RNAmining: A machine learning stand-alone and web server tool for RNA coding potential prediction

Thaís A. R. Ramos^{1,2,3}, Nilbson R. O. Galindo², Raúl Arias-Carrasco³, Cecília F. da Silva², Vinicius Maracaja-Coutinho^{1,3,4,*}, Thaís G. do Rêgo^{1,2,*}

¹ Programa de Pós-Graduação em Bioinformática, Bioinformatics Multidisciplinary Environment (BioME), Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Brazil.

² Departamento de Informática, Centro de Informática, Universidade Federal da Paraíba, João Pessoa, Brazil.

³ Advanced Center for Chronic Diseases (ACCDiS), Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago, Chile.

⁴ Instituto Vandique, João Pessoa, Brazil.

* To whom correspondence should be addressed. Tel: +55 83 3216-7506; Email: gaudenciothais@gmail.com . Correspondence may also be addressed to Vinicius Maracaja-Coutinho. Tel: +56 2 2978 2826; Email: vinicius.maracaja@uchile.cl .

ABSTRACT

Non-coding RNAs (ncRNAs) are important players in the cellular regulation of organisms from different kingdoms. One of the key steps in ncRNAs research is the ability to distinguish coding/non-coding sequences. We applied 7 machine learning algorithms (Naive Bayes, SVM, KNN, Random Forest, XGBoost, ANN and DL) through 15 model organisms from different evolutionary branches. Then, we created a stand-alone and web server tool (RNAmining) to distinguish coding and non-coding sequences, selecting the algorithm with the best performance (XGBoost). Firstly, we used coding/non-coding sequences downloaded from Ensembl (April 14th, 2020). Then, coding/non-coding sequences were balanced, had their tri-nucleotides counts analysed and we performed a normalization by the sequence length. Thus, in total we built 180 models. All the machine learning algorithms tests were performed using 10-folds cross-validation and we selected the algorithm with the best results (XGBoost) to implement at RNAmining. Best F1-scores ranged from 97.56% to 99.57% depending on the organism. Moreover, we produced a benchmarking with other tools already in literature (CPAT, CPC2, RNAcon and Transdecoder) and our results outperformed them, opening opportunities for the development of RNAmining, which is freely available at <https://rna mining.integrativebioinformatics.me/> .

INTRODUCTION

Non-coding RNAs (ncRNAs) are key functional players on different biological processes in organisms from all domains of life (1,2). Its investigation is already routine in almost every transcriptome or genome project. Dysregulations in these molecules may lead to different types of human disease, including cancers (3), neurological disorders (4) and cardiovascular infirmities (5).

The genome of eukaryotic (6) organisms is, in general, composed majoritary by non-coding transcripts, with complex organisms estimated to transcribe more than 75% of their genomes (7). Besides strong evidence associating these ncRNAs to key functions in the cell, their majority are not

yet associated with a functional mechanism. In a transcriptome project exists an important step in the computational identification of ncRNAs, which is the evaluation of their potential to be translated into proteins using different bioinformatics approaches (8,9). To computationally evaluate the coding potential of a set of transcripts, available tools or algorithms normally analyse specific characteristics available in its primary sequences (e.g. nucleotides counts, the existence of a trustful open reading frame).

For instance, RNAcon implements a SVM-based model for the discrimination between coding and non-coding sequences (10). Coding Potential Assessment Tool (CPAT) (11) assesses the coding potential through an alignment-free method, which uses a logistic regression model built based on different characteristics of the sequence open reading frame (ORF), which includes length, coverage and nucleotides compositional bias. TransDecoder identifies candidate coding transcripts based on other distinctive features from predicted ORFs (e.g. a minimum length ORF, a log-likelihood score, encapsulated ORF) (12). CPC2 (13) trained a support vector machine (SVM) model using Fickett TESTCODE score, open reading frame (ORF) length, ORF integrity and isoelectric point as features. The LIBSVM (14) package was employed by training a SVM model using the standard radial basis function kernel (RBF kernel) with the training dataset containing 17,984 high-confident human protein-coding transcripts and 10,452 non-coding transcripts (11).

Here, we applied and benchmarked seven different machine learning algorithms (Random Forest, Gradient boosting (XGBoost), Naive Bayes, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Deep Learning (DL)) through 15 organisms from different evolutionary branches, in order to evaluate their performance in distinguishing coding and non-coding RNA sequences. Next, we developed a stand-alone and web server tool, called RNAmining (<http://rnaming.integrativebioinformatics.me/>), by selecting and implementing the algorithm with the best performance in all organisms (XGBoost). RNAmining was evaluated through 24 organisms from the eukaryotic tree of life and its results outperformed public available tools commonly used for that purpose.

MATERIAL AND METHODS

Machine learning classifier algorithms selection

In the classification process there is a division related to the learning paradigm, with classification algorithms divided into: (i) *Symbolic*, which seeks to learn by constructing symbolic representations of a concept through the analysis of examples and counterexamples (e.g. Decision Trees and Rule-based System); (ii) *Statistical*, which looks for statistical methods and use models to find a good approximation of the induced concept (e.g. Bayesian learning); (iii) *Based on Examples* (lazy systems), which aims to classify examples never seen using similar known examples, assuming that the new example will belong to the same class as the similar example (e.g. K-Nearest neighbor); (iv) *Based on Optimization*, which consists of maximizing (or minimizing) an objective function or finding an optimal hyperplane that best divides two classes (e.g. Support Vector Machine (SVM) and Neural Networks); (v) *Connectionist Representation*, which represents simplified mathematical constructions

inspired by the biological model of the nervous system (e.g. Neural Networks). In this benchmarking, we decided to evaluate the performance of selected algorithms from each paradigm type in the coding potential prediction of RNA sequences: Random Forest, XGBoost, Naive Bayes, K-NN, SVM and Neural Networks (Artificial Neural Networks and Convolutional Neural Networks (CNN)).

