Hindawi
BioMed Research International
Volume 2020, Article ID 2468789, 11 pages
https://doi.org/10.1155/2020/2468789

Research Article

Its2vec: Fungal Species Identification Using Sequence Embedding and Random Forest Classification

Chao Wang,[1] **Ying Zhang**,[2] **and Shuguang Han**[3]

[1]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2]*Department of Pharmacy, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China*
[3]*Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 60054, China*

Correspondence should be addressed to Ying Zhang; zhangying_hmu@163.com and Shuguang Han; shughan@uestc.edu.cn

Received 3 March 2020; Revised 20 March 2020; Accepted 25 March 2020; Published 29 May 2020

Guest Editor: Qin Ma

Fungi play essential roles in many ecological processes, and taxonomic classification is fundamental for microbial community characterization and vital for the study and preservation of fungal biodiversity. To cope with massive fungal barcode data, tools that can implement extensive volumes of barcode sequences, especially the internal transcribed spacer (ITS) region, are necessary. However, high variation in the ITS region and computational requirements for processing high-dimensional features remain challenging for existing predictors. In this study, we developed Its2vec, a bioinformatics tool for the classification of fungal ITS barcodes to the species level. An ITS database covering more than 25,000 species in a broad range of fungal taxa was assembled. For dimensionality reduction, a word embedding algorithm was used to represent an ITS sequence as a dense low-dimensional vector. A random forest-based classifier was built for species identification. Benchmarking results showed that our model achieved an accuracy comparable to that of several state-of-the-art predictors, and more importantly, it could implement large datasets and greatly reduce dimensionality. We expect the Its2vec model to be helpful for fungal species identification and, thus, for revealing microbial community structures and in deepening our understanding of their functional mechanisms.

1. Introduction

Metabarcoding is among the most promising approaches in the study of microbial communities [1–3] and has provided new insights into microbial impacts on crop yields [4], human health [5], and ecology [6]. Fungi are immensely diverse; the latest best estimate within this kingdom suggests that their total species number is somewhere between 2.2 and 2.8 million [7]. To date, only 144,000 (less than 7%) fungal species have been named and classified, while the vast majority are currently unknown to science [7]. Fungi play essential roles in many ecological processes as organic matter decomposers, mutualists with algae and plants [4, 8], plant pathogens, and components of the food chain [9–11]. Taxonomic classification is fundamental for microbial community characterization [12] and is vital for the study and preservation of fungal biodiversity [13]. However, it is difficult to identify specimens when their morphological characters are lacking or incomplete [14]. Several rRNA genes have been successfully employed for fungal species identification, including the small ribosomal subunit, the large ribosomal subunit, the RNA polymerase II binding protein, and the internal transcribed spacer (ITS). Among these, the ITS (including ITS1 and ITS2 separated by the 5.8S genic region) has been widely adopted as a marker for fungal identification and diversity exploration [15–19] because this region is ubiquitous and shows great variation in sequence and length [9].

Several ITS reference databases for fungal species identification have been developed. The UNITE database [20] and Warcup training set [9] are the most commonly used. The UNITE database (https://unite.ut.ee/) was first released in 2003 and focused on ectomycorrhizal fungi in north Europe [21], and it has been under successive development since then. UNITE aims to collect and disseminate all fungal ITS metadata from all geographical regions [20]. The latest version of the UNITE database (version 8.0) comprises approximately 1,000,000 public fungal ITS sequences (~459,000 species) for reference and provides valuable data

for metabarcoding software pipelines [14]. The Warcup training set was developed from the UNITE database and includes only sequences with authoritative taxonomic or lineage information [9]. In addition, ITS barcodes in the BOLD database (http://www.boldsystems.org/) [22] and the ITS1 database comprising sequences of NCBI GenBank (http://www.ncbi.nlm.nih.gov/) have been used for fungal species identification [10, 15]. DNA barcode-based taxonomic assignment can be achieved by using similarity-based or prediction-based (alignment-free) methods. Similarity-based methods (e.g., BLAST) align the query sequence with all sequences in the reference database, which is time-consuming and inefficient when compared to alignment-free methods [1, 10]. Several prediction-based methods for fungal species prediction, including RDP classifier [9, 23], SINTAX [12], Mycofier [10], Mothur [24], and funbarRF [15], using various machine learning algorithms, have been developed in the past few years. To generate feature vectors, k-mer and its derivative, spaced k-mer, have been used for sequence encoding. The RDP classifier, which implements a naïve Bayes algorithm for taxonomy assignment, uses 8-mers as features [9]. SINTAX and Mothur, which use a non-Bayesian [12] and the k-nearest neighbor (kNN) algorithm [24], respectively, also use 8-mers. The naïve Bayes classifier Mycofier uses 5-mer features [10]. The random forest- (RF-) based predictor funbarRF uses spaced k-mer features for taxonomy classification [15].

Although great progress has been made in fungal species identification using machine learning algorithms, there still is room for further improvement. First, the species in the abovementioned datasets are only a small fraction of the species that have been named and classified. For example, the Warcup training set (version 2) covers 8,551 species of 1,461 genera [9], the ITS database of BOLD includes 3,674 species of 777 genera [15], and the ITS1 database of NCBI comprises 1,794 species of 510 genera [10]. A larger dataset that covers a broad range of fungal taxa would provide more valuable insights into microbial community compositions. Second, the k-mer-based representation method counts the frequency of all possible subsequences of length k of a sequence, which usually yields high-dimensional (i.e., 4^8 for RDP classifier) and sparse vectors [25]. To address this issue, a representation method that can reduce dimensionality and encode each sequence into a dense, numeric vector is required.

Feature extraction is very important for constructing a computational predictor [4, 26–40]. Recently, a new efficient method for nucleotide sequence representation was proposed using a word embedding algorithm [41], such as word2vec [42]. Word embedding was originally developed for natural language processing [41]. In this model, each word is characterized by its context, i.e., neighboring words, and embedded in a predefined n-dimensional vector, where similar words have close vectors. This word representation method has been successfully employed to generate features from biological sequences. Asgari and Mofrad [43] applied the word2vec framework to represent and extract features of DNA sequences and protein families and achieved an average classification accuracy of 93% based on the classification of 7,027 protein families. Subsequently, a number of studies using this approach for the distributed representation of biological sequences were reported, including analyses of DNA [44–46], non-coding RNA [47], long non-coding RNA [48–50], and 16S rRNA [25].

