

Received March 10, 2020, accepted March 28, 2020, date of publication April 2, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985114

ItLnc-BXE: A Bagging-XGBoost-Ensemble Method With Comprehensive Sequence Features for Identification of Plant lncRNAs

GUANGYAN ZHANG^{1,2}, ZIRU LIU¹, JICHEN DAI³,
ZILAN YU¹, SHUAI LIU¹, AND WEN ZHANG^{1,4}

¹College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

²College of Public Management, Huazhong Agricultural University, Wuhan 430070, China

³Schaefer School of Engineering and Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA

⁴Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan 430070, China

Corresponding author: Wen Zhang (zhangwen@hza.u.edu.cn)

This work was supported in part by the National Key Research and Development Program under Grant 2018YFC0407904, in part by the National Natural Science Foundation of China under Grant 61772381 and Grant 61572368, in part by the Fundamental Research Funds for the Central Universities under Grant 2042017kf0219 and Grant 2042018kf0249, and in part by the Huazhong Agricultural University Scientific and Technological Self-innovation Foundation.

ABSTRACT Since long non-coding RNAs (lncRNAs) have involved in a wide range of functions in cellular and developmental processes, an increasing number of methods have been proposed for distinguishing lncRNAs from coding RNAs. However, most of the existing methods are designed for lncRNAs in animal systems, and only a few methods focus on the plant lncRNA identification. Different from lncRNAs in animal systems, plant lncRNAs have distinct characteristics. It is desirable to develop a computational method for accurate and robust identification of plant lncRNAs. Herein, we present a plant lncRNA identification method ItLnc-BXE, which utilizes comprehensive features and the ensemble learning strategy. First, a diversity of sequence features is collected and filtered by feature selection to represent transcripts. Then, several base learners are trained and further combined into a single meta-learner by ensemble learning, and thus an ItLnc-BXE model is constructed. ItLnc-BXE models are evaluated on datasets of six plant species, the results show that ItLnc-BXE outperforms other state-of-the-art plant lncRNA identification methods, achieving better and robust performance (AUC>95.91%). We also perform some experiments about cross-species lncRNA identification, and the results indicate that dicots-based and monocots-based models can be used to accurately identify lncRNAs in lower plant species, such as mosses and algae. In addition, source codes and supplementary data are available at <https://github.com/BioMedicalBigDataMiningLab/ItLnc-BXE>.

INDEX TERMS Plant lncRNA identification, ensemble learning, lncRNA-related features, feature selection.

I. INTRODUCTION

The recent improvements in high-throughput sequencing and the application of machine learning methods have led to the identification of numerous novel gene sequences [1]–[5]. As a consequence, the source of coding and non-coding RNAs has been greatly enlarged. Long non-coding RNAs (lncRNAs) are a class of RNA molecules that not encode proteins, with lengths exceeding 200 nucleotides [6]. Although lncRNAs were thought to be transcriptional noise at first, increasing works demonstrate that they exert significant impacts on many biological

processes, such as tissue development and external stimuli response [7]–[9].

Since only a few lncRNAs have been annotated, many machine learning-based methods have been proposed for lncRNA identification, such as CPC2 [10], CPAT [11], PLEK [12] and etc. CPC2 employed an SVM model using the RBF kernel to distinguish coding RNAs from non-coding RNAs. CPAT used the logistic regression (LR) for novel lncRNA identification. PLEK applied a computational pipeline based on an improved k-mer scheme and an SVM algorithm. These methods were all alignment-free, which implied that they only made use of features derived directly from sequences. For example, CPC2 constructed a feature set composed of four intrinsic features, which were

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

peptide length, isoelectric point, Fickett TESTCODE score and open reading frame (ORF) integrity, while CPAT adopted ORF length, ORF coverage, Fickett TESTCODE score and hexamer score.

In the history of lncRNA identification, the focus has always been human and animals, but few methods can be used for plants. Plant lncRNAs are different from animal lncRNAs and may have distinct characteristics. Most of the plant lncRNAs regulate gene expression through multiple mechanisms, such as target mimicry, transcription interference, histone methylation, and DNA methylation. They also play essential roles in flowering, male sterility, nutrition metabolism, biotic and abiotic stress, and other biological processes as regulators in plants [6]. The insufficiency of lncRNAs remains one of the major problems in plants, and most popular databases have a preference for collecting animal lncRNAs. With increasing demands in plant lncRNAs, several databases, such as RNAcentral, Ensembl Plants and CANTATAdb, began to collect plant lncRNAs. Still, there are many plant lncRNAs remain to be annotated. Therefore, it is desirable to develop a computational method for the accurate identification of plant lncRNAs.

To our knowledge, two methods have been proposed for plant lncRNA identification: PLncPRO [13] and PLIT [14]. PLncPRO used some software, such as BLASTX to extract features. Based on a total of 71-dimensional features, PLncPRO then employed the random forest algorithm for RNA identification. Using PLncPRO models, they discovered some high-coincidence lncRNAs in rice and chickpea under abiotic stress conditions. In PLIT, seven ORF and sequence-based features, and six codon bias features were extracted from training data. PLIT adopted a feature selection process that combined the least absolute shrinkage and selection operator (LASSO) with iterative random forests (IRF) to identify a list of optimal features. After that, a random forest classifier was used for plant lncRNA identification. More comprehensive studies are in demand for the plant lncRNA identification.