All the machine learning methods were executed using scikit-learn (Version 0.21.3) (15), except for Neural Network and Deep Learning models which were implemented using Keras API with Tensorflow as backend (Version 2.3.0) (<https://github.com/keras-team/keras>) and XGBoost algorithm which was executed using XGBoost Library (version 1.2.0) (16) in Python Language (Version 3.8). The Random Forest model was implemented using empirical tests and the best result was selected for training the model. We considered the default parameters with the exception of the number of trees used (150 estimators) and the criterion parameter setted to 'entropy' for information gain. KNN and Naive Bayes models were trained with the default values. The SVM parameters were obtained through grid search method and the resulting model was trained with the Radial Basis Function (RBF) kernel, with the Regularization parameter (C) and Kernel coefficient (Gamma) defined in 1000 and 0.8, respectively. Artificial Neural Networks and Deep Learning were performed with different architectures according to grid search and empirical tests. The first ANN experiment was composed of three hidden layers consisting of 32-16-8 neurons, respectively; the second ANN experiment was performed with 64-32-16-8 neurons; and the third experiment was executed with 32-32-16-8 neurons. Next, we produced four experiments with Deep Learning using 2 CNN layers, followed by 2 fully connected (dense) layers: the first experiment had 512(CNN)-512(CNN) filters and 28(Dense)-1(Dense) neurons; the second was created with 64(CNN)-64(CNN) filters and 128(Dense)-1(Dense) neurons; the third was performed with 32(CNN)-32(CNN)-128(Dense)-1(Dense) neurons; and the last was built with 128(CNN)-128(CNN)-128(Dense)-1(Dense) neurons. These layers received as input the total number of attributes (i.e. combination of tri-nucleotides counts, described in the next topics). The hyperparameters used to execute the DL and ANN approaches are made available in Supplementary File S1.

Datasets selection and filtering criteria

We compared the algorithms performances using different sets of coding and non-coding RNA sequences from Ensembl (April 14th 2020) (17) database, covering 15 organisms of distinct representative Chordata clades (Figure 1A): *Anolis carolinensis* (Sauria, Squamata), *Chrysemys picta bellii* (Sauria, Testudines), *Crocodylus porosus* (Archosauria, Pseudosuchia), *Danio rerio* (Actinopterygii, Teleostei), *Eptatretus burgeri* (Agnatha, Myxiniidae), *Gallus gallus* (Archosauria, Theropoda), *Homo sapiens* (Placentalia), *Latimeria chalumnae* (Sarcopterygii, Coelacanth), *Monodelphis domestica* (Marsupialia), *Mus musculus* (Placentalia), *Notechis scutatus* (Sauria, Squamata), *Ornithorhynchus anatinus* (Monotremata), *Petromyzon marinus* (Agnatha, Petromyzontiformes), *Sphenodon punctatus* (Sauria, Rhynchocephalia), *Xenopus tropicalis* (Amphibia). All non-coding RNA sequences for each organism were downloaded from Ensembl transcripts. In order to obtain a balanced set of sequences (i.e. equal number of coding and non-coding), the group of coding RNAs were randomly selected in order to obtain the same number of

ncRNAs for each species. Moreover, before generating the models, the sequences were normalized through their length (*i.e.* each tri-nucleotide count was divided by the total size of the given sequence). All sequences in FASTA format can be retrieved at RNAMining website (<https://rnamining.integrativebioinformatics.me/download>).

Training and testing datasets, model building and quality measuring for coding potential evaluation

Sequences were randomly divided into training and testing datasets, using 80% of the data for training and 20% for testing. For ANN and CNN experiments, sequences were splitted into 60% of the data for training and 20% for validation. The testing dataset was the same used in the other machine learning algorithms. The number of sequences used for each organism for the training and test sets can be observed in Table 1. Next, we generated 180 models (*i.e.* one per algorithm for each organism, whereas three experiments for ANN models and four experiments for CNN models), which were further evaluated in this work.

After selection of the best model, it was applied and evaluated in other nine organisms (Figure 1A), including five Chordata and four phylogenetically distant species. Among the chordates, the models were tested in *Carassius auratus* (Actinopterygii, Teleostei), *Gorilla gorilla gorilla* (Placentalia), *Pseudonaja textilis* (Sauria, Squamata), *Rattus norvegicus* (Placentalia) and *Terrapene carolina triunguis* (Sauria, Testudines). Within non-chordates species, we evaluated the model in *Arabidopsis thaliana* (Plantae, Eudicots), *Caenorhabditis elegans* (Nematoda), *Drosophila melanogaster* (Insecta, Diptera) and *Saccharomyces cerevisiae* (Fungi, Ascomycota). Finally, it was evaluated using artificial sequences containing the same nucleotides composition of the ncRNAs for each species of the testing dataset (Table 1). Ten sets of random sequences containing the same number of ncRNAs per species were generated using MEME suite Version 5.1.1 with default parameters (18). All sequences in FASTA format can be retrieved at RNAMining website (<https://rnamining.integrativebioinformatics.me/download>).

Table 1. Set of sequences used in the training and testing datasets. List of organisms and the total number of sequences used for testing and training both coding and non-coding RNAs. The numbers are separated into training / testing values. All sequences can be retrieved at RNAmMining website (<https://rnaminig.integrativebioinformatics.me/download>).

