Published online 25 April 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 11* ***e93***
doi: 10.1093/nar/gku325

Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm

Supatcha Lertampaiporn[1], Chinae Thammarongtham[2], Chakarida Nukoolkit[3], Boonserm Kaewkamnerdpong[1] and Marasri Ruengjitchatchawalya[4,5,*]

[1]Biological Engineering Program, Faculty of Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd, Bangmod, Thung Khru, Bangkok 10140, Thailand, [2]Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology Thonburi (Bang Khun Thian Campus), 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok 10150, Thailand, [3]School of Information Technology, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd, Bangmod, Thung Khru, Bangkok 10140, Thailand, [4]Biotechnology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian Campus), 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok 10150, Thailand and [5]Bioinformatics and Systems Biology Program, King Mongkut's University of Technology Thonburi (Bang Khun Thian Campus), 49 Soi Thian Thale 25, Bang Khun Thian Chai Thale Rd, Tha Kham, Bangkok 10150, Thailand

Received February 10, 2014; Revised April 02, 2014; Accepted April 7, 2014

ABSTRACT

To identify non-coding RNA (ncRNA) signals within genomic regions, a classification tool was developed based on a hybrid random forest (RF) with a logistic regression model to efficiently discriminate short ncRNA sequences as well as long complex ncRNA sequences. This RF-based classifier was trained on a well-balanced dataset with a discriminative set of features and achieved an accuracy, sensitivity and specificity of 92.11%, 90.7% and 93.5%, respectively. The selected feature set includes a new proposed feature, SCORE. This feature is generated based on a logistic regression function that combines five significant features—structure, sequence, modularity, structural robustness and coding potential—to enable improved characterization of long ncRNA (lncRNA) elements. The use of SCORE improved the performance of the RF-based classifier in the identification of Rfam lncRNA families. A genome-wide ncRNA classification framework was applied to a wide variety of organisms, with an emphasis on those of economic, social, public health, environmental and agricultural significance, such as various bacteria genomes, the *Arthrospira* (*Spirulina*) genome, and rice and human genomic regions. Our framework was able to identify known ncRNAs with sensitivities of greater than 90% and 77.7% for prokaryotic and eukaryotic sequences, respectively. Our classifier is available at http://ncrna-pred.com/HLRF.htm.

INTRODUCTION

Non-coding RNAs (ncRNAs) are involved in a variety of important biological functions in the cell, including the control of chromosome dynamics, RNA splicing, RNA editing, translational inhibition and mRNA destruction (1). ncRNAs have recently been acknowledged to be diverse and significantly more important than previously thought (2–3). In human transcriptome analysis, >70% of the human genome is likely transcribed into ncRNAs, whereas protein-coding transcripts account for only ∼2–3% of the genome (4–6). NcRNAs can be roughly classified into short ncRNAs (such as microRNA (miRNAs), short-interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and short hairpin RNAs (shRNAs)) or long ncRNAs (lncRNAs), depending on the transcript size (3,7–9). Short ncRNAs are shorter than 200 nucleotides (nt), while lncRNAs are longer than 200 nt (10). LncRNAs can range in size from 200 to 100 000 nt (3,11). In contrast to well-studied short ncRNAs such as miRNAs and snoRNAs, lncRNAs are relatively uncharacterized, but accumulating evidence indicates that they likely have a broad range of functions (7–9,12). lncRNAs have been proposed to constitute the major fraction of eukaryotic transcriptomes; the transcriptome is involved in the regulation of many important cellular processes as well as the epigenetic control of complex mechanisms (5,13). Dysfunctions

[*]To whom correspondence should be addressed. Tel: +66 2 470 7481; Fax: +66 2 452 3455; Email: marasri.rue@kmutt.ac.th

of lncRNAs have been associated with various diseases, including cancers, cardiovascular diseases and neurodegenerative diseases (14–16). The range of ncRNAs is expanding rapidly as new ncRNAs continue to be discovered via high-throughput sequencing. However, a large number of ncRNAs likely remain to be identified (1,17). Thus, the identification and annotation of ncRNAs are important steps for the elucidation of various regulatory mechanisms in the cell.

Current experimental methods have yielded promising results but are subject to certain limitations. The expression of most ncRNAs is lower than that of mRNAs, and ncRNAs display tissue/stage-specific expression patterns (1,18,19). Moreover, high-throughput sequencing causes enormous informatics, and requires extensive computational analysis (15). Thus, computational identification methods may complement experimental methods to quickly identify ncRNAs in new genomes, particularly ncRNAs that are transcribed under specific conditions in specific cell types. Various computational methods have been proposed to predict ncRNAs, including comparative (20–23) and non-comparative methods (24–31).

Most ncRNA identification algorithms are designed to identify structural ncRNAs based on low-energy structures. However, the use of secondary structure feature alone is generally not sufficiently statistically robust enough to detect all types of ncRNA (32) because a random RNA with high GC content can also fold into a low-energy structure. Computational methods utilizing thermodynamic stability structure features have successfully identified highly structured ncRNAs such as tRNAs, snoRNAs and miRNAs but cannot identify less densely structured classes of ncRNAs, such as lncRNAs (11). LncRNAs share some features with protein-coding genes: they can be spliced, 5′-capped and/or polyadenylated (13). However, unlike protein-coding genes, lncRNAs are generally expressed at low levels and lack strong evolutionary conservation across species compared to protein-coding sequences and small RNAs (e.g. miRNAs and snoRNAs). To identify lncRNAs, ncRNA gene identification method must incorporate other signal features that can be used to characterize lncRNAs. The recognition of a wide range of ncRNAs that exhibit heterogeneity among different species and different ncRNA families remains computationally challenging. In addition, some ncRNAs have weak structure signals.

In this work, we developed a generalized classifier for ncRNAs based on an ensemble of multiple decision trees, a random forest (RF), to discriminate ncRNAs from genomic sequences. Due to the complex and heterogeneous nature of ncRNAs, the classification of ncRNAs based on structures or sequences may not be appropriate. To classify ncRNAs efficiently, we take into account various characteristics of the ncRNAs (such as modularity elements, structural robustness scores, base-pair features, sequence compositions and structural features). We hypothesized that certain combinations of significant features would improve the characterization of heterogeneous ncRNAs. Thus, we propose a new composite feature based on a logistic regression model, SCORE, to increase the sensitivity of the RF prediction. We used our framework to scan for ncRNAs in a wide range of genomes, including both bacterial and eukaryotic genomes. In particular, we focused on genomes of economic, social, public health, environmental and agricultural importance.

MATERIALS AND METHODS

Dataset

Two training datasets were used: training 1 and training 2. The training 1 dataset was composed of all ncRNA seed alignment sequences obtained from the Rfam database, version 11.0 (33). The positive data were randomly selected from 32 300 ncRNAs from various organisms. Then, ncRNA sequences shorter than 50 nt were excluded, and CD-HIT (34) was used to eliminate redundant sequences with sequence similarities above 80%. Thus, the positive data consisted of 6649 non-redundant sequences. The negative data included RefSeq (35) coding sequences and shuffled sequences of both coding and ncRNA sequences. A set of 2147 non-redundant coding sequences was extracted from the coding region sequences of human RefSeq genes, and these were used to validate the lack of annotated or unannotated ncRNA sequences. A set of 4502 shuffled sequences obtained from 2147 coding and 2355 non-coding sequences was shuffled while preserving both the mono- and di-nucleotide frequencies. The balance between the number of positive and negative training sets was maintained because an imbalanced distribution could affect the performance of the classifier. The training 2 dataset included all lncRNAs obtained from the lncRNAdb database (36), a comprehensive database of lncRNAs, as the positive training data. The negative training data were randomly selected from the negative samples of the training 1 dataset.