In this work, we present a plant lncRNA identification method ItLnc-BXE, based on comprehensive sequence features and ensemble learning strategy. We collect 23 types of features that fall into four categories, and the ReliefF-GA feature selection method is adopted to determine an optimal feature subset for a specific species. The subset is used to represent lncRNAs. After that, we construct the ItLnc-BXE model. We compile n data subsets by sampling data from the training dataset and accordingly build n base learners using extreme gradient boosting (XGBoost). Base learners are then combined using LR to develop the final ItLnc-BXE model. The performance of ItLnc-BXE models are evaluated on the datasets of six plant species with different lncRNAs/protein-coding transcripts (pcts) ratios. When compared with PLIT and PLncPRO, ItLnc-BXE produces better results, which results from three aspects: (1) comprehensive features provide diverse information about plant lncRNAs, (2) ReliefF-GA method reduces redundancy among features

and (3) ensemble learning strategy utilizes strengths from base learners.

Next, we summarize the contributions of the manuscript:

(1) Animal and human long non-coding RNAs (lncRNAs) have been widely annotated, while plant lncRNAs are still understudied. Therefore, we present a novel method ItLnc-BXE for plant lncRNA identification, which yields better performance than art-of-the-state methods.

(2) We collect comprehensive plant sequence features from previous works.

(3) We adopt a novel feature selection method ReliefF-GA to generate the optimal feature set.

(4) Ensemble learning strategy is applied to improve the performance and robustness of models.

(5) We discuss features after feature selection and find some shared features.

(6) We also conduct cross-species identification experiments and find we can train ItLnc-BXE models based on the higher plant lncRNAs to predict lower plant lncRNAs.

II. MATERIALS AND METHODS

A. DATASETS

Here, we collect plant transcripts from three databases, i.e., CANTATAdb version 2.0 [15], Ensembl Plants [16] and RNAcentral [17]. CANTATAdb is an authoritative and comprehensive database of computationally identified plant lncRNAs, and currently contains 239,631 lncRNAs from 39 species. lncRNAs in it are qualified because CANTATAdb specifically removes transcripts shorter than 200nt in the lncRNA discovery process, such as microRNAs and siRNAs. All lncRNAs of six plant species (*Arabidopsis thaliana*, *Solanum tuberosum*, *Oryza sativa*, *Hordeum vulgare*, *Physcomitrella patens*, *Chlamydomonas reinhardtii*) are downloaded from CANTATAdb, and used as positive instances. The cDNAs in Ensembl Plants are sequences for protein-coding genes, which are the contrary to non-coding RNAs. Thus, we download the cDNA sequences of all six species as pct or negative data from Ensembl Plants, which is a database containing highly credible annotation of plant genomes. As the pct of *Hordeum vulgare* and *Physcomitrella patens* are much more than those of other species, we randomly select 60000 *Hordeum vulgare* pct and 20000 *Physcomitrella patens* pct. For the other four species, all pct in the database are downloaded. In addition, the lengths of lncRNAs and pct that we collect are in the same range, to guarantee the robustness of the identification. To construct reliable datasets, we take three steps to preprocess raw data.

Step1. removing invalid sequences

First, we remove lncRNAs and pct that lack annotations from raw data. Second, pct in raw data may also include non-coding RNAs, such as lncRNAs, rRNAs, and tRNAs, and we remove these non-coding RNAs, according to the annotations in RNAcentral.

TABLE 1. Benchmark plant transcript datasets of six species.

Species	Main datasets		Independent datasets	
	lncRNAs	pcts	lncRNAs	pcts
Arabidopsis thaliana	3357	28944	372	372
Solanum tuberosum	3926	37150	436	436
Oryza sativa	2059	34224	228	228
Hordeum vulgare	5260	29039	584	584
Physcomitrella patens	1056	14851	117	117
Chlamydomonas reinhardtii	2618	16733	290	290

Note that we abbreviate six species by using their initials: Arabidopsis thaliana (A), Solanum tuberosum (S), Oryza sativa (O), Hordeum vulgare (H), Physcomitrella paten (P) and Chlamydomonas reinhardtii (C).

Step2. removing redundant sequences

CD-hit [18] is a widely used program for clustering protein or nucleic acid sequences with high efficiency, helping remove the highly similar sequences [19]. We use CD-hit as a filter to remove redundant lncRNAs and pcts with a similarity threshold of 80%.

Step3. constructing datasets

For each species, we randomly choose 10% of the data as the independent dataset for feature selection [20]. The rest, 90% of data is taken as the main dataset for cross-validation. Finally, the benchmark datasets of six species are constructed (Table 1).

For each species, we take all lncRNAs as positive instances, and randomly select pcts as negative instances with the ratios (lncRNAs/ pcts) of 1:1, 1:3 and 1:5, respectively.

B. FEATURE EXTRACTION

As obtaining information directly from RNA sequences is difficult, we consider transferring each sequence into a vector of digital features. So, we collect comprehensive plant lncRNA features from the published scientific literature [10, 11, 21]. All collected features are classified into four categories: sequence-based features, ORF-based features, codon-based features, and alignment-based features. All features are summarized (Table 2), and we will give a brief description of each feature.

1) SEQUENCE-BASED FEATURES

Sequence-based features are directly extracted from transcripts or indirectly calculated by them.

Transcript length ('Length') is one of the most fundamental features used to distinguish lncRNAs from pcts as lncRNAs' length exceeding 200 nucleotides. GC content is the percentage of guanine (G) and cytosine (C) in four kinds of nitrogenous bases, including adenine (A) and thymine (T). The study [13] reported that the GC contents in lncRNAs are less rich than those in pcts. Hexamer score ('Hexamer') [21] is calculated based on the occurrence of hexamer along

TABLE 2. Summary of features in this work.

