PLOS ONE

RESEARCH ARTICLE

LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature

Cong Pian☯, Guangle Zhang☯, Zhi Chen, Yuanyuan Chen, Jin Zhang, Tao Yang, Liangyun Zhang*

Department of Mathematics, College of Science, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China

☯ These authors contributed equally to this work.
* zlyun@njau.edu.cn

Citation: Pian C, Zhang G, Chen Z, Chen Y, Zhang J, Yang T, et al. (2016) LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature. PLoS ONE 11(5): e0154567. doi:10.1371/journal.pone.0154567

Editor: Vinod Scaria, CSIR Institute of Genomics and Integrative Biology, INDIA

Received: September 18, 2015

Accepted: April 15, 2016

Published: May 26, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work is supported by the National Natural Science Foundation of China (11571173, 11401311, 31301229) and the Natural Science Foundation of Jiangsu Province (BK20141358). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that no competing interests exist.

Abstract

As a novel class of noncoding RNAs, long noncoding RNAs (lncRNAs) have been verified to be associated with various diseases. As large scale transcripts are generated every year, it is significant to accurately and quickly identify lncRNAs from thousands of assembled transcripts. To accurately discover new lncRNAs, we develop a classification tool of random forest (RF) named LncRNApred based on a new hybrid feature. This hybrid feature set includes three new proposed features, which are MaxORF, RMaxORF and SNR. LncRNApred is effective for classifying lncRNAs and protein coding transcripts accurately and quickly. Moreover,our RF model only requests the training using data on human coding and non-coding transcripts. Other species can also be predicted by using LncRNApred. The result shows that our method is more effective compared with the Coding Potential Calculate (CPC). The web server of LncRNApred is available for free at http://mm20132014.wicp.net:57203/LncRNApred/home.jsp.

Introduction

More and more studies have indicated that protein coding genes account for less than 2% of the mammalian genome over the past decades[1–11]. A huge mass of genome that was previously regarded as "dark matter" is transcribed to non-coding RNAs (ncRNAs) [12–16]. Moreover, an increasing number of studies shows that ncRNAs have crucial and essential regulatory functions, even if it doesn't encode proteins [17]. According to the size of transcripts, ncRNAs fall into two categories, short and long ncRNAs (lncRNAs). Short ncRNAs roughly consist of small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), short-interfering RNAs (siRNAs) and short hairpin RNAs (shRNAs) [18–21]. In general, the length of short ncRNAs is shorter than 200 nt. In contrast the length of lncRNAs is longer than 200 nt [22]. As the major part of eukaryotic transcriptomes, lncRNAs have been verified to be associated with various diseases like cancers[23–30], heart failure [31–34], AIDS [35–41].

LncRNADisease database was constructed by Chen et al. [42], and contains more than 1000 lncRNA-disease entries, including 321 lncRNAs and 221 diseases from nearly 500 publications. Therefore, the identification and annotation of lncRNAs are crucial steps for understanding various regulatory mechanisms.

With the development of current experimental technology, a large number of lncNRAs have been annotated in the transcriptome. However, experimental methods have certain limits, such as the poor expression of most lncRNAs and the difficulty of enormous experimental data analysis [15,43]. Thus, it is essential to develop computational methods to identify lncRNAs from the transcriptome accurately and quickly.

There are many methods to identify ncRNAs [44–55]. For instance, Liu et al. introduced a tool called CONC (coding or non-coding) based on support vector machines (SVM) to classify transcripts according to a hybrid feature set [56]. This feature set consists of alignment entropy, amino acid composition, predicted percentage of exposed residues, predicted secondary structure content, number of homologs from database searches, compositional entropy and peptide length. However, CONC is slow for abundant datasets, and its web server is not available. Moreover, the outputs of CONC does not provide related information. Thus, Lei et al. developed a online software called Coding Potential Calculator (CPC) to identify the protein-coding potential of transcripts based on six biologically meaningful sequence features [57]. Compared with CONC, CPC is more accurate and run faster. It also has a more friendly web interface. Lin et al. present a software named PhyloCSF to distinguish protein coding by analyzing a multi-species nucleotide sequence alignment. It is a method of comparative genomics [58]. Their results indicate PhyloCSF is applicable for evaluating the protein-coding potential of transcript models or individual exons. Lei Sun et al. [59] develop a tool named LncRScan-SVM by integrating features derived from gene structure, transcript sequence, potential codon sequence and conservation. Kun Sun et al. [60] use one conservation, two Open Reading Frame (ORF) and seven nucleotide sequence features to construct a support vector machine classifier (iSeeRNA) for the identification of long intergenic non-coding RNAs (lincRNAs). Liguo Wang et al. [61] build a tool named Coding Potential Assessment Tool (CPAT), which can rapidly identify coding and non-coding transcripts. CPAT uses a logistic regression model built with four sequence features: open reading frame coverage hexamer usage bias, Fickett TESTCODE statistic and open reading frame size. However, the above tools are not suitable for classifying lncRNAs, which contain long putative Open Reading Frame (ORF) or short protein-like subsequences [62,63]. To overcome the challenge, Liang Sun et al. [64] develop the Coding-Non-Coding Index (CNCI) software, a powerful tool, by profiling adjoining nucleotide triplets (ANT), to effectively recognize protein-coding and non-coding sequences.