To verify the results, genomic data for the following species were downloaded from the NCBI GenBank genome database (ftp://ftp.ncbi.nih.gov/genomes/): *Escherichia coli K12* (U00096), *Acholeplasma laidlawii PG-8A* (CP000896), *Acidovorax sp. JS42* (CP000539), *Brucella suis 1330* Chromosome (chr.) 1 (AE014291), *Candidatus methanoregula boonei 6A8* (CP000780), *Oryza sativa japonica* chr.1 (CM000138), *Arthrospira (Spirulina) platensis NIES-39* (AP011615), *Penaeus monodon mitochondrial* genome (AF217843), *Mycobacterium tuberculosis H37Rv* (AL123456), *Pseudomonas aeruginosa* (CP000438) and *Homo sapiens* genomic regions containing ncRNAs. These genomes and genomic regions were used to test the method. For each of the test genomes, with the exception of the *A. platensis* and *P. monodon* mitochondrial genomes, sets of known ncRNAs were downloaded from the Rfam database version 11.0 in generic feature format. The files were then parsed to retrieve the start and end positions of all types of ncRNAs in the test genomes. For *H. sapiens*, we extracted five regions containing known ncRNAs: (i) five known lncRNAs (GNAS_AS1_1–5) in the 57 417 000–57 426 000 bp region of chr. 20; (ii) nine known miRNAs in the 49 767 700–49 779 500 bp region of chr. X; (iii) six known lncRNAs (MAT2A_A-F) in the 85 770 900–85 772 300 bp region of chr. 2; (iv) five known lncRNAs (HOTAIRM1_1–5) in the 27 135 000–27 139 900 bp region of chr. 7; and (v) six known lncRNAs (HOXA11_AS1_1–6) in the 27 225 000–27 228 900 bp region of chr. 7. Furthermore, northern blot verified ncRNA candidates from the literature (37) were used to test the ability of the proposed method for the identification of novel ncRNAs.

Features

Various types of features were extracted that could be important for ncRNA prediction. In total, 369 features were considered in this work, which can be divided into five categories, as summarized in Table 1. Descriptions of the features are listed in the Supplementary Material and Method 1.

Feature selection

Due to the large number of features, which may include redundant features, a feature selection process is needed to filter out irrelevant, redundant and uninformative features and select only the most informative set for ncRNAs identification. In this work, correlation-based feature selection (CFS) (38) and a genetic algorithm (GA) search method were used to select the discriminative feature subset. The CFS+GA combination has been used for pre-miRNA data to yield a compact feature subset and an improvement in performance (31). The CFS+GA method selected a feature subset with the highest merit criterion, and the combination of those features exhibited good predictive power in ncRNA identification. A set of 20 features from the 369 ncRNA features was selected as the discriminatory set by CFS+GA, including Prob, zG, SC_score, SCxdP, Bits, CM_score, Bits2, SCORE, AAAA, AGGA, CAAC, CAGU, CUAC, CUGA, GAUA, GCAU, GUUC, UAAG, UACA and UUUU. To visualize the spread of the training data 1 for each selected feature, graphical boxplots are shown in Figure 1.

Logistic regression model

A logistic regression model (39) was used to define the complex relationships among important characteristic factors and to describe the ncRNA elements in the lncRNAs. The model was trained with training set 2. A combination of features was used to fit a logistic regression model represented by the following equation:

$$P(Y = \text{ncRNA}|X) = \text{logistic}(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}.$$

In this study, the logit transformation (the logarithm of the odds ratio or likelihood ratio) was used for the link function with logistic regression. The logit function, referred to as the SCORE feature, is defined as

$$\text{SCORE} = \text{logit}(X) = \log\left(\frac{X}{1-X}\right) = \log\left(\frac{P(Y = \text{ncRNA}|X = X)}{P(Y = \text{other}|X = X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n,$$

where β_i represents the regression coefficient of the explanatory variable X_i. The logistic regression model was implemented by using the generalized linear model (GLM) function in R (40). We used GLM to describe a function of related factors to predict the possibility that a sequence was lncRNA. We defined all relevant features that might be involved in predicting lncRNAs, and ANOVA tests were used to identify the statistically significant features associated with lncRNAs. Various regression models were explored, and the optimal model was validated based on a 10-fold cross validation. A stepwise forward variable selection was used to select the significant factors based on their *p*-values.

The logistic model was based on five significant features: sequence similarity (Bits), structural motif similarity (CM_score), modularity sequence profile score (Bits2), coding potential (logodds) and structural robustness score (SCxabsZG). These features were included in the logistic function to describe the lncRNAs because they are statistically significant based on the Wald test ($p < 0.001$). The sequence similarity is a Bit score obtained from BLAST (Basic Local Alignment Search Tool) (41). Bits2 is a modular sequence similarity obtained from BLAST search against the modularity sequence profile database. To build the modularity sequence profile database, we collected the modularity pivot sequences from Rfam seed alignment by searching for portions of common RNA sequences in long noncoding alignments and curating them as custom libraries for BLAST. CM_score is an RNA secondary structure similarity score obtained by using infernal (42) to search against the covariance models (CM) (43), which are RNA secondary profiles from a custom modular CM library. In brief, our modular CM database was created in three steps: (i) manual extraction of the critical conserved substructures from selected lncRNA seed alignments and CM model construction, (ii) searches for short motifs RNA in the Rfam seed alignments using CMfinder (44) and CM model construction, and (iii) scans for and removal of redundant models. The structural robustness score was the product of a self-containment (SC) score (45) and the absolute value of the z-score (zG) (46). The coding potential based-feature (logodds) was extracted from the framefinder s/w (47,48) to identify the longest reading frame in the three forward frames of the sequences. Finally, the logistic model was used as a composite feature to discriminate between lncRNAs and other sequences in the machine learning (ML) model. All analyses were performed in the R statistical environment (40).

Machine learning algorithms

We trained an RF as the main classifier. An RF is an ensemble method that uses decision trees as its base classifiers (49,50). Each of the individual decision trees was trained on a random subset of the total features to maximize the classification criteria at each node, and the different classification hypothesis trees were then combined to form the ensemble (51,52). The predictions of the RF model represent a consensus of the predictions made by all of the individual trees. We used Weka (53) and the randomForest R package (40,51) with 10 trees (ntree = 10), and 5 randomly