The aim of this work was to develop a machine learning-based classifier for classifying fungal DNA barcodes. The filtered UNITE database covering broad range of fungal taxa was constructed, and a word embedding algorithm was employed to represent ITS sequences as dense, low-dimensional vectors. We demonstrate that this novel tool name "Its2vec" achieved an accuracy comparable to that of state-of-the-art predictors. We expect that Its2vec can aid in the computational classification of fungal species.

2. Materials and Methods

2.1. ITS Database and Preprocessing. One of our goals was to gain a deeper insight into microbial community structures by developing a database that contains as much representative fungal species as possible. The UNITE_public database contains fungal ITS sequences from both the International Nucleotide Sequence Database Collaboration (INSDC) and UNITE dataset [16, 51]; the latest UNITE_public (INSDC+UNITE) v8.0 includes 887,397 sequences. The dataset was clustered at several similarity thresholds to obtain species-level operational taxonomic units, referred to as species hypotheses (SHs); for each SH, if two or more ITS sequences are available, a representative sequence was chosen randomly to represent the SH [16], which resulted in the sh_general database, including 35,667 sequences (https://unite.ut.ee/repository.php) [20].

In this study, an ITS database was developed starting from the sh_general database. Sequences assigned a SH in the UNITE_public were extracted using sequences in sh_general as a query, resulting in a dataset with 513,953 sequences belonging to 32,523 SHs (at least 2 representative sequences for each SH). Then, any sequences without clear taxonomic information at the genus and species levels (i.e., g__unidentified; s__uncultured) were discarded. Thus, 429,494 sequences confined to 27,520 SHs were obtained (Figure 1(a)). As some SHs were represented by hundreds of sequences, to reduce the heterogeneity in sequence numbers, 10 sequences were randomly selected for SH that contained more than 10 sequences. Finally, 126,388 sequences belonging to 27,520 SHs were retained for analysis.

2.2. Distributed Representation of ITS Sequences. A distributed representation of the ITS sequence was generated in two major steps. First, the ITS sequences were lexically represented as a large set of k-mers (Figure 1(c)). For a sequence of length N, $N - k + 1$ $k-$ mers were generated by moving a window of size k along the sequence. These k-mers [52] are similar to the words belonging to a corpus in natural language processing. Ten datasets with k-mer lengths ranging from 3 to 12 were evaluated. Given this k-mer set, the second step was to train the distributed representation. k-mer embedding training was processed using the skip-gram model of word2-vec [41, 53, 54] implemented in Gensim 3.40 (https://radimrehurek.com/gensim/apiref.html) (Figure 1(b)). Wet set min_count = 1, epochs = 5, and the window size was varied

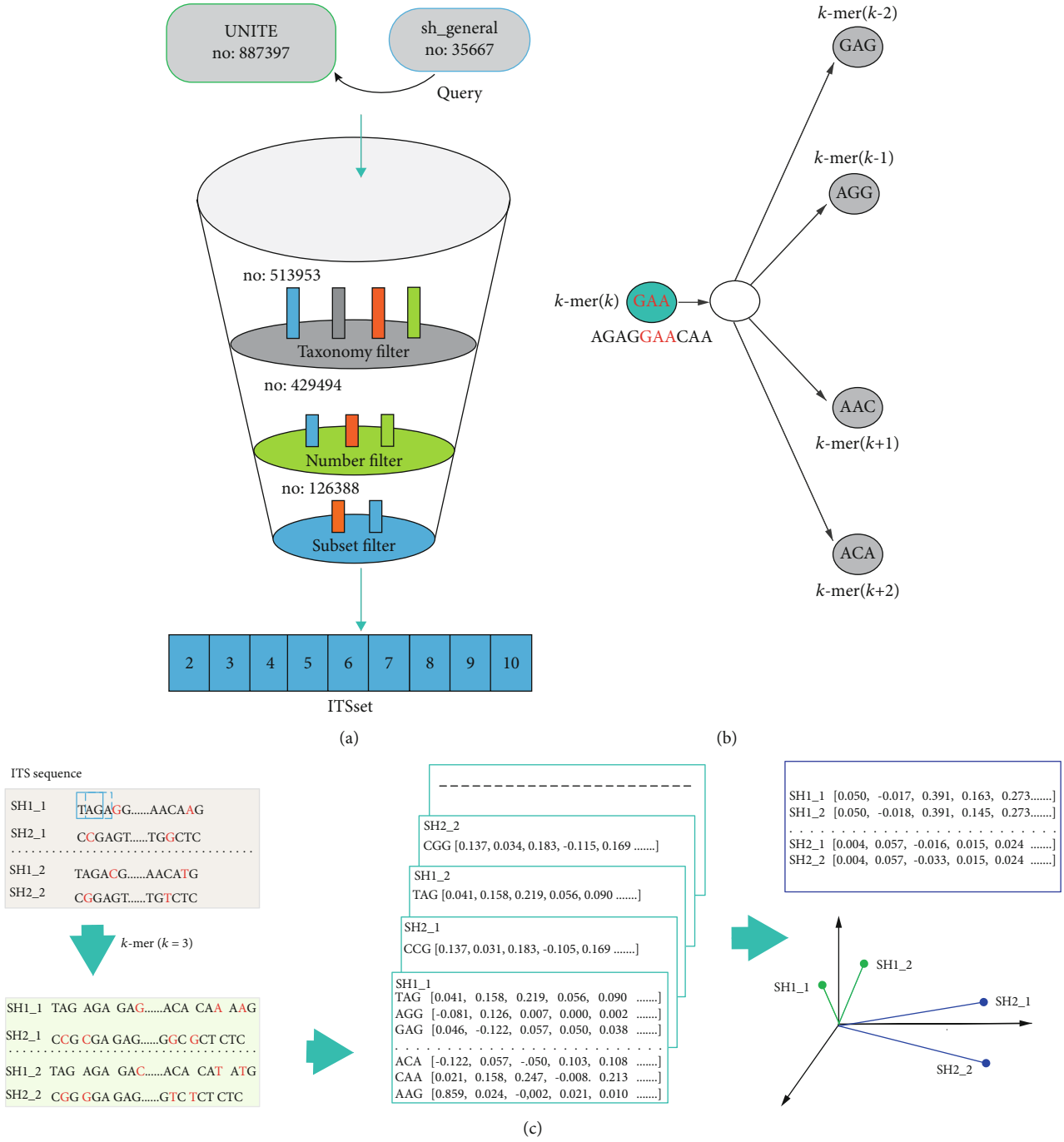

Figure 1: Continued.

