Functional & Integrative Genomics
https://doi.org/10.1007/s10142-021-00769-w

ORIGINAL ARTICLE

LncMachine: a machine learning algorithm for long noncoding RNA annotation in plants

H. Busra Cagirici[1,2] · S. Galvez[3] · Taner Z. Sen[1] · Hikmet Budak[4]

Received: 13 August 2020 / Revised: 20 January 2021 / Accepted: 25 January 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

Following the elucidation of the critical roles they play in numerous important biological processes, long noncoding RNAs (lncRNAs) have gained vast attention in recent years. Manual annotation of lncRNAs is restricted by known gene annotations and is prone to false prediction due to the incompleteness of available data. However, with the advent of high-throughput sequencing technologies, a magnitude of high-quality data has become available for annotation, especially for plant species such as wheat. Here, we compared prediction accuracies of several machine learning algorithms using a 10-fold cross-validation. This study includes a comprehensive feature selection step to refine irrelevant and repeated features. We present a crop-specific, alignment-free coding potential prediction tool, LncMachine, that performs at higher prediction accuracies than the currently available popular tools (CPC2, CPAT, and CNIT) when used with the Random Forest algorithm. Further, LncMachine with Random Forest performed well on human and mouse data, with an average accuracy of 92.67%. LncMachine only requires either a FASTA file or a TAB separated CSV file containing features as input files. LncMachine can deploy several user-provided algorithms in real time and therefore be effortlessly applied to a wide range of studies.

Keywords LncRNA · Machine learning · Random Forest · Plants

Introduction

With current advances in high-throughput sequencing technologies, a vast number of transcripts have been experimentally determined for a plethora of different species, including a number of plants, animals, insects, and microbes (Szymański and Barciszewski 2002; Claverie 2005; Mercer et al. 2011; Cagirici et al. 2017; IWGSC 2018). Transcriptomic and genomic studies have revealed that although the lengths of many of these transcripts are greater than 200 nucleotides, the majority do not code for functional proteins (Pennisi 2012; Budak et al. 2020). Such transcripts have been defined as long noncoding RNAs (lncRNAs). Initially, the lack of evidence for their function and evolutionary conservation raised concerns about the potential importance of lncRNAs (Struhl 2007). However, many of these concerns have now been experimentally addressed by the functional characterization of lncRNAs in many important biological processes (i.e., COOLAIR/COLDAIR) (Heo and Sung 2011). Studies in the last decade have revealed diverse regulatory functions, including biologically significant interactions such as between lncRNA:RNA and lncRNA:chromatin (Chekanova 2015) and involvement in several important biological processes, such as vernalization (Swiezewski et al. 2009), photo morphogenesis (Wang et al. 2014), reproduction (Ding et al. 2012), nodulation (Campalans 2004), and environmental stress adaptation (Liu et al. 2012).

Furthermore, lncRNAs appear to exhibit tissue-specific expression and functional conservation (Cabili et al. 2011; Ulitsky et al. 2011). Although sequence conservation almost always accounts for the functionality of a sequence, vice versa is not always true (Shannon et al. 2003). Instead of full-length sequence conservation, lncRNAs may have conserved small binding sites at the structural level to maintain functional

✉ Hikmet Budak
hikmet.budak@icloud.com

[1] US Department of Agriculture - Agricultural Research Service, Crop Improvement Genetics Research Unit, Western Regional Research Center, 800 Buchanan St, Albany, CA 94710, USA

[2] Faculty of Engineering and Natural Sciences, Sabanci University, Tuzla, Istanbul, Turkey

[3] ETSI Informatica, University of Malaga Andalucía Tech., 29071 Malaga, Spain

[4] Montana BioAgriculture Inc., Missoula, MT, USA

Published online: 26 February 2021

interactions with proteins or other DNA/RNAs (Militti et al. 2014). Therefore, the understanding of the diverse functions of lncRNAs has the potential to provide insights into the different constraints that also drive conservation of other RNA classes, such as messenger RNAs (mRNAs) and micro RNAs (miRNAs) (Hezroni et al. 2015).

Despite their importance, computational identification of lncRNA during genome annotation is challenging. To distinguish lncRNAs from classes of small noncoding RNAs, such as miRNAs, the size of the transcript can be used. But discrimination based on length is not sufficient for identification: for example, both lncRNAs and mRNAs are long and share similar splicing and poly-A tailed structures, and in this case, other discriminants such as structural and functional features need to be used (Ulitsky and Bartel 2013). Additionally, lncRNA transcripts cannot be identified solely through homology, as the sequences are less conserved between species than protein-coding genes (Pang et al. 2006), and the presence of open reading frames in lncRNAs adds another layer of complexity. Another challenge is the growing evidence suggesting that some lncRNAs may not be noncoding but in fact code for short functional peptides. The best known example is the lncRNA known as early nodulin 40 (ENOD40) (Campalans 2004), whose conserved nucleotide sequence at the 5′ end encodes two short peptides with lengths of 12 and 24 amino acids (Rohrig et al. 2002). Proteogenomic and mass spectrometry have also been carried out to identify peptides using small ORFs (Andrews and Rothnagel 2014; Zhu et al. 2018).

In recent years, several predictive tools have been developed to distinguish between lncRNAs and coding RNAs using a range of different features and algorithms. The most popular of these tools are also among the most accurate and informative: Coding Potential Calculator (CPC) (Kong et al. 2007), Coding Noncoding Index (CNCI) (Sun et al. 2013), and Coding Potential Assessment Tool (CPAT) (Wang et al. 2013).