Species	Total	Coding	ncRNAs
Models Generation (training / testing):			
<i>Anolis carolinensis</i>	12,542 / 3,136	6,243 / 1,596	6,299 / 1,540
<i>Chrysemys picta bellii</i>	11,260 / 2,816	5,626 / 1,412	5,634 / 1,404
<i>Crocodylus porosus</i>	7,388 / 1,848	3,700 / 918	3,688 / 930
<i>Danio rerio</i>	12,984 / 3,246	6,527 / 1,588	6,457 / 1,658
<i>Eptatretus burgeri</i>	1,742 / 436	867 / 222	875 / 214
<i>Gallus gallus</i>	16,851 / 4,213	8,426 / 2,106	8,425 / 2107
<i>Homo sapiens</i>	92,844 / 23,212	46,575 / 11,453	46,269 / 11,759
<i>Latimeria chalumnae</i>	4,668 / 1,168	2,344 / 574	2,324 / 594
<i>Monodelphis domestica</i>	34,336 / 8,584	17,113 / 4,347	17,223 / 4,237
<i>Mus musculus</i>	35,272 / 8,818	17,668 / 4,377	17,604 / 4,441
<i>Notechis scutatus</i>	2,705 / 677	1,351 / 340	1,354 / 337
<i>Ornithorhynchus anatinus</i>	12,604 / 3,152	6,280 / 1,598	6,324 / 1,554
<i>Petromyzon marinus</i>	4,243 / 1,061	2,107 / 545	2,136 / 516
<i>Sphenodon punctatus</i>	1,456 / 364	723 / 187	733 / 177
<i>Xenopus tropicalis</i>	2,224 / 556	1,120 / 270	1,104 / 286
RNAmMining Evaluation:			
<i>Arabidopsis thaliana</i>	11,308	5,654	5,654
<i>Caenorhabditis elegans</i>	50,558	25,279	25,279
<i>Carassius auratus</i>	15,004	7,502	7,502
<i>Drosophila melanogaster</i>	31,808	15,904	15,904
<i>Gorilla gorilla gorilla</i>	15,978	7,989	7,989
<i>Pseudonaja textilis</i>	1,486	743	743
<i>Rattus norvegicus</i>	18,662	9,331	9,331
<i>Saccharomyces cerevisiae</i>	848	424	424
<i>Terrapene carolina triunguis</i>	2,054	1,027	1,027

Figure 1. A. Taxonomic tree according to the used organisms for the models building (black color) and validation (red color). B. Pipeline used to perform the benchmarking and create the tool. Firstly, we download the coding and non-coding sequences from Ensembl; Next, we perform the trinucleotides counts and sequence normalization. After this, we created a machine learning

benchmarking within the 7 algorithms and selected the one with the best performance to be implemented in the RNAMining tool (XGBoost algorithm), which was again evaluated using sequences from 9 other different species and sets of artificially generated ones. Finally, we performed a novel benchmarking with RNAMining against the public available tools for coding potential prediction.

Comparisons with publicly available tools

The performance of all algorithms in the coding potential evaluation was compared with publicly available tools commonly employed for this purpose (RNAcon (10), CPAT (11), Transdecoder (12) and CPC2 (13)), using default parameters. It is worth to note that CPAT only made available models for *H. sapiens* with a coding probability (CP) cutoff of 0.364 (*i.e.* CP $\geq$ 0.364 indicates coding sequence); *M. musculus* with a CP cutoff of 0.44; *D. melanogaster* with a CP cutoff of 0.39; and *D. rerio* with CP cutoff 0.38. Therefore, for the other organisms we built new models using our training sets and we used the statistical method provided by the authors to calculate the cutoffs probability for coding prediction: *A. carolinensis* (0.4); *C. picta bellii* (0.57); *C. porosus* (0.38); *E. burgeri* (0.35); *G. gallus* (0.42); *L. chalumnae* (0.365); *M. domestica* (0.51); *N. scutatus* (0.15); *O. anatinus* (0.28); *P. marinus* (0.34); *S. punctatus* (0.18); *X. tropicalis* (0.25). The whole workflow of RNAMining development can be visualized in Figure 1B.

Comparisons with publicly available tools

The XGBoost method was implemented using XGBoost Library (version 1.2.0) in Python Language (Version 3.8) and the models for each species were saved using pickle Python's library. The web server interface was developed using HTML and CSS. The connection within the front and back-end was implemented through Javascript. The control of files and the connection with Python's scripts was performed through PHP language. RNAMining user friendly tool and its stand-alone version can be accessed at <https://rnamining.integrativebioinformatics.me/>. Instructions on how to use it and a whole documentation are made available. Its source code with a Docker platform can be freely obtained at <https://gitlab.com/integrativebioinformatics/RNAMining>.

RESULTS

Using machine learning algorithms to improve the coding potential prediction of RNA sequences

It is known that the algorithms performance in predictive analysis is influenced by particularities available in the genomes sequences of the organisms used in the training set (19), and it should be taken into account when developing novel tools for nucleotides coding prediction. Thus, it is necessary to test several methods to observe which ones can have a good prediction for specific species from evolutionary branches. Similar to Panwar *et al.* (10), we used the tri-nucleotides counts to distinguish coding and non-coding sequences. We evaluated the performance of seven machine learning algorithms using representative organisms from different different branches of the Chordata

clade. For that, we used a training and testing set composed by sequences from the same species. The algorithm with best performance within all evaluated organisms, according to F1-scores metric, was XGBoost, as one can see in the following: *A. carolinensis* (98.79); *C. picta bellii* (98.00); *C. porosus* (98.15); *D. rerio* (97.98); *E. burgeri* (97.56); *G. gallus* (99.24); *H. sapiens* (98.50); *L. chalumnae* (99.57); *M. domestica* (98.84); *M. musculus* (97.73); *N. scutatus* (96.51); *O. anatinus* (97.61); *P. marinus* (99.42); *S. punctatus* (99.20); *X. tropicalis* (99.13) (Table 2). As observed, XGBoost algorithm presented F-score values above 97.00, with the worst performance obtained for *Eptatretus burgeri* with a F-score of 97.56. The best performance was obtained for *Petromyzon marinus* with 99.42. All detailed performances with sensitivity, specificity, precision, accuracy, F1-score and the confusion matrix from each algorithm is listed in Supplementary File S2. Based on these results, XGBoost was selected to be implemented in a novel web server and stand-alone tool for RNA coding potential prediction called RNAmMining.