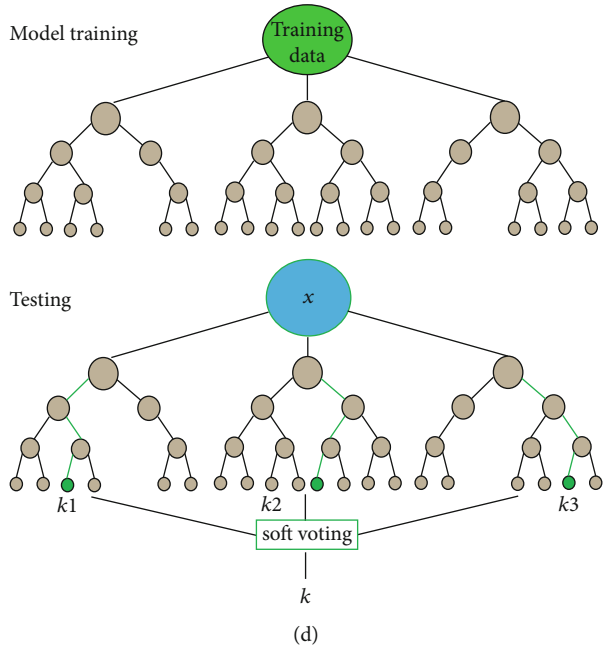

Figure 1: Schematic view of Its2vec. (a) Pipeline scheme of ITS dataset construction. (b) The skip-gram architecture of word2vec, which predicts surrounding k-mers (GAA, AGG, AAG, and ACA) based on a given center word (GAA). (c) Pipeline scheme of distributed representation of ITS sequences. For example, the ITS sequence SH1_1 (length N) was first represented by an N-2 3-mer set (TAG, AGA, GAG,…, AAG). Then, for each k-mer, we generated a distributed vector representation based on the skip-gram model with a vector of size 100, i.e., TAG [0.041, 0.158, 0.219…]. Thus, sequence SH1_1 was represented by the average of all n-2 k-mers, which also is a vector of size 100, i.e., SH1_1 [0.050, –0.017, 0.391…]. Similar words have close vectors; in this figure, SH1_1 and SH2_1 are close to SH1_2 and SH2_2, respectively. (d) Flow diagram showing model training and testing using the RF classifier.

from 2 to 9; other parameters were set to their default value. Thus, each k-mer was presented as a numeric vector of size 100, and each sequence (length k) was represented by the average of all vectors of $N - k + 1$ k – mers, which also is a vector of size 100 (Figure 1(a)).

2.3. Classifier and Dataset for Training and Validation. Several supervised learning techniques, such as naïve Bayes [9, 10, 12, 55], kNN [24], and RF [15, 35, 56–61], have been used for predicting ITS sequences. In this study, RF was selected for the modeling of ITS sequences because it is a powerful machine-learning algorithm that is nonparametric, robust to noise, and suitable for large datasets [62] (Figure 1(d)). For each SH, the class label was assigned to an integer and the number of classes was equal to the number of SHs, namely, 25,720.

The filtered database contained more than 25,000 SHs, and each SH was represented by at least 2 sequences. Given the extensive dataset (including more than 120,000 sequences) and the heterogeneity in sequence numbers among species, training and validation on the whole dataset would be arduous. Therefore, the ITS dataset was divided into 9 subdatasets, termed ITSset_2 to ITSset_10. Each subdataset contained species represented by a specific number of sequences, i.e., ITSset_2 contained species with 2 representative sequences. Detailed information on sequences and species in each subdataset is presented in Table 1. For the ITSset with k ($k \geq 2$) sequences per SH, k-fold cross validation (CV) was employed for model evaluation. The ITSset was split into k smaller subsets, and a model was first trained by k-1 subsets and then validated on the remaining subset. Note that the ITSser_9 contains 5,374 SHs (10 representative sequences per SH); thus, training on this subset on RF is challenging and therefore, 5 sequences were randomly selected for each SH of ITSset_9. Hence, for ITSset_9, the model was evaluated on 26,870 sequences of 5,374 SHs (5 representative sequences per SH).

We introduced 4 standard metrics, Accuracy, Recall, Precision and Mathew's correlation coefficient (MCC) [31, 50, 63–77], to evaluate the performance of the proposed models:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},
\end{aligned} \tag{1}$$

where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

2.4. Comparison with Other Fungal Classification Methods. The performance of the Its2vec model was compared with

Table 1: Taxonomic coverage of the ITS database established in this study.

Taxonomy level	ITS subsets									Number of taxa
	ITSset_2	ITSset_3	ITSset_4	ITSset_5	ITSset_6	ITSset_7	ITSset_8	ITSset_9	ITSset_10	
Phyla	15	13	9	8	7	8	8	7	10	18
Classes	58	46	41	34	31	32	28	30	43	63
Order	165	142	128	106	97	98	78	78	133	187
Family	516	424	381	317	280	263	220	201	418	626
Genus	2073	1432	1116	875	693	633	497	404	1598	3385
Species	8586	4141	2503	1684	1236	929	701	566	5374	25720
Sequences	17172	12423	10012	8420	7416	6503	5608	5094	53740	126388

Table 2: Accuracy of the models constructed with different k-mers and subsets.