In this paper, we introduce a generalized classifier based on an integrated algorithm called random forest (RF) to distinguish lncRNAs from protein-coding transcripts. Besides, we propose three new features, which are MaxORF, RMaxORF and SNR. A new hybrid feature set with 89 dimension can be formed by combining 86 sequence features and the three new features just mentioned together. The results show that the first three important features are MaxORF, SNR and RMaxORF. At the same time, we develop a user-friendly web server named LncRNApred and compare the LncRNApred with Coding Potential Calculator(CPC). LncRNApred demonstrates better performance compared with CPC.

Materials and Methods

Datasets

The NONCODE version 3.0 [65] (http://www.noncode.org/NONCODERv3/) currently contains 33665 non-redundant lncRNA sequences of human. In this paper, 33665 lncRNAs are

selected as positive samples. For the negative samples, protein-coding transcripts are extracted from UCSC database [66] (http://hgdownload.soe.ucsc.edu/downloads.html), from which 38268 mRNAs can be obtained. After removing the mRNAs with length of <20000 and >200, 38229 mRNA sequences are retained.

In order to avoid over-fitting, some redundant samples should be removed. Therefore, we select 2033 lncRNAs and 2031 mRNAs from 33665 lncRNAs and 38229mRNAs respectively as the training dataset by Self Organizing Feature Map (SOM) [67]. These training samples can effectively describe the whole data. The remaining samples are used to assess our model.

In order to test the generalization of our RF classifier, 35851 lncRNAs and 27728 mRNAs of mouse are obtained from the database of NONCODE version 3.0 and UCSC database respectively [65,66]. In addition, 2551 lncRNAs of other species are downloaded from NONCODE version 3.0. Repetitive sequences and those with other letters except for 'A', 'a', 'C', 'c', 'G', 'g', 'T', 't', 'U', 'u' are removed. The remaining 2113 lncRNAs of other species and above samples of mouse are also used to evaluate our classifier.

The selection of training samples

The accuracy of a RF classifier depends highly on the selection of training samples. So we should select representative samples to construct training dataset. In this paper, we use a clustering method to obtain representative samples. In order to find an appropriate clustering method, we analysis four different cases: (1) k-means clustering (2) hierarchical clustering (3) SOM (Self Organizing Feature Map) clustering (4) non-clustering. In the first three cases, we use three different clustering methods to select 2000 lncRNAs from 33665 lncRNAs and 2000 mRNAs from 38229 mRNAs as the training dataset. In the fourth case, we randomly select 2000 lncRNAs from 33665 lncRNAs and 2000 mRNAs from 38229 mRNAs as the training dataset of RF. Therefore, four RF models can be constructed respectively. As shown in Table 1, the classification performance after the pretreatment of clustering is better than that without the pretreatment of clustering. Besides, the results also show that SOM clustering algorithm outperforms the other three cases. According to the above discussion, Self Organizing Feature Map (SOM) is used to select representative samples in our paper.

SOM is a type of Artificial Neural Network (ANN). In 1990, Teuvo Kohonen proposed SOM [67] and effectively used it to classify input vectors according to the way they are grouped in the input space. SOM is different from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as Back Propagation Artificial Neural Network), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

Like most artificial neural networks, SOMs operate in two modes: training and mapping. "Training" builds the map using input examples (a competitive process), while "mapping" automatically classifies a new input vector.

A SOM consists of components called neurons. Associated with each node is a weight vector of the same dimension as the input data vector. The self-organizing map describes a mapping

Table 1. The classification performance after the pretreatment of clustering.

Method	S_p (%)	S_n (%)	ACC (%)
RF	91.2	90.4	90.8
K-means+RF	92.4	91.2	91.8
Hierarchical+RF	92.6	91.4	92.0
SOM+RF	**93.4**	**92.5**	**92.9**

