Indian J. Genet., (2022); 82(3): 280-288
doi: 10.31742/ISGPB.82.3.2

ISSN: 0975-6906
www.ijgpb.com

RESEARCH ARTICLE

Identification of efficient learning classifiers for discrimination of coding and non-coding RNAs in plant species

Priyanka Guha Majumdar[1], A. R. Rao[2,*], Amit Kairi, P. K. Meher and Sarika Sahu

Abstract

Though the non-coding RNAs (ncRNAs) do not encode for proteins, they act as functional RNAs and regulate gene expression besides their involvement in disease-causing mechanisms and epigenetic mechanisms. Thus, discriminating ncRNAs from coding RNAs (cRNAs) is important in transcriptome studies. Several machine learning-based classifiers, including deep learning classifiers, have been employed for discriminating cRNAsfrom ncRNAs. However, the performance comparison of such classifiers in plant species is yet to be ascertained. Thus, in the present study, the performance of the classifiers such as Deep Neural Network (DNN), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were evaluated for classifying cRNAs and ncRNAsby using the datasets of plant species including crops such as rice, wheat, maize, cotton, sunflower, barley, banana, grape, papaya. Further, the performance of classifiers was assessed by following the cross-validation process as well as by considering an independent test data set of 3,997 cRNAs and 4,110 ncRNAs. The results revealed that Random Forest classifier exhibited highest performance accuracy (99.803%) among the machine learning classifiers, followed by DNN (99.519%), SVM (97.364%) and ANN (99.260%). The present study is expected to help computational and experimental biologists for easy discrimination between coding and non-coding RNAs.

Keywords: Coding RNAs, deep learning, machine learning, non-coding RNAs

Introduction

The RNAs are broadly classified into two classes, namely, coding RNAs (cRNAs) and non-coding RNAs (ncRNAs). Advances in the transcriptome studies reveal the presence of numerous types of ncRNAs such as lncRNAs, lincRNAs, circRNAs, piRNAs, SRPs, tmRNAs, Rnase P, Rnase MRP, besides commonly known ones like tRNA, rRNA, snRNA, snoRNA, miRNA in the transcriptome. Recent studies divulged that majority of the genomes of higher organisms are composed of ncRNAs (Davidson et al. 1977; Mattick and Gagen 2001; Shabalina et al. 2001). The ncRNAs perform various roles like gene expression regulation (Meister et al. 2004), mRNA splicing (Padgett 2001; Valadkhan 2005), RNA modification (Liang et al. 1995; Kiss et al. 2002; Falaleeva et al. 2016), chromatin modulation (Shevchenko 2018) and others. With the reports on various types of ncRNAs like lncRNAs, circRNAs, the distinction between the two classes have become more difficult as many of them are even longer than the coding RNAs, thereby mimicking cRNAs. Thus, accurate classification of cRNAs and ncRNAs need to be reassessed from the viewpoint of understanding the epigenetic mechanisms in phenotypic variation and evolution.

In the recent past, with the advent of Next Generation Sequencing (NGS) technologies, there was a surge in the availability of plant transcriptome data in the public domain (https://plants.ensembl.org/info/data/ftp/index.html). So, there is a need to analyse such huge amount of data and distinguish ncRNAs from cRNAs as well as to investigate their function and evolution. Since the wet-lab based experiments consume lot of resources both in terms of time and money for distinguishing cRNAs from ncRNAs, use of computational techniques could be a better alternative. Many computational and experimental strategies have been deployed so far to classify coding and non-coding RNAs. The CST miner (Castrignanò et al. 2004), an alignment-

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, [1]P.G. School, ICAR-IARI, New Delhi 110 012, India.

[2]Indian Council of Agricultural Research, New Delhi 110 001, India.

***Corresponding Author:** A. R. Rao, Indian Council of Agricultural Research, New Delhi 110 001, India., E-Mail: raocshl.word@gmail.com

How to cite this article: Majumdar P. G., Rao A. R., Kairi A., Meher P. K. and Sahu S. 2022. Identification of efficient learning classifiers for discrimination of coding and non-coding RNAs in plant species. Indian J. Genet. Plant Breed., **82**(3): 280-288.

Source of support: ICAR, New Delhi

Conflict of interest: None.