ITSset	3-mer	4-mer	5-mer	6-mer	7-mer	8-mer	9-mer	10-mer	11-mer	12-mer
ITSset_2	70.04	68.08	68.32	69.23	69.64	70.37	71.37	71.23	70.53	69.34
ITSset_3	83.96	82.70	83.53	83.88	83.71	83.94	84.71	84.22	83.41	82.64
ITSset_4	89.15	89.13	89.538	89.63	89.79	89.85	90.40	90.12	89.07	87.95
ITSset_5	92.36	92.71	93.17	93.02	92.84	92.96	93.37	93.15	92.47	91.78
ITSset_6	93.34	93.68	93.99	94.22	93.82	94.05	94.12	94.39	93.31	92.97
ITSset_7	94.99	95.40	95.92	95.73	95.76	96.02	96.03	95.77	94.99	94.88
ITSset_8	96.09	96.20	96.47	96.45	96.34	96.31	96.43	96.36	96.31	95.74
ITSset_9	95.84	96.47	96.54	96.78	96.62	96.58	96.51	96.37	95.90	95.78
ITSset_10	84.37	84.57	85.78	86.37	86.36	87.19	87.96	86.26	86.06	84.93

that of 3 other predictors, namely, RDP classifier, funbarRF, and Mothur. The executable source codes of RDP classifier (https://sourceforge.net/projects/rdp-classifier/), funbarRF (https://cran.r-project.org/web/packages/funbarRF/) and Mothur (https://github.com/mothur/mothur/releases/tag/v1.40.5) were downloaded to a local machine and applied to 2 fungal datasets. The performance of the 4 methods was first evaluated on the ITSset_5, which contained 1,684 SHs (5 sequences per SH). The training datasets Warcup of RDP and ITSset_5 were extracted from the UNITE database. Sequences in the datasets of funbarRF and the Warcup training set were approximately 97% nonredundant. Hence, the second dataset Fold-10, containing 1,084 species (10 sequences per species) after removing the 5.8S sequences and ITS1 sequences of the original dataset of funbarRF, was further used for model evaluation in this study as recommended by Meher et al. [15]. Accuracies were calculated over 5-fold CV for ITSset_5 and 10-fold CV for dataset Fold-10.

3. Results

3.1. Evaluation of Sequence Embedding on Fungal ITS. We first evaluated the performance of sequence embedding. All 126,388 sequences were represented by a k-mer corpus. Then, the k-mer embedding space was obtained by training a skip-gram model of word2vec on the corpus. Thus, each ITS sequence was represented as a numeric vector of size 100. In this process, two parameters, the length k of the k-mer and the window size w of the skip-gram model, were optimized. The accuracy of models constructed with different k is shown in Table 2. The classification accuracy improved by 1–3% when k ranged from 3 to 12 (Table 2). It can be seen that the accuracy was the highest when k was near to 9; i.e., the accuracy reached a maximum at 9-mer for 7 subsets and at 8-mer and 10-mer for the remaining 2 (Table 2). Subsets with a larger number of sequences per SH (species) yielded a higher accuracy, ranging from 68% for 2 sequences per SH to 97% for 9 sequences per SH. Similar results were obtained for the others 3 metrics, recall, precision, and MCC, where the maximum value was obtained at 9-mer for most subsets. Detailed results are provided in Supplementary Tables S1, S2, and S3. The optimum value of k was set as 9 in following experiments.

The window size w of the skip-gram model was varied from 1 to 7, and the classification accuracy was higher when w was near to 4 for datasets having a rather low number (2–4) of sequences per SH, whereas a higher accuracy was obtained at $w = 2$ for subsets containing more than 5 sequences per species (Table 3). For w larger than the above thresholds, the accuracy slightly decreased or stabilized. The accuracy score was 71.65% for ITSset_2 (2 sequences per SH), and it gradually increased with the number of representative sequences for each SH and reached 97.02% in ITSset_9 (9 sequences per SH) (Table 3). For other metrics (precision, recall, and MCC), findings were similar; detailed results are provided in Supplementary Tables S4, S5, and S6. Considering the improvement in the accuracy for SHs

Table 3: Accuracy of the models constructed with different window sizes and subsets.

ITSset	Size = 1	Size = 2	Size = 3	Size = 4	Size = 5	Size = 6	Size = 7
ITSset_2	66.96	70.60	71.19	71.65	71.30	70.75	70.84
ITSset_3	81.26	84.10	84.73	85.20	84.88	84.47	84.15
ITSset_4	87.74	90.26	90.14	90.32	90.39	89.92	89.87
ITSset_5	91.50	93.47	93.69	93.30	93.33	93.10	93.17
ITSset_6	93.03	94.62	94.62	94.30	94.27	94.39	94.30
ITSset_7	95.42	96.28	95.99	95.96	95.85	95.77	95.71
ITSset_8	96.06	97.02	96.63	96.50	96.63	96.67	96.54
ITSset_9	96.43	97.02	96.90	96.84	96.76	96.60	97.02
ITSset_10	85.62	88.00	87.73	87.71	88.06	87.42	96.91