Table 1. Summary of the 369 ncRNA features

Feature group	No. of features	Feature symbol
Sequence-based features	277	%G+C,% A+U,%AA,%AC,%AG,%AU,%CA,%CC,%CG,%CU,%GA,%GC,%GG,%GU,%UA,%UC,%UG,%UU,%AAAA -%UUUU (256 of 4-Mer), Blast_bits score (Bits), Modular_Bits (Bits2), Coding Potential (logodds)
Secondary structure features	23	MFE, efe, MFEI1, MFEI2, MFEI3, MFEI4, dG, dQ, dD, dF, Prob, zG, zQ, zD, zF, nefe, Freq, diff, dH, dS, Tm, CM score (CM), SCORE
Base-pair features	28	dP, zP, div, tot_bp, stem, loop, A-U/L, G-U/L, G-C/L,%A-U/Stem,%G-C/Stem,%G-U/Stem, Probpair1–10, Avg_PP, NonBP_A, NonBP_C, NonBP_G, NonBP_U, Non_BPP
Triplet sequence-structure	32	A(((, A((., A(.., A(.(,A.((,A.(.,A..(, A. . . , C(((, C((., C(.., C(.(, C.((, C.(., C..(, C. . . , G(((, G((., G(.., G(.(, G.((, G.(., G..(, U. . . , U(((, U((., U(.., U(.(, U.((, U.(., U..(, U. . .
Structural robustness features (SC-derived features)	9	SC, SCxMFE/Mean_dG, SCxdP, SCxabsZG, SC/(1-dP), SC/NonBP_A, SC/NonBP_C, SC/NonBP_G, SC/NonBP_U
Total	369	

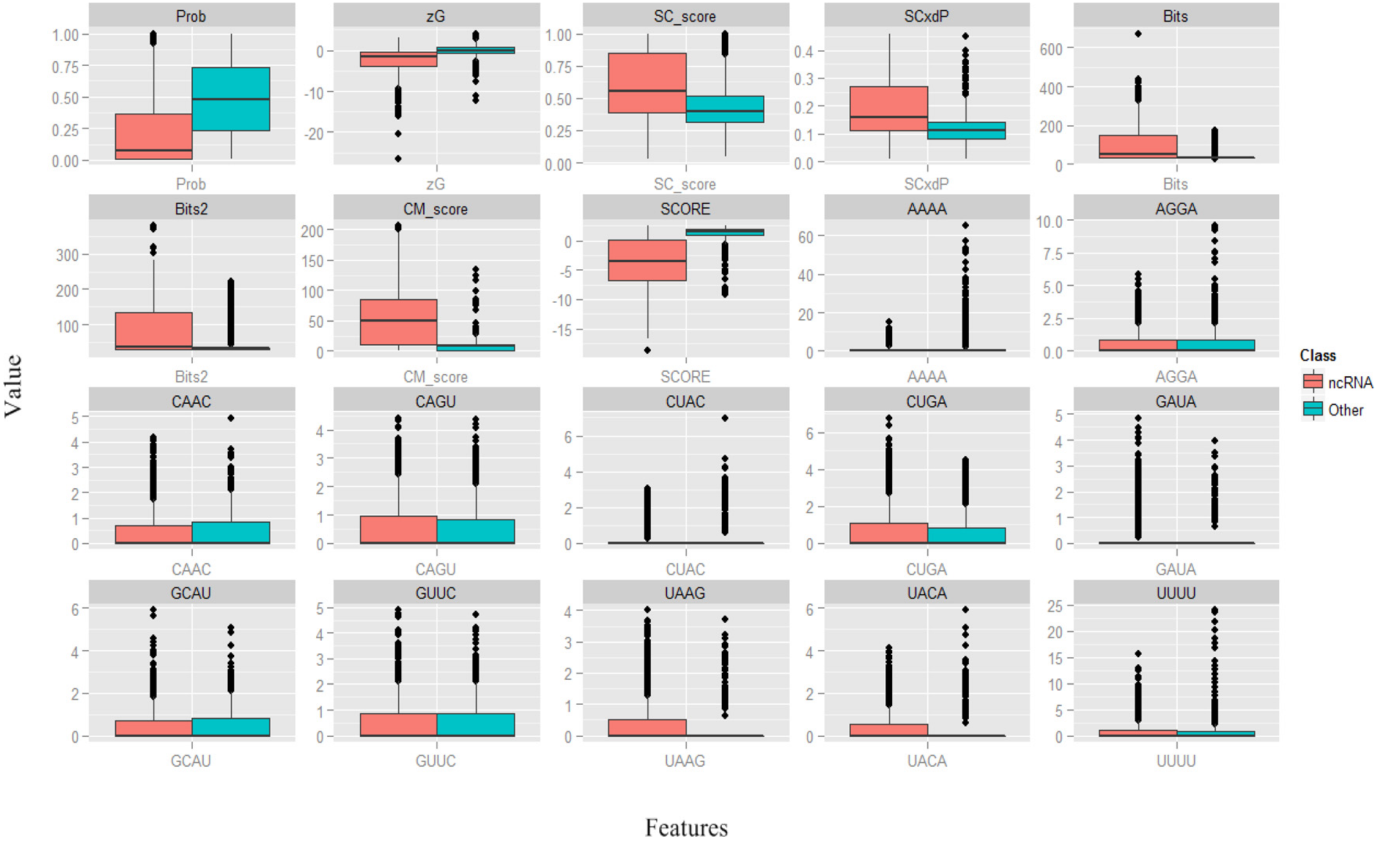

Figure 1. Boxplots showing the spread of the training data for each of the 20 selected features: Prob, zG, SC_score, SCxdP, Bits, CM_score, Bits2, SCORE, AAAA, AGGA, CAAC, CAGU, CUAC, CUGA, GAUA, GCAU, GUUC, UAAG, UACA and UUUU. For each plot, the left side represents the ncRNA class, and the right side represents other classes.

sampled features as candidates at each split (mtry = 5). Detailed implementation of other ML methods was used in this work, including Naïve Bayes (NB), support vector machine (SVM), K-nearest neighbors (kNN), neural network (MLP), decision tree (DT), rule induction (RIPPER) and RF. The algorithm selection procedure is described in the Supplementary Material and Method 2.

Classification framework for ncRNAs

The process of computational ncRNA gene identification is illustrated in Figure 2. The input was a sequence of any length, preferably with a size between 75 and 200 nt. For the entire genome, a sliding window module was used to split the genome into multiple overlapping sub-sequences with a size of 120 nt and a step of 40 nt. A window size of 120 nt, as used in RNAz (21) screens, is an appropriate input size because it is large enough to detect most ncRNAs, ranging from small ncRNAs to at least a substructure of a lncRNA (43). The window size was primarily tested and trained as the performance was optimized (data not shown). In the next step, a feature extraction module was used to extract the 20 most informative features from each input sequence.