Received: Dec. 2021 **Revised:** May 2022 **Accepted:** June 2022

based method, uses cross-species genome comparison to identify coding and non-coding conserved sequence tags. Likewise, QRNA (Rivas and Eddy 2001) uses Pair Hidden Markov Models (PHMMs), and CRITICA (Badger and Olsen 1999) uses comparative sequence analysis for identifying coding regions. The EST scan (Iseli et al. 1999) employs HMM to identify coding sequences from ESTs, and the RNA code (Washietl et al. 2011) uses multiple sequence alignment to discriminate between coding and non-coding regions. Now with the availability of a large number of coding and non-coding RNA sequences, machine learning-based approaches have become more prevalent. Several *in silico* tools such as CONC (Liu et al. 2006), CNCI (Sun et al. 2013), CPC (Kong et al. 2007), PLEK (Li et al. 2014), RNAcon (Panwar et al. 2014), CPC2 (Kang et al. 2017) have been developed for discrimination of coding and non-coding RNAs by using the Support Vector Machines. Further, LncRNA-ID (Achawanantakun et al. 2015) and LncRNApred (Pian et al. 2016) use Random Forest for classification of ncRNAs and cRNAs. The PlncPRO (Khemka and Singh 2017) uses plant RNA sequences to train a Random Forest classifier to classify plant coding and long non-coding RNAs (lncRNAs). Of late, deep learning algorithms are being widely used for classification tasks, as deep learning yielded excellent results in speech recognition (Hinton et al. 2012), computer vision (Krizhevsky et al. 2012), and natural language processing (Mikolov et al. 2011). In this direction, DeepLNC (Tripathi et al. 2016) has been developed for classifying long non-coding RNAs (lncRNAs) from coding RNAs using Deep Neural Network (DNN). However, there is a need to assess the performance of various machine learning algorithms and deep learning algorithms for the classification of plant coding and ncRNAs. Here, in the present investigation, the performances of four learning classifiers, *viz*., Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Deep Neural Networks (DNN) were assessed and compared for identification of efficient classifiers for discriminating coding and non-coding RNAs in 63 plant species including important crops.

Materials and methods

Data collection and preparation

The coding and non-coding transcript sequences of 63 plant species, covering cereals, pulses, oilseeds, fruits, and forestry trees, were downloaded from the EnsemblePlants database (Bolser et al. 2016; https://plants.ensembl.org/info/data/ftp/index.html). Since the size of the datasets is large, representative sample of coding and non-coding sequences were drawn by keeping the respective proportion of sequences from each crop species intact. Then, the sequences with more than 80% similarity and ambiguous nucleotides were removed within each set of coding and non-coding sequences by using CD-HIT (http://weizhongli-lab.org/cd-hit/). Finally, a total of 19987 and 20550 sequences were found in coding and non-coding sets. To avoid the prediction bias towards the class having higher number of sequences, balanced dataset comprising of 15990 cRNAs and 16440 ncRNAs were used to train the model and 3997 and 4110 respective sequences of coding and non-coding RNAs were utilized as independent test set.

Feature extraction

The sequence data cannot be used as such as input for training and testing of machine learning classifiers. Thus, information from each sequence was converted into numeric form by feature extraction. A total of 1472 sequence features were generated (Table 1).

Table 1. Summary of the feature set used to map the input sequences into numeric feature vectors

S. No.	Feature Name	Description	#Feature
1	Transcript length	Length of the transcript sequences.	1
2	ORF length	Length of the longest putative open reading frame (ORF) among the 6 frames.	1
3	ORF coverage	Ratio of ORF length to transcript length.	1
4	Peptide length	Length of the protein sequence coded by the longest ORF.	1
5	K-mer frequencies	Frequencies of k-mers (k= 1, 2, 3, 4, 5).	1364
6	BLAST features	By considering Swiss-Prot protein sequences as database and transcripts as query sequence (i) number of blast hits, (ii) average length of alignment of the blast hits, (iii) average percent identity of the blast hits, (iv) average of e-values and (v) average of bit-scores of the blast hits.	5
7	Amino acid composition	Percentage of 20 amino acids in the putative protein sequences.	20
8	Molecular weight	Molecular weight of the putative proteins of the transcripts.	1
9	Isoelectric point	The theoretical isoelectric point (pI) values of the putative proteins.	1
10	GC%	Per cent Guanine and Cytosine content calculated with a PERL script.	1
11	Codon Bias Indices	Percentage of nucleotides (A, C, G, T) in third position of codons, Codon Adaptation Index (CAI), Codon Bias Index (CBI), number of synonymous codons, gravy, AROMO.	12
12	RSCU	Relative Synonymous Codon Usage for 64 codons.	64

For the extraction of features: transcript length, ORF length, ORF coverage, Peptide length, amino acid composition, molecular weight, isoelectric point and GC%, in-house perl scripts were developed. Whereas, for k-mer features, BLAST features, codon bias indices and RSCU, respective R-scripts were used. The transcripts for which no BLASTX hits were found, the values of number of hits, average length of alignment, average percent identity and average bit-score were treated as zero. Similarly, the corresponding e-values were given the values as (mode+1) of the e-values in the set. The features which were dependent on the length of the transcript sequences were normalized.