CPC uses a Support Vector Machine (SVM) algorithm with a standard radial basis function kernel to differentiate coding RNAs from ncRNAs based on both the extend and quality of the ORFs and the evidence of sequence similarity to proteins (Kong et al. 2007). In 2017, the CPC algorithm was updated to an alignment-free CPC2 (Kang et al. 2017), which has increased the speed and accuracy of identification. As an alignment-free tool, CPC2 has become species neutral that does not require training for different species. Selected features were evolved in CPC2 to include ORF length, ORF integrity, isoelectric point, and Fickett score. Fickett score was adapted from CPAT and refers to the asymmetrical distribution of each base favored in a sequence (Wang et al. 2013).

Another algorithm, CPAT, evaluates coding potential using an alignment-free Logistic Regression model (Wang et al. 2013). Its features include ORF length, Fickett score, and hexamer score. Hexamer score captures the score for codon usage bias of adjacent amino acids in a sequence (Wang et al. 2013). CPAT has an advantage over CPC2 as it allows users to create a model with their own data.

In comparison, CNCI is an alignment-free tool using SVM with a radial basis function kernel. It differentiates coding RNAs and ncRNAs based on the intrinsic composition of the sequence (Sun et al. 2013). Similar to hexamer score in CPAT, CNCI estimates the codon bias using unequal distribution of adjoining nucleotide triplets (ANTs) via a sliding window approach. The most likely coding domain sequence (MLCDS) is selected after scanning each sequence six times within each potential reading frame. Although this quantity shows similarities with the hexamer score, the ANT approach performs a more comprehensive downstream analysis to include the classification of partial transcripts (Han et al. 2016). CNCI was later upgraded to CNIT (Coding-Noncoding Identifying Tool) to provide faster and more accurate evaluation of sequences using the same features (Guo et al. 2019).

There are several other, but less popular lncRNA prediction tools, which use different prediction models and feature sets. Some of these include PLEK (Li et al. 2014), BASiNET (Ito et al. 2018), LncRNA-ID (Achawanantakun et al. 2015), and DeepLNC (Tripathi et al. 2016). In short, PLEK facilitates Support Vector Machine using k-mer-based features to distinguish lncRNAs from coding RNAs (Li et al. 2014). BASiNET uses decision tree algorithms trained with alignment-free features (Ito et al. 2018). In comparison, DeepLNC facilitates deep learning (Tripathi et al. 2016), where LncRNA-ID uses Random Forests (Achawanantakun et al. 2015). Some tools even construct an ensemble of models such as gradient boosting and Random Forests for the prediction of plant lncRNAs (Simopoulos et al. 2018).

Although current computational methods have yielded encouraging results, certain limitations are yet to be overcome. Predictions are highly dependent on training data, and while many tools aim to achieve high overall accuracy across several species, some focus on a narrow set of species. Recent studies have shown that species-specific predictions are optimally obtained from training data of the same or a closely related species (Singh et al. 2017). Singh et al. showed that PLncPRO, a model built specifically for monocots, achieved higher accuracy for lncRNA prediction when applied to monocots, rather than dicots and vice versa.

We developed a lncRNA prediction model, LncMachine, for crop plants and analyzed its performance for a wide range of crop species, including wheat. Wheat is a major crop across the globe, ranking second in human consumption worldwide (FAO 2019). To accurately identify both lncRNA transcripts and coding transcripts, we developed an alignment-free prediction workflow that includes several machine learning algorithms, which users can train for their species of interest. We evaluated several features included in other studies and performed feature selection algorithms to extract the best set of features to distinguish coding and noncoding sequences. Using this feature set and comprehensive training data, we first obtained 10-fold cross-validation accuracies for nine different algorithms, including Support Vector Machines, Logistic Regressions, and Random Forests. We then compared the prediction accuracies for two independent wheat transcript datasets for hexaploid and tetraploid wheat species, as well as for the plant lncRNAs that are available at the GreeNC database. Last, we included the comparison of prediction accuracies using the test data provided in the CPC2, CPAT, and CNIT tools to show that the LncMachine accuracies are not biased for a specific dataset.

Material and methods

Datasets

Training datasets were collected from two databases: lncRNA sequences from CANTATAdb v2 (Szcześniak et al. 2019) and mRNA sequences from Ensembl Plants (v37). CANTATAdb v2 contained lncRNAs for a wide range of plant species, and these lncRNAs were based on genome assemblies deposited in Ensembl Plants v37. We selected lncRNAs from monocotyledons and eudicotyledons, which had corresponding cDNAs deposited in Ensembl Plants (v37). Sequences with >90% N stretches and <200 bp in length were removed. Redundant sequences having a sequence identity of at least 90% were also removed using CD-HIT (Fu et al. 2012) at its default settings. After filtering, an equal number of cDNAs and lncRNAs were randomly selected for each species. A total of 90,104 lncRNA sequences and 90,104 cDNA sequences were included in the training dataset.

Test datasets included (1) lncRNAs from monocotyledons and eudicotyledons deposited at GreeNC database (Gallart et al. 2016), (2) an equal number of lncRNAs from IWGSC wheat RefSeq v1.0 annotation and high-confidence CDSs from IWGSC wheat RefSeq v1.1 annotation (IWGSC 2018) available through https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations, (3) an equal number of lncRNAs and high-confidence CDSs of tetraploid wheat (Svevo) (Maccaferri et al. 2019) available through https://www.interomics.eu/durum-wheat-genome, as well as at GrainGenes (https://wheat.pw.usda.gov) (Blake et al. 2019), and (3) the datasets used for testing in CPC2 (Kang et al. 2017), CPAT (Wang et al. 2013), and CNIT (Guo et al. 2019). These datasets were used for comparisons of accuracies independent of the datasets chosen.