Table 2. Benchmarking machine learning methods for coding potential prediction based on tri-nucleotides counts. F1-score for each one of the 15 species in which the algorithms were tested. Other metrics (sensitivity, specificity, precision, accuracy and the confusion matrix) used for the comparison of the algorithm's performance were made available at the Supplementary File S2.

Species	ANN	CNN	KNN	NAIVE	RANDOM	SVM	XGBoost
				BAYES	FOREST		
<i>Anolis carolinensis</i>	98.47	98.31	93.55	95.50	98.30	98.03	98.79
<i>Chrysemys picta bellii</i>	96.54	96.02	93.54	93.13	96.89	96.04	98.00
<i>Crocodylus porosus</i>	96.74	96.48	93.67	93.93	97.26	96.35	98.15
<i>Danio rerio</i>	97.54	97.77	95.44	94.55	97.56	97.27	97.98
<i>Eptatretus burgeri</i>	94.88	95.69	92.24	94.57	97.35	95.82	97.56
<i>Gallus gallus</i>	98.47	98.27	96.87	95.11	98.91	98.06	99.24
<i>Homo sapiens</i>	98.01	97.66	96.63	86.00	98.30	96.83	98.50
<i>Latimeria chalumnae</i>	99.05	98.72	91.61	98.23	99.56	99.24	99.57
<i>Monodelphis domestica</i>	98.39	98.09	97.11	95.31	98.67	98.01	98.84
<i>Mus musculus</i>	96.67	96.96	95.95	91.56	97.66	96.10	97.73
<i>Notechis scutatus</i>	95.90	94.10	87.77	89.81	94.94	95.73	96.51
<i>Ornithorhynchus anatinus</i>	97.23	96.59	93.59	91.45	96.99	96.38	97.61
<i>Petromyzon marinus</i>	98.40	98.26	88.10	95.99	98.79	97.49	99.42
<i>Sphenodon punctatus</i>	97.83	96.97	78.41	96.70	96.46	95.29	99.20
<i>Xenopus tropicalis</i>	98.28	98.81	85.53	97.14	98.88	97.20	99.13

Using RNAmMining in evolutionary related and unrelated organisms

To demonstrate the generalization of the model built in our tool, we evaluated its performance using the following nine Chordata and non-Chordata organisms that were not used in our training step: *A. thaliana*; *C. elegans*; *C. auratus*; *D. melanogaster*; *G. gorilla gorilla*; *P. textilis*; *R. norvegicus*; *S. cerevisiae*; *Terrapene carolina triunguis*. In the training set described in the previous topic, we used sequences from representative species from amphibians, birds, mammals, fishes and reptiles. In this

new experiment we executed tests using other chordates, but also covering other evolutionary groups such as plants, fungi, insects and nematodes. The F1-score obtained values varying from 86.25 to 98.10. The worst performance was when we used the training set from *L. chalumnae* (Sarcopterygii, Coelacanth) to predict the coding potential of known coding genes and ncRNAs from *D. melanogaster* (Insecta, Diptera). However, the best performance was obtained when we applied the training set from *C. picta bellii* (Sauria, Testudines) in coding and ncRNA sequences from *Terrapene carolina triunguis* (Sauria, Testudines). The F1-score for each organism, together with the respective training set evaluated, can be found in Table 3, meanwhile the confusion matrix and the other metrics can be visualized in Supplementary File S3.

Even without using any plant in the original training set, we applied the different models to predict the coding potential of known coding and ncRNA sequences from *A. thaliana* (Plantae, Eudicots). The lowest F1-score that RNAmMining obtained was 93.31 using a fish model (*Petromyzon marinus*, Agnatha, Petromyzontiformes). The best F1-score was obtained with a marsupial model (*M. domestica*, Marsupialia) that reached 97.40. Thus, this experiment demonstrated the efficiency of the method and the models created even when applied in organisms phylogenetically distant from those used in training.

Table 3. Evaluation (F1-score) of the models generated by XGBoost, the method implemented in RNAmMining, according to evolutionary related and unrelated organisms. Each line comprises the model for each one of the trained species, meanwhile the columns represent the set of 9 evolutionary related and unrelated organisms in which the method was evaluated. Other metrics (sensitivity, specificity, precision, accuracy and the confusion matrix) used for the comparisons were made available at the Supplementary File S3.

Testing Training	Arabidopsis thaliana	Caenorhabditis elegans	Carassius auratus	Drosophila melanogaster	Gorilla gorilla	Pseudonaja textilis	Rattus norvegicus	Saccharomyces cerevisiae	Terrapene carolina triunguis
<i>Anolis carolinensis</i>	95.35	89.97	94.77	97.16	95.17	96.56	96.74	93.07	95.83
<i>Chrysemys picta bellii</i>	97.24	97.79	95.97	98.13	97.01	97.73	97.15	96.09	98.10
<i>Crocodylus porosus</i>	96.19	96.76	95.73	97.87	97.01	96.90	97.25	95.07	97.56
<i>Danio rerio</i>	96.64	90.50	95.29	97.96	97.24	96.89	96.42	93.96	96.62
<i>Eptatretus burgeri</i>	94.90	95.57	94.80	96.73	95.34	95.43	95.76	91.49	95.51
<i>Gallus gallus</i>	97.60	97.89	95.76	98.02	97.93	97.79	97.59	96.48	97.69
<i>Homo sapiens</i>	95.71	81.25	92.19	96.44	97.73	96.24	94.60	93.57	95.65
<i>Latimeria chalumnae</i>	93.71	96.78	91.63	86.25	96.30	93.39	94.37	95.47	95.63
<i>Monodelphis domestica</i>	97.40	97.91	95.69	98.04	97.90	97.53	97.46	93.54	97.31
<i>Mus musculus</i>	96.44	87.68	94.66	97.17	97.57	97.31	96.67	94.32	96.30
<i>Notechis scutatus</i>	97.16	97.54	95.22	97.46	97.35	97.37	96.79	94.96	97.22
<i>Ornithorhynchus anatinus</i>	97.39	97.48	95.39	87.74	97.32	97.86	97.29	94.67	97.53
<i>Petromyzon marinus</i>	93.31	94.48	92.07	87.74	95.81	93.47	94.72	92.48	95.56
<i>Sphenodon punctatus</i>	94.00	97.07	91.94	86.89	96.60	93.95	94.12	95.02	95.81
<i>Xenopus tropicalis</i>	93.46	96.65	91.53	84.86	95.51	93.68	93.16	94.42	95.02

