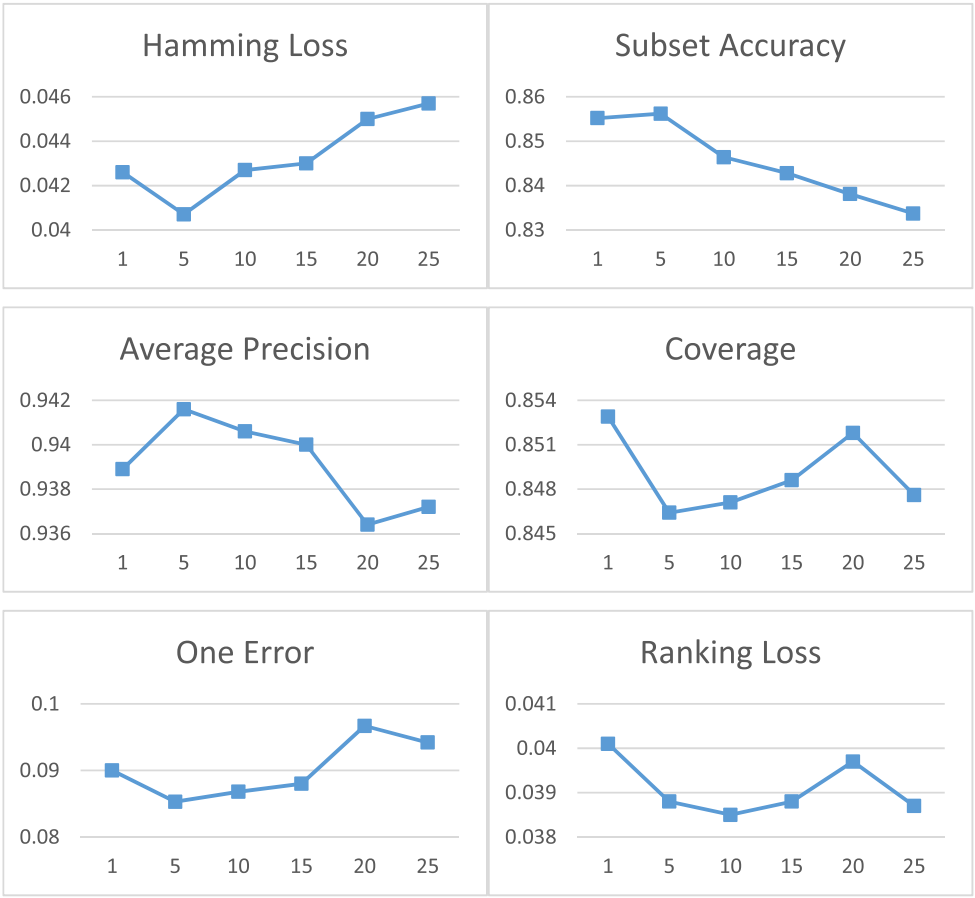


FIGURE 1. Six measurements of MF-EFP averaged over the 30 runs on the dataset I with different number of neighbors K. The horizontal axis is the number of neighbors, and the vertical axis represents the average measurement values by five-fold cross-validation.

model) [25]. Table 3 shows the average precision, subset accuracy, one error, coverage, hamming loss, and ranking loss for the dataset obtained for each approach in multi-label enzyme functional classification. All performance measurements of MF-EFP are outperformed other three models. Our new design compromises the merits of PseAAC and

FIGURE 2. Six measurements of MF-EFP averaged over the 30 runs on the dataset I with different number of weights α. The horizontal axis is the weights of the two scores based on features and neighbors, and the vertical axis represents the average measurement values by five-fold cross-validation.

SAAC increased capacity for predicting enzyme multifunctional classes compared to the using PseAAC, SAAC, and GM2 only.

To test the classification performance of the multifunctional enzyme further and compare to the existed predictors, the same benchmark dataset listed in Table 2 as investigated in Che *et al.* [10] and exactly the same cross-validation approach were adopted for demonstration. The results of proposed predictor and classifiers IBLR_ML, BRkNN, RakEL, HOMER are presented in Table 4. It is not difficult to find that the proposed predictor in this paper remarkably outperformed all those classifiers in all metrics from the Table 4, and outperformed those classifiers by about 6% in average Precision specially.

Our novel multi-label learning algorithm hML-KNN is designed based on hMuLab and MLKNN, we also compared three algorithm performance in predicting the multifunctional classes of enzymes. Listed in Table 5 are the results obtained with hML-KNN, hMuLab and MLKNN on the aforementioned benchmark dataset I and dataset II by jackknife test. As we can see form Table 5, for such stringent and complicated benchmark datasets, the subset accuracy achieved by hML-KNN is over 85.62% and 89.73% in dataset I and dataset II respectively, which is 7% and 6% higher than by hMuLab and 12% and 12% higher than by MLKNN.