Feature extraction

Initially, we extracted 93 features based on sequence intrinsic properties (File S1) which later were subject to feature selection. The initial features were composed of the following:

1	ORF length
2	ORF coverage
3	Sequence length
4	GC%
5–8	k-mer (*k*=1) frequencies; monomer frequencies of the four nucleotides
9–24	k-mer (*k*=2) frequencies; dimer frequencies of the four nucleotides
25–88	k-mer (*k*=3) frequencies; trimer frequencies of the four nucleotides
89	Fickett score from full length sequence
90	Fickett score from CDSs
91	Hexamer score
92	ORF integrity
93	Isoelectric point

To remove irrelevant and collinear features, we applied several feature selection methodologies available through python scikit-learn package (Pedregosa et al. 2011). These included the following:

- Variance threshold
- Univariate feature selection with ANOVA F-test
- Random Forest Classifier
- Recursive feature elimination
- Lasso regularization
- Pearson correlation

Each methodology provided a list of best features, including collinear features. For feature selections, collinearity was reduced by Pearson pairwise correlation of the best features.

Model construction and evaluation

LncMachine can build prediction models using several machine learning algorithms although the default was set to Random Forest. The script was provided in supplementary files (File S2) and available at GitHub at https://github.com/hbusra/lncMachine.git. Prediction models were built using nine machine learning algorithms from the python scikit-learn package: (1) LogisticRegression, (2) RandomForest, (3) Multilayer Perceptron (NeuralNet), (4) NearestNeighbors, (5) DecisionTree, (6) Gaussian Naïve Bayes (NaiveBayes), (7) AdaBoost, (8) Quadric Discriminant Analysis (QDA), and (9) Support Vector Machines with linear kernel (LinearSVM). Training accuracies were calculated by a 10-fold cross-validation. Using the

StratifiedKFold function of the scikit-learn package, training data were split into 10 different training/test sets. For each training/test set, the prediction accuracies for prediction models were calculated using the different feature sets suggested for each algorithm. Training performance was assessed by the mean and the standard deviations of the accuracy scores. Due to their computational cost, Support Vector Machine (SVM) algorithms were not included for further analyses after cross-validation (Table 1).

Testing of the prediction models on several plant datasets was then completed. The prediction performance was evaluated based on statistic metrics: accuracy (ACC), precision (PRE), sensitivity/recall (SN), specificity (SP), and F-score (Powers 2007). These were defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Fscore} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

TP	true positive
TN	true negative
FP	false positive
FN	false negative

Additional performance assessments were performed through plotting the Receiver Operating Characteristic (ROC) curve for visualization and calculating the respective Area Under the Curve (AUC) score from the ROC curve (Bradley 1997).

To compare the prediction performance of the LncMachine against other tools, CPC2, CPAT, and CNIT were utilized. All three tools were updated recently and have been considered among the most popular coding prediction tools. CPC2 is a species-neutral tool that does not provide a training option for different species; therefore, it was run at its default settings without training. CNIT was run in plant mode using 20 threads. CPAT was trained with the same training data used in the current study. The cutoff for the identification of coding and noncoding transcripts was determined as described in its manual (Wang et al. 2013).

Table 1 Performance of prediction models using training data with a 10-fold cross-validation

Algorithm	Training accuracy (%)	Std (%)
RandomForest	94.09	± 0.18
AdaBoost	93.62	± 0.16
NearestNeighbors	93.40	± 0.16
NeuralNet	93.39	± 0.30
LinearSVM	91.85	*NA
LogisticRegression	91.69	± 0.20
DecisionTree	90.86	± 0.23
QDA	88.10	± 0.44
NaiveBayes	87.38	± 0.28

*NA: not available because Support Vector Machine (SVM) with linear kernel was only run at 1-fold due to its computational cost

Results

Experimental setup and model construction

Using a set of publicly available lncRNA and mRNA sequences for 18 plant species, we constructed a plant-based lncRNA prediction tool, LncMachine, by evaluating the performance of eight machine learning algorithms and selecting the best features. For a total of 93 sequence intrinsic features, we assessed five feature selection methodologies: Lasso, Random Forest, recursive feature elimination, variance threshold, and univariate feature selection, all followed by Pearson pairwise correlation in addition to elimination of features by Pearson correlation alone (Table S1). After 10-fold cross-validations, the feature set was selected by variance threshold followed by Pearson correlation resulted in the highest accuracy and AUC score of ROC (Area Under the Receiver Operating Characteristics) (Bradley 1997) for the Random Forest Classifier (Table S1). Random Forest Classifier and the features selected by variance threshold followed by Pearson correlation were selected for further analysis.

First, we applied a variance threshold to select features that showed more than 80% of variance. Variance scores varied between 0 and 21,496,995 among 93 features. We selected the top ranking features with a score of at least 4. This highlighted 15 features: sequence length, ORF length, ORF coverage, GC content, T content, A content, C content, G content, CG content, GC content, isoelectric point, TT content, AA content, AT content, and TA content. Ranking of the features was provided in the Supplementary Table S2. Later, we applied Pearson pairwise correlation and selected only the highest scoring features based on correlation coefficient. The final set of features were sequence length, ORF length, GC content (GC%), and isoelectric point (pI) (Table S2). The three features except pI score were slightly higher in coding sequences on average, whereas pI score showed a similar distribution among coding and noncoding sequences but slightly higher among noncoding sequences on average (Fig. 1). Figure 2