Feature Selection

Feature selection is an important aspect of model building to achieve maximum accuracy, as all the extracted features do not contribute to the model accuracy. In fact, some features may negatively affect the model accuracy. To select important features, two algorithms, namely, Random Forest (RF) (Breiman 1996, 2001; Boruta Kursa and Rudnicki 2010) were employed. In case of RF, variable importance measure: the prediction error on the out-of-bag data, is recorded for each variable before and after permuting the variable for each classification tree. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences. The more is the value of mean decrease in accuracy, the higher is the importance. Boruta (Kursa and Rudnicki 2010) algorithm is a variant of RF variable importance measure. It searches for a minimal subset of features from the dataset that are important in performance of the model with regard to classification accuracy, unlike ranking of features in RF algorithm. The Boruta algorithm works by adding another layer of randomness to the given data set by creating shuffled copies of all features, known as shadow features. At each iteration, Boruta check whether a real feature has higher importance than the best of its shadow features and constantly removes the irrelevant features. Finally, the algorithm stops either when all features get confirmed or rejected or reaches a specified limit of random forest runs. Here, the feature selection was done by considering results of RF and Boruta algorithms. The accepted features from both Boruta and RF together were finally considered for downstream analysis.

Prediction with machine learning and deep learning algorithms

The standard machine learning techniques, *viz.*, ANN (McCulloch and Pitts 1943), SVM (Boser et al. 1992), RF (Breiman 1 2001) and deep learning algorithm: DNN (Ivakhnenko 1967) were used for the classification of cRNAs and ncRNAs. The machine learning and deep learning techniques were implemented through respective packages of R-software. The trees.h2o.randomForest function of the "h2o" R-package was utilized to execute the random forest (RF) model. A total of 100 classification trees (*ntree*) were constructed and 21 features (*mtry*) were considered for each split of a node for each classification tree. The SVM is a modern classification technique and the performance of which depends upon the choice of kernel function. A highly referred R-package "e1071" was adopted to train the SVM model with Radial Basis Function (RBF) as the kernel function with default parametric values of cost and gamma. The parameters of the ANN classifiers are the number of units in the hidden layer, number of hidden layers and activation function. The "neuralnet" R-package was used for training the ANN. Single hidden layer with three units was considered here, where the activation function used was *tanh*.

In general, machine learning algorithms learn from the data representations that are to be selected carefully to capture the underlying intrinsic characters of the data as well as to map to the outputs correctly. A major problem in machine learning is the choice of data representations, also referred as "representation learning". Deep learning is a sub field of representation learning and it solves complex problems by representations that are expressed in terms of simpler representations. Deep learning algorithms (DLAs) learn complex concepts by decomposing them into simpler concepts. DLAs do not need feature engineering as they can learn the features and decide the features that effectively map the input to the output. DLAs are data driven techniques that need huge amount of data for learning and the parameters like number of hidden layers, number of units in each hidden layer, epoch, activation function, learning rate, input layer dropout ratio *etc.* are called hyper parameters. R package h2o was used to train the DNN classifier for binary classification with (i) three hidden layers, (ii) *tanh* activation function (iii) 200 epochs, (iv) learning rate 0.002, and (v) input drop out ratio 0.5.

k-fold cross-validation

Inferring the results from single training and testing sets is often biased (Rafaeilzadeh et al. 2009). Thus, *K*-fold cross validation technique was employed by splitting the data *K*-times and taking the averages of performance metrics over *K*-folds. A 10-fold cross validation was used where each dataset was divided into ten equal parts and each time nine parts were used to train the model and one part was kept for testing. The classification accuracy was obtained as an average over 10-folds.

The source code developed for execution of machine learning and deep learning algorithms along with k-fold cross validation are given in Supplementary Table S1: model-codes.R. The workflow showing the steps for assessing the performance of binary classifiers is given in Fig. 1.

Model testing and performance metrices

Performance metrics such as accuracy, sensitivity, specificity, precision, F-1 score and Matthews Correlation Coefficient (MCC) were measured to evaluate the performance of the

classifiers. The metrics are as follows:

$$\text{Accuracy} = (TP+TN)/(TP+FN+FP+TN)$$

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{Precision} = TP/(TP+FP)$$

F-1 score = 2*recall*precision/(recall+precision), where recall is same as sensitivity for the binary classification

MCC = [(TP*TN)-(FP*FN)]/sqrt[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]

Here, TP = True positive, TN = True negative, FP = False positive and FN = False negative

Results and discussion

A total of 15990 coding RNA and 16,440 non-coding RNA sequences were selected for training the classifiers, after removing sequences with more than 80% similarity and ambiguous nucleic acids in each of the classes. Further, a total of 3997 and 4110 sequences of coding and non-coding RNAs, respectively were set aside as independent test dataset to evaluate the performances of the trained models. To assess the level of similarity between the sequences of binary classes of the training dataset, CD-HIT-2D was used at percent similarity levels, *viz.*, 90, 80, 70, 60 (Table 2). Results revealed that 21.8% of the total training sequences share at least 60% between class similarity, which in turn ascertains the discriminating power of the feature sets to distinguish the two classes of RNAs by various classifiers. This ensures accurate classification of coding and non-coding RNAs by classifiers, even if there is a good amount of similarity exists at the sequence level.