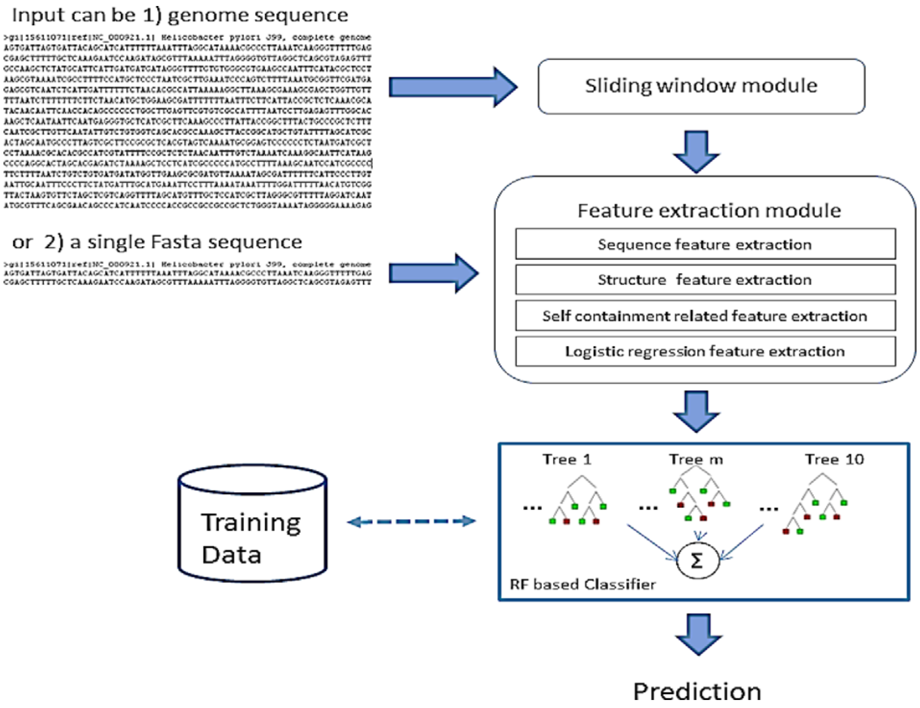
Input can be 1) genome sequence
or 2) a single Fasta sequence
Sliding window module
Feature extraction module
Sequence feature extraction
Structure feature extraction
Self containment related feature extraction
Logistic regression feature extraction
Training Data
Tree 1
Tree m
Tree 10
...
Σ
RF based Classifier
Prediction

Figure 2. Process of computational ncRNA gene prediction.

In the final step, the trained RF classifier was used to classify the input sequences.

Performance assessment

To precisely evaluate the predictive power of our classification model, several standard performance measures were used:

$$\text{Accuracy(ACC)} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad \text{Sensitivity(Sn)} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{Positive Predictive Value(PPV)} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad \text{Specificity(Sp)} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{False Positive Rate(FPR)} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{TN}+\text{FP}}.$$

We evaluated the area under the ROC curve (AUC), in which the sensitivity is plotted as a function of the false positive rate (FPR) at different decision thresholds. A greater AUC value indicates a better classification result. In case of genome screening, False Discovery Rate (FDR) was calculated using the same approach as previously described in (29), which involves steps in calculating FPR in the shuffled genome. The shuffling method ensured that all unknown ncRNAs located within the genome were fragmented (30).

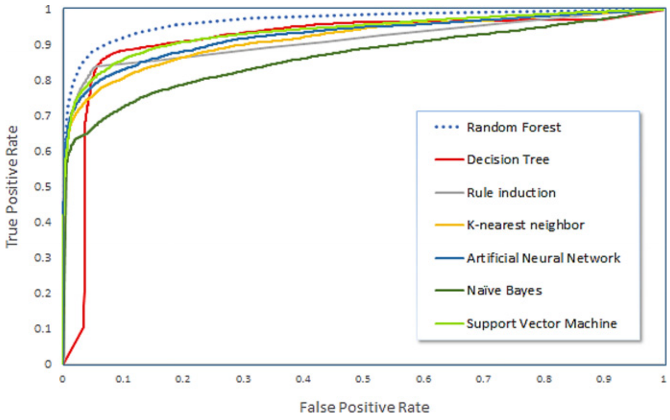

Figure 3. ROC curves showing the performances of the classifiers in 10-fold cross validation.

RESULTS AND DISCUSSION

Predictive power of the newly proposed model

The performances of ML algorithms usually depends on the task to which they are applied. Different algorithms are able to take advantage of different characteristics and relationships in a given dataset, and thus we constructed seven classifiers based on seven algorithms trained on the same training data and the same features and evaluated their performances using 10-fold cross validation (see the Algorithm selection in the Supplementary Method 2). The RF method outperformed the six other techniques for the prediction of ncRNAs. To visualize the performance of the difference algorithms, ROC curves were plotted (Figure 3). On average, RF yielded the highest AUC (of 0.972), which reflects a better balance compared to the other algorithms. The better performance of RF could be due to the robustness provided by the bootstrap aggregating (bagging) technique and the random feature selection process for building the ensemble of the DT model. Our data included a broad range of ncRNA families; thus, RF is more suitable than other methods because the heterogeneity of the ncRNA subfamilies can be captured by an ensemble of trees.

To assess the overall sensitivity of our RF method for screening previously annotated ncRNAs in various prokaryotic genomes, we began with *E. coli* because it is a well-studied prokaryotic model organism (with a genome size of 4 686 137 nt and 265 known ncRNA families) and has been previously used as a test genome for several *de novo* ncRNA prediction methods (24–26). A neural network (NN)-based method (24) outperforms all other reported methods when applied to the *E. coli* genome. Therefore, we compared the genome-wide performance of our method in identifying known ncRNAs in *E. coli* to that of NN (24). Our RF model performed significantly better than NN in terms of both sensitivity and specificity for identifying ncRNAs on both strands of the *E. coli* genome (Table 2).

We also performed an independent test to compare the prediction performance of our method with the performances of two recently developed approaches that used SVMs: smyRNA (29) and ncRNAscout (30). The smyRNA method uses primary sequence motif features to detect ncRNAs, and the ncRNAscout uses both a primary sequence motif and secondary structure-based features for ncRNA discovery. The results of these methods for various compact microbial genomes with varying lengths and %GC contents are summarized in Table 3. Our RF-based method performed very well in identifying well-characterized, previously annotated ncRNAs with a broad range of GC contents and nucleotide lengths.

The high FPR is another important aspect of computational ncRNA identification in genomes. It is impossible to calculate the FPR in real genomes because knowledge about ncRNAs is incomplete, and real genome sequences may contain some unknown novel ncRNAs. Thus, to evaluate the FDR, a shuffled genome was used to ensure that all unknown ncRNAs in the genome were broken up (29,30). An FDR of 16.9% was achieved by our method for the detection of known ncRNAs in the shuffled *Acidovorax sp. JS42* genome, which indicates that the model can discriminate real ncRNAs from shuffled genomic sequences.

Newly proposed feature improving prediction accuracy of our model

To determine which features play the most important roles in prediction, we plotted the importance of each variable based on permutation and gini criteria. The RF method provides the importance of each feature measure, which is called the relative variable importance. As shown in Figure 4, the SCORE feature ranked as the top discriminative feature, demonstrating the effectiveness of this feature for ncRNA identification. However, it has been suggested that the importance of an RF variable may be misleading if it is biased toward variables with many categories and continuous predictor variables (54,55). A revised RF model (cforest) based on conditional inference trees using subsampling without replacement has been proposed to provide a more reliable measure of variable importance. Thus, a conditional feature importance strategy was also plotted by using the cforest function (56) as shown in Supplementary Figure S1. The revised RF variable importance was in agreement with the original RF variable importance.

Different features are useful for identifying different classes of ncRNAs; thus the integration of multiple types of data can improve the sensitivity of computational methods for ncRNA identification (57). To address the diversity in ncRNA families, our feature selection method selected various types of features: sequence, structure, modularity, robustness and composite features. Sequence- and structure-based features are used extensively for ncRNA identification, including features such as MFE, zG (a structure or thermodynamic stability-based feature) and k-mer (a sequence-based feature). However, these commonly used features are not sufficiently general to encompass all ncRNA families. For example, algorithms that use only thermodynamic stability appear to have an advantage in detecting simple hairpin structural RNAs (58). Certain RNA families do indeed exhibit thermodynamic stability. However, this concept cannot be applied to all ncRNA families (46,59). Some features, such as nucleotide composition (k-mer), have been used with some success for ncRNA gene prediction (e.g. in smyRNA and ncRNAscout). A statistically significant signal based on CG content can be applied restrictively in AT-rich hyper-thermophiles (60–63). However, these sequence features are limited to compact bacterial organisms with base-composition biases, and their associated signals are generally insufficient for generalized ncRNA gene identification in most organisms (58).