Type	Feature	Dimension	Annotation
Sequence Based Features	Length	1	used
	GC content	1	used
	Hexamer	1	used
	Fickett	1	used
	CTD	30	new
	PI	1	new
ORF Based Features	GRAVY	1	new
	Instability	1	new
	ORF-length	1	used
	ORF-integrity	1	new
Codon Based Features	ORF-coverage	1	used
	FF-score	1	used
	FOP	1	used
Alignment Based feature	CUB	1	used
	RCBS	1	used
	EW	1	used
	SCUO	61	used
	Trimers	64	used
Alignment Based feature	Numbers of hits	1	used
	Significance score	1	used
	Total bit score	1	used
	Frame entropy	1	used

Note that Annotations 'used'/'new' mean that features have/ haven't been used in the plant lncRNA identification.

a sequence. Fickett score ('Fickett') is a simple linguistic feature that distinguishes protein-coding from non-coding transcripts according to the combinational effect of nucleotide composition and codon usage bias. Composition, transition and distribution features ('CTD') consider the nucleotide composition (descriptor 'C'), transition (descriptor 'T') and distribution (descriptor 'D') of RNA sequences [21]. 'C' describes the content of four nucleotides among the sequence. 'T' represents the percent frequency with the conversion of four nucleotides between adjacent positions, which means the content of AG (or GA, vice versa), AC, TG, TC and GC along a sequence. 'D' indicates five relevant positions (0, 25%, 50%, 75%, 100%) among the transcripts of four nucleotides. Isoelectric point ('PI') [10] is the theoretical isoelectric point of a predicted peptide calculated by the ProtParam module in BioPython. The grand average of hydropathy ('GRAVY') [10] value means the grand average of hydropathicity, a predicted peptide of which is calculated by the ProtParam module in BioPython. Instability provides an estimate of the stability of the protein in a test tube with a weight value of instability to different dipeptides.

2) ORF-BASED FEATURES

ORF is a portion of a gene sequence that contains a sequence of bases and could potentially encode a protein. ORF features are fundamental ones to distinguish lncRNA from pcts.

ORF length is the maximum length of the ORF. Studies [11], [22] revealed that protein-coding genes usually have long ORFs (>100 codons), while putative long ORFs in

non-coding genes can hardly be observed. ORF integrity indicates whether the ORF begins with a start codon and ends with an in-frame stop codon [23]. ORF coverage is the ratio of ORF length to transcript length. It is reported that ORF coverage is much lower in non-coding RNAs than that in protein-coding RNAs [11]. ORF score (FF-score), is extracted using Framefinder software [13].

3) CODON-BASED FEATURES

Codon-based features are related to the different usage frequencies of codons that occur in the pcts as one specific amino acid usually can be translated from several synonymous codons.

The frequency of the optimal codons ('FOP') [14] is the ratio of the number of optimal codons to a total number of synonymous codons. The codon usage bias ('CUB') [14] is the index that estimates the differences of codon bias between test set sequences and reference set sequences. The strength of relative codon bias ('RCBS') [24] is an overall score of a gene that indicates the influence of RCB of each codon. RCB reflects the level of gene expression. The weighted sum of relative entropy ('EW') [14] evaluates the degree of deviation from equal codon usage. The synonymous codon usage order ('SCUO') [14] is also a measure related to entropy-based codon bias. The relative synonymous codon usage ('RSCU') [25] refers to the relationship between observed codon frequencies and the number of times codon. We also calculated frequencies of 64 trimers ('Trimers') among A, C, G and T to capture potential codon usage bias.

4) ALIGNMENT-BASED FEATURES

Alignment-based features are obtained by aligning all the sequences to curated sequences to observe the similarity between unpredicted transcripts and labeled ones. Different from the intrinsic properties of each transcript itself, alignment-based features are necessary.

BLAST is a useful tool for finding regions of similarity between nucleotide or protein sequences [26]. The basic idea of BLAST is to align the query sequence with sequences in a database. Then it generates satisfying aligned word pairs, and each pair is called a 'hit'. We use the BLAST program to assess whether lncRNAs have significant similarity to pcts in the SWISS-PROT database [27]. The following four features are extracted by parsing the BLAST outputs [13]. The number of hits is as a fundamental indicator in BLAST, and the number of hits for pcts is expected higher than that for lncRNAs. However, many sequences show random unimportant matches to a BLAST database, so the quality of the hit is considered using three more features: significance score, total bit score, and frame entropy. The significance score establishes an intuitive relationship between the e-value in BLAST and the quality of hits of a given sequence. The total bit score simply sums up all the bit scores which are a normalized measure evolved from raw alignment score in BLAST. Frame entropy indicates the way of the hits distributed in different reading frames.

C. FEATURE SELECTION

Feature selection is a process of selecting the most discriminative features from a set of features [28]. This method can be used to identify and remove redundant features that do not contribute to or even decrease the accuracy of predictive models. There are two types of commonly used feature selection methods: filter methods and wrapper methods. However, filter methods ignore dependencies among features, whereas wrapper methods are inefficient in time cost [14]. Here, we adopt a novel feature selection method [29], called ReliefF-GA, by combining the ReliefF [30] with the Genetic Algorithm (GA) [31]. Integration of ReliefF and GA overcomes the weaknesses of a single filter or wrapper method, thus lead to an effective selection scheme.