A total of 1,472 features from an exhaustive search of literature were considered to discriminate coding and non-coding RNAs. Feature selection was done by assessing the importance of each feature by employing two different algorithms: Random Forest variable importance measure and Boruta. Random forest ranked each variable based on their decrease in mean accuracy whereas Boruta used mean importance, to rank the features. RF showed lowest decrease in mean accuracy when penta-mers were excluded from the feature set. A similar result has also been revealed by Boruta algorithm, which has labelled penta-mer features as "rejected". So based on these results from both RF and Boruta, penta-mer features were not used for training the classification models. Thus, it was found that mono-, di-, tri- and tetra-mers mostly contribute to the variations on genome as well as for binary classification. After discarding the penta-mer features, a total of 448 features were selected for training the models. The classifiers, namely, RF, SVM, ANN, DNN were trained with the same training dataset and with the selected features. In all the models, a 10-fold cross-validation was performed and the average accuracy over 10-folds for different classifiers are presented in Fig. 2.

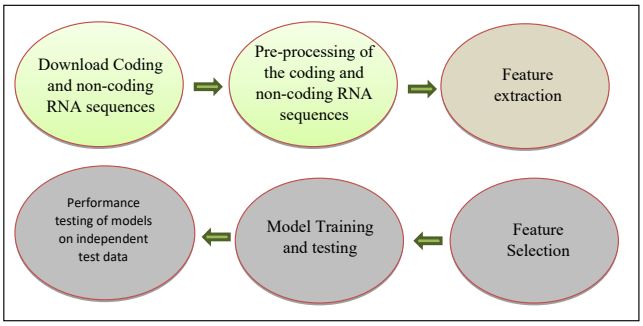

Fig. 1. Workflow for development of binary models

Table 2. Similarity between coding and non-coding classes

Between class similarity (%)	No. of similar sequences
90	28
80	86
70	1,169
60	7,064

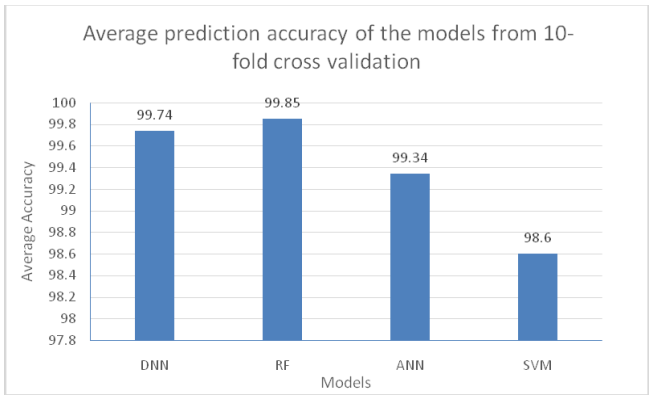

Fig. 2. Average prediction accuracy of the models from 10-fold cross validation accuracy

The average accuracy from 10-fold cross-validation of the classifiers revealed in Fig. 2 that RF showed highest accuracy (99.85%), followed by DNN (99.74%), ANN (99.34%) and SVM (98.60%). Liu et al. (2006) trained a SVM based classifier with cDNAs collected from GenBank, corresponding to all the eukaryotic protein sequences present in Swiss-Prot database, and ncRNAs from RNADB and NONCODE databases. After filtering out similar sequences, finally, used 5,610 and 2,670 sequences as coding and non-coding data sets and extracted 180 protein features as input for classification. They reported higher specificity and sensitivity of 97% and 98%, respectively for SVM classifier, built-in in the developed classifier CONC when compared to naïve Bayes classifier under 10-fold cross-validation. Kong et al. (2007) trained SVM classifier by using 5,610 coding cDNAs and 2,670 ncRNAs, that were used in training CONC tool, with 6 sequence features and reported the performance of 10-fold cross-validation accuracy as 95.77%.

Table 3. Comparison of classifiers based on performance metrics using independent test data

Method	Accuracy (%)	Sensitivity	Specificity	Precision	F1-score	MCC
DNN	99.519	0.994	0.996	0.996	0.995	0.990
RF	99.803	0.997	0.999	0.999	0.998	0.996
SVM	97.364	0.979	0.922	0.992	0.985	0.854
ANN	99.260	0.989	0.996	0.996	0.993	0.985

Table 4. Confusion matrix of the classifiers from independent test dataset

Method	TP	FP	FN	TN
RF	4098	4	12	3996
DNN	4087	16	23	3984
SVM	4077	33	88	392
ANN	4065	15	45	3985