Finally, in order to show that the results obtained were not by chance, we created 10 datasets of artificial sequences containing the same number, length and nucleotides composition of the coding and ncRNA sequences from the 15 species used in our testing shown in Table 1. The F1-score mean, minimum and maximum values of the 10 datasets from each organism can be visualized in Table 5.

The confusion matrix and all the other metrics (accuracy, specificity, sensitivity and precision) can be found in Supplementary File S4. As we can visualize, the F1 measurement mean remained below 38.00 for all artificial sequences created for the tested organisms, with the exception of *P. marinus* (F1-score equals to 64.13), which still had a F1-score below to the values obtained with the other organisms tested for the coding potential prediction (Table 4).

Table 4. Evaluation of RNAmMining performance according to different sets of artificial sequences from each trained model. F1-score metrics for 10 datasets of artificial sequences randomly generated for each species. The mean, minimum and maximum values are displayed separated by organism. Other metrics (sensitivity, specificity, precision, accuracy and the confusion matrix) used for the comparisons were made available at the Supplementary File S4.

Species	MEAN	MINIMUM	MAXIMUM
<i>Anolis carolinensis</i>	1.66	0.86	2.44
<i>Chrysemys picta bellii</i>	1.08	0.70	1.40
<i>Crocodylus porosus</i>	0.95	0.43	1.72
<i>Danio rerio</i>	1.25	0.12	2.21
<i>Eptatretus burgeri</i>	2.31	0.90	3.51
<i>Gallus gallus</i>	2.48	1.88	2.89
<i>Homo sapiens</i>	11.15	10.53	11.52
<i>Latimeria chalumnae</i>	24.86	21.95	27.03
<i>Monodelphis domestica</i>	1.34	1.00	1.18
<i>Mus musculus</i>	6.64	5.74	7.58
<i>Notechis scutatus</i>	1.80	0.58	3.99
<i>Ornithorhynchus anatinus</i>	3.62	2.67	5.04
<i>Petromyzon marinus</i>	64.13	62.99	65.76
<i>Sphenodon punctatus</i>	37.43	31.72	41.84
<i>Xenopus tropicalis</i>	23.26	17.65	28.21

Comparing RNAmMining performance with publicly available tools

Next, we compared RNAmMining performance with other four tools commonly used for nucleotides coding potential prediction: CPAT, CPC2, RNAcon and Transdecoder. We used as input all coding and ncRNA sequences from the testing dataset used in the 15 species listed in Table 1. According to the F1-score metric, RNAmMining outperformed all the tools in all organisms with the exception of CPAT for *L. chalumnae*, in which both tools presented an equal F1-score of 99.57. The comparative performance of all tools can be observed in Table 5. The detailed results regarding their accuracy, sensitivity, specificity, precision, F1-score and the confusion matrix can be found in Supplementary File S2. Finally, we used the t-student test to compare the results from RNAmMining and the other tools, revealing that our software presented significantly better results in performing coding potential predictions based on known coding genes and ncRNAs. The p-values obtained in these comparisons were: 0.0026 (vs CPAT); 1.57e-05 (vs CPC2); 2.69e-05 (vs RNAcon); and 2.89e-05 (vs Transdecoder).

Table 5. Benchmarking results from RNAmMining and the other tools already described in the literature according to organisms from different evolutionary branches. F1-score metric for CPAT, CPC2, RNAcon, Transdecoder and RNAmMining, based on the predictions using models provided by each tool or generated according to their instructions. The bold numbers are the best values regarding F1-score metric. The results for other metrics were made available at the Supplementary File S2.

Species	CPAT	CPC2	RNAcon	Transdecoder	RNAmMining
<i>Anolis carolinensis</i>	94.55	86.87	83.03	88.26	98.79
<i>Chrysemys picta bellii</i>	92.56	89.01	82.36	84.80	98.00
<i>Crocodylus porosus</i>	94.07	92.48	84.32	87.63	98.15
<i>Danio rerio</i>	94.64	87.17	80.97	87.74	97.98
<i>Eptatretus burgeri</i>	95.59	78.82	75.84	76.26	97.56
<i>Gallus gallus</i>	96.95	90.69	75.81	83.50	99.24
<i>Homo sapiens</i>	95.20	75.85	71.73	76.02	98.50
<i>Latimeria chalumnae</i>	99.57	91.60	97.45	98.86	99.57
<i>Monodelphis domestica</i>	96.24	91.44	80.90	85.22	98.84
<i>Mus musculus</i>	95.48	81.40	76.78	80.80	97.73
<i>Notechis scutatus</i>	85.19	86.29	84.83	83.44	96.51
<i>Ornithorhynchus anatinus</i>	87.47	72.04	84.73	84.63	97.61
<i>Petromyzon marinus</i>	96.59	75.14	95.11	96.68	99.42
<i>Sphenodon punctatus</i>	97.61	91.91	97.86	95.24	99.20
<i>Xenopus tropicalis</i>	99.07	97.92	98.70	97.77	99.13