First, the ReliefF is applied to remove features that are not contributive or even counterproductive to classification. ReliefF is the extension of the Relief algorithm [32]. It is more robust and can deal with incomplete data compared with the original Relief algorithm. Similar to Relief, the key idea of the ReliefF is to estimate the quality of features according to how well their values distinguish among sequences that are near to each other. After performing the ReliefF algorithm, features with an importance score less than zero are removed because the threshold of zero implies whether this feature is contributed.

Next, we perform GA to obtain the optimal feature subset for each species. GA is a heuristic optimization method inspired by natural evolution. For feature selection by GA, it starts with a set of candidate individuals called the population. Each individual, also a combination of selected features in the population indicates a solution to the selection problem. With the initial population, it starts iterations to produce better approximations. In each generation, individuals in the population may undergo crossovers, mutations and then being selected according to their levels of fitness. To simplify our problems, binary encoding is adopted to represent the feature combination. A bit of '1' means the corresponding feature is selected, whereas '0' indicates not. Hence, a solution is converted into a binary string with length equal to the total number of features. We initialize the first generation based on feature candidates produced by ReliefF. For each combination in the population, the XGBoost classifier is built on the training set and tested on the test set. After that, we consider models' predictive AUC score to represent the level of fitness of corresponding feature combinations. Eventually, GA obtains the optimal feature subset after a series of iterative computations.

D. ENSEMBLE LEARNER CONSTRUCTION

In machine learning, an ensemble learner consists of several base learners, and each base learner will have its own classification strengths, resulting in stronger and more accurate predictions than individual base learners, and ensemble models have many successful applications in bioinformatics [33]–[38]. The bagging algorithm is a commonly used ensemble strategy [39]. It has an effective application of

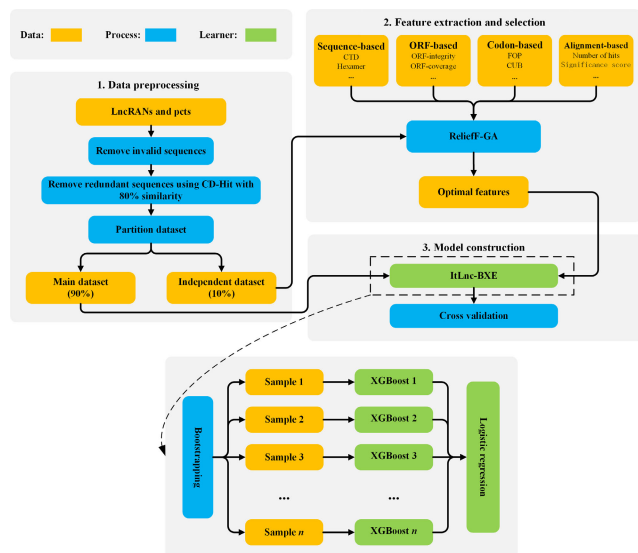


FIGURE 1. Workflow of ItLnc-BXE that involves the following steps: construction of benchmark dataset, extraction of four categories of features and feature selection, and construction of ItLnc-BXE model. (1) We exclude wrongly annotated sequences from raw data and then use CD-Hit to remove similar sequences. Subsequently, data are divided into main datasets and independent datasets. (2) We collect 175-dimensional features from four categories and adopt ReliefF-GA to select optimal features. This process is based on independent datasets. (3) Sequences in the main datasets are transformed into feature vectors according to optimal features in the step (2), on which ItLnc-BXE models are constructed. We sample training datasets into n subsets, based on which n XGBoost base learners are built. Then, all learners are combined using logistic regression.

reducing the variance and improving the classification ability of the base learners in supervised learning. Here, we propose an ensemble model for the plant lncRNA identification using a bagging algorithm. First, we use the bootstrapping algorithm to generate multiple data subsets from the training dataset. The XGBoost is a scalable tree boosting system that has superior performance in supervised learning [9], [40], and we apply it to build multiple base learners based on data subsets. How to combine these base learners is critical and challenging work. Popular ways, such as arithmetic mean and majority voting are usually utilized. We adopt an LR meta-learner to reduce the information redundancy between base learners. The LR meta-learner uses the outputs from base learners as inputs, and then produce a score indicating the probability of being a plant lncRNA.

E. WORKFLOW OF ItLnc-BXE

A workflow describes the process of ItLnc-BXE (Figure 1). First, we construct the datasets for each species, including main datasets and independent datasets, and they have no overlap. Next, the feature selection is implemented for each species using the independent datasets, and the optimal feature subsets are determined for the model construction. After that, we perform cross-validation experiments based on the main datasets.

In each fold of cross-validation, we divide training data into two parts: 7/10 of data for training base learners and 3/10 of

data for the LR learner. Using the bootstrapping algorithm, n data subsets are sampled from 7/10 of training data, based on which n XGBoost learners are trained. These base learners are applied to the prediction of the rest 3/10 of data, the results of which are regarded as training data for the LR learner. In this way, the ItLnc-BXE model is constructed.

III. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETTING

We evaluate all ItLnc-BXE models on the datasets of six species: *Arabidopsis thaliana* (A), *Solanum tuberosum* (S), *Oryza sativa* (O), *Hordeum vulgare* (H), *Physcomitrella patens* (P) and *Chlamydomonas reinhardtii* (C).