doi:10.1371/journal.pone.0154567.t001

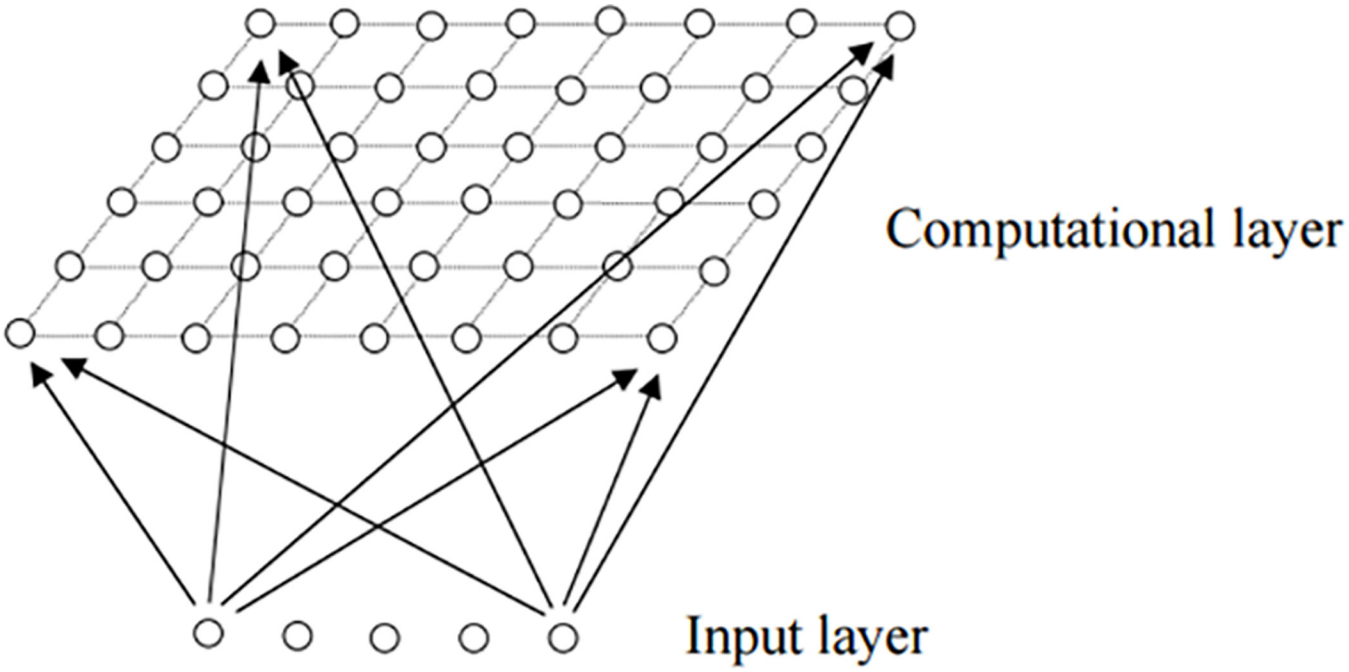


Fig 1. Two dimensional SOM neural network model.

doi:10.1371/journal.pone.0154567.g001

from a higher-dimensional input space to a lower-dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector. Fig 1 describes two dimensional SOM neural network model. All neurons in the competition layer are fully connected.

The main SOM learning algorithm can be described as follows:

Let $X = [x_1, x_2, \cdots, x_m]$, be the input vector. We construct two-dimensional network with n output node. Set w_{ij} be the weight vector connecting the ith input node and the jth output nodes.

(1) Initialization of weights.

The weights (w_{ij}) should be initialized randomly. The value of every weight must be different.

(2) Calculate the distance between the input vector and weight vector.

$$d_j = \sum_{i=1}^{m} (x_i(t) - w_{ij}(t))^2. \quad (1)$$

$x_i(t)$ represents the value of input vector x at time t.

(3) Select the winning neuron $i(x)$.

Select the nearest unit as winner. The neuron i is the winning neuron.

$$i(x) = \min_j (d_j). \quad (2)$$

(4) Adjust the connection weight vector of the output node.

Update weight vector of the SOM according to the update function:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) h_{j,i(x)}(t)(x(t) - w_{ij}(t)). \quad (3)$$

where $\eta(t)$ is a learning efficiency function. To ensure the convergence of the learning process, $\eta(t)$ is monotonically decreasing. $h_{j,i(x)}$ is a neighborhood function of the winning neuron.

(5) Repeat the step (2) to (4), and update the learning parameters, until a certain stopping criterion is met.

We use the following steps to select the training dataset.

Given a dataset $Q = \{x_i \mid x_i \in R^n, i = 1,\ldots, N\}$, K is the number of neurons in the competitive layers.

Step 1: The N samples are imported to the input layer of SOM.

Step 2: Calculate the number of training samples for every neurons in the competitive layers and record them as $w = [w_1, w_2, \cdots, w_K]$,

Step 3: Let L be the number of training dataset. Randomly select O_i samples from the ith neuron as the training samples. O_i can be calculated by the following formula

$$O_i = \left\lceil \frac{w_i}{N} \times L \right\rceil, \tag{4}$$

where $\lceil A \rceil$ rounds the element of A to the nearest integers greater than or equal to A.

Step4: The $(O_1+O_2+\cdots+O_K)$ samples of training dataset can be obtained.

In this study, we choose 8×8 neurons in the competitive layers and 2000 training samples. Fig 2 shows the distribution in the 64 neurons of lncRNAs or mRNAs. Each hexagon represents one neuron and there are 64 neurons in total. Every digit inside the hexagon is the number of lncRNAs (or mRNAs) which belong to the corresponding neuron. All neurons in the competition layer are fully connected. We use above steps to choose training samples. For example, neuron node in the lower right corner of Fig 2 is 385 and the total number of mRNAs is 38229. Thus, we should randomly select $2000\times\lceil 385/38229 \rceil$ samples from that neuron node. The final number of mRNA training samples N_{mRNA} and LncRNA training samples $N_{\mathrm{ln}\ cRNA}$