RNAmMining stand-alone and web server tool

RNAmMining tool was made available in both stand-alone and web server versions. The tools only require the nucleotide sequences of the RNAs in which the user intends to perform the coding potential prediction in FASTA format, together with the species name in a standardized format related to the model to be used. Besides our tool presented good results even when using phylogenetically distant organisms, we recommend to always use the most closely-related species to the one the user wants to perform the predictions. Furthermore, RNAmMining documentation presents all the guidelines on how to generate a model for a particular set of sequences and organisms of interest. The web interface of RNAmMining tool was developed to allow users to quickly perform the coding potential prediction without the need of installing any specific program and using only a generic internet browser. The only requirement for running the tool is a FASTA file containing the nucleotide sequences and the organism model that the user wants to use, which can be selected in a drop-down menu containing all 15 organisms used in the training step (Figure 2A). There is no limit of the number of sequences, but the web server supports files up to 20Mb. For files bigger than that, we recommend to use the stand-alone RNAmMining tool. RNAmMining will automatically classify the FASTA sequences used as input and identify which of them are coding or non-coding RNAs. Finally, as a result it offers a table with the sequences' IDs and its classification as coding or non-coding, which

can also be downloaded in tabular format, together with two separate FASTA files containing both the coding and non-coding sequences separately (Figure 2B).

RunRnaming

[Run](#)
[About](#)
[Tutorial](#)
[Download](#)
[Contact](#)

Upload your file

Dataset

.fasta - examples

```

>RNA_Sequence_1
GTCTCCCTTAGAGTTCCCTTGACCACTTCACGTGGGACCTTCTCAATTATAATGACTTC
CTACTGAAGTGTGTTGGGGGAACCTCTGGTGCCAT
>RNA_Sequence_2
ATCGCTTCTCGGCTCTTTTGGCTAAGATCAAGGTGAGGAACAAATATATTGAGTTTTAA
TAACCTTTGTTTTGAAATTAATGTGTTGGTGTGACAGATCACAACTTCTTTTCAGTAAT
TTCTAGGGAATATTCTCA
>RNA_Sequence_3
CCTGGCCCCAAGACATTGCAAGTGTGGCTGCTGTGCACATGTATGTCAAGTGGCA
GAGAGGAGAGAGGCTGTGACTACACTCAGTTTCTGACCTGTGAACATCTGAATGATTA
TTTACTTAACACT
>RNA_Sequence_4
GATCAGCTGATGAGTGCCTAATAGAACGAACGTGTGTCTTGGGT
>RNA_Sequence_5
CGTTCGCGTAGACGGGAGAGGGACACCTTCGGGTGTGCTGGTTATCTACCAAGTCGG
CGAACCTTTCTCGTTCTGCCACATA

```

Escolher arquivo

Nenhum arquivo selecionado

Organisms

Organisms List:

Homo sapiens

Run RNAMining

Example of results are made available here.

B

Laboratory of Artificial Intelligence Applications - Universidade Federal do Rio de Janeiro

Run About Tutorial Download Contact

Results

[Download All Files](#)

[Download Coding Sequences](#)

[Download Non-coding Sequences](#)

Show 10 entries

Search:

Sequence ID	Coding Potential Classification
ENSMUST00000239379.1 ncma chromosome GRCh38:9:88521066:88522889-1 gene:ENSMUSG00000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nuclear RNA host gene 5 [Source:MGI SymbolAcc:MGI9191905]	non-coding
ENSMUST00000239362.1 ncma chromosome GRCh38:11:116672917:116675793.1 gene:ENSMUSG00000020812.18 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Snhg5 description:small nuclear RNA host gene 16 [Source:MGI SymbolAcc:MGI913543]	non-coding
ENSMUST00000239344.1 ncma chromosome GRCh38:9:88521052:88522890-1 gene:ENSMUSG00000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nuclear RNA host gene 5 [Source:MGI SymbolAcc:MGI9191905]	non-coding
ENSMUST00000239311.1 ncma chromosome GRCh38:2:128237095:128240585-1 gene:ENSMUSG00000074813.14 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Morbid description:myeloid RNA regulator of BCL2L1 induced cell death [Source:MGI SymbolAcc:MGI365219]	non-coding
ENSMUST00000239294.1 cds chromosome GRCh38:3:707665311:107741638.1 gene:ENSMUSG000000036528.16 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:Pfifb2 description:PTFR interacting protein, binding protein 2 (liprin beta 2) [Source:MGI SymbolAcc:MGI894849]	coding
ENSMUST00000239292.1 cds chromosome GRCh38:7:19424775:19458821-1 gene:ENSMUSG000000030397.11 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:Mark4 description:MAP/microtubule affinity regulating kinase 4 [Source:MGI SymbolAcc:MGI1920955]	coding
ENSMUST00000239290.1 ncma chromosome GRCh38:9:88521066:88522890-1 gene:ENSMUSG00000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nuclear RNA host gene 5 [Source:MGI SymbolAcc:MGI9191905]	non-coding
ENSMUST00000239275.1 ncma chromosome GRCh38:2:128336729:128333688-1 gene:ENSMUSG00000074813.14 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Morbid description:myeloid RNA regulator of BCL2L1 induced cell death [Source:MGI SymbolAcc:MGI365219]	non-coding
ENSMUST00000239252.1 cds chromosome GRCh38:8:45953608:45975440-1 gene:ENSMUSG000000071103.10 gene_biotype:protein_coding transcript_biotype:sense_intronic_decay gene_symbol:T00029J07Rik description:RIKEN cDNA T00029J07 gene [Source:MGI SymbolAcc:MGI1916729]	coding
ENSMUST00000239232.1 ncma chromosome GRCh38:17:1214364:12815341.1 gene:ENSMUSG000000078247.5 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Aim description:antisense lgr2r RNA [Source:MGI SymbolAcc:MGI353471]	non-coding

Show 1 to 10 of 8,818 entries

Previous 1 2 3 4 5 ... 882 Next

Figure 2- RNAmMining web server overview. A. Job launcher screen (Run tab). The user only needs to upload the nucleotide sequences in FASTA format and select the model to be used based on the evolutionary close related species. **B. Results web page screen.** General report containing the list of coding and non-coding sequences in a dynamic table, in which the user can search for a particular sequence or filter only those coding or non-coding RNAs by using a free text form that will filter the results in the table dynamically. The user can download the complete table in tabular format and two FASTA files containing the set of coding and non-coding RNAs separately.