As ItLnc-BXE samples sub-datasets from training data and build several base learners on them, the number of base learners used in ensemble models ought to be determined. We perform 10-fold cross-validation to evaluate ItLnc-BXE models with different numbers of base learners using the independent datasets and determine to use five base learners for ItLnc-BXE according to the experimental results.

10-fold cross-validation is performed on the main sets to evaluate the performance of ItLnc-BXE and compared methods. We adopt popular evaluation metrics, including the area under the ROC curve (AUC), the area under the precision-recall curve (AUPR), accuracy (ACC), sensitivity (SN), specificity (SP), precision (PRE), f1-score and Matthews correlation coefficient (MCC).

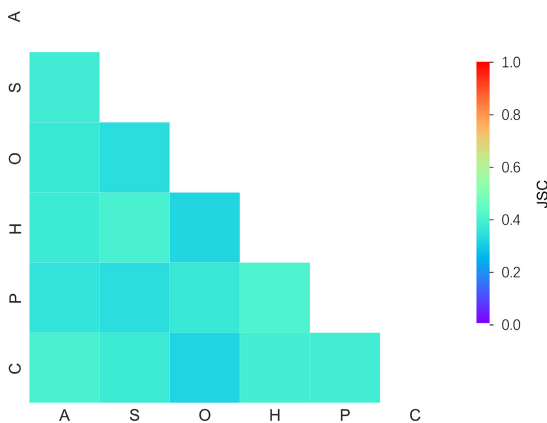
B. FEATURE DISCUSSION

Features are critical for distinguishing lncRNAs from pcts, thus we consider a variety of features for the plant lncRNA identification. However, these features may make different contributions to the identification of lncRNAs from different species, and some are redundant. To make analysis, we apply the ReliefF method to score the importance of collected features, and we obtain scores of all features for each species. Then, we calculate the Pearson correlation coefficient (PCC) between scores of all features for every two species (Table 3). PCC is used to measure the correlation of the feature ranking lists for different species. The results show that some species have relatively high correlations (>0.8) and others have comparably low correlations (<0.6), indicating that the species have the preference for features. Therefore, it is necessary to determine optimal feature subsets from candidates to build species-specific models.

We adopt ReliefF-GA to select optimal feature subsets for the model construction. Since there are six species, we implement the feature selection using their independent datasets respectively. The results show that ReliefF-GA greatly reduces 175-dimensional features for all species to lower dimensions (“A”: 89, “S”: 93, “O”: 87, “H”: 95, “P”: 90, “C”: 88), and refer to Supplementary Table S0 for detail. However, the optimal feature subsets are different for each species. To explore commonly used features for

TABLE 3. Pearson correlation coefficients (PCCs) of Relief importance score between every two species.

	A	S	O	H	P	C
A	1					
S	0.908	1				
O	0.821	0.870	1			
H	0.856	0.808	0.827	1		
P	0.775	0.834	0.792	0.667	1	
C	0.552	0.523	0.483	0.532	0.316	1

**FIGURE 2.** Heatmap of Jaccard similarity coefficients (JSCs) between every two species. JSC ranges from 0 to 1, and the bigger JSC means two species share more features.

two species, we respectively calculated the Jaccard similarity coefficient (JSC). JSC is defined as follows:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

where S_1 and S_2 are two sets, $|S_1 \cap S_2|$ means the card of the intersection and $|S_1 \cup S_2|$ means the card of the union. JSC indicates the similarity between the two sets. Then, we calculate JSC for every two species (Figure 2). The results show that JSCs range from 0.32 to 0.40, which means that these species share some features but use more different features. Further, we pay attention to those commonly used features. We find several features (Instability, one dimension from CTD, two from RSCU and three from Trimers) are shared by six species, indicating that these features are preferred in the plant lncRNA identification. Moreover, some features (Fop, Frame Entropy, three dimensions from CTD, four from Trimers and eight from RSCU) are shared by five species at most, indicating that they can be commonly used in plant lncRNA identification. For more commonly used features share by more than two species, refer to Supplementary Table S1. We also calculate the percentages about how many new features (CTD, PI, GRAVY, Instability, and ORF-integrity) are included in the optimal feature subsets. For species “A”, “S”, “O”, “H”, “P” and “C”, the results

are 60%, 80%, 60%, 60%, 60%, and 80% respectively. To be more specific, CTD is included in all species; PI is included in “H” and “C”; GRAVY is included in “A”, “S”, “P” and “C”; Instability is included in all species; ORF-integrity is included in “S” and “O”. As more than half of newly added features are involved in the final model of all species, we consider these features are useful.

C. COMPARISON WITH OTHER METHODS

Two machine learning methods: PLIT and PLncPRO have been presented to the plant lncRNA identification, and we adopt them for comparison. Source codes of PLIT and PLncPRO are publicly available, so we can correctly build PLIT and PLncPRO models, and then compare the performance of ItLnc-BXE with them on six different species (“A”, “S”, “O”, “H”, “P”, “C”). For each species, we consider the datasets with different lncRNAs/pcts ratios (1:1, 1:3 and 1:5). All prediction models are evaluated by using 10-fold cross-validation.

Here, we take the results on species “A” for analysis (see in Table 4). ItLnc-BXE produces better AUC scores than PLIT and PLncPRO in terms of the AUC scores for all species, and improvements in accuracy, SP, f1-score, MCC, and PRE are also observed. Moreover, we explore how the difference between lncRNAs/pcts ratios in datasets influences the performance of prediction models. Results show that ItLnc-BXE produces similarly AUC scores and accuracy on datasets with different ratios, and the conclusion can also be drawn from the results of compared methods. Further, we calculate standard deviations of AUC scores of each model on these datasets, and it seems that ItLnc-BXE has lower standard deviations (0.017) than PLIT (0.273) and PLncPRO (0.071), indicating that ItLnc-BXE is robust to the data imbalance.