To address the diversity in ncRNA families, our selected feature subset was composed of various complementary features that could enhance RF performance by capturing most possible aspects of heterogeneity in ncRNAs. These different features are useful for identifying different types of ncRNAs. For example, the CM score is advantageous for identifying ncRNAs with a high degree of conservation in sequence and structure. The BLAST score is advantageous for identifying ncRNAs with a high degree of sequence similarity. For example, SRP RNAs, U5 RNAs and U3 RNAs are well conserved at the sequence level (64). The SC_score or SC feature is advantageous for the detection of pre-miRNAs, which have high structural robustness and remain stable through two cleavage steps during their biogenesis (45). The pre-miRNAs generally exhibit a greater SC value, between 0.85 and 0.98. Moreover, pre-miRNAs are also anticipated to have higher SCxdP values compared to other ncRNA sequences (31). The Prob and zG features are useful for identifying ncRNAs with high thermodynamic stability. However, some classes of ncRNAs appear to be more difficult to detect than others (58), such as those with complex structural and highly variable lncRNAs. lncRNAs lack strong conservation (only 5% of lncRNA bases are evolutionarily constrained (8,17)) because they consist of multiple short regions that possess functional modules. The distinct modules of lncRNAs, which have a variety of secondary structure elements, interact with proteins, DNAs or RNAs to achieve specific regulatory outcomes (9–10,65–

Table 2. Performance comparison of our prediction method and the neural network (NN) method (24) for identifying known ncRNAs on both strands of the *E. coli* genome

Algorithm	Strand	Prediction performance measurement		
		Sensitivity	Specificity	FPR
NN (24)	+	0.72	0.66	0.34
Our model (RF)	+	0.89	0.86	0.14
NN (24)	–	0.65	0.73	0.27
Our model (RF)	–	0.93	0.87	0.13

Table 3. ncRNAs detected in four genomes: *Acholeplasma laidlawii PG-8A* (CP000896), *Acidovorax sp. JS42* (CP000539), *Brucella suis 1330* Chromosome (chr.) 1 (AE014291) and *Candidatus methanoregula boonei 6A8* (CP000780)

Genome source	Nucleotide length (nt)	GC content (%)	Based on Rfam 10				Based on Rfam 11	
			No. of known ncRNAs	Detected known ncRNAs (%)			No. of known ncRNAs	Detected by our model (%)
				ncRNA scout	smyRNA	Our model		
A. laidlawii PG-8A	1 496 992	31.92	42	92.857	95.238	95.238	64	96.875
Acidovorax sp. JS42	4 448 856	66.17	134	96.269	70.149	97.015	181	95.580
B. suis 1330 chromosome1	2 107 794	57.21	49	89.796	71.429	100	63	93.650
C. methanoregula boonei 6A8	2 542 943	54.51	23	73.913	56.522	100	40	97.500

Our method was compared to ncRNAscout (30) and smyRNA (29).

67). Thus, we captured the functional elements of lncRNAs in most possible ways by combining the power and advantages of various significant features: sequence, structure, robustness, coding potential and modular characteristics. RNA structural motifs are building blocks of the complex RNA architecture, and the recurrence of RNA structural motifs implies their high modularity and functional importance (10,68). On this basis, we hypothesized that lncRNAs containing various sequences and structural elements could be captured by scoring for these modular sequences and structural motifs. We generated the database containing various motifs, both structural and sequence-based (see the Materials and Methods section for more details). Furthermore, we hypothesized that certain combinations of significant features can reliably distinguish ncRNA elements from other sequences. To combine unrelated features with different scales, a unified statistical framework was needed. Therefore, we developed a scoring scheme based on the logistic regression function of the five significant features and used this scheme to discriminate lncRNA elements from other sequences. We generated several feature composition models based on LDA (Linear Discriminant Analysis), PCA (Principal Component Analysis) and LR (Logistic Regression) techniques and compared them by 10-fold cross validation. Based on their performances, we defined the logistic regression model as a composite feature, SCORE, because the method has a higher AUC compared to other techniques (R-squared = 0.7258, AUC = 0.9015). Moreover, logistic regression is relatively robust and easily updated, and allows meaningful interpretation. It is a hypothesis-driven model that can provide more useful information for ncRNA identification.

To examine the spread of SCORE values in the training data and compare to the well-known MFE feature, we plotted boxplots and histograms using rattle (69) in R. As shown in Figure 5, a significant difference was observed between ncRNAs and other classes in the SCORE feature. To explore whether the SCORE feature is able to capture lncRNA elements and improve the performance of the classifier, we compared the lncRNA prediction results obtained with our model framework with and without the SCORE feature. We randomly selected 10 lncRNA families from the Rfam database and used them as testing data. The performances of the model with and without the SCORE feature are compared in Table 4. These results provide preliminary evidence to support our hypothesis that the SCORE feature facilitates the identification of lncRNA regions. Table 4 shows that both of our RF frameworks successfully recovered most of the lncRNAs. In terms of sensitivity, the results suggest that our RF model can be applied generally to lncRNA elements. Moreover, these results indicate that our features can be used effectively in ML models to detect lncRNAs.

Genome-wide screen performance

We used our framework to scan for ncRNAs in a wide range of genomes (Table 5). To identify known annotated ncR-