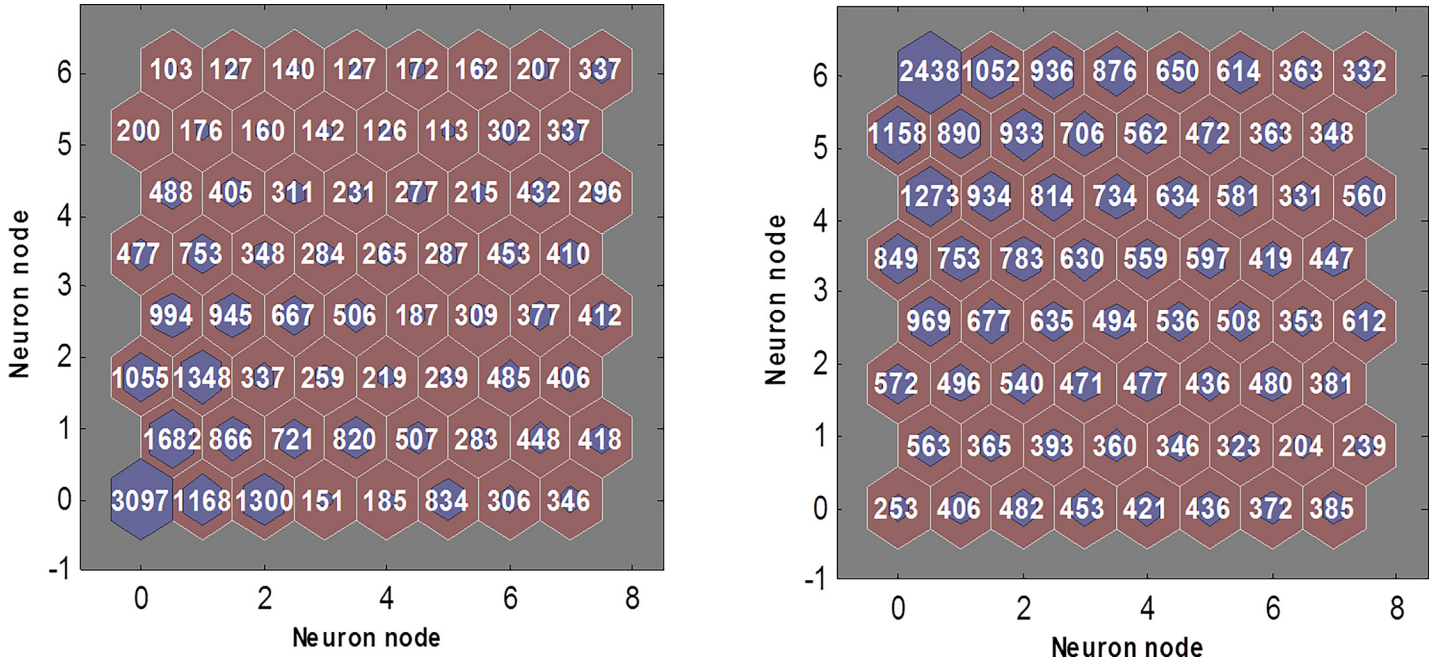


Fig 2. The result of SOM clustering. The left side represents the distribution in the 64 neurons of lncRNAs. Every digit of the hexagon is the number of lncRNAs which belong to one class. The right side represents the distribution in the 64 neurons of mRNAs, and every digit of hexagon is the number of mRNAs which belong to one class.

doi:10.1371/journal.pone.0154567.g002

are as follows:

$$N_{mRNA} = 2000 \times \left(\left\lceil \frac{2438}{38229} \right\rceil + \left\lceil \frac{1052}{38229} \right\rceil + \cdots + \left\lceil \frac{372}{38229} \right\rceil + \left\lceil \frac{385}{38229} \right\rceil \right) = 2031 \quad (5)$$

$$N_{\mathrm{lncRNA}} = 2000 \times \left(\left\lceil \frac{103}{30740} \right\rceil + \left\lceil \frac{127}{30740} \right\rceil + \cdots + \left\lceil \frac{306}{30740} \right\rceil + \left\lceil \frac{346}{30740} \right\rceil \right) = 2033 \quad (6)$$

Feature

Signal to noise ratio (SNR). Let $s[n]$ be a sequence of length N. Let $I = \{A,G,C,T\}$, for any $b \in I$.

$$u_b[n] = \begin{cases} 1, & S[n] = b \\ 0, & S[n] \neq b \end{cases} \quad n = 0, 1, 2, \cdots, N-1 \quad (7)$$

There are four binary indicator sequence $\{u_b[k]\}$, $b \in I$, which is called Voss mapping [68]. For instance, given a DNA sequence as follows:

$$5' \ldots ATCTCACTGGT \ldots 3'$$

the Voss mapping of this DNA sequence can be represented as

$$u_T = \{\ldots 0,1,0,1,0,0,0,1,0,0,1 \ldots\}, \quad u_A = \{\ldots 1,0,0,0,0,1,0,0,0,0,0 \ldots\},$$
$$u_C = \{\ldots 0,0,1,0,1,0,1,0,0,0,0 \ldots\}, \quad u_G = \{\ldots 0,0,0,0,0,0,0,0,0,1,1,0 \ldots\}.$$

Using Discrete Fourier Transform (DFT) on the indicator sequences respectively, we get for $b \in I$,

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-i\frac{2\pi nk}{N}}, \quad k = 0, 1, \cdots, N-1. \quad (8)$$

There are four complex sequences ($\{U_b[k]\}$, $b \in I$) in total. The power spectrum of the whole sequence is defined as $\{P[k]\}$:

$$P[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_G[k]|^2 + |U_C[k]|^2, \quad k = 0, 1, \cdots, N-1 \quad (9)$$

Given a sequence, the power spectrum curve can be obtained by (9). In Fig 3, an obvious peak appeared at $N/3$ in the power spectrum curve of the mRNA sequence, while there is no peak in the lncRNA sequence. This statistical phenomenon is known as the period-3 behavior [69]. It was proved that the 3-base periodicity is mainly caused by the unbalanced nucleotide distributions in a DNA sequence [70,71,72,73]. The nucleotide distribution in the three codon positions is unbalanced in a coding sequence, while in a non-coding sequence, the nucleotides distribute uniformly in the three codon positions. The main reason of this phenomenon is that proteins prefer special amino acid and thus nucleotide usage in a coding region is highly biased.