There are three parameters in hML-KNN as parameter K for the neighbor score, parameter α to adjust the weights of the two scores based on features and neighbors, parameter λ is the tradeoff of the square of error and regularization term, just like hMuLab. The parameters were optimized by a standard five-fold cross-validation based on the subset accuracy in the dataset I. Fig.1 shows the mean values of different evaluation measurements with K=1, 2, 5, …, 25. As we can see, at the beginning, all the six performance measurements are improved significantly along with an increasing number of nearest neighbors. The measurement gets worse after K=5. Fig.2 shows the mean values of different evaluation measurements with $\alpha = 0, 0.25, 0.5, 0.75, 1$. When $\alpha = 0.25$, it is not difficult to find that the evaluation measurement is the best. We choose the default parameters $\lambda = 1$ just like the course of choosing parameter K and α. Upon optimization, the parameters were fixed and remained the same throughout all experiments.

Why could the proposed model be so powerful? This is because many key features, which are deeply hidden in complicated enzyme sequence, can be extracted via the approach of mixing together PseAAC and SAAC, hMuLab algorithm utilized both the feature-based information and the sample-based neighbor information, and a self-adapted label assigning threshold.

V. CONCLUSION

In this work, we proposed a new method of multifunctional enzymes prediction, which is a hybrid model by integrating the feature and neighbor label information, a query enzyme is formulated into the general pseudo amino acid composition (PseAAC) merging split amino acid composition (SAAC), and adopt a self-adapted label assigning threshold. hML-KNN is unique predictor that can predict sever functional classes at present.

The jackknife and cross-validation test results indicate that our method demonstrates better versatility and effectiveness. A user-friendly web-server for the new predictor has been established at http://www.jci-bioinfo.cn/MF-EFP, where users can easily get their desired results. It is anticipated that predictor will become a very useful high throughout tool for identifying multifunctional enzymes, and the novel approach and technique can also be used to investigate many other protein related problems.

REFERENCES

[1] A. Roy, J. Yang, and Y. Zhang, "COFACTOR: An accurate comparative algorithm for structure-based protein function annotation," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W471–W477, Jul. 2012.

[2] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER suite: Protein structure and function prediction," *Nature Methods*, vol. 12, no. 1, pp. 7–8, Jan. 2015.

[3] S. Amidi, A. Amidi, D. Vlachakis, N. Paragios, and E. I. Zacharaki, "Automatic single- and multi-label enzymatic function prediction by machine learning," *PeerJ*, vol. 5, Mar. 2017, Art. no. e3095.

[4] N. Kumar and J. Skolnick, "EFICAz2.5: Application of a high-precision enzyme function predictor to 396 proteomes," *Bioinformatics*, vol. 28, no. 20, pp. 2687–2688, Oct. 2012.

[5] S. Quester and D. Schomburg, "EnzymeDetector: An integrated enzyme function prediction tool and database," *BMC Bioinf.*, vol. 12, no. 1, p. 376, Dec. 2011.

[6] M. M. Sharif, A. Thrwat, I. I. Amin, A. Ella, and H. A. Hefeny, "Enzyme function classification based on sequence alignment," *Adv. Intell. Syst. Comput.*, vol. 340, pp. 409–418, Jan. 2015.

[7] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multifunctional enzyme by hierarchical multi-label classifier," *J. Comput. Theor. Nanosci.*, vol. 10, no. 4, pp. 1038–1043, Apr. 2013.

[8] Y.-C. Wang, X.-B. Wang, Z.-X. Yang, and N.-Y. Deng, "Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature," *Protein Peptide Lett.*, vol. 17, no. 11, pp. 1441–1449, Nov. 2010.

[9] Y.-C. Wang, Y. Wang, Z.-X. Yang, and N.-Y. Deng, "Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context," *BMC Syst. Biol.*, vol. 5, no. 1, p. S6, 2011.

[10] Y. Che, Y. Ju, P. Xuan, R. Long, and F. Xing, "Identification of multifunctional enzyme with multi-label classifier," *PLoS ONE*, vol. 11, no. 4, 2016, Art. no. e0153503.

[11] L. De Ferrari, S. Aitken, J. van Hemert, and I. Goryanin, "EnzML: Multi-label prediction of enzyme classes using InterPro signatures," *BMC Bioinf.*, vol. 13, no. 1, p. 61, 2012.

[12] W.-L. Huang, H.-M. Chen, S.-F. Hwang, and S.-Y. Ho, "Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method," *Biosystems*, vol. 90, no. 2, pp. 405–413, Sep. 2007.