DISCUSSION

The coding potential prediction of nucleotides is a key step in the definition of the repertoire of non-coding RNAs in a genome or transcriptome project, especially when dealing with non-model organisms. Sometimes, predictive tools for the computational characterization of RNA molecules, in analyses like the prediction of specific RNA families (19) or the estimation of a network of RNA-RNA (20) or protein-RNA interactions (21), have their performance affected according to the training organism, increasing the number of false positives when applied in evolutionarily distant species. In this work, we evaluated the performances of seven different supervised machine learning algorithms, using eukaryotic species from a variety of evolutionary clades, revealing their potential to be used in the development of novel and improved computational tools for the coding potential prediction of RNA sequences. Artificial intelligence has been widely used in computational biology (22,23), but its application to characterize ncRNAs has been limited.

In this benchmarking, we opted to analyze the tri-nucleotides counting as the main feature to be evaluated for the coding potential prediction, followed by a normalization considering the sequences length (*i.e.* each tri-nucleotide count was divided by the total size of the given sequence). Panwar *et al.* (10) used nucleotides counting successfully for this purpose. They considered 40,905 non-coding RNAs from Rfam release 10.0 database and 62,473 coding RNA sequences from Human RefSeq database, divided into 50% of training and 50% of test (*i.e.* the training and test sets were composed of 20,453 non-coding and 31,237 coding sequences). They used the counts of mono-, di-, tri-, tetra- and penta-nucleotides and a combination of all counts using the SVM method, and showed that using tri-nucleotides counts is enough to predict the coding potential of ncRNAs with better accuracies. Our comparisons of the machine learning algorithms revealed XGBoost as the algorithm with better performance, presenting efficiency in predicting the coding potential of RNA sequences even when using the models of distantly related organisms. This latter shows the usefulness of this approach for performing coding predictions in non-model organisms.

We implemented XGBoost in RNAmMining, a stand-alone and web server tool flexible to be used in genome or transcriptome projects focused in both model and non-model eukaryotic organisms. Our tool outperformed similar approaches, such as CPAT (11), CPC2 (13), RNAcon (10) and Transdecoder (12). Both versions of the software are easy to use, with the web version providing a simple report and FASTA format files that can be used in downstream analysis. It provides 15 models generated from eukaryotic from different evolutionary clades. Other models can be generated by the user using the stand-alone version, which can be used with simple command line operations. These

features facilitate its usage for experienced users and, especially, for those without any programming experience, which can easily perform large-scale predictions of the coding potential of nucleotide sequences in both genome or transcriptome initiatives.

DATA AVAILABILITY

Ensembl is an open access genome browser for vertebrate genomes in the Ensembl website (<https://www.ensembl.org/index.html>).

RNAmining is a tool for coding potential prediction which is freely available at (<https://rna mining.integrativebioinformatics.me/download>).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Savio Torres de Farias for the helpful discussions during the preparation of this manuscript.

FUNDING

This work was funded in part by grants from ANID-FONDECYT (11161020), ANID-PAI (PAI79170021) and ANID-FONDAP (15130011) to VMC. TARR received a Master and a PhD fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. RAC received a post-doctoral fellowship from ACCDiS.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

REFERENCES

1. Mattick JS. The central role of RNA in the genetic programming of complex organisms. *An Acad Bras Cienc.* 2010;82: 933–939.
2. Gelsinger DR, DiRuggiero J. The Non-Coding Regulatory RNA Revolution in Archaea. *Genes.* 2018;9. doi:10.3390/genes9030141.
3. Croce CM. Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet.* 2009;10: 704–714.
4. Schaefer A, O'Carroll D, Tan CL, Hillman D, Sugimori M, Llinas R, et al. Cerebellar neurodegeneration in the absence of microRNAs. *J Exp Med.* 2007;204: 1553–1558.
5. Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, Muth AN, et al. Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell.* 2007;129: 303–317.

6. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. *Nature*. 2012;489: 101–108.
7. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A*. 2014;111: 6131–6138.
8. Arias-Carrasco R, Vásquez-Morán Y, Nakaya HI, Maracaja-Coutinho V. StructRNAfinder: an automated pipeline and web server for RNA families prediction. *BMC Bioinformatics*. 2018;19: 55.
9. Torres F, Arias-Carrasco R, Caris-Maldonado JC, Barral A, Maracaja-Coutinho V, De Queiroz ATL. LeishDB: a database of coding gene annotation and non-coding RNAs in *Leishmania braziliensis*. *Database*. 2017;2017. doi:10.1093/database/bax047.
10. Panwar B, Arora A, Raghava GPS. Prediction and classification of ncRNAs using structural information. *BMC Genomics*. 2014;15: 127.
11. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 2013;41: e74.
12. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8: 1494–1512.
13. Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*. 2017;45: W12–W16.
14. Chang C-C, Lin C-J. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*. 2011. pp. 1–27. doi:10.1145/1961189.1961199.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12: 2825–2830.
16. Python API Reference — xgboost 1.3.0-SNAPSHOT documentation. [cited 14 Oct 2020]. Available: https://xgboost.readthedocs.io/en/latest/python/python_api.html.
17. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46: D754–D761.
18. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37: W202–8.
19. Aguiar RR, Ambrosio LA, Sepúlveda-Hermosilla G, Maracaja-Coutinho V, Paschoal AR. miRQuest: integration of tools on a Web server for microRNA research. *Genet Mol Res*. 2016;15. doi:10.4238/gmr.15016861.
20. Sinan Uğur Umu PPG. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics*. 2017;33: 988.
21. Nithin C, Mukherjee S, Bahadur RP. A non-redundant protein-RNA docking benchmark version 2.0. *Proteins*. 2017;85. doi:10.1002/prot.25211.
22. de Brito DM, Maracaja-Coutinho V, de Farias ST, Batista LV, do Rêgo TG. A Novel Method to Predict Genomic Islands Based on Mean Shift Clustering Algorithm. *PLoS One*. 2016;11: e0146352.