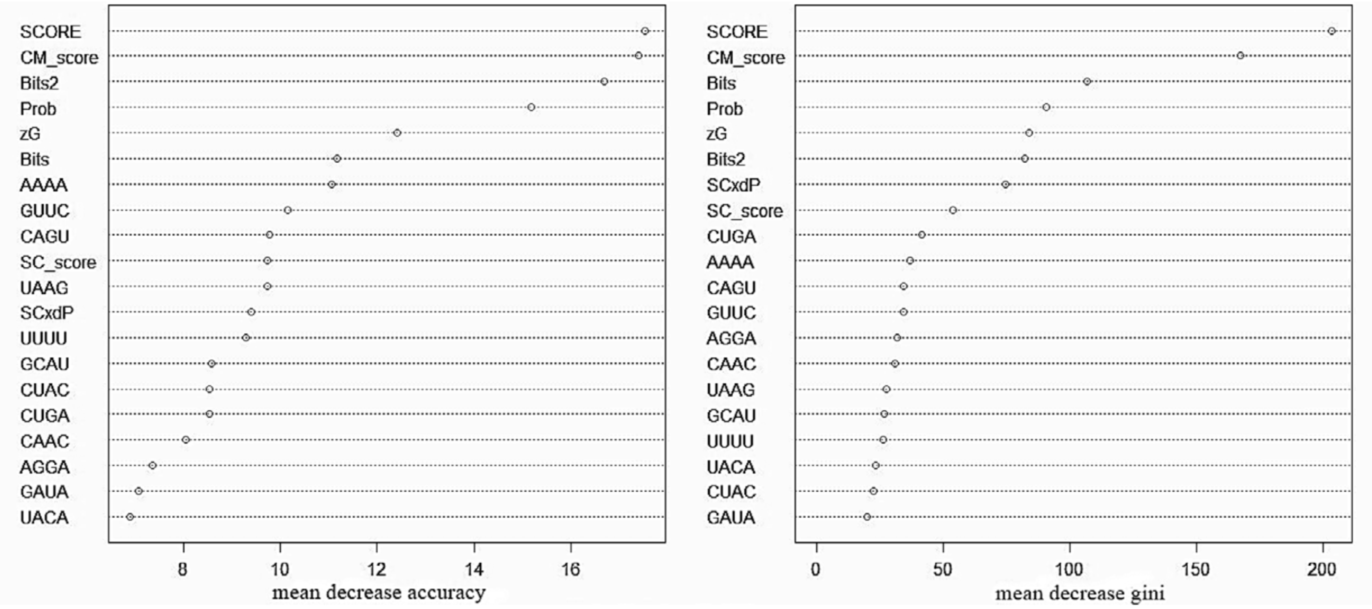

Figure 4. The variable importance based on the RF method (randomforest package). Left: permutation importance. The RF algorithm estimates the importance of a variable based on the increases in prediction error when the out-of-bag (OOB) error for that variable is permuted while all others are left unchanged. Right: gini importance. The RF algorithm estimates the decrease in impurity in the splitting criterion produced by each variable.

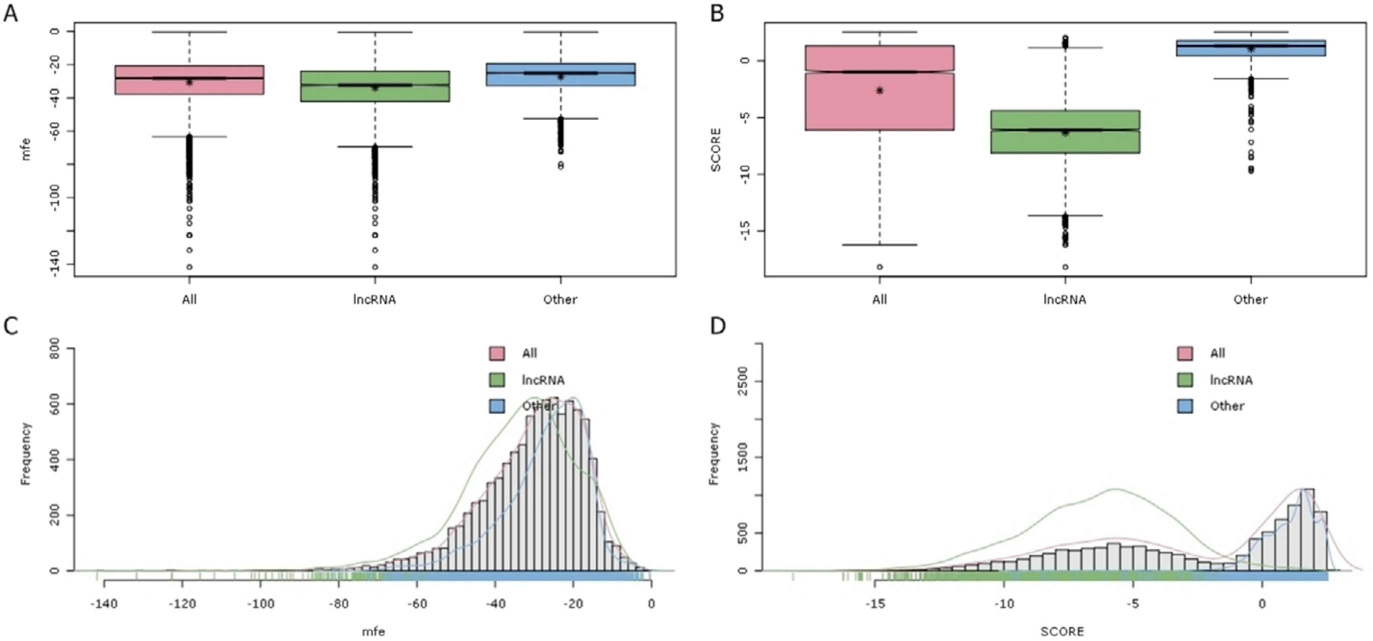

Figure 5. (**A**) and (**B**) The boxplots illustrate the spread of the training 2 dataset for the MFE feature (left) and the SCORE feature (right). The y-axis shows the values of the feature, and the x-axis shows the class of data. (**C**) and (**D**) Histogram plots of the MFE and SCORE features.

NAs in bacterial genomes, we used *M. tuberculosis H37Rv* and *P. aeruginosa PA14* as testing genomes. Our framework was able to identify known ncRNAs with a sensitivity of >90%.

Earlier methods have been applied to various prokaryotic genomes, but we wanted to test the ability of our method to identify lncRNAs and miRNAs, which are usually found in higher eukaryotes. Therefore, we obtained the following eukaryotic sequences from GenBank: *O. sativa* (chr.1) genomic sequence and five *H. sapiens* genomic regions (Region1: the 57 417 000–57 426 000 bp of chr. 20; Region2: the 49 767 700–49 779 500 bp of chr. x; Region3: the 85 770 900–85 772 300 bp of chr.2; Region4: the 27 135 000–27 139 900 bp of chr. 7; Region5: the 27 225 000–27 228 900 bp of chr. 7). Our framework also correctly identified most of the known ncRNAs in the eukaryotic genomic regions, with a sensitivity of >77%.

Table 4. Performance comparison of the model with and without the SCORE feature for detecting various Rfam lncRNA regions

Rfam	Description	No. of sequences	No. of correctly predicted sequences (%)	
			with SCORE	without SCORE
RF01890	LincRNA-p21 conserved region 2	10	9 (90)	6 (60)
RF01905	HOTAIR intron conserved region2	68	65 (95.5)	62 (91.2)
RF01909	CDKN2B antisense RNA1 intronic conserved region	59	59 (100.0)	59 (100)
RF01977	HOX antisense intergenic RNA myeloid conserved 3	76	74 (97.3)	68 (89.5)
RF02090	DAOA antisense RNA1 conserved region 1	66	58 (87.8)	50 (75.7)
RF02124	JPX transcript, XIST activator conserved region 1	78	71 (91.0)	63 (80.8)
RF02132	HOXB13 antisense RNA 1 conserved region 1	53	50 (94.3)	49 (92.5)
RF02138	HOXA11 antisense RNA 1 conserved region 2	83	81 (97.5)	78 (93.9)
RF02143	Hydatidiform mole associated and imprint region	56	53 (94.6)	48 (85.7)
RF02148	MEST intronic transcript 1	51	49 (96.1)	47 (92.2)

Table 5. Performance of our model in a genome-wide screen

Genome	Strand	Sensitivity accuracy
Mycobacterium tuberculosis H37Rv (AL123456.2)	+	90.7%
Mycobacterium tuberculosis H37Rv (AL123456.2)	–	91.9%
Pseudomonas aeruginosa PA14 (CP000438.1)	+	97.5%
Pseudomonas aeruginosa PA14 (CP000438.1)	–	94.3%
Oryza sativa, Japonica, chromosome 1 (CM000138.1)	+	92.3%
Oryza sativa, Japonica, chromosome 1 (CM000138.1)	–	86.5%
Homo sapiens 1 (Containing five known lncRNAs: GNAS_AS-1-5)	+	80% (4/5)
Homo sapiens 2 (Containing nine known miRNAs)	+	77.7% (7/9)
Homo sapiens 3 (Containing six known lncRNAs: MAT2A_A-F)	+	83.3% (5/6)
Homo sapiens 4 (Containing five known lncRNAs: HOTAIRM1_1-5)	+	100% (5/5)
Homo sapiens 5 (Containing six known lncRNAs: HOXA11_AS1_1-6)	+	83.3% (5/6)
Arthrospira platensis NIES-39 (AP011615.1, region: 1–1 200 000)	+	413[a]
Penaeus monodon mitochondria (AF217843)	+	39[a]

[a] Need to be experimentally verified.