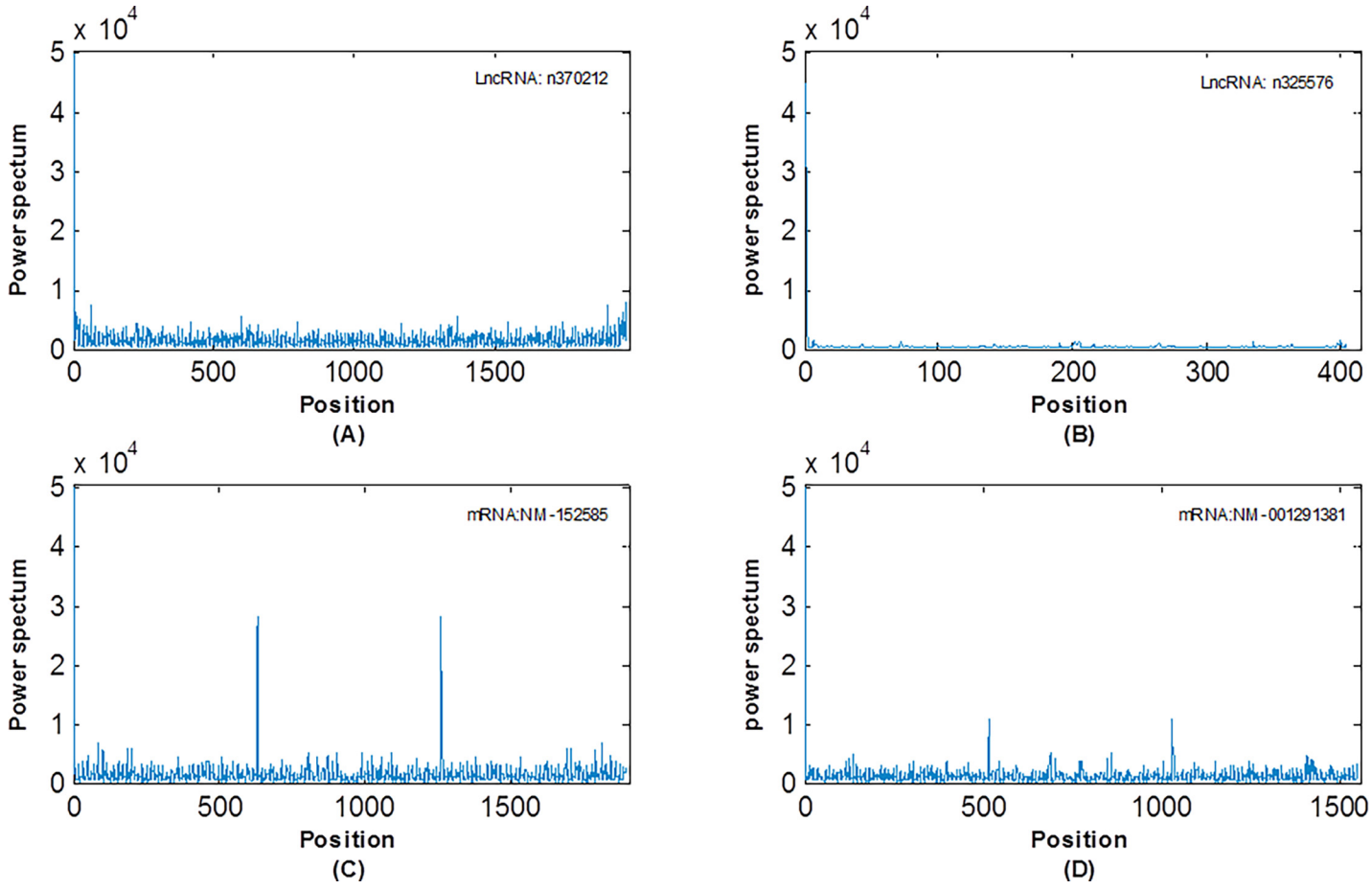


Fig 3. Power spectrum of mRNAs and lncRNAs. (A) and (B) represent the power spectrum of two different lncRNAs, and (C) and (D) represent power spectrum of two different mRNAs.

doi:10.1371/journal.pone.0154567.g003

Signal to noise ratio (SNR) is defined as following:

$$SNR = \frac{P[\frac{N}{3}]}{\bar{E}} \tag{10}$$

$$\bar{E} = \frac{\sum_{k=0}^{N-1} P[k]}{N}, \tag{11}$$

where $\bar{E}$is the mean of the total power spectrum of the whole sequence [69].

SNR not only shows the relative height of the spectrum peak, but also reflects the 3-periodic property. As shown in Fig 4, the white boxes on the bar graph represent the number of mRNA (or lncRNA) in each bar. The mean of SNR of mRNAs and lncRNAs are 7.43 and 2.06 respectively. Besides, we calculate that 72.7% (24488/33665) SNR of lncRNAs are less than 2. On the contrary, 89% (34020/38229) SNR of mRNAs are greater than 2. The P-value is 7.3123e-115 by Student's t-test. The result shows that there are obvious differences in the SNR between the positive samples and negative samples. Therefore, SNR can be used to distinguish lncRNA and mRNA as an important feature.