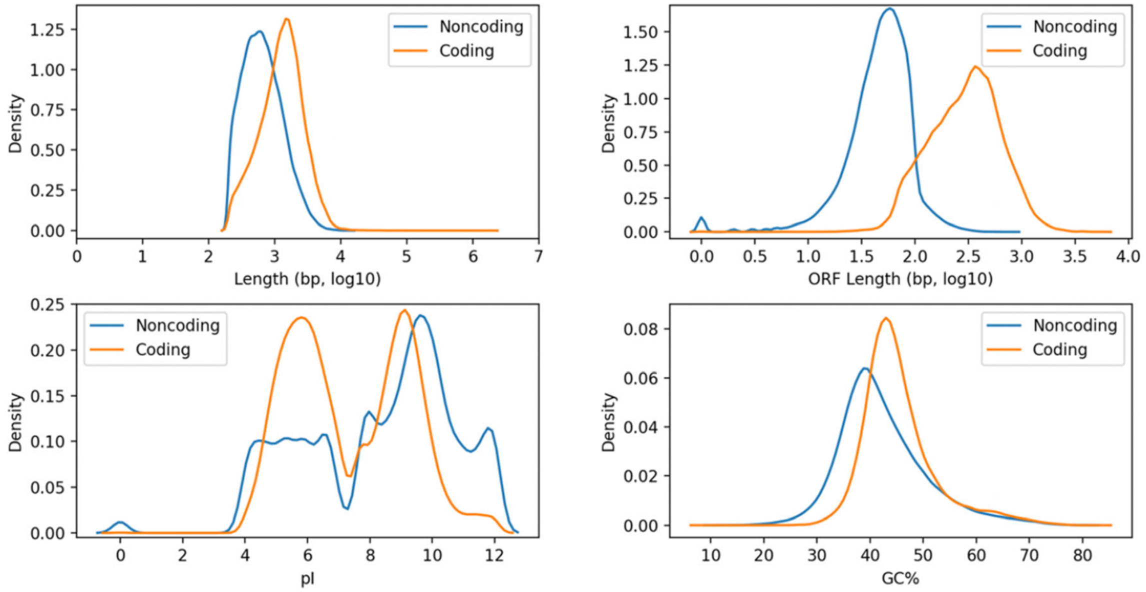


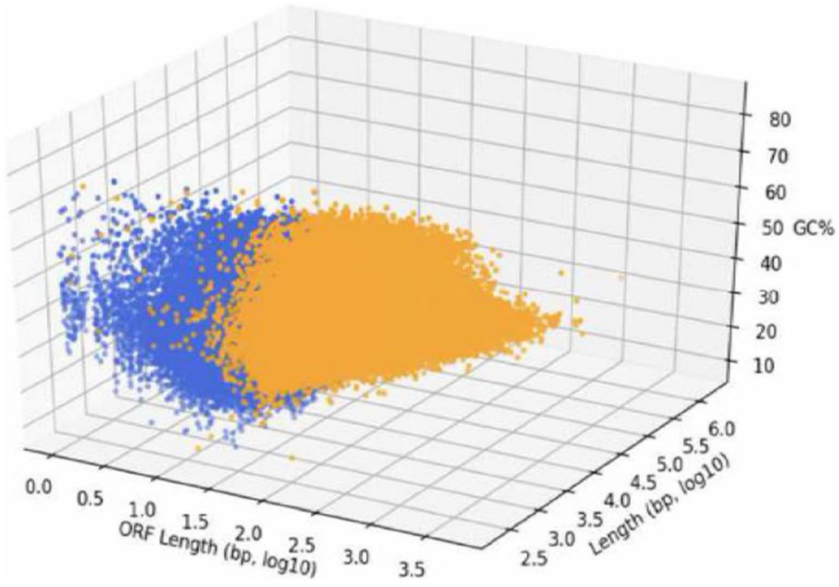
Fig. 1 The density distribution of the selected features to build the prediction model for coding (orange) and noncoding (blue) sequences. (**a**) Sequence length, (**b**) ORF length, (**c**) pI score, and (**d**) GC content

shows these three major features (sequence length, ORF length, and GC%) in three-dimensional space to assess the separation of coding and noncoding sequences based on the features selected. Our results showed that sequence length and ORF length have high separation power for coding and noncoding transcripts as even the largest noncoding sequences tend to contain small ORFs. Table 1 shows 10-fold cross-validation accuracies of the nine machine learning algorithms using these four features. All of the algorithms resulted in over 87% cross-validation accuracy (Table 1), indicating a good fit of the selected features in the prediction models. Given that Support Vector Machine (SVM) algorithm with linear kernel was not among the top performing algorithms, cross-validation was only performed by 1-fold due to the computational cost of SVM algorithms. Our results showed that LncMachine performs best with the RandomForest algorithm on the training data.

Performance evaluation against other plant datasets

We evaluated the performance of machine learning models on lncRNAs based on sensitivity in 6 plant species from the GreeNC database (Table 2). Our results showed that some algorithms perform very poorly on certain species. For example, lncRNAs of *Oryza sativa Japonica* were identified with only a range of 13–30 % sensitivity by five of the algorithms,

Fig. 2 The three-dimensional plot of the three features: sequence length, ORF length, and GC %, on coding (orange) and noncoding (blue) sequences



whereas QDA and Naïve Bayes provided a 99% sensitivity. Interestingly, among all nine algorithms, QDA and Naïve Bayes provided >96% sensitivity for all of the species. These results suggest that QDA and Naïve Bayes algorithms are better suited for the identification of lncRNAs for a wide range of plant species.