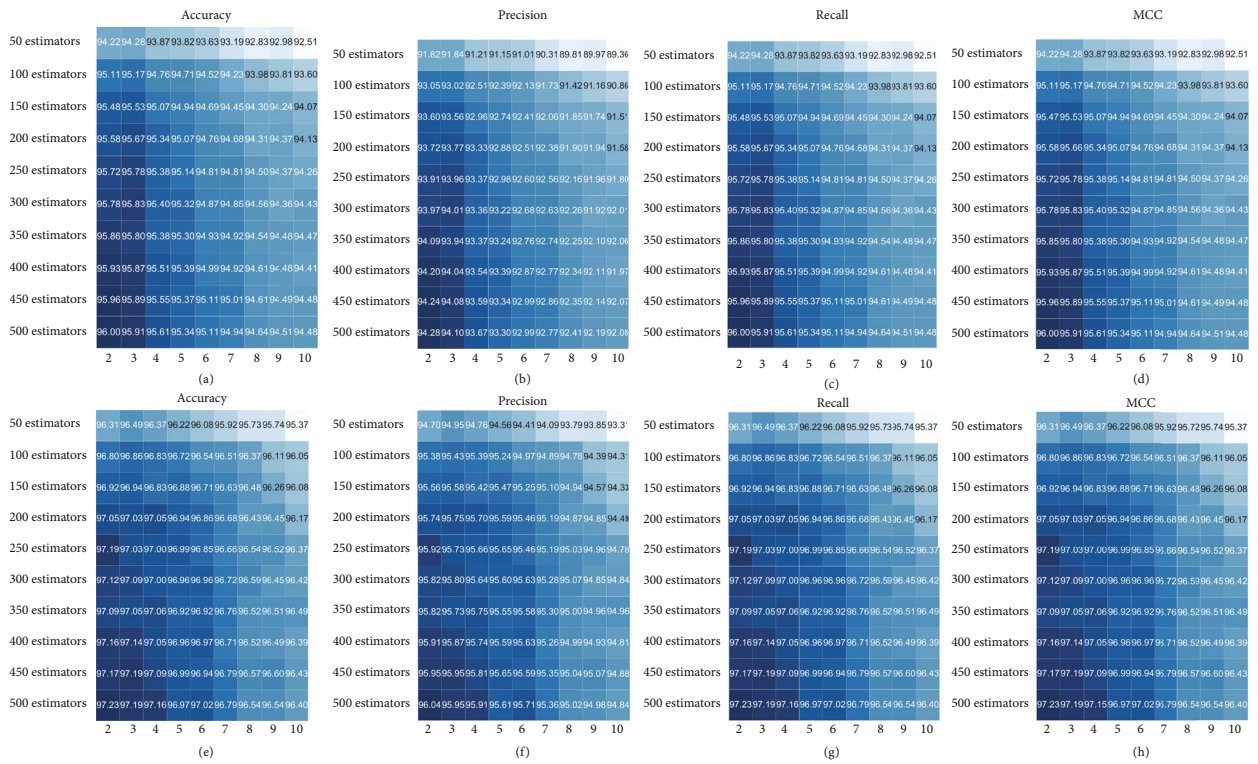

Figure 2: Accuracy, precision, recall, and MCC values of the RF model constructed using different numbers of features and estimators. (a–d) represent evaluation results of ITSset_5; (e–h) represent the evaluation results of ITSset_7.

represented by low number of sequences, w was set to 4. As the 4 evaluation metrics showed similar variation tendencies in the 9 subsets, subsequent experiments were conducted using ITSset_5 and ITSset_7, for simplicity.

3.2. RF Classifier. Random forest (RF) was widely employed in the bioinformatics researches [78–81]. Two key parameters of the RF classifier were optimized, namely, the number of features considered for splitting at each leaf node (max_features) and the number of trees in the forest (n_estimators). By default, max_features and n_estimators were set to 10 (square root of features) and 100, respectively. We varied the two parameters to generate 90 models, where max_features was varied from 2 to10 in intervals of 1 and the n_estimators was varied from 50 to 500 in intervals of 50. Figure 2 shows heat maps of accuracy, recall, MCC, and precision across the parameter combinations. In the accuracy heat map of ITSset_7 shown in Figure 2(e), it can be observed that the values gradually increase from the right upper region to the lower left region of the heat map, where the number of estimators linearly increased and the number of features decreased. Maximum values were obtained for the model based on 2 features and 500 estimators. The recall, MCC, and precision values are shown in Figures 2(f)–2(h); regions with more intense color in the three panels largely correspond to those in Figure 2(e). The evaluation results for ITSset_5 are shown in Figures 2(a)–2(d). It can be noted that higher values in the heat maps are observed in the similar regions of heat maps for ITSset_7. The values of the 4 metrics obtained for the

Table 4: Performance of the Its2vec model on 9 ITSsets based on optimized parameters.

ITSset	Accuracy	Precision	Recall	MCC
ITSset_2	78.62	72.10	78.62	0.79
ITSset_3	89.70	85.96	89.70	0.90
ITSset_4	93.36	90.69	93.36	0.96
ITSset_5	95.51	93.58	95.51	0.96
ITSset_6	95.95	94.09	95.95	0.96
ITSset_7	96.96	95.62	96.96	0.97
ITSset_8	97.50	96.37	97.50	0.98
ITSset_9	97.53	96.37	97.53	0.98
ITSset_10	90.23	86.48	90.23	0.90

Table 5: Comparison of the accuracy of the Its2vec and other existing predictors.

ITS dataset	Classifier	Accuracy	Significance
ITSset_5	Its2vec	95.51 ± 1.55	a*
	RDP	98.68 ± 0.55	b
	Mothur	97.97 ± 0.62	Bc
	funbarRF	91.00 ± 2.656	c
Fold-10	Its2vec	89.80 ± 1.92	a
	RDP	89.36 ± 2.21	a
	Mothur	85.54 ± 2.54	b
	funbarRF	84.94 ± 4.65	b

*Different letters indicate significant differences among the methods according to Tukey's HST test at $P < 0.05$.

model based on 2 features and 100 estimators were very close to the maximum values in the lower left region. Thus, considering the computational resources and the extensive dataset in this study, max_features was set to 2 and n_estimators was set to 100.

3.3. Performance Analysis Based on Optimized Parameters and Comparison with Other Predictors. The performance of the classifier was evaluated on all 9 subsets based on the optimized parameters ($k = 9$, window = 4, max_features = 2, and n_estimators = 100). Table 4 shows the values of accuracy, precision, recall, and MCC for the datasets. The accuracy was 78.62% for ITSset2, which contains only 2 sequences per SH. With an increasing number of sequences in the SHs, the accuracy increased, and it reached a maximum value of 97.53% for ITSset_9. The precision, recall, and MCC showed similar trends, reaching maximum values of 96.37%, 97.53%, and 0.98 on ITSset_9, respectively.

The predictive power of Its2vec was compared with that of three state-of-the-art predictors, using two benchmark datasets. The results are presented in Table 5. The accuracy of Its2vec (95.51%), Mothur (97.80%), and RDP (98.68%) was significantly higher than that of funbarRF (91.0%) for dataset ITSset_5. For the Fold-10 dataset, Its2vec achieved a better performance than the other approaches; its accuracy was 0.54%, 4.26%, and 4.86% higher than that of RDP, Mothur, and funbarRF, respectively. Thus, Its2vec had an accuracy comparable to that of RDP and Mothur for ITSset_5 of the UNITE database and outperformed the other 3 predictors when applied to the Fold-10 dataset of BOLD.

4. Discussion

Fungi play essential roles in many ecological processes. Taxonomic classification is fundamental in functional investigations and endangered species conservation. The ITS region has been widely used as a DNA barcode for fungal species classification as it has a high PCR amplification success rate and species discriminatory power within the fungal kingdom [10]. Commonly used alignment-based methods often assign unidentified barcodes to species based on information on the cluster they are of in the barcode tree [82]. However, sequence alignment may be difficult for distantly related species due to the variability in nucleic acid base pairs and sequence length. Further, alignment-based methods are not suitable for metabarcoding analysis. In this study, we developed a new fungal ITS classification approach that uses a distributed representation technique to generate features of ITS sequences and applies RF for species identification.