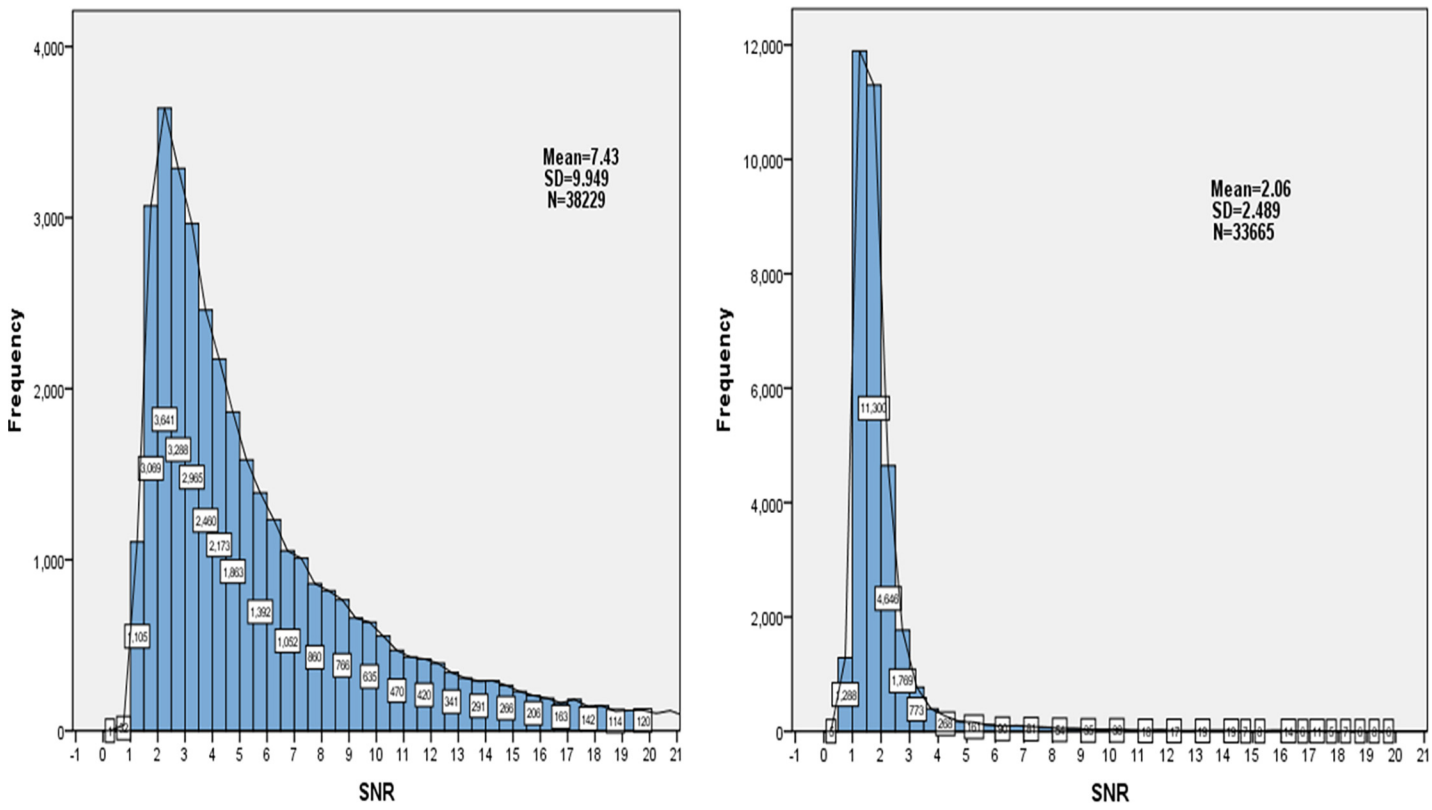


Fig 4. The distribution of SNR. The left side represents the SNR distribution of 38229 mRNAs, and the right side represents SNR distribution of 33665 lncRNAs.

doi:10.1371/journal.pone.0154567.g004

Open reading frame (ORF). Compared with long non-coding transcripts, protein coding transcripts are more likely to have a long ORF. Therefore, we select two ORF features to distinguish lncRNAs and protein coding transcripts. One is the length of the longest ORF (MaxORF) in the three forward frames, and the other is the normalized MaxORF (RMaxORF).

$$RMaxORF = \frac{MaxORF}{L}, \tag{12}$$

where L is the length of sequence.

Sequence features. In this work, 4 1-mer strings, 16 *2*-mer strings and 64 *3*-mer strings are used to identify lncRNA and mRNA. Besides, the length of sequence (Length) and (G+C)% are selected as two sequence features.

Feature selection

For a lncRNA sequence or mRNA sequence, we combine the 1 dimensional SNR feature, 2 dimensional ORF features and 86 dimensional sequence features to get a hybrid feature vector with 89 dimension. However, not every feature contributes to the classification accuracy. Golub et al. [74] use the feature score criterion (FSC) to calculate the score of each feature, and rank them in descending order. The first p features are selected as the information features. Setting $p<n$ (n is the dimension of features), we need to determine the optimal p value by the experimental results. As shown in Table 2, the second line represents the performance of RF model with the top 5 features. The Sensitivity (Sn) and Specificity (Sp) are 91.2% and 90.2%

Table 2. Effect of the number of features on the classification accuracy rate of V-ELM.

Number of features (p)	Sn (%)	Sp (%)
4	91.0	89.1
5	91.2	90.2
10	91.5	90.7
15	92.4	90.9.
20	92.6	91.0
25	93.1	91.6
30	**93.4**	**92.5**
35	93.2	92.1
40	93.1	92.0
45	93.4	92.2
50	93.3	92.3
55	93.4	92.1
60	93.2	92.4
86	92.9	92.3

doi:10.1371/journal.pone.0154567.t002

respectively. The experimental results show that the performance of RF model is relatively stable while p>30. At the same time, the accuracy of RF classifier reaches maximum when p = 30, and the Sensitivity (Sn) and Specificity (Sp) are 93.4% and 92.5% respectively. Therefore, we choose p = 30 as the information feature set of RF classifier.

On the premise of the optimal classification accuracy, the minimum value of p is selected. The score of each feature can be obtained by the following formula.

$$FSC(g_i) = \left| \frac{\mu_i^+ - \mu_i^-}{\sigma_i^+ + \sigma_i^-} \right|, \tag{13}$$

where $\mu_i^+(\mu_i^-)$ and $\sigma_i^+(\sigma_i^-)$ are the mean and standard deviation respectively of the feature of g_i in the positive (negative) class samples. The higher the *FSC* score is, the stronger classification ability the feature has.