Additionally, the prediction models with QDA, Naïve Bayes, and Logistic Regression outperformed the three most popular tools, CPC2, CPAT, and CNIT, for all of the species tested from the GreeNC database. Among the popular tools, CPC2 predicts lncRNAs with a sensitivity ranged between 77 and 93%, CPAT between 41 and 75%, and CNIT between 43 and 63%. It was interesting to observe that although CNIT and CPAT were trained specifically for plants (i.e., in contrast to CPC2), they provided the lowest sensitivities for identifying plants lncRNAs.

We also evaluated the prediction performances on two real-life case studies: hexaploid and tetraploid wheat datasets. An equal number of lncRNAs and high-confidence CDS sequences were retrieved for both hexaploid (Chinese Spring) and tetraploid (Svevo) wheats (Table 3). For the full set of coding and noncoding sequences of the two wheat species, all nine algorithms provided >92% accuracy. Overall, our default RandomForest model outperformed all the remaining tools and algorithms for the wheat datasets with 98.65% and 99.25% accuracies for hexaploid and tetraploid wheats, respectively (Fig. 3 and Table 4). CPC2, CPAT, and CNIT provided accuracies >94%. For the wheat species, CPAT performed better than both CPC2 and CNIT.

These results show that even when the same algorithm is used, different parameters, platforms, feature sets, and training data can affect prediction accuracies. For example, CNIT, which uses LogisticRegression, and LncMachine, which uses LogisticRegression, showed drastic differences in the prediction accuracies for various test datasets. Across the test datasets, LncMachine with LogisticRegression outperformed CNIT (96.58% vs 98.06% for hexaploid wheat; 97.77% vs 97.43% for tetraploid wheat). There was a larger difference in the prediction of GreeNC plant lncRNAs, where LncMachine with LogisticRegression provided an average of 91.44% accuracy as opposed to an average of 50.59% from CNIT. Since CNIT was run in plants mode, our results suggest that its prediction capability is highly dependent on the dataset used. Interestingly, CPC2, which was by default trained on human data, provided better prediction accuracies than CNIT for both wheat datasets and GreeNC lncRNAs (Table 2 and Table 4).

Performance evaluation against CPC2, CPAT, and CNIT test datasets

To prevent any bias introduced by the selected datasets and to test species other than plants, we used the test datasets provided by CPC2, CPAT, and CNIT for comparison of prediction accuracies. It is important to note that the datasets provided by these tools are mostly unbalanced datasets, which might introduce a bias to prediction accuracies in the case when the prediction model favors either coding or noncoding sequences. Test datasets of CPC2 and CPAT were mostly non-plant species. However, although our training data only included plant sequences, the LncMachine with LinearSVM performed 93.59% (±2) accuracy for non-plant datasets (Table S3). The LncMachine with LogisticRegression was also shown to be efficient at identifying non-plant coding and noncoding

Table 2 Performance comparison of prediction models on GreeNC lncRNAs in terms of sensitivity

Model	GreeNC (lncRNAs)					
	Arabidopsis thaliana	*Brachypodium distachyon*	*Oryza sativa Japonica*	*Sorghum bicolor*	*Triticum aestivum*	*Zea mays*
QDA	*99.14*	*98.30*	*99.27*	*98.11*	*96.65*	*99.65*
NearestNeighbors	63.56	74.82	27.06	74.63	82.01	69.71
DecisionTree	58.58	68.46	29.69	66.69	76.07	66.64
RandomForest	61.24	74.70	26.98	72.80	80.79	70.31
NeuralNet	44.78	48.14	13.40	44.18	54.20	55.09
AdaBoost	60.44	65.20	25.70	64.13	74.10	66.11
NaiveBayes	98.80	98.05	99.10	97.83	96.25	99.60
LogisticRegression	87.30	93.02	87.89	92.44	94.43	93.53
LinearSVM	74.57	85.24	66.64	84.98	90.35	83.34
Other tools						
CPC2	80.46	88.76	77.42	88.60	93.19	88.57
CPAT	73.50	61.80	41.47	58.76	75.36	70.12
CNIT	46.86	50.56	51.81	47.38	43.95	62.96

Highest accuracy scores for each species were italicized

Table 3 Description of datasets used in training and validation of the wheat lncRNA prediction model

Dataset	Source	Reference	# of mRNA	# of lncRNA
Chinese Spring	Hexaploid wheat	IWGSC et al. (2018)	87,511	87,511
Svevo	Tetraploid wheat	Maccaferri et al. (2019)	115,437	115,437

sequences, with an average accuracy of 93.40 % (±2) and an average F1-score of 0.92 (Table S3). The LncMachine with its default RandomForest algorithm provided very similar results, with an average accuracy of 92.67% (±3) and an average F1-score of 0.91. These results suggest that LncMachine would also work well for non-plant species, such as human and mouse.

CNIT, on the other hand, provided several plant coding and noncoding sequences for the test set. These plant datasets were unbalanced, such that there were five lncRNA sequences for Sorghum bicolor and 39,045 coding sequences (Table S3). Based on the CNIT plant test datasets, LncMachine with Neural Networks outperformed all the algorithms and the tools CPAT, CPC2, and CNIT in terms of specificity and accuracy on average. Sensitivity was >0.996 on average for the LncMachine with Logistic Regression and with LinearSVM and CPC2. However, CPAT provided a better F1 score (0.44 on average as opposed to 0.41 of LncMachine with Neural Networks). It should be noted that the low number of lncRNAs available for most of the plant species in CNIT datasets may have resulted in unreliable F1 scores. For the species with more than 2500 lncRNA sequences in their test sets, F1 scores should be considered more accurate and reliable (Table S3). Overall, our results suggest that there is not any single solution that can fit all the datasets available. The users have the responsibility to select the best fitting algorithm for their specific study.