Table 5. Comparison of the developed models with the existing models/tools

Method	Accuracy	Sensitivity	Specificity
RF (present approach)	99.80	99.70	99.90
LncRNA-ID (RF)	95.78	96.28	95.28
LncRNApred (RF) (mouse)	94.30	95.27	93.48
PLncPRO (RF)	83-99.5	-	-
SVM (present approach)	97.36	97.90	92.20
CPC1 (SVM)	93.20	99.50	87.30
CPC2 (SVM)	96.10	95.20	97.00
DNN (present approach)	99.52	99.40	99.60
DeepLNC (DNN)	98.07	98.98	97.19

The performance of the classifiers was also assessed by using the independent test data set of 8,110 sequences. 23.07% out of total 8110 sequences were having more than 60% similarity while comparing the sequences between the two classes. Performances of the classifiers were assessed based on the metrics like accuracy, sensitivity, specificity, precision, F1-score and MCC. Table 3 represents the comparative performance measures of the classifiers and Table 4 gives the confusion matrix based on which the performance metrics were calculated. It is evident from Tables 3 and 4 that RF has highest performance accuracy for classification of coding and non-coding RNAs. A comparison of the proposed approach involving RF with the existing models using similar techniques, in terms of performance metrics, is given in Table 5. It is observed that the proposed approaches: involving RF, SVM and DNN individually, have exhibited higher accuracy over the existing models which used similar machine learning techniques. A sample list of coding and non-coding RNAs of various plant species along with their identification number, probability of classification and predicted classes are shown in Table 6. The probability p_0 indicates the observation being classified as non-coding RNA and p_1 indicates the observation being classified as coding RNA (Table 6).

Kong et al. (2007) tested SVM based classifier, CPC, with 3 independent datasets *viz.*, Rfam and RNADB non-coding database and EMBL CDS, and reported the performance accuracies as 98.62%, 91.50% and 99.08%, respectively. Kong et al. (2017) used human 17,984 coding and 10,452 non-coding sequences, extracted 4 sequence intrinsic features by Random Forest to develop a SVM based tool called CPC2. CPC2 reported 96.1%, testing accuracy 97% specificity and 95.2% sensitivity. In the present study, the SVM based model exhibited 97.364% accuracy, 92.2% specificity and 97.9% sensitivity. The present findings in the results corroborate the earlier findings of Kong et al. (2007). In a similar way, Pian et al. (2016) trained a Random Forest classifier using 33,665 lncRNAs and 38,229 mRNA sequences that showed 97.38% accuracy as compared to SVM's 96.21% and ANN's 96.49% accuracy, which are in line with the findings of the present study where RF exhibited higher accuracy of 99.803% as compared to SVM (97.364%) and ANN (99.260%). Tripathi *et al.* (2016) classified lncRNAs from coding RNAs using deep neural network model by DeepLNC tool. They trained DeepLNC with 80,214 lncRNAs collected from LNCipedia and 99,395 coding sequences from RefSeq with 1104 k-mer features as model input. DeepLNC reported 98.07 % accuracy, sensitivity of 98.98 %, and specificity of 97.19 % and out-performed other tools based on machine learning algorithms like RF, SVM. On the contrary, the present results revealed better performance of RF over DNN (99.519%), SVM and ANN.

Deep Gene (Yuan et al. 2016) a deep learning-based cancer type classifier when compared with machine learning classifiers like SVM, showed higher accuracy of classification by at least 24%. In fields like drug discovery, DNN out-performed SVM (Korotcov et al. 2017). Their findings are similar to that of the results being reported from the present study, where in DNN has 99.519% accuracy and SVM has 97.364% accuracy.

Cho et al. (2018) developed classifiers for source tracking of chemical leaks reported that three classifiers trained with RF out-performed other six classifiers trained with DNN, based on accuracy and error. In the present study we also came out with similar result where RF performed better than DNN. Although, Cho et al. (2018) concluded that RF may not always out-perform DNN in leak detection, however, with the wide variety of different structural models of deep learning like CNN, RNN, Auto-encoders, the performance of the deep learning methods can be improved.

Table 6. Sample list of cRNAs ('1") and ncRNAs ("0") of different plant species along with their probabilities classification and predicted classes