The results for all species are included in Supplementary Table S2-S9. In general, ItLnc-BXE produces better results than PLIT and PLncPRO on the benchmark datasets of all six species. The superiority of ItLnc-BXE is owing to several factors. First, we consider a variety of representative sequence features that have proved to be useful in lncRNA identification. They bring enrich information for building high-accuracy models. Second, the feature selection method helps to determine the most informative features and reduce redundancy. Third, the ensemble learning strategy makes use of the strengths of base learners, thus leads to robust performance.

D. PERFORMANCE OF CROSS-SPECIES IDENTIFICATION

As discussed above, we can build species-specific ItLnc-BXE models based on datasets of species, and it is very interesting to examine the performance of ItLnc-BXE in cross-species identification. We conduct the following experiments based on the datasets of six species with ratio 1:1, and all models are built on the main sets. It is important to point out that, in cross-species identification between every two species, we test a specific model based on all data (main set and independent

TABLE 4. Performance comparison between ItLnc-BXE and compared methods on species Arabidopsis Thaliana.

lncRNAs/pcts	Method	AUC (%)	ACC (%)	AUPR (%)	SN (%)	SP (%)	MCC (%)	F1-score (%)	PRE (%)
1:1	ItLnc-BXE	99.27	96.60	99.08	97.68	95.53	93.23	96.64	95.63
	PLIT	95.78	87.98	95.65	87.94	88.03	75.99	87.97	88.03
	PLncPRO	99.09	95.73	98.85	98.66	92.79	91.61	95.85	93.19
1:3	ItLnc-BXE	99.31	96.60	97.72	93.57	97.61	90.95	93.22	92.88
	PLIT	96.29	90.33	90.38	76.14	95.07	73.57	79.73	83.74
	PLncPRO	99.25	96.66	97.47	95.12	97.18	91.23	93.45	91.83
1:5	ItLnc-BXE	99.30	97.07	96.25	90.71	98.34	89.40	91.15	91.63
	PLIT	96.41	92.55	86.30	68.46	97.37	71.58	75.38	83.94
	PLncPRO	99.23	96.97	95.92	91.12	98.14	89.10	90.92	90.72

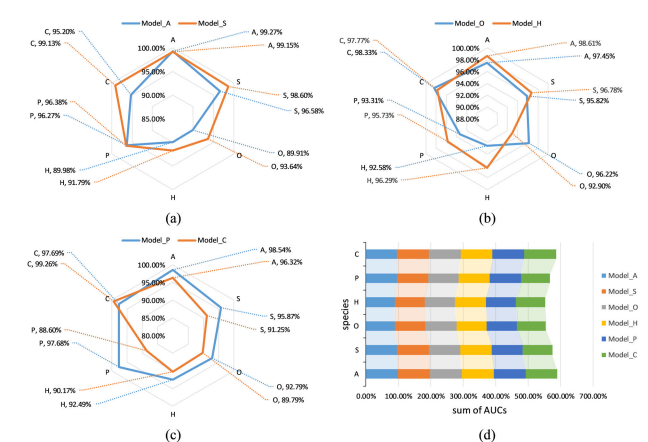


FIGURE 3. Performance comparison (AUC scores) of ItLnc-BXE in cross-species lncRNA identification. We train species-specific models on training data of six species and test them on test data of six species respectively. "Model_A" means the model built on data of species "A". "A", "S", "O", "H", "P" and "C" are abbreviations of six species and they are summarized in Table 1. (a) Predicting results of dicots-based models on data of six species, (b) results of monocots-based model and (c) results of lower plant-based models. (d) Sum of AUC scores of six species.

set) of another species. As for species self-identification, we just take the results of 10-fold cross-validation based on the main sets.

To present results, we draw radar and bar figures (Figure 3). Six plant species can be classified into four categories: dicots (*Arabidopsis thaliana* and *Solanum tuberosum*), monocots (*Oryza sativa* and *Hordeum vulgare*), the moss (*Physcomitrella patens*) and the alga (*Chlamydomonas reinhardtii*). Thus, the species in a category are visualized in a sub-figure for comparison. In general, ItLnc-BXE produces AUCs ranging from 88.60% to 99.27%. The performance under other metrics are also provided in Supplementary Table S10 - S17.

ItLnc-BXE models constructed on dicots produce relatively high AUCs on the moss (96.27%-96.38%) and the alga (95.20%-99.13%), but comparably low on monocots (89.91%-93.64%) (Figure 3a). Dicots may conserve plenty of

biological commonness with mosses and algae in the process of evolution. More specifically, some lncRNAs in mosses and algae may have the same composition and function as those in dicots, which leads to good prediction performance of dicots-based models on the moss and the alga. Species 'A' and 'S' are both dicots, and they produce similar and very high AUCs in cross-species identification. This indicates lncRNAs in different dicot species are closely similar to each other.