Additional application

Although our main purpose was to develop a lncRNA prediction tool specifically for crop plants, LncMachine can be modified to be used of any kind of species. When FASTA files for coding and noncoding sequences are provided by a user, LncMachine extracts features to distinguish lncRNAs from mRNAs and performs prediction of the coding potential of the sequences provided. Otherwise, if a features file in CSV format is provided, this new tool can be run to train a specific model other than lncRNA prediction using several machine learning algorithms which can be subsequently used for prediction of the test sets. The required columns for training a prediction model include “class” and “features” as separate columns. The samples can be specified as “readID” in the CSV file.

Discussion

Genome annotation can be an arduous task, particularly when distinguishing coding sequences from lncRNAs which

Table 4 Performance comparison of prediction models on wheat datasets

Model	Hexaploid wheat				Tetraploid wheat			
	ACC	PRE	SN	F-score	ACC	PRE	SN	F-score
QDA	92.79	0.88	1.00	0.93	94.27	0.90	1.00	0.95
NearestNeighbors	98.40	1.00	0.97	0.98	98.93	1.00	0.98	0.99
DecisionTree	96.19	0.99	0.94	0.96	97.03	0.99	0.95	0.97
RandomForest	*98.65*	*1.00*	*0.98*	*0.99*	*99.25*	*1.00*	*0.99*	*0.99*
NeuralNet	96.66	1.00	0.93	0.97	97.64	1.00	0.95	0.98
AdaBoost	97.84	1.00	0.96	0.98	98.75	1.00	0.98	0.99
NaiveBayes	93.00	0.88	1.00	0.93	94.27	0.90	1.00	0.95
LogisticRegression	96.58	0.94	0.99	0.97	97.77	0.96	1.00	0.98
LinearSVM	97.37	0.96	0.98	0.97	98.62	0.97	1.00	0.99
Other tools								
CPC2	96.64	0.99	0.94	0.97	97.93	1.00	0.96	0.98
CPAT	98.06	0.98	0.98	0.98	99.03	0.98	1.00	0.99
CNIT	94.82	0.95	0.95	0.95	97.48	0.99	0.96	0.97

ACC accuracy, *PRE* precision, *SN* sensitivity, *SP* specificity. Highest values of the metrics were shown in italics

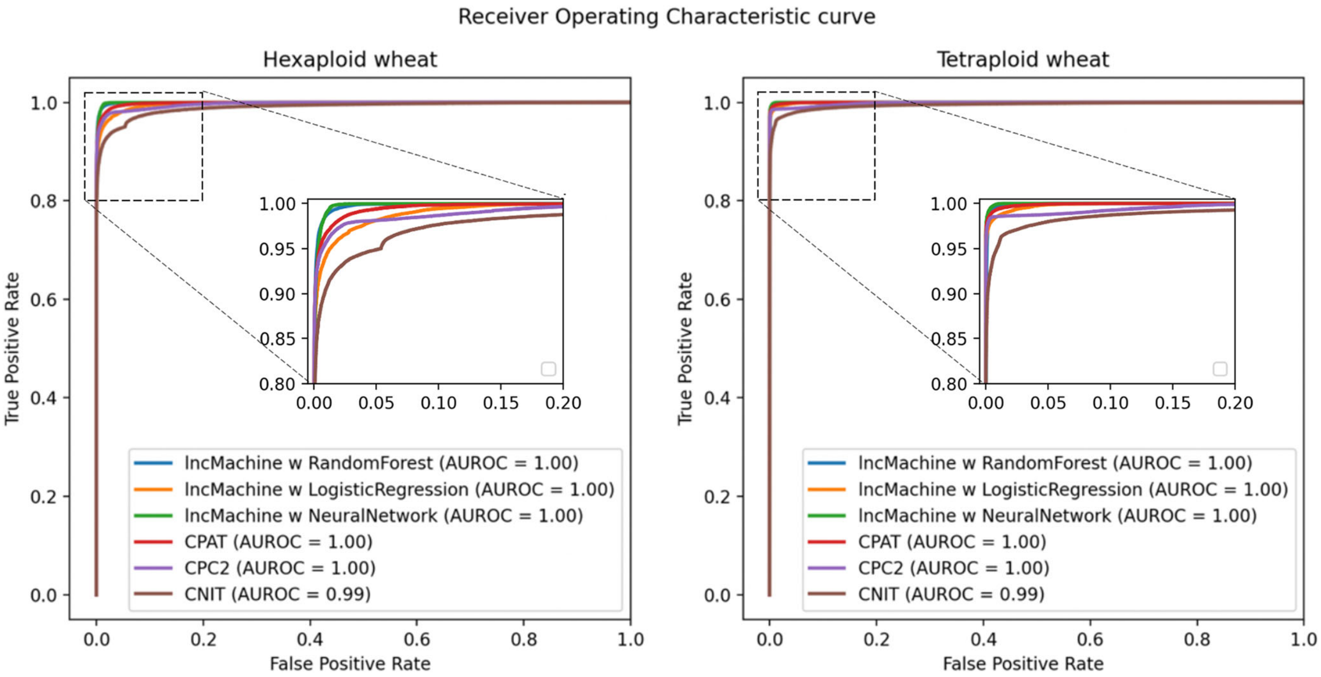


Fig. 3 The Receiver Operating Characteristic (ROC) curves of hexaploid and tetraploid wheat datasets using the LncMachine with the best three algorithms and other tools (CPAT, CPC2, and CNIT)

resemble coding sequences. Homology and alignment-based methods are highly dependent on the availability of prior data and the evidence of sequence conservation between species. However, this is insufficient when considering species-specific sequences and/or non-conserved sequences. Therefore, it is crucial to find better methods of classification to promote accurate identification of coding and noncoding sequences.