As shown in Fig 5, a set of 30 features from the 89 features was selected by FSC, including MaxORF, RMaxORF, Length, SNR, CG%, CGG%, GC%, CCG%, GCG%, CGC%, GCC%, G%, (G+C)%, TCG%, CGA%, A%, GGC%, TAG%, CC%, TCT%, CCC%, C%, T%, TAA%, GG%, TA%, ATA%, ACG%, CGT%, and AT%. We find that the FSC differences of 30 features between lncRNAs and mRNAs are apparent, especially the features of MaxORF, RMaxORF, SNR and Length. In addition, except for MaxORF, RMaxORF, SNR and Length, the Sn and Sp for top four features are 91% and 89.1% respectively. We mark the following 8 features (CG%, CGG%, GC%, CCG%, GCG%, CGC%, GCC%, G%, (G+C)%) in red. We find that these features only relate to the nucleotide of 'C' or 'G'. In order to visualize the spread of the lncRNAs and mRNAs for the top 13 features, graphical boxplots are shown in Fig 6.

Prediction System Assessment

For a prediction problem, a classifier can classify an individual instance into the following four categories: false positive (F_P), true positive (T_P), false negative (F_N) and true negative (T_N). The total prediction accuracy (*ACC*), Specificity (S_p), Sensitivity (S_n) and Mathew's correlation

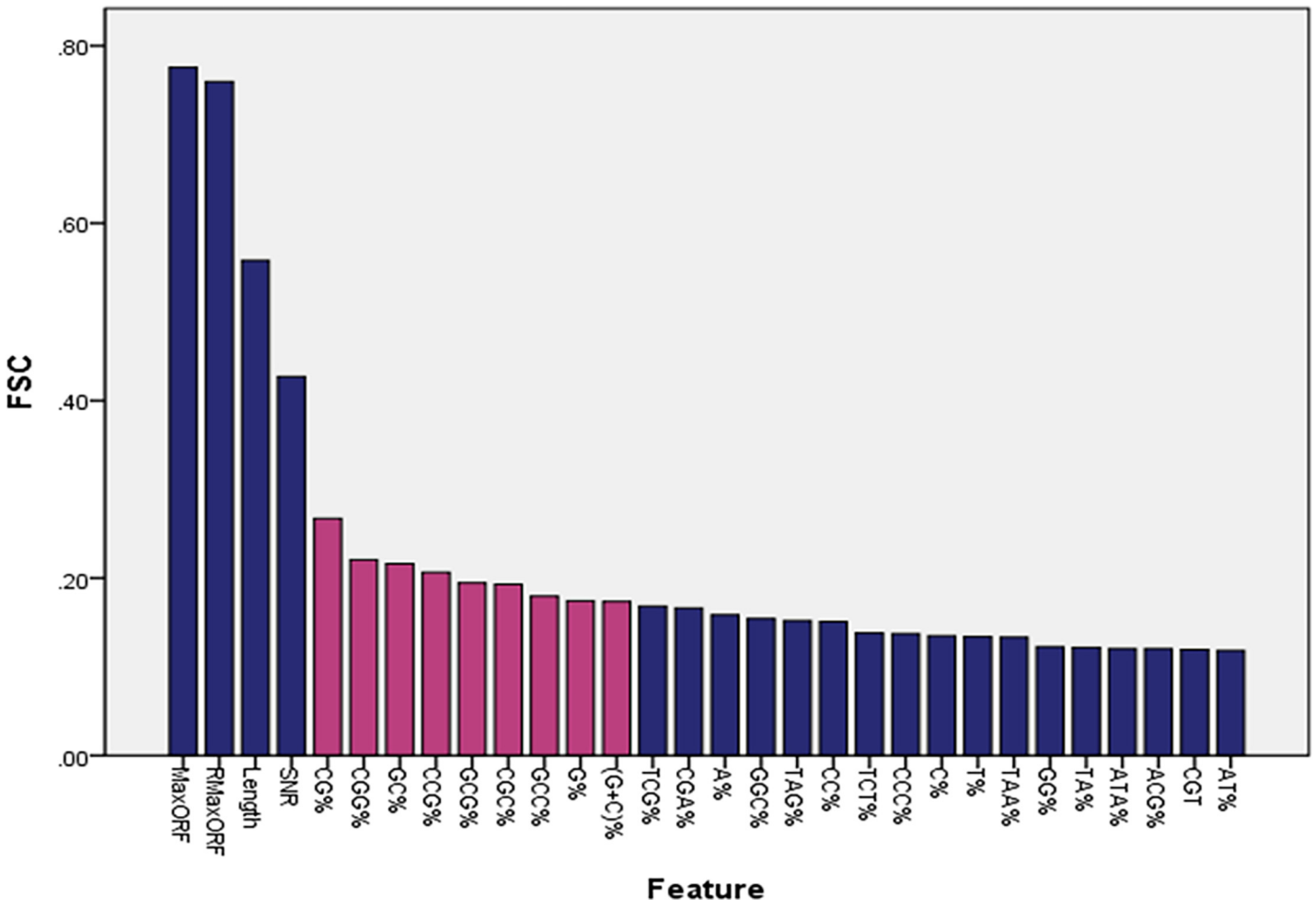


Fig 5. The bar chart shows the top 30 of FSC score.

doi:10.1371/journal.pone.0154567.g005

coefficient (*MCC*) [75] for assessment of the prediction system are given by

$$S_n = \frac{T_P}{T_P + F_N} \tag{14}$$

$$S_p = \frac{T_N}{T_N + F_P} \tag{15}$$

$$ACC = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{16}$$

$$MCC = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P) \times (T_N + F_N) \times (T_P + F_N) \times (T_N + F_P)}} \tag{17}$$

where T_P is the number of lncRNAs identified correctly, F_N the number of lncRNAs identified incorrectly, T_N the number of mRNAs identified correctly, and F_P the number of mRNAs identified incorrectly.