[13] A. Mohammed and C. Guda, "Application of a hierarchical enzyme classification method reveals the role of gut microbiome in human metabolism," *BMC Genomics*, vol. 16, no. 7, p. S16, Dec. 2015.

[14] M. H. Osman, C. Y. Liong, and I. Hashim, "Hybrid learning algorithm in neural network system for enzyme classification," *Int. J. Adv. Soft Comput. Appl.*, vol. 2, pp. 209–220, Jul. 2010.

[15] Y. Li, S. Wang, R. Umarov, B. Xie, M. Fan, L. Li, and X. Gao, "DEEPre: Sequence-based enzyme EC number prediction by deep learning," *Bioinformatics*, vol. 34, no. 5, pp. 760–769, Mar. 2018.

[16] H.-B. Shen and K.-C. Chou, "EzyPred: A top–down approach for predicting enzyme functional classes and subclasses," *Biochem. Biophys. Res. Commun.*, vol. 364, no. 1, pp. 53–59, Dec. 2007.

[17] W. Tian, A. K. Arakaki, and J. Skolnick, "EFICAz: A comprehensive approach for accurate genome-scale enzyme function inference," *Nucleic Acids Res.*, vol. 32, no. 21, pp. 6226–6239, 2004.

[18] H.-L. Zou and X. Xiao, "Classifying multifunctional enzymes by incorporating three different models into Chou's general pseudo amino acid composition," *J. Membrane Biol.*, vol. 249, no. 4, pp. 551–557, Aug. 2016.

[19] Z. Zou, S. Tian, X. Gao, and Y. Li, "MlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning," *Frontiers Genet.*, vol. 9, p. 714, Jan. 2019.

[20] K.-C. Chou and H.-B. Shen, "Predicting protein subcellular location by fusing multiple classifiers," *J. Cellular Biochem.*, vol. 99, no. 2, pp. 517–527, Oct. 2006.

[21] T. H. Afridi, A. Khan, and Y. S. Lee, "Mito-GSAAC: Mitochondria prediction using genetic ensemble classifier and split amino acid composition," *Amino Acids*, vol. 42, pp. 1443–1454, Apr. 2012.

[22] C. Huang and J.-Q. Yuan, "A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types," *J. Membrane Biol.*, vol. 246, pp. 327–334, Apr. 2013.

[23] P. Wang, R. Ge, X. Xiao, M. Zhou, and F. Zhou, "HMuLab: A biomedical hybrid MUlti-LABel classifier based on multiple linear regression," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1173–1180, Sep. 2017.

[24] M. Arif, M. Hayat, and Z. Jan, "iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 442, pp. 11–21, Apr. 2018.

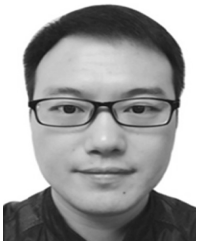
[25] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "IDNA-prot: Identification of DNA binding proteins using random forest with grey model," *PLoS ONE*, vol. 6, no. 9, 2011, Art. no. e24756.

XUAN XIAO received the M.S. degree from the Tianjing University of Science and Technology, Tianjin, China, in 2002, and the Ph.D. degree from Donghua University, Shanghai, China, in 2006. He is currently a Professor with the Department of Computer, Jingdezhen Ceramic Institute, Jingdezhen, China. His research interests include pattern recognition, bioinformatics, and sensory evaluation.

LI-WEN DUAN was born in Jiangxi, China, in 1995. He is currently pursuing the M.S. degree in control theory and control engineering with the Jingdezhen Ceramic Institute, Jingdezhen, China. His research interests include machine learning, data mining, and bioinformatics.

GUANG-FU XUE received the B.S. degree in computer science and technology from the Jingdezhen Ceramic Institute, Jingdezhen, China, in 2017, where he is currently pursuing the M.S. degree. His research interest includes bioinformatics statistics.

GANG CHEN was born in Jiujiang, Jiangxi, China, in 1995. He is currently pursuing the master's degree in computer science and technology with the Jingdezhen Ceramic College. His current research interests include machine learning and data mining in bioinformatics.



PU WANG received the M.S. degree from the Jingdezhen Ceramic Institute, Jingdezhen, China, in 2009, and the Ph.D. degree from the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, in 2017. He is currently an Associate Professor with the Computer School, Hubei University of Arts and Science, Xiangyang, China. His research interests include machine learning and bioinformatics.

WANG-REN QIU received the M.S. degree from Dalian Maritime University, Dalian, China, in 2006, and the Ph.D. degree from the University of Technology, China, in 2012. He is currently a Professor with the Department of Computer, Jingdezhen Ceramic Institute, Jingdezhen, China. His research interests include fuzzy sets and its applications, bioinformatics, and data mining.