Previously assembled fungal datasets are rather small; with 8,551 species of 1,461 genera, the Warcup training set currently is the largest. The latest version of the UNITE database (version 8.0) covers ~459,000 species and thus provides valuable data for metabarcoding software pipelines [20]. One of the main aims of this study was to develop an ITS database that covers a broad range of fungal taxa. After data filtering, 126,388 sequences belonging to 27,520 SHs (species)—which is three times the number in the Warcup training set—were retained for analysis. Generally, the sequence identities in a dataset are kept <80% to avoid overestimation. However, as the number of sequences for each SH is very small (2–9) and the numbers of classes are rather large (more than 250,000), this preprocessing step was not feasible in this study. To the best of our knowledge, none of the existing species identification studies using DNA barcodes [1, 9, 10, 12, 15, 24] have reported such a preprocessing step. When applied to large datasets, classifiers directly using k-mer-based features (commonly, 8-mer, 4^8 features) are constrained by computational power. A word embedding algorithm was employed to represent each ITS barcode sequence as a dense, 100-dimensional vector.

To optimize the distributed representation of the ITS sequence, the length k of k-mer was optimized. We found that 8- and 9-mers resulted in the best performance. This implied that a k near 9 might be more informative, whereas a k larger than 10 may result in redundancy as evidenced by our results (Table 2). Similarly, RDP [9], SINTAX [12], and MOTHUR [24] use 8-mers. As for the window size of the skip-gram model, we noticed that a smaller window resulted in a higher accuracy, especially for SHs represented by more than 5 sequences ($w = 2$), which suggests that numbers of neighboring words predicted by the input (center) word are related to sequence abundance. Concerning RF

classifiers, more trees always lead to a better performance and a more robust model, as shown in Figure 2, but can be associated with an excessively long training time and high computer memory demands. In our study, the computational capacity was exceeded when the model was applied to the ITSset_3 (12,423 sequences) with n_estimators = 450. Two parameters, max_features and n_estimators, were varied to determine the best split of the tree. Larger features will decrease the accuracy (Figure 2), indicating that a higher accuracy is obtained when the trees in the RF show more differences. Further, large features also need more computational power. Therefore, considering the large dataset, max_features was set to 2, and n_estimators was set to 100 for the RF classifier, which resulted in an accuracy close to that of the model constructed with 500 estimators (0.43–0.81 smaller) (Figure 2).

The Its2vec model was compared with other three predictors in terms of performance. Although the classification accuracy of our model was ~3% lower than that of RDP and Mothur for ITSset_5, it should be noted that the number of features in RDP and Mothur are larger than that in Its2vec. For instance, RDP and Mothur take 8-mer frequency as input, generating 65,536 features, whereas the average number of features used by Its2vec was 700 (the average length of ITS sequences was 693 in this study), which was further reduced to 100 after distributed representation by the word embedding method. Because of this dimensionality reduction, Its2vec can be applied to large databases, while RDP is not suitable for such large datasets because of the computational power requirement. In the Fold-10 dataset, Its2vec showed the best performance. The funbarRF predictor uses g-spaced features as input; the number of features of the model was 90 ($g = 1 + 2 + 3 + 4 + 5$), which is close to the number of features generated by word2vec. However, our model achieved significantly higher accuracy than funbarRF in both the ITSset_5 and Fold-10 datasets (Table 5). It should be pointed out that the accuracy might be further improved in several ways, i.e. taking the pseudo components and tertiary structure of the ITS into consideration.

5. Conclusion

We presented Its2vec, a bioinformatics tool for the classification of fungal ITS barcodes to the species level. To cover a broad range of fungal taxa, an ITS database covering more than 25,000 species was constructed. For dimensionality reduction, the word embedding algorithm was used to represent the fungal DNA barcode sequences as a dense, 100-dimensional vector. Its2vec achieved an accuracy comparable to those of state-of-the-art predictors. We expect that the Its2vec model will be helpful for the identification of fungal species and, thus, for furthering our understanding of their functional mechanisms and guiding their application in agriculture. Also, computational intelligence such as neural networks [83–85], evolutionary algorithms [86, 87], and unsupervised learning [2, 88, 89] can be applied in this field.

Data Availability

All data and source code were available at http://lab.malab.cn/~wangchao/softwares/software.html.

Disclosure

A brief abstract of this manuscript has been submitted to the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, for the purpose of academic salon.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Nos. 61771331 and 61922020).

Supplementary Materials

Supplementary Table S1: precision of the models constructed with different k-mers and subsets. Supplementary Table S2: recall of the models constructed with different k-mers and subsets. Supplementary Table S3: MCC of the models constructed with different k-mers and subsets. Supplementary Table S4: precision of the models constructed with different window sizes and subsets. Supplementary Table S5: recall of the models constructed with different window sizes and subsets. Supplementary Table S6: MCC of the models constructed with different window sizes and subsets. *(Supplementary Materials)*

References

[1] N. Chaudhary, A. K. Sharma, P. Agarwal, A. Gupta, and V. K. Sharma, "16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets," *PLoS One*, vol. 10, no. 2, article e0116106, 2015.

[2] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, no. D1, pp. D554–D560, 2020.

[3] J. Fu, J. Tang, Y. Wang et al., "Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification," *Frontiers in Pharmacology*, vol. 9, p. 681, 2018.

[4] J. Di, B. Zheng, Q. Kong et al., "Prioritization of candidate cancer drugs based on a drug functional similarity network constructed by integrating pathway activities and drug activities," *Molecular Oncology*, vol. 13, no. 10, pp. 2259–2277, 2019.

[5] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "Human gut microbes associated with obesity," *Nature*, vol. 444, no. 7122, pp. 1022-1023, 2006.

[6] A. Konopka, "What is microbial community ecology?," *The ISME Journal*, vol. 3, no. 11, pp. 1223–1230, 2009.

[7] K. Willis, *State of the world's fungi 2018*, Royal Botanic Gardens, Kew, 2018.

[8] C. Wang and W. Y. Zhuang, "Evaluating effective trichoderma isolates for biocontrol of rhizoctonia solani causing root rot of Vigna unguiculata," *Journal of Integrative Agriculture*, vol. 18, no. 9, pp. 2072–2079, 2019.

[9] V. Deshpande, Q. Wang, P. Greenfield et al., "Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences," *Mycologia*, vol. 108, no. 1, pp. 1–5, 2016.