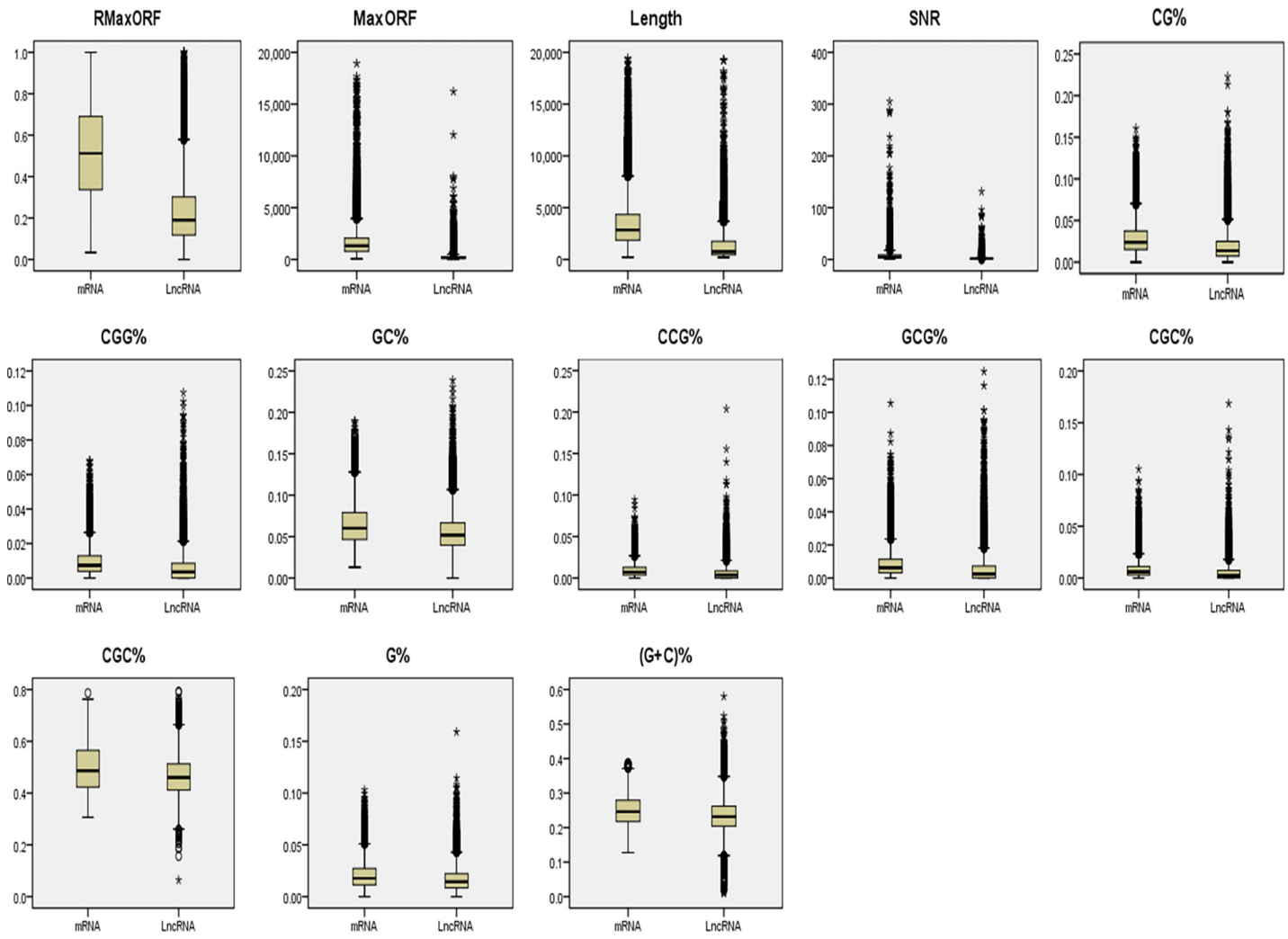


Fig 6. Boxplots of the top 13 features: MaxORF, RMaxORF, SNR, Length, CG%, CGG%, GC%, CCG%, GCG%, CGC%, GCC%, G% and (G+C)%. For each plot, the left side represents the mRNA, and the right side represents lncRNA.

doi:10.1371/journal.pone.0154567.g006

Results and Discussion

Identification framework for lncRNAs

The statistical results show that the smallest MaxORF of 38268 mRNAs and 33665 lncRNAs are 54 and 0 respectively. However, the sequences with short ORF usually do not encode proteins. Therefore, we consider that the sequence with MaxORF<54 is regarded as a lncRNA. The workflow of lncRNAs identification model is illustrated in Fig 7. First, 30 dimension features of lncRNAs and mRNAs can be extracted. The lncRNAs with MaxORF>54 are selected as positive dataset. The mRNAs with length ≥200 and length < 20000 are selected as negative dataset. Second, we select representative 2033 lncRNAs and 2031 mRNAs as training samples by the SOM algorithm. The remaining data are used to test the model. Finally, a RF model is constructed based on the training dataset. In addition, we also use other species besides human beings with MaxORF>54 to test our RF classifier except for human. The sequences with MaxORF<54 are directly predicted to be lncRNAs.