Plant spp.	Classification coding RNA				
	Gene Id	Observed Class	P0	P1	Predicted class
Ostreococcus lucimarinus CCE9901 (green algae)	ABO99537	1	0.001	0.999	1
Aegilops tauschii subsp. *strangulata* (monocots)	AET7Gv20347800.3	1	0.085	0.915	1
Arabidopsis thaliana	AT1G56225.1	1	0.288	0.712	1
Arabidopsis thaliana	AT4G38401.1	1	0.629	0.371	0
Brassica oleracea var. *oleracea* (wild cabbage)	Bo1g082480.1	1	0.001	0.999	1
Brassica rapa	Bra037076.1	1	0.070	0.930	1
Chondrus crispus (carragheen)	CDF39906	1	0.011	0.989	1
Brassica napus	CDY65930	1	0.111	0.889	1
Dioscorea cayenensis subsp. *rotundata* (Guinea yam)	Dr00025.1	1	0.001	0.999	1
Galdieria sulphuraria (red algae)	EME27915	1	0.000	1.000	1
Theobroma cacao	EOX91046	1	0.011	0.989	1
Amborella trichopoda	ERM99382	1	0.001	0.999	1
Amborella trichopoda	ERN03221	1	0.052	0.948	1
Phaseolus vulgaris	ESW09909	1	0.022	0.978	1
Phaseolus vulgaris	ESW19603	1	0.110	0.890	1
Musa acuminata subsp. *malaccensis* (wild Malaysian banana)	GSMUA_Achr11T03940_001	1	0.031	0.969	1
Musa acuminata subsp. *malaccensis* (wild Malaysian banana)	GSMUA_Achr8T29260_001	1	0.560	0.440	0
Gossypium raimondii (eudicots)	KJB50375	1	0.021	0.979	1
Beta vulgaris subsp. *vulgaris* (Sugar beet)	KMT20152	1	0.060	0.940	1
Oryza longistaminata (monocots)	KN539468.1_FGT001	1	0.000	1.000	1
Vigna angularis (adzuki bean)	KOM56766	1	0.000	1.000	1
Brachypodium distachyon	KQJ99783	1	0.275	0.725	1
Setaria italica	KQK87785	1	0.001	0.999	1
Setaria italica	KQL27937	1	0.569	0.431	0
Glycine max	KRG92471	1	0.001	0.999	1
Glycine max	KRH76837	1	0.000	1.000	1
Sorghum bicolor	KXG39011	1	0.021	0.979	1
Daucus carota subsp. *sativus* (carrot)	KZM94941	1	0.001	0.999	1
Daucus carota subsp. *sativus* (carrot)	KZN11081	1	0.458	0.542	1
Manihot esculenta	OAY52937	1	0.022	0.978	1
Oryza brachyantha	OB07G28880.1	1	0.031	0.969	1
O. barthii	OBART03G05550.1	1	0.001	0.999	1
Oryza glumaepatula	OGLUM01G13690.1	1	0.001	0.999	1
Nicotiana attenuata (eudicots)	OIT40279	1	0.028	0.972	1
Lupinus angustifolius (narrow-leaved blue lupine)	OIV99006	1	0.200	0.800	1
Lupinus angustifolius (narrow-leaved blue lupine)	OIW04334	1	0.213	0.787	1
Oryza meridionalis	OMERI12G03230.1	1	0.001	0.999	1

(Contd. .....)

Corchorus capsularis (jute)	OMO53783	1	0.000	1.000	1
Corchorus capsularis (jute)	OMP00050	1	0.030	0.970	1
Prunus persica	ONI35610	1	0.257	0.743	1
Oryza nivara	ONIVA08G10880.1	1	0.115	0.885	1
Oryza punctata	OPUNC01G38110.1	1	0.101	0.899	1
Sorghum bicolor	OQU91703	1	0.001	0.999	1
Oryza glaberrima	ORGLA05G0029800.1	1	0.000	1.000	0
Oryza rufipogon (monocots)	ORUFI12G21890.1	1	0.560	0.440	1
Oryza sativa, *Japonica* Group (Japanese rice)	Os02t0759500-01	1	0.011	0.989	1
Helianthus annuus (common sunflower)	OTF90324	1	0.001	0.999	1
Helianthus annuus (common sunflower)	OTG28704	1	0.041	0.959	1
solanum tuberosum	PGSC0003DMT400096199	1	0.319	0.681	1
Populus trichocarpa (black cottonwood)	PNS96742	1	0.222	0.778	1
Populus trichocarpa (black cottonwood)	PNT48868	1	0.000	1.000	1
Chlamydomonas reinhardtii	PNW79620	1	0.080	0.920	1
Physcomitrium patens (mosses)	Pp3c20_10750V3.2	1	0.010	0.990	1
Glycine max	RCW18896	1	0.051	0.949	1
Triticum aestivum	TraesCS6 A02G182100.1	1	0.014	0.986	1
Triticum uratu	TRIUR3_16468-T1	1	0.050	0.950	1
Eragrostis curvula	TVU51707	1	0.067	0.933	1
Homo sapiens (human)	VIT_18s0001g02630.t01	1	0.041	0.959	1
Vigna radiculata	Vradi03g00460.1	1	0.001	0.999	1
Prunus dulcis (almond)	VVA40890	1	0.050	0.950	1
Zea mays	Zm00001d053935_T011	1	0.010	0.990	1
	Classification of ncRNAs				
Amborella trichopoda	AMTR_s00172t00274680	0	0.960	0.040	0
Amborella trichopoda	AMTR_s00222t00273490	0	0.920	0.080	0
Arabidopsis thaliana	AT1G03987.1	0	1.000	0.000	0
Arabidopsis thaliana	AT5G03985.1	0	0.493	0.507	1
Corchorus capsularis	CCACVL1_07549-1	0	0.960	0.040	0
Chondrus crispus	CHC_970-1	0	0.960	0.040	0
Beta vulgaris subsp. *vulgaris* (sugar beet)	ENSRNA049434646-T1	0	0.910	0.090	0
Os-Nipponbare	EPlORYSAT000373629	0	0.920	0.080	0
Daucus carota subsp. *sativus* (carrot)	EPlT00049782094	0	0.980	0.020	0
Daucus carota subsp. *sativus* (carrot)	EPlT00050810916	0	1.000	0.000	0
Daucus carota subsp. *sativus* (carrot)	EPlT00050864937	0	0.432	0.568	1
Daucus carota subsp. *sativus* (carrot)	EPlT00050875324	0	0.492	0.508	1
Galdieria sulphuraria (red algae)	Gasu_nc0086-1	0	0.900	0.100	0
Musa acuminata subsp. *malaccensis* (wild Malaysian banana)	GSMUA_Achr10T27833_001	0	0.920	0.080	0
Helianthus annuus (common sunflower)	HannXRQ_Chr12g0366031-1	0	1.000	0.000	0
Helianthus annuus (common sunflower)	HannXRQ_Chr12g0367421-1	0	0.970	0.030	0
Medicago truncatula (barrel medic)	MTR_4g025680	0	0.958	0.042	0
Malus domestica (apple)	ncRNA:MD06G0075000	0	1.000	0.000	0
Malus domestica (apple)	ncRNA:MD06G0107900	0	0.929	0.071	0
Theobroma cacao (cacao)	Tc01v2_t005220.1	0	0.980	0.020	0