23. Ramos TAR, Maracaja-Coutinho V, Ortega JM, do Rêgo TG. CORAZON: a web server for data normalization and unsupervised clustering based on expression profiles. BMC Res Notes. 2020;13: 338.

A

B

Upload your file

Dataset

.fasta - examples

>RNA_Sequence_1
GTCTCCCTAGAGTCCCTTGACCACCTCACTGGGGACCTTCTCTAATTATAATGACTTC
CTACTGAAGTGTTTGGGGAACTCTGGTGCCAT
>RNA_Sequence_2
ATCGCTTCTGGCCTTTGGCTAAGATCAAGGTAGGAACAATATATTGAAGTTTTTA
TAACTTTGTTTTTGAATTAATGTTTGGTGTGACAGATCACAATTTCTTTCAGTAAT
TTCTAGGAATATCTCA
>RNA_Sequence_3
CCTTGGCCCCAAGACATTGCAAGGTGGCTGCTGTGTGCACATATGTCTACTAGGTGGCA
GAGAGGAGAGAGGCCTGTGACTACACTCACTGTTTCTGACCTGTGAACATCTGAATGATTAA
TTTTACTAACCT
>RNA_Sequence_4
GATCAGCTGATGAGTCGAAATAGAACGAAACGTGTGCTCTGGGT
>RNA_Sequence_5
CGTTGCGGTAGGACGGGAGAGGACACTTCCTGGGTGTGCTGTTTATCTACCAAGTCGG
CGAACCTTCTCTGTTGCCACCATA

Escolher arquivo Nenhum arquivo selecionado

Organisms

Organisms List:

Homo sapiens

Run RNAMining

Example of results are made available here.

Copyright © 2020

Laboratory of Integrative Bioinformatics - Universidad de Chile

Bioinformatics Multidisciplinary Environment - Universidade Federal do Rio Grande do Norte

Laboratory of Artificial Intelligence Applications - Universidade Federal da Paraíba

Results

Download All Files

Download Coding Sequences

Download Non-coding Sequences

Show 10 entries

Search:

Sequence ID	Coding Potential Classification
ENSMUST000002393791 ncma chromosome:GRCm38:9:88521066:88522889-1 gene:ENSMUSG000000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nucleolar RNA host gene 5 [Source:MGI Symbol:Acc:MGI1919905]	non-coding
ENSMUST000002393621 ncma chromosome:GRCm38:11:16672917:166757931 gene:ENSMUSG00000020812.18 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Snhg16 description:small nucleolar RNA host gene 16 [Source:MGI Symbol:Acc:MGI1913543]	non-coding
ENSMUST000002393441 ncma chromosome:GRCm38:9:88521052:88522890-1 gene:ENSMUSG000000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nucleolar RNA host gene 5 [Source:MGI Symbol:Acc:MGI1919905]	non-coding
ENSMUST000002393111 ncma chromosome:GRCm38:2:128237095:128240685-1 gene:ENSMUSG000000074813.14 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Morbid description:myeloid RNA regulator of BCL2L1 induced cell death [Source:MGI Symbol:Acc:MGI3652191]	non-coding
ENSMUST000002392941 cds chromosome:GRCm38:7:10765311:1077146381 gene:ENSMUSG000000036528.16 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:Pfifb2 description:PTPRF interacting protein, binding protein 2 (lpirin beta 2) [Source:MGI Symbol:Acc:MGI894649]	coding
ENSMUST000002392921 cds chromosome:GRCm38:7:19424775:19458821-1 gene:ENSMUSG000000030397.11 gene_biotype:protein_coding transcript_biotype:protein_coding gene_symbol:Mark4 description:MAP1 microtubule affinity regulating kinase 4 [Source:MGI Symbol:Acc:MGI1920955]	coding
ENSMUST000002392901 ncma chromosome:GRCm38:9:88521066:88522890-1 gene:ENSMUSG000000097195.10 gene_biotype:ncRNA transcript_biotype:retained_intron gene_symbol:Snhg5 description:small nucleolar RNA host gene 5 [Source:MGI Symbol:Acc:MGI1919905]	non-coding
ENSMUST000002392751 ncma chromosome:GRCm38:2:128316297:128333668-1 gene:ENSMUSG000000074813.14 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Morbid description:myeloid RNA regulator of BCL2L1 induced cell death [Source:MGI Symbol:Acc:MGI3652191]	non-coding
ENSMUST000002392521 cds chromosome:GRCm38:8:45953608:45975440-1 gene:ENSMUSG00000007103.10 gene_biotype:protein_coding transcript_biotype:nonsense_mediated_decay gene_symbol:I700029J07Rik description:RIKEN cDNA 1700029J07 gene [Source:MGI Symbol:Acc:MGI1916729]	coding
ENSMUST000002392321 ncma chromosome:GRCm38:17:12741364:12815341 gene:ENSMUSG000000078247.5 gene_biotype:ncRNA transcript_biotype:ncRNA gene_symbol:Aim description:antisense lgl2r RNA [Source:MGI Symbol:Acc:MGI353471]	non-coding

Showing 1 to 10 of 8,818 entries

Previous 1 2 3 4 5 ... 882 Next