[10] L. Delgado-Serrano, S. Restrepo, J. R. Bustos, M. M. Zambrano, and J. M. Anzola, "Mycofier: a new machine learning-based classifier for fungal ITS sequences," *BMC Research Notes*, vol. 9, no. 1, p. 402, 2016.

[11] J. Yin, W. Sun, F. Li et al., "VARIDT 1.0: variability of drug transporter database," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1042–D1050, 2020.

[12] R. Edgar, *Sintax: a simple non-bayesian taxonomy classifier for 16s and its sequences*, BioRxiv, CHS, 2016.

[13] D. L. Hawksworth, "Fungal diversity and its implications for genetic resource collections," *Studies in Mycology*, vol. 50, pp. 9–18, 2004.

[14] A. D. Roe, A. V. Rice, S. E. Bromilow, J. E. Cooke, and F. A. Sperling, "Multilocus species identification and fungal DNA barcoding: insights from blue stain fungal symbionts of the mountain pine beetle," *Molecular Ecology Resources*, vol. 10, no. 6, pp. 946–959, 2010.

[15] P. K. Meher, T. K. Sahu, S. Gahoi, R. Tomar, and A. R. Rao, "FunbarRF: DNA barcode-based fungal species prediction using multiclass random Forest supervised learning model," *BMC Genetics*, vol. 20, no. 1, p. 2, 2019.

[16] U. Kõljalg, R. H. Nilsson, K. Abarenkov et al., "Towards a unified paradigm for sequence-based identification of fungi," *Molecular Ecology*, vol. 22, no. 21, pp. 5271–5277, 2013.

[17] Y. Wang, S. Zhang, F. Li et al., "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Research*, vol. 48, no. D1, pp. D1031–D1041, 2019.

[18] W. Xue, F. Yang, P. Wang et al., "What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation," *ACS Chemical Neuroscience*, vol. 9, no. 5, pp. 1128–1140, 2018.

[19] P. Wang, X. Zhang, T. Fu et al., "Differentiating physicochemical properties between addictive and nonaddictive ADHD drugs revealed by molecular dynamics simulation studies," *ACS Chemical Neuroscience*, vol. 8, no. 6, pp. 1416–1428, 2017.

[20] R. H. Nilsson, K. H. Larsson, A. F. S. Taylor et al., "The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications," *Nucleic Acids Research*, vol. 47, no. D1, pp. D259–D264, 2019.

[21] U. Kõljalg, K. H. Larsson, K. Abarenkov et al., "UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi," *New Phytologist*, vol. 166, no. 3, pp. 1063–1068, 2005.

[22] I. N. Sarkar and M. Trizna, "The barcode of life data portal: bridging the biodiversity informatics divide for DNA barcoding," *PLoS One*, vol. 6, no. 7, article e14689, 2011.

[23] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[24] P. D. Schloss, S. L. Westcott, T. Ryabin et al., "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[25] S. Woloszynek, Z. Zhao, J. Chen, and G. L. Rosen, "16S rRNA sequence embeddings: meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses," *PLoS Computational Biology*, vol. 15, no. 2, article e1006721, 2019.

[26] B. Liu, "BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1280–1294, 2019.

[27] W. Li, J. Yu, B. Lian et al., "Identifying prognostic features by bottom-up approach and correlating to drug repositioning," *PLoS One*, vol. 10, no. 3, article e0118672, 2015.

[28] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[29] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: similarity kernel fusion for accurately discovering mirna-disease association," *Frontiers in Genetics*, vol. 9, no. 618, pp. 1–13, 2018.

[30] L. Xu, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "An efficient classifier for alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, p. 3140, 2018.

[31] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, p. 158, 2018.

[32] L. Cheng, J. Sun, W. Xu, L. Dong, Y. Hu, and M. Zhou, "OAHG: an integrated resource for annotating human genes with multi-level ontologies," *Scientific Reports*, vol. 6, no. 1, pp. 1–9, 2016.

[33] L. Cheng, Y. Jiang, H. Ju et al., "InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk," *BMC Genomics*, vol. 19, Suppl 1, p. 919, 2018.

[34] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: pseudouridine sites identification based on RNA sequence information," *BMC Bioinformatics*, vol. 19, no. 1, p. 306, 2018.

[35] Q. Xu, Y. Xiong, H. Dai et al., "PDC-SGB: prediction of effective drug combinations using a stochastic gradient boosting algorithm," *Journal of Theoretical Biology*, vol. 417, pp. 1–7, 2017.

[36] B. Li, J. Tang, Q. Yang et al., "NOREVA: normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Research*, vol. 45, no. W1, pp. W162–W170, 2017.

[37] Y. Hu, T. Zhao, N. Zhang, T. Zang, J. Zhang, and L. Cheng, "Identifying diseases-related metabolites using random walk," *BMC Bioinformatics*, vol. 19, Suppl 5, p. 116, 2018.

[38] J. Zhang and B. Liu, "A review on the recent developments of sequence-based protein feature extraction methods," *Current Bioinformatics*, vol. 14, no. 3, pp. 190–199, 2019.

[39] L. Yu, X. Sun, S. Tian, X. Shi, and Y. Yan, "Drug and nondrug classification based on deep learning with various feature selection strategies," *Current Bioinformatics*, vol. 13, no. 3, pp. 253–259, 2018.

[40] Z. Liao, S. Wan, Y. He, and Q. Zou, "Classification of small GTPases with hybrid protein features and advanced machine

learning techniques," *Current Bioinformatics*, vol. 13, no. 5, pp. 492–500, 2018.

[41] T. Mikolov et al., "Efficient estimation of word representations in vector space," 2013, http://arxiv.org/abs/1301.3781.

[42] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, 2019.

[43] E. Asgari and M. R. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLoS One*, vol. 10, no. 11, article e0141287, 2015.

[44] P. Feng, H. Ding, H. Yang, W. Chen, H. Lin, and K. C. Chou, "iRNA-PseColl: identifying the occurrence sites of different rna modifications by incorporating collective effects of nucleotides into PseKNC," *Molecular Therapy Nucleic Acids*, vol. 7, no. C, pp. 155–163, 2017.

[45] G. Pan, L. Jiang, J. Tang, and F. Guo, "A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties," *International Journal of Molecular Sciences*, vol. 19, no. 2, p. 511, 2018.

[46] J. Tang, J. Fu, Y. Wang et al., "Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains," *Molecular & Cellular Proteomics*, vol. 18, no. 8, pp. 1683–1699, 2019.