(Contd. .....)

Theobroma cacao (cacao)	Tc04v2_t006820.1	0	0.990	0.010	0
Zea mays	Zm00001d000809_T001	0	0.860	0.140	0
Zea mays	Zm00001d022982_T001	0	0.860	0.140	0
Zea mays	Zm00001d023098_T001	0	0.445	0.555	1
Zea mays	Zm00001d027201_T001	0	0.990	0.010	0

In case of DeepLNC, DNN out-performed the machine learning models like RF, SVM, ANN but in the present study, we found that performance of DNN was next to RF with a negligible difference in performance. The reason could be that ncRNA data considered in our investigation has different types of ncRNAs with varying lengths and features. The present study also found that there remains a room for improvement in performance of deep learning models by increasing the training instances, different deep learning architecture and features. Form the present study, it is observed that for the classification of coding and non-coding RNAs of plants, RF can be used with highest accuracy followed by DNN, ANN, SVM classifiers.

Author's contribution

Conceptualization of research (PGM, ARR); Designing of the experiments (PGM,ARR); Contribution of experimental materials (Public domain data); Execution of lab experiments (PGM, ARR); Analysis of data and interpretation (PGM, ARR, AK, PKM, SS); Preparation of the manuscript (PGM, ARR, AK, PKM, SS).

Acknowledgments

The research is part of a Ph.D. Thesis of the first author submitted to Post-Graduate School, ICAR-IARI, New Delhi 110 012, India. The first author gratefully acknowledges the Indian Council of Agricultural Research for providing post-graduate scholarship during the period of study.

References

Achawanantakun R., Chen J., Sun Y. and Zhang Y. 2015. LncRNA-ID: Long non-coding RNA Identification using balanced random forests, Bioinformatics, **31**(24): 3897–3905.

Badger J. H. and Olsen G. J. 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol. Biol. Evol., **16**(4): 512-24.

Bolser D., Staines D. M., Pritchard E. and Kersey P. 2016. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. Methods Mol. Biol., **1374**: 115-40.

Boser B.E., Guyon I.M. and Vapnik V.N. 1992. A Training Algorithm for Optimal Margin Classifiers. Proc. 5th Annual Workshop on Computational Learning Theory (COLT'92), Pittsburgh, 27-29 July 1992, pp.144-152.

Breiman L. 1996. Bagging predictors. Machine Learning, **26**(2): 123–140.

Breiman L. 2001. Random Forests. Machine Learning, **45**: 5–32.

Castrignanò T., Canali A., Grillo G., Liuni S., Mignone F. and Pesole G. 2004. CSTminer: a web tool for the identification of coding and non-coding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res., 32 (Web Server issue): W624–W627.

Cho J., Kim H., Gebreselassie A. L. and Shin D. 2018. Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data. J. Loss Prevent. Process Indust., **56**: 548-558.

Davidson E.H., Klein W.H. and Britten R.J. 1977. Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript. Dev. Biol., **55**: 69–84.