Similarly, monocots models produce relatively high AUCs on both dicots (95.82%-98.61%), the moss (93.31%-95.73%) and the alga (97.77%-98.33%) (Figure 3b). This also means lncRNAs in monocots conserve close similarity to those in dicots, mosses, and algae. However, there seems to be a contradiction. As mentioned before, dicots-based models produce comparably low AUCs on monocots, indicating lncRNAs in dicots are not very similar to those in monocots. This is actually explicable, probably because lncRNAs in dicots are less abundant than those in monocots. As a lack of information on monocot lncRNAs, dicots-based models will not perform equally well on monocots. Besides, within monocots, the AUCs between species 'O' and 'H' do not exceed 92.90%. We suspect that although monocots root from similar ancestors, they may gradually obtain specific biological properties, resulting in one always has defects to identify lncRNAs of other monocots.

Since the moss and the alga are both lower plants, we put them together for analysis (Figure 3c). The moss-based model produces relatively high AUCs on dicots (95.87%-98.54%) and the alga (97.69%), but comparably low AUCs on monocots (92.49%-92.79%). This demonstrates again that lncRNA properties in mosses are similar to those in dicots and algae, and lncRNAs in monocots are possibly far more abundant than those in mosses. The alga-based model produces comparably low AUCs on both monocots (89.79%-90.17%) and the moss (88.60%) and ranging AUCs on dicots (91.25%-96.32%). It is probably because the alga owns fewer similarities to other plants.

From another angle, we sum up the AUCs of ItLnc-BXE models for each tested species (Figure 3d). The results show that dicots produce the highest sums (575%-589%), the moss and the alga produce relatively high sums (568%-587%), and monocots produce comparably low sums (553%-555%). The AUC sums can describe how easy lncRNAs of one species are predicted by ItLnc-BXE models based on other species. Therefore, lncRNAs of dicots, mosses, and algae are more predictable than monocots; even if lack of data of lower plants, we can utilize dicots or monocots data to build models to identify lower plant lncRNAs.

IV. CONCLUSION

In this work, we propose an ItLnc-BXE based on the ensemble learning and bagging algorithm to identify plant lncRNAs. ItLnc-BXE makes use of diverse features that have been proven to be useful in lncRNA identification or related works, and the feature selection method is used to select the optimal feature subset. The frame of ensemble learning further improves performance. ItLnc-BXE constructs species-specific models on datasets of six plants respectively, and cross-validation experiments show that these models produce good performance. When compared with PLIT and PLncPRO, ItLnc-BXE yields better results on datasets of six species. ItLnc-BXE is a promising method for identifying lncRNAs from transcripts.

The studies on the sequence features reveal that plant lncRNAs in different species have a preference for different features but still share some features. Moreover, the studies on the cross-species lncRNA identification of six species suggest that: (1) cross-species models achieve good performance, (2) lncRNAs in dicots, mosses and algae are easy to be identified using models based on other species. Therefore, we can build ItLnc-BXE models of species with abundant data sources to identify lncRNAs in species lack of data.

ACKNOWLEDGMENT

(Guangyan Zhang and Zirui Liu contributed equally to this work.)

REFERENCES

- [1] F. Chalmel, A. Lardenois, B. Evrard, A. D. Rolland, O. Sallou, M.-C. Dumargne, I. Coiffec, O. Collin, M. Primig, and B. Jégou, "High-resolution profiling of novel transcribed regions during rat Spermatogenesis," *Biol. Reproduction*, vol. 91, no. 1, p. 5, and 1–13 Jul. 2014.
- [2] A. G. Matera, R. M. Terns, and M. P. Terns, "Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs," *Nature Rev. Mol. Cell Biol.*, vol. 8, no. 3, pp. 209–220, Mar. 2007.
- [3] T. Ravasi, "Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome," *Genome Res.*, vol. 16, no. 1, pp. 11–19, 2006.
- [4] H. Wei and B. Liu, "iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization," *Briefings Bioinf.*, Jun. 2019.
- [5] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jul. 29, 2019, doi: [10.1109/TCBB.2019.2931546](https://doi.org/10.1109/TCBB.2019.2931546).
- [6] X. Liu, L. Hao, D. Li, L. Zhu, and S. Hu, "Long non-coding RNAs and their biological roles in plants," *Genomics, Proteomics Bioinf.*, vol. 13, no. 3, pp. 137–147, Jun. 2015.
- [7] J. A. Chekanova, "Long non-coding RNAs and their functions in plants," *Current Opinion Plant Biol.*, vol. 27, pp. 207–216, Oct. 2015.
- [8] E.-D. Kim and S. Sung, "Long noncoding RNA: Unveiling hidden layer of gene regulatory networks," *Trends Plant Sci.*, vol. 17, no. 1, pp. 16–21, Jan. 2012.
- [9] Y.-C. Zhang and Y.-Q. Chen, "Long noncoding RNAs: New regulators in plant development," *Biochem. Biophys. Res. Commun.*, vol. 436, no. 2, pp. 111–114, Jun. 2013.
- [10] Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao, "CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W12–W16, Jul. 2017.
- [11] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, "CPAT: Coding-potential assessment tool using an alignment-free logistic regression model," *Nucleic Acids Res.*, vol. 41, no. 6, p. e74, Apr. 2013.
- [12] A. Li, J. Zhang, and Z. Zhou, "PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme," *BMC Bioinf.*, vol. 15, no. 1, p. 311, 2014.
- [13] U. Singh, N. Khemka, M. S. Rajkumar, R. Garg, and M. Jain, "PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea," *Nucleic Acids Res.*, vol. 45, no. 22, p. e183, Dec. 2017.
- [14] S. Deshpande, J. Shuttleworth, J. Yang, S. Taramonli, and M. England, "PLIT: An alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets," *Comput. Biol. Med.*, vol. 105, pp. 169–181, Feb. 2019.
- [15] M. W. Szcześniak, W. Rosikiewicz, and I. Makołowska, "CANTATAdb: A collection of plant long non-coding RNAs," *Plant Cell Physiol.*, vol. 57, no. 1, p. e8, Jan. 2016.
- [16] D. Bolser, D. M. Staines, E. Pritchard, and P. Kersey, "Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomics data," in *Plant Bioinformatics*. New York, NY, USA: Humana Press, 2016, pp. 115–140.
- [17] K. P. Williams, "RNAcentral: An international database of ncRNA sequences," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D123–D129, Jan. 2015.
- [18] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.
- [19] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, 2020.
- [20] H. Wei, Q. Liao, and B. Liu, "lncRNADis-FB: Identify lncRNA-disease associations by fusing biological feature blocks through deep neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 6, 2020, doi: [10.1109/TCBB.2020.2964221](https://doi.org/10.1109/TCBB.2020.2964221).
- [21] X. Tong and S. Liu, "CPPred: Coding potential prediction based on the global description of RNA sequence," *Nucleic Acids Res.*, vol. 47, no. 8, p. e43, May 2019.
- [22] M. C. Frith, T. L. Bailey, T. Kasukawa, F. Mignone, S. K. Kummerfeld, M. Madera, S. Sunkara, M. Furuno, C. J. Bult, J. Quackenbush, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, G. Pesole, and J. S. Mattick, "Discrimination of non-protein-coding transcripts from protein-coding mRNA," *RNA Biol.*, vol. 3, no. 1, pp. 40–48, Jan. 2006.
- [23] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao, L. Wei, and G. Gao, "CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine," *Nucleic Acids Res.*, vol. 35, no. 2, pp. W345–W349, Jul. 2007.
- [24] U. Roymondal, S. Das, and S. Sahoo, "Predicting gene expression level from relative codon usage bias: An application to escherichia coli genome," *DNA Res.*, vol. 16, no. 1, pp. 13–30, 2009.
- [25] P. M. Sharp, T. M. F. Tuohy, and K. R. Mosurski, "Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes," *Nucleic Acids Res.*, vol. 14, no. 13, pp. 5125–5143, 1986.
- [26] S. Altschul, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
- [27] C. O'Donovan, "High-quality protein knowledge resource: SWISS-PROT and TrEMBL," *Briefings Bioinf.*, vol. 3, no. 3, pp. 275–284, Jan. 2002.
- [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, Jan. 2016.