Table 6. Performance of our model in detecting unknown verified ncRNAs in *P. aeruginosa PA14*

Regulatory class of sRNAs	No. of sRNAs	
	Tested candidates	Detected by our method
I intergenic sRNAs	19	17
II 5′-UTR sRNAs	10	9
III asRNAs	19	14
IV intergenic contains CRISPR	1	1
V 3′-UTR sRNAs	3	3
Total	52	44 (84.6%)

The tool was also evaluated for its ability to identify putative ncRNA candidates in unannotated genomes based on known ncRNA databases (33,70,71). The *P. monodon* mitochondria and the *A. platensis* genomes were used for testing. The model framework was applied to the *A. platensis* genome (NC_016640.1, genomic region: 1–1 200 000 nt; strand +) and 413 putative ncRNA candidates were identified and reported (results are shown in the Supplementary data 1). A total of 49 putative ncRNA candidates were identified as ncRNAs with high confidence (>0.85 Confidence, highlighted in bold in the Supplementary data 1). We also clustered the putative ncRNA candidates with Rfam sequences using RNAClust (72) and compared them to Rfam CM to suggest their putative functions via sequence-structure homology. The RNAClust (72) and Locarna (73) pipelines were used to detect similarity to Rfam sequences based on sequence-structure similarity (clustering trees provided upon request). To achieve maximal sensitivity with our method, the screening should be repeated with different window sizes. Consistent with a previous study (74), we also suggest using three different sizes of sliding windows (one short, 55–100 nt; one long, 125–200 nt; and one intermediate, 120 nt) to increase sensitivity because ncRNAs vary in size and structure.

Performance with unknown ncRNAs

To verify that our method has the ability to detect novel unknown ncRNAs in addition to known ncRNAs in the database, 52 sRNA candidates (in bacteria, the regulatory RNAs are called ncRNAs or sRNAs) validated by northern blot (37) were used. These candidate sequences have not been submitted to the current Rfam (ver.11). These test data are completely independent from the training data and are not part of our training dataset. To test the ability of our method to detect unknown ncRNAs, we obtained the genome sequence of *PA14* from GenBank (accession number CP000438.1) and used our method to identify sRNAs in its genome. The prediction results were compared with the experimental sRNA identification data (37), as summarized in Table 6. The sRNA candidates were classified according to functional/structural categories for regulatory RNA in bacteria (75): Class I represents trans-encoded or intergenic RNAs; Class II represents 5′-untranslated region (5′-UTR) sense RNAs; Class III represents antisense RNAs (asRNAs) or cis-encoded; Class IV represents intergenic containing CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-like array; and Class V represents 3′-untranslated region (3′-UTR) sense RNAs.

Our approach was able to predict various types of sRNA transcripts derived from 3′-UTR or 5′-UTR sense regions, antisense regions and intergenic regions. Most of the missed predictions were in the ‘antisense sRNA’ class (five of eight missed). This result is consistent with an earlier study (76) that addressed the observation that computational approaches are likely to focus on identifying ncRNAs in intergenic regions and are likely to miss small RNAs expressed from the non-coding strands of known genes and small RNAs of <50 nt. In addition, asRNAs may function mainly by complementing the coding sequences rather than being expressed as a specific sequence and/or structure-based feature (77). Moreover, our method depends on the fixed size of a sliding window (120 nt). As discussed earlier, to increase the overall sensitivity of the framework, multiple sizes of sliding windows are desirable. However, the high sensitivity of multiple sliding windows comes at the expense of computational time. We demonstrate here that our method can also detect novel ncRNA candidates. In contrast to the sequences used in the preceding section, these sRNA candidates appear to represent unannotated novel ncRNA genes identified by our tool. Taken together, we have demonstrated that our framework, in combination with the selected discriminating features, can identify both known and unknown ncRNAs.

CONCLUSION

We describe a hybrid method that uses a logistic regression model as a composite feature in an RF-based classifier model to detect various ncRNAs. The RF model has the advantage of robustness due to a bagging process and random selection of features to build the ensemble of trees. Moreover, the RF method can cope with the heterogeneous characteristics of diverse ncRNAs because its algorithm combines multiple decision trees with multiple classification rules. In addition to the robustness of the model, the composite feature incorporated as a statistical model in the RF-based classification method enhances the performance of the model for identifying lncRNA elements. The lncRNA is a challenging set of ncRNAs because of limited knowledge of lncRNA characteristics. The ncRNA identification framework proposed here exhibits high performance not only in recognizing known ncRNAs in a wide range of genomes but also in identifying novel ncRNAs in unannotated genomes. Both known and unknown ncRNAs can be identified with high accuracy. However, our scheme has some limitations. First, the size of the sliding window affects the performance of the classifier. Second, because ncRNA knowledge is limited, the framework may not be completely accurate and may need to be further refined to incorporate new ncRNA families. The logistic regression can be easily updated to incorporate new data. This composite feature may be improved in the future by using a non-linear combination of features based on a GA. We are also interested in building various logistic regression models, including separate logistic regression models for long, intermediate and short ncRNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [77–86].

FUNDING

National Research University Project of Thailand’s Office of the Higher Education Commission [54000318]; King Mongkut’s University of Technology Thonburi. Funding for open access charge: King Mongkut’s University of Technology Thonburi; National Research University Project of Thailand’s Office of the Higher Education Commission [54000318].

Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. Hum. Mol. *Genet.*, **15**, R17–R29.
2. Weinberg,Z. and Ruzzo,W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, **22**, 35–39.
3. Brosnan,C.A. and Voinnet,O. (2009) The long and the short of noncoding RNAs. *Curr. Opin. Cell Biol.*, **21**, 416–425.
4. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
5. Costa,F.F. (2010) Non-coding RNAs: Meet thy masters. *BioEssays*, **32**, 599–608.
6. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Pauli,A., Rinn,J.L. and Schier,A.F. (2011) Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.*, **12**, 136–149.
8. Managadze,D., Rogozin,I.B., Chernikova,D., Shabalina,S.A. and Koonin,E.V. (2011) Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.*, **3**, 1390–1404.
9. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. . *Nat. Rev. Genet.*, **10**, 155–159.
10. Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.
11. Gorodkin,J. and Hofacker,I.L. (2011) From structure Prediction to genomic screens for novel non-coding RNAs. *PLoS Comput. Biol.*, **7**, e1002100.