Falaleeva M., Pages A., Matuszek Z., Hidmi S., Agranat-Tamir L., Korotkov K., et al. (2016). Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. PNAS, USA, **113**: E1625–34.

Hinton G., Deng L., Yu D., Mohamed A.R., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Dahl G. and Kingsbury B. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Signal Process Mag IEEE, **29**(6): 82–97.

Iseli C., Jongeneel C. V. and Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol., **1**: 38-48.

Ivakhnenko A. G. and Lapa V. G. 1967. Cybernetics and Forecasting Techniques. American Elsevier Publishing Co. ISBN 978-0-444-00020-0.

Kang Y.J., Yang D. C., Kong L., Hou M., Meng Y.Q., Wei L. and Gao G. 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res., **3**(45): W12-W16.

Khemka N. and Singh U. 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**: 10.

Kiss T. 2002. Small nucleolar RNAs: an abundant group of non-coding RNAs with diverse cellular functions. Cell, **109**: 145–8.

Kong L., Zhang Y., Ye Z.Q., Liu X.Q., Zhao S.Q., Wei L. and Gao G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res., **36**: W345-349.

Korotcov A., Tkachenko V., Russo D. P. and Ekins S. 2017. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. Mol. Pha., **14**(12): 4462-4475.

Krizhevsky A., Sutskever I. and Hinton G. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. Curran Associates, Inc., **25**: 1106–1114.

Kursa M. B. and Rudnicki W. R. 2010. Feature Selection with Boruta Package. J. Statistical Software, **36**: 1-13.

Li A., Zhang J. and 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**: 311.

Li W. and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, **22**(13): 1658–1659.

Liang W. Q. and Fournier M. J. 1995. U14 base-pairs with 18S rRNA: a novel snoRNA interaction required for rRNA processing. Genes Dev., **9**: 2433–43.

Liu J., Gough J. and Rost B. 2006. Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines. PLoS Genet., **2**(4): e29.

Mattick J. S. and Makunin I. V. 2006. Non-coding RNA. Human Molecular Genetics, **15**(1): R17–R29.

Mattick J.S. and Gagen M.J. 2001. The evolution of controlled multitasked gene networks: the role of introns and other non-coding RNAs in the development of complex organisms. Mol. Biol. Evol., **18**: 1611–1630.

McCulloch W. S. and Pitts W. 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Mathemat. Biophysics, **5**(4): 115-133.

Meister G. and Tuschl T. 2004. Mechanisms of gene silencing by double-stranded RNA. Nature, **431**: 343–349.

Mikolov T., Deoras A., Kombrink S., Burget L. and Cernock`y J. 2011. Empirical evaluation and combination of advanced language modeling techniques. INTERSPEECH. ISCA, 605–608.

Padgett R. A. 2001. mRNA Splicing: Role of snRNAs. Encyclopedia of Life Sciences, doi: 10.1038/npg.els.0000879

Panwar B., Arora A. and Raghava G. 2014. Prediction and classification of ncRNAs using structural information, BMC Genomics, **15**: 127.

Pian C., Zhang G., Chen Z., Chen Y., Zhang J., Yang T. and Zhang L. 2016. LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature. PloS One, **11**(5): e0154567.

Refaeilzadeh P., Tang L. and Liu H. 2009. Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Rivas E. and Eddy S.R. 2001. Non-coding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, **2**: 8.

Shabalina S.A., Ogurtsov A.Y., Kondrashov V.A. and Kondrashov A.S. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet., **17**: 373–376.

Shevchenko A. I., Grigor'eva E. V., Medvedev S. P., Zakharova I. S., Dementyeva E. V., Elisaphenko E. A., Malakhova A. A., Pavlova S. V. and Zakian S. M. 2018. Impact of Xist RNA on chromatin modifications and transcriptional silencing maintenance at different stages of imprinted X chromosome inactivation in vole Microtus levis. Chromosoma, **127**(1): 129-139.

Sun L., Luo H., Bu D., Zhao G., Yu K., Zhang C., Liu Y., Chen R. and Zhao Y. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res., **41**(17): e166.

Tripathi R., Kumari V., Patel S., Singh Y. and Varadwaj. 2016. DeepLNC, a tool for the prediction of Long-non coding RNAs using a Deep Neural Network,Netw Model Anal Health Inform Bioinforma, **5**: 21.

Valadkhan S. 2005. snRNAs as the catalysts of pre-mRNA splicing. Curr. Opin. Chem. Biol., **9**(6): 603-8.

Washietl S., Findeiss S., Müller S. A., Kalkhof S., Von Bergen M., Hofacker I. L., Stadler P. F. and Goldman N. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA, **17**(4): 578-94.

Yuan Y., Shi Y., Li C., Kim J., Cai W., Han Z. and Feng D. D. 2016. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics, **17**: 476.