- [29] B. Li, P. Zhang, G. Ren, and Z. Xing, "A two stage feature selection method for gear fault diagnosis using ReliefF and GA-wrapper," in *Proc. Int. Conf. Measuring Technol. Mechatronics Autom.*, vol. 1, 2009, pp. 578–581.
- [30] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.
- [31] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. White Plains, NY, USA: Longman, 1989, p. 372.
- [32] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Workshop Mach. Learn.* Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.
- [33] W. Zhang, K. Jing, F. Huang, Y. Chen, B. Li, J. Li, and J. Gong, "SFLN: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions," *Inf. Sci.*, vol. 497, pp. 189–201, Sep. 2019.
- [34] Y. Gong, Y. Niu, W. Zhang, and X. Li, "A network embedding-based multiple information integration method for the MiRNA-disease association prediction," *BMC Bioinf.*, vol. 20, no. 1, p. 468, Dec. 2019.
- [35] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLoS Comput. Biol.*, vol. 14, no. 12, 2018, Art. no. e1006616.
- [36] S. Wan, Y. Duan, and Q. Zou, "HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source," *PROTEOMICS*, vol. 17, nos. 17–18, Sep. 2017, Art. no. 1700262.
- [37] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, and Z. Hong, "Improving tRNAscan-SE annotation results via ensemble classifiers," *Mol. Inform.*, vol. 34, nos. 11–12, pp. 761–770, Nov. 2015.
- [38] X. Ru, P. Cao, L. Li, and Q. Zou, "Selecting essential microRNAs using a novel voting method," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 16–23, Dec. 2019.
- [39] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, vol. 19, no. 9, pp. 1090–1099, Jun. 2003.
- [40] B. Liu, Z. Luo, and J. He, "SgRNA-PSM: Predict sgRNAs on-target activity based on position-specific mismatch," *Mol. Therapy-Nucleic Acids*, vol. 20, pp. 323–330, Jun. 2020.



JICHEN DAI is currently pursuing the degree with the Schaefer School of Engineering and Science, Stevens Institute of Technology, USA. His research interests include machine learning and data mining.



ZILAN YU is currently pursuing the bachelor's degree with the College of Informatics, Huazhong Agricultural University, China. Her research interests include machine learning and data mining.



SHUAI LIU is currently pursuing the bachelor's degree with the College of Informatics, Huazhong Agricultural University, China. His research interests include machine learning and data mining.



GUANGYAN ZHANG is currently pursuing the bachelor's degree with the College of Public Management, Huazhong Agricultural University, China. His research interests include machine learning and data mining.



ZIRU LIU is currently pursuing the bachelor's degree with the College of Informatics, Huazhong Agricultural University, China. His research interests include machine learning and data mining.



WEN ZHANG received the bachelor's and master's degrees in computational mathematics and the Ph.D. degree in computer science from Wuhan University, in 2003, 2006, and 2009, respectively. He is currently a Professor with the College of Informatics, Huazhong Agricultural University, China. His research interests include machine learning and bioinformatics.

...