Although there have been many proposed lncRNA coding prediction tools (Han et al. 2016; Ventola et al. 2017), only limited sources are available for crop plants (Singh et al. 2017; Guo et al. 2019; Negri et al. 2019). Additionally, most plant-specific tools that have been developed are difficult to implement for further studies. The most trusted tools for the coding potential predictions include CPAT (Wang et al. 2013), CPC2 (Kang et al. 2017), and CNIT (Guo et al. 2019), which have all been improved and updated recently. Here, we investigated several features to distinguish coding and noncoding sequences in crop plants, compared several algorithms for their efficiencies with different sets of data, and provided performance measures for these tools.

The performances of machine learning models highly depend on the training data and the selected features. Several features proposed by different studies have been shown to be informative in the classification of coding and noncoding transcripts (Wang et al. 2013; Kang et al. 2017; Ito et al. 2018; Guo et al. 2019). These features include k-mers, basic structural features like length and GC content, Fickett score, hexamer score, ORF integrity, and isoelectric point. Although each of these features was proposed as good representatives of the differences between coding and noncoding sequences, no single feature has been proposed as the most superior. A combination of several features has typically been used in previous studies (Simopoulos et al. 2018; Negri et al. 2019). After collecting the features suggested by the most commonly used prediction tools, we compiled a list of 93 features, most of which were collinear. Our results showed that various feature selection algorithms, which proposed different sets of features, did not necessarily result in better classifications (Table S1). By comparing all feature selection strategies, we were able to obtain the best representation of coding and noncoding sequences. Overall, our results show that the final set of features (sequence length, ORF length, GC%, and pI) are suitable for the most algorithms. This combination of features has not previously been used, although individually each feature has been included in several other studies (Wang et al. 2013; Kang et al. 2017; Simopoulos et al. 2018).

In this study, we proposed an accurate model, LncMachine with RandomForest, for lncRNA and mRNA identification in wheat and other crop species. As training data, we used the comprehensive set of plant lncRNAs deposited in CANTATAdb v2. Among many other lncRNA databases, CANTATAdb was updated recently and receives regular support and maintenance (Szcześniak et al. 2019). As for the feature set, we incorporated a final set of four features to achieve better prediction accuracies. With comprehensive training data and a substantial list of features, we compared nine different algorithms for their prediction performances using the same training data and the same feature sets. Interestingly, training accuracies were over 87% for all the algorithms with the top performing at 94%

accuracy (Table 1), indicating a good fit between the selected features and the training data.

Comparison of these algorithms and the most trusted tools like CPAT, CPC2, and CNIT in various test sets showed that no single prediction model outperformed all the other tools and models in every setting. Instead, it was observed that each tool performed differently for different datasets (Table S3). Depending on the purpose of the study, we suggest using LncMachine with different algorithms for different species: Random Forest, as default algorithm, suitable for both plants (Table 4) and non-plant species (Table S3), and Logistic Regression and Neural Networks for unbalanced datasets or to introduce a bias for lncRNA predictions. Finally, LncMachine is able to implement several algorithms, providing the best adaptable model. As it is highly customizable, it can be applied across a wide range of studies. LncMachine will be a valuable contribution to the rapidly growing field of biological machine learning.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10142-021-00769-w.

Acknowledgements HB acknowledges Montana BioAg Inc. and this work is dedicated to Mr. Lyle Benjamin (past president of the MGGA), who recently left the farm industry but remains an avid supporter of science. His advocacy for innovation, and unwavering belief, has been invaluable in integrating agricultural research into Montana farms. SG acknowledges project P18-RT-992 from Junta de Andalucía (Andalusian Regional Government), Spain (co-funded by FEDER). HBC was supported by TUBITAK for her PhD and the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy (DOE) and the US Department of Agriculture (USDA) Agricultural Research Service (ARS). This research was supported by the USDA-ARS Project Number 2030-21000-024-00-D (HB and TZS). USDA is an equal opportunity provider and employer.

Data availability The datasets supporting the conclusions of this article are included within the article as Supplementary Material. LncMachine together with prebuilt prediction models and the training/test datasets are available through GitHub at https://github.com/hbusra/lncMachine.git.

Compliance with ethical standards

Competing interests The authors declare no competing interests.

References

Achawanantakun R, Chen J, Sun Y, Zhang Y (2015) LncRNA-ID: long non-coding RNA IDentification using balanced random forests. Bioinformatics 31:3897–3905. https://doi.org/10.1093/bioinformatics/btv480.

Andrews SJ, Rothnagel JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet. https://doi.org/10.1038/nrg3520

Blake VC, Woodhouse MR, Lazo GR, Odell SG, Wight CP, Tinker NA et al (2019) GrainGenes: centralized small grain resources and digital platform for geneticists and breeders. Database (Oxford):2019. https://doi.org/10.1093/database/baz065

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. https://doi.org/10.1016/S0031-3203(96)00142-2

Budak H, Kaya SB, Cagirici HB (2020) Long non-coding RNA in plants in the era of reference sequences. Front Plant Sci 11:276. https://doi.org/10.3389/fpls.2020.00276

Cabili M, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. https://doi.org/10.1101/gad.17446611

Cagirici HB, Biyiklioglu S, Budak H (2017) Assembly and annotation of transcriptome provided evidence of miRNA mobility between wheat and wheat stem sawfly. Front Plant Sci 8:1653

Campalans A (2004) *Enod40*, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in *Medicago truncatula*. The Plant Cell 16:1047–1059. https://doi.org/10.1105/tpc.019406

Chekanova JA (2015) Long non-coding RNAs and their functions in plants. Curr Opin Plant Biol 27:207–216. https://doi.org/10.1016/j.pbi.2015.08.003.