[47] G. Aoki and Y. Sakakibara, "Convolutional neural networks for classification of alignments of non-coding RNA sequences," *Bioinformatics*, vol. 34, no. 13, pp. i237–i244, 2018.

[48] L. Cheng, P. Wang, R. Tian et al., "LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Research*, vol. 47, no. D1, pp. D140–D144, 2019.

[49] J. Tang, J. Fu, Y. Wang et al., "ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings in Bioinformatics*, vol. 21, no. 2, pp. 621–636, 2020.

[50] M. Zhang, F. Li, T. T. Marquez-Lago et al., "MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics*, vol. 35, no. 17, pp. 2957–2965, 2019.

[51] I. Karsch-Mizrachi, T. Takagi, G. Cochrane, and on behalf of the International Nucleotide Sequence Database Collaboration, "The international nucleotide sequence database collaboration," *Nucleic Acids Research*, vol. 46, no. D1, pp. D48–D51, 2018.

[52] H. Lin, Z. Y. Liang, H. Tang, and W. Chen, "Identifying sigma70 promoters with novel pseudo nucleotide composition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1316–1321, 2019.

[53] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, Neural Information Processing Systems Foundation, Inc., 2013.

[54] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Research*, vol. 47, no. 20, p. e127, 2019.

[55] P. M. Feng, H. Ding, W. Chen, and H. Lin, "Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 530696, 6 pages, 2013.

[56] H. Lv, Z.-M. Zhang, S.-H. Li, J.-X. Tan, W. Chen, and H. Lin, "OUP accepted manuscript," *Briefings in Bioinformatics*, 2019.

[57] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *Journal of Proteome Research*, vol. 16, no. 5, pp. 2044–2053, 2017.

[58] B. Liu, F. Yang, D. S. Huang, and K. C. Chou, "iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, 2018.

[59] Y. Ding, J. Tang, and F. Guo, "Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information," *International Journal of Molecular Sciences*, vol. 17, no. 10, p. 1623, 2016.

[60] W. Chen, H. Ding, P. Feng, H. Lin, and K. C. Chou, "iACP: a sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, pp. 16895–16909, 2016.

[61] L. Cheng, H. Zhuang, H. Ju et al., "Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study," *Frontiers in Genetics*, vol. 10, p. 94, 2019.

[62] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[63] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, 2019.

[64] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artificial Intelligence in Medicine*, vol. 83, pp. 82–90, 2017.

[65] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine*, vol. 83, pp. 67–74, 2017.

[66] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, 2018.

[67] J. Zhang, Q. Chen, and B. Liu, "DeepDRBP-2L: a new genome annotation predictor for identifying DNA binding proteins and RNA binding proteins using convolutional neural network and long short-term memory," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.

[68] B. Liu, S. Chen, K. Yan, and F. Weng, "iRO-PsekGCC: identify DNA replication origins based on Pseudo k-tuple GC composition," *Frontiers in Genetics*, vol. 10, p. 842, 2019.

[69] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, "FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association," *BMC Genomics*, vol. 19, no. S10, pp. 911–925, 2018.

[70] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, 2019.

[71] L. Xu, G. Liang, C. Liao, G. D. Chen, and C. C. Chang, "K-Skip-n-Gram-RF: a random forest based method for alzheimer's disease protein identification," *Frontiers in Genetics*, vol. 10, no. 33, 2019.

[72] W. Sun, Y. Han, S. Yang et al., "The assessment of interleukin-18 on the risk of coronary heart disease," *Medicinal Chemistry*, vol. 15, 2019.

[73] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, 2018.

[74] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae," *Briefings in Functional Genomics*, 2019.

[75] X. Wang, X. Zhu, M. Ye et al., "STS-NLSP: a network-based label space partition method for predicting the specificity of membrane transporter substrates using a hybrid feature of structural and semantic similarity," *Frontiers in Bioengineering and Biotechnology*, vol. 7, no. 306, 2019.

[76] X. Shan, X. Wang, C. D. Li et al., "Prediction of CYP450 enzyme-substrate selectivity based on the network-based label space division method," *Journal of Chemical Information and Modeling*, vol. 59, no. 11, pp. 4577–4586, 2019.

[77] 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, p. 2571, 2018.

[78] X. Ru, L. Li, and Q. Zou, "Incorporating distance-based top-n-gram and random forest to identify electron transport proteins," *Journal of Proteome Research*, vol. 18, no. 7, pp. 2931–2939, 2019.

[79] Z. Lv, S. Jin, H. Ding, and Q. Zou, "A random forest sub-Golgi protein classifier optimized via dipeptide and amino acid composition features," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 215, 2019.

[80] R. Su, X. Liu, L. Wei, and Q. Zou, "Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response," *Methods*, vol. 166, pp. 91–102, 2019.

[81] J. Song, C. Li, C. Zheng, J. Revote, Z. Zhang, and G. I. Webb, "MetalExplorer, a Bioinformatics Tool for the Improved Prediction of Eight Types of Metal-Binding Sites Using a Random Forest Algorithm with Two- Step Feature Selection," *Current Bioinformatics*, vol. 12, no. 6, pp. 480–489, 2017.

[82] E. Weitschek, G. Fiscon, and G. Felici, "Supervised DNA Barcodes species classification: analysis, comparisons and results," *BioData mining*, vol. 7, no. 1, p. 4, 2014.

[83] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural p systems with colored spikes," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 1106–1115, 2018.

[84] F. G. C. Cabarle, H. N. Adorna, M. Jiang, and X. Zeng, "Spiking neural p systems with scheduled synapses," *IEEE Transactions on Nanobioscience*, vol. 16, no. 8, pp. 792–801, 2017.

[85] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "deepDR: a network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, 2019.

[86] H. Xu, W. Zeng, D. Zhang, and X. Zeng, "MOEA/HD: a multiobjective evolutionary algorithm based on hierarchical decomposition," *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 517–526, 2019.

[87] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on minkowski distance for many-objective optimization," *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 3968–3979, 2019.

[88] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Transactions on Cybernetics*, pp. 1–12, 2019.

[89] X. Zeng, W. Lin, M. Guo, and Q. Zou, "A comprehensive overview and evaluation of circular RNA detection tools," *PLoS Computational Biology*, vol. 13, no. 6, article e1005420, 2017.