12. Ponjavic,J., Ponting,C.P. and Lunter,G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.
13. Sati,S., Ghosh,S., Jain,V., Scaria,V. and Sengupta,S. (2012) Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Res.*, **40**, 10018–10031.
14. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
15. Moran,V.A., Perera,R.J. and Khalil,A.M. (2012) Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res.*, **40**, 6391–6400.
16. Zhang,H., Chen,Z., Wang,X., Huang,Z., He,Z. and Chen,Y. (2013) Long non-coding RNA: a new player in cancer. *J. Hematol. Oncol.* **6**, 37.
17. Marques,A.C. and Ponting,C.P. (2009) Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.*, **10**, R124.
18. Bernhart,S.H. and Hofacker,I.L. (2009) From consensus structure prediction to RNA gene finding. Brief Funct. *Genomic Proteomic*, **8**, 461–471.
19. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, **22**, 1775–1789.
20. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
21. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2454–2459.
22. Coventry,A., Kleitman,D.J. and Berger,B. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 12102–12107.
23. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
24. Tran,T., Zhou,F., Marshburn,S., Stead,M., Kushner,S. and Xu,Y. (2009) De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics*, **25**, 2897–2905.
25. Saetrom,P., Sneve,R., Kristiansen,K.I., Snøve,O., Grünfeld,T., Rognes,T. and Seeberg,E. (2005) Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. *Nucleic Acids Res.*, **33**, 3263–3270.
26. Wang,C., Ding,C., Meraz,R.F. and Holbrook,S.R. (2006) PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, **22**, 2590–2596.
27. Washietl,S., Findeiß,S., Müller,S.A., Kalkhof,S., Bergen,M.V., Hofacker,I.L., Stadler,P.F. and Goldman,N. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578–594.
28. Raasch,P., Schmitz,U., Patenge,N., Vera,J., Kreikemeyer,B. and Wolkenhauer,O. (2010) Non-coding RNA detection methods combined to improve usability, reproducibility and precision. *BMC Bioinformatics*, **11**, 491.
29. Salari,R., Aksay,C., Karakoc,E., Unrau,P.J., Hajirasouliha,I. and Sahinalp,S.C. (2009) smyRNA: A Novel Ab Initio ncRNA Gene Finder. *PLoS ONE*, **4**, e5433.
30. Bao,M., Cervantes-Cervantes,M., Zhong,L. and Wang,J.T.L. (2012) Searching for non-coding RNAs in genomic sequences using ncRNAscout. *Genomics, Proteomics Bioinformatics*, **10**, 114–121.
31. Lertampaiporn,S., Thammarongtham,C., Nukoolkit,C., Kaewkamnerdpong,B. and Ruengjitchatchawalya,M. (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic Acids Res.*, **41**, e21.
32. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
33. Burge,S., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E., Eddy,S., Gardner,P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
34. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
35. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
36. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.
37. Ferrara,S., Brugnoli,M., De Bonis,A., Righetti,F., Delvillani,F., Deho,G., Horner,D., Briani,F. and Bertoni,G. (2012) Comparative profiling of Pseudomonas aeruginosa strains reveals differential expression of novel unique and conserved small RNAs. *PLoS One*, **7**, e36553.
38. Hall,M.A. (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA. pp. 359–366.
39. Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.
40. R Development Core Team. (2006) *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria.
41. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Libman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
42. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
43. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
44. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
45. Lee,M.T. and Kim,J. (2008) Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Comput. Biol.*, **4**, e1000150.
46. Freyhult,E., Gardner,P.P. and Moulton,V. (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.
47. Kong,L., Zhang,Y., Ye,Z., Liu,X., Zhao,S., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
48. Slater,G.C. (2000) *Algorithms for the Analysis of Expressed Sequence Tags*. University of Cambridge, Cambridge.
49. Breiman,L. (2001) Random forests. *Mach. Learning*, **45**, 5–32.
50. Larrañaga,P., Calvo,B., Santana,R., Bielza,C., Galdiano,J., Inza,I., Lozano,J.A., Armañanzas,R., Santafé,G., Pérez,A. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinform.*, **7**, 86–112.
51. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
52. Yang,P., Hwa Yang,Y., Zhou,B. and Zomaya,A. (2010) A review of ensemble methods in bioinformatics. *Curr. Bioinformatics*, **5**, 296–308.
53. Frank,E., Hall,M., Trigg,l. E., Holmes,G. and Witten,I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.
54. Strobl,C., Boulesteix,A., Zeileis,A. and Hothorn,T. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25
55. Strobl,C., Boulesteix,A., Kneib,T., Augustin,T. and Zeileis,A. (2008) Conditional variable importance for random forests. *BMC Bioinformatics*, **9**, 307.
56. Hothorn,T., Hornik,K. and Zeileis,A. (2006) party: a laboratory for recursive part(y)tioning. [R package version 0.9-0].
57. Lu,Z.J., Yip,K.Y., Wang,G., Shou,C., Hillier,L.W., Khurana,E., Agarwal,A., Auerbach,R., Rozowsky,J., Cheng,C. *et al.* (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.*, **21**, 276–285.
58. Babak,T., Blencowe,B.J. and Hughes,T.R. (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, **8**, 33.
59. Clote,P., Ferré,F., Kranakis,E. and Krizanc,D. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.

60. Klein,R.J., Misulovin,Z. and Eddy,S.R. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 7542–7547.
61. Schattner,P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.*, **30**, 2076–2082.
62. Larsson,P., Hinas,A., Ardell,D.H., Kirsebom,L.A., Virtanen,A. and Söderbom,F. (2008) De novo search for non-coding RNA genes in the AT-rich genome of Dictyostelium discoideum: performance of Markov-dependent genome feature scoring. *Genome Res.*, **18**, 888–899.
63. Gardner,P.P. (2009) The use of covariance models to annotate RNAs in whole genomes. *Brief. Funct. Genomic Proteomic*, **8**, 444–450.
64. Menzel,P., Gorodkin,J. and Stadler,P.F. (2009) The tedious task of finding homologous noncoding RNA genes. *RNA*, **15**, 2075–2082.
65. Zhong,C. and Zhang,S. (2012) Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment. *Nucleic Acids Res.*, **40**, 1307–1317.
66. Miler,T.L. (2009) Modular organization and composability of RNA. *Publicly Accessible Penn Dissertations*. Paper 244.
67. Bhartiya,D., Pal,K., Ghosh,S., Kapoor,S., Jalali,S., Panwar,B., Jain,S., Sati,S., Sengupta,S., Sachidanandan,C. *et al.* (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database*, **11**, doi: 10.1093/database/bat034.
68. Novikova,I.V., Hennelly,S.P. and Sanbonmatsu,K.Y. (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.*, **40**, 5034–5051.
69. William,G.J. (2009) Rattle: a data mining GUI for R. *R J.* **1**, 45–55.
70. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbø,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.
71. Kin,T., Yamada,K., Terai,G, Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.
72. Kaczkowski,B., Torarinsson,E., Reiche,K., Havgaard,J., Stadler,P. and Gorodkin,J. (2009) Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*, **25**, 291–294.
73. Will,S., Joshi,T. and Hofacker,I. (2012) locaARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, **18**, 900–914.
74. Kavanaugh,L.A. and Dietrich,F.S. (2009) Non-coding RNA prediction and verification in Saccharomyces cerevisiae. *PLoS Genet.*, **5**, e1000321.
75. Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.
76. Kawano,M., Reynolds,A.A., Miranda-Rios,J. and Storz,G. (2005) Detection of 5′- and 3′-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic Acids Res.*, **33**, 1040–1050.
77. Georg,J., Voss,B., Scholz,I., Mitschke,J., Wilde,A. and Hess,W.R. (2009) Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol. Syst. Biol.*, **5**, 305.