Claverie J-M (2005) Fewer genes, more noncoding RNA. Science 309: 1529–1530. https://doi.org/10.1126/science.1116800.

Ding J, Shen J, Mao H, Xie W, Li X, Zhang Q (2012) RNA-directed DNA methylation is involved in regulating photoperiod- sensitive male sterility in rice. Mol Plant 5:1210–1216. https://doi.org/10.1093/mp/sss095.

FAO (2019) FAO Statistics. FAOSTAT Stat. Database. Available at: http://www.fao.org/faostat/en/#data/QC. Accessed March 11, 2019

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts565

Gallart, A. P., Pulido, A. H., De Lagrán, I. A. M., Sanseverino, W., and Cigliano, R. A. (2016). GREENC: a Wiki-based database of plant IncRNAs. Nucleic Acids Res doi:https://doi.org/10.1093/nar/gkv1215.

Guo JC, Fang SS, Wu Y, Zhang JH, Chen Y, Liu J et al (2019) CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz400

Han S, Liang Y, Li Y, Du W (2016) Long noncoding RNA identification: comparing machine learning based tools for long noncoding transcripts discrimination. Biomed Res Int 2016. https://doi.org/10.1155/2016/8496165

Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331:76–79. https://doi.org/10.1126/science.1197349

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep 11: 1110–1122. https://doi.org/10.1016/j.celrep.2015.04.023

Ito EA, Katahira I, Vicente FF d R, Pereira LFP, Lopes FM (2018) BASiNET — Biological sequences network: a case study on coding and non-coding RNAs identification. Nucleic Acids Res 46. https://doi.org/10.1093/nar/gky462.

IWGSC, IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361: eaar7191. https://doi.org/10.1126/SCIENCE.AAR7191

Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L et al (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx428

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L et al (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35. https://doi.org/10.1093/nar/gkm391.

Li A, Zhang J, Zhou Z (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15. https://doi.org/10.1186/1471-2105-15-311.

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L et al (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 24:4333–4345. https://doi.org/10.1105/tpc.112.102855.

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M et al (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895. https://doi.org/10.1038/s41588-019-0381-3.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA et al (2011) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30:99–104. https://doi.org/10.1038/nbt.2024

Militti C, Maenner S, Becker PB, Gebauer F (2014) UNR facilitates the interaction of MLE with the lncRNA roX2 during Drosophila dosage compensation. Nat Commun 5:4762. https://doi.org/10.1038/ncomms5762.

Negri TDC, Alves WAL, Bugatti PH, Saito PTM, Domingues DS, Paschoal AR (2019) Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants. Brief Bioinform. https://doi.org/10.1093/bib/bby034

Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22:1–5. https://doi.org/10.1016/j.tig.2005.10.003.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pennisi E (2012) ENCODE project writes eulogy for junk DNA. Science 337:1159–1161. https://doi.org/10.1126/science.337.6099.1159

Powers DMW (2007) Evaluation: from precision, recall and f-factor. Tech Rep SEI-07-001

Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M (2002) Soybean ENOD40 encodes two peptides that bind to sucrose synthase. Proc Natl Acad Sci 99:1915–1920. https://doi.org/10.1073/pnas.022664799.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Simopoulos CMA, Weretilnyk EA, Golding GB (2018) Prediction of plant lncRNA by ensemble machine learning classifiers. BMC Genomics 19. https://doi.org/10.1186/s12864-018-4665-2.

Singh U, Khemka N, Rajkumar MS, Garg R, Jain M (2017) 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 45. https://doi.org/10.1093/nar/gkx866.

Struhl K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 14:103–105. https://doi.org/10.1038/nsmb0207-103.

Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C et al (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res 41. https://doi.org/10.1093/nar/gkt646.

Swiezewski S, Liu F, Magusin A, Dean C (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature 462:799–802. https://doi.org/10.1038/nature08618.

Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I (2019) CANTATAdb 2.0: expanding the collection of plant long noncoding RNAs. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-9045-0_26

Szymański M, Barciszewski J (2002) Beyond the proteome: non-coding regulatory RNAs. Genome Biol 3:reviews0005. https://doi.org/10.1186/gb-2002-3-5-reviews0005

Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK (2016) DeepLNC, a long non-coding RNA prediction tool using deep neural network. Netw Model Anal Health Inform Bioinforma 5:21. https://doi.org/10.1007/s13721-016-0129-2

Ulitsky I, Bartel DP (2013) XLincRNAs: genomics, evolution, and mechanisms. Cell. https://doi.org/10.1016/j.cell.2013.06.020

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. https://doi.org/10.1016/j.cell.2011.11.055

Ventola GMM, Noviello TMR, D'Aniello S, Spagnuolo A, Ceccarelli M, Cerulo L (2017) Identification of long non-coding transcripts with feature selection: a comparative study. BMC Bioinformatics. https://doi.org/10.1186/s12859-017-1594-z

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res 41. https://doi.org/10.1093/nar/gkt006.

Wang Y, Fan X, Lin F, He G, Terzaghi W, Zhu D et al (2014) Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. Proc Natl Acad Sci 111:10359–10364. https://doi.org/10.1073/pnas.1409457111.

Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M et al (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun. https://doi.org/10.1038/s41467-018-03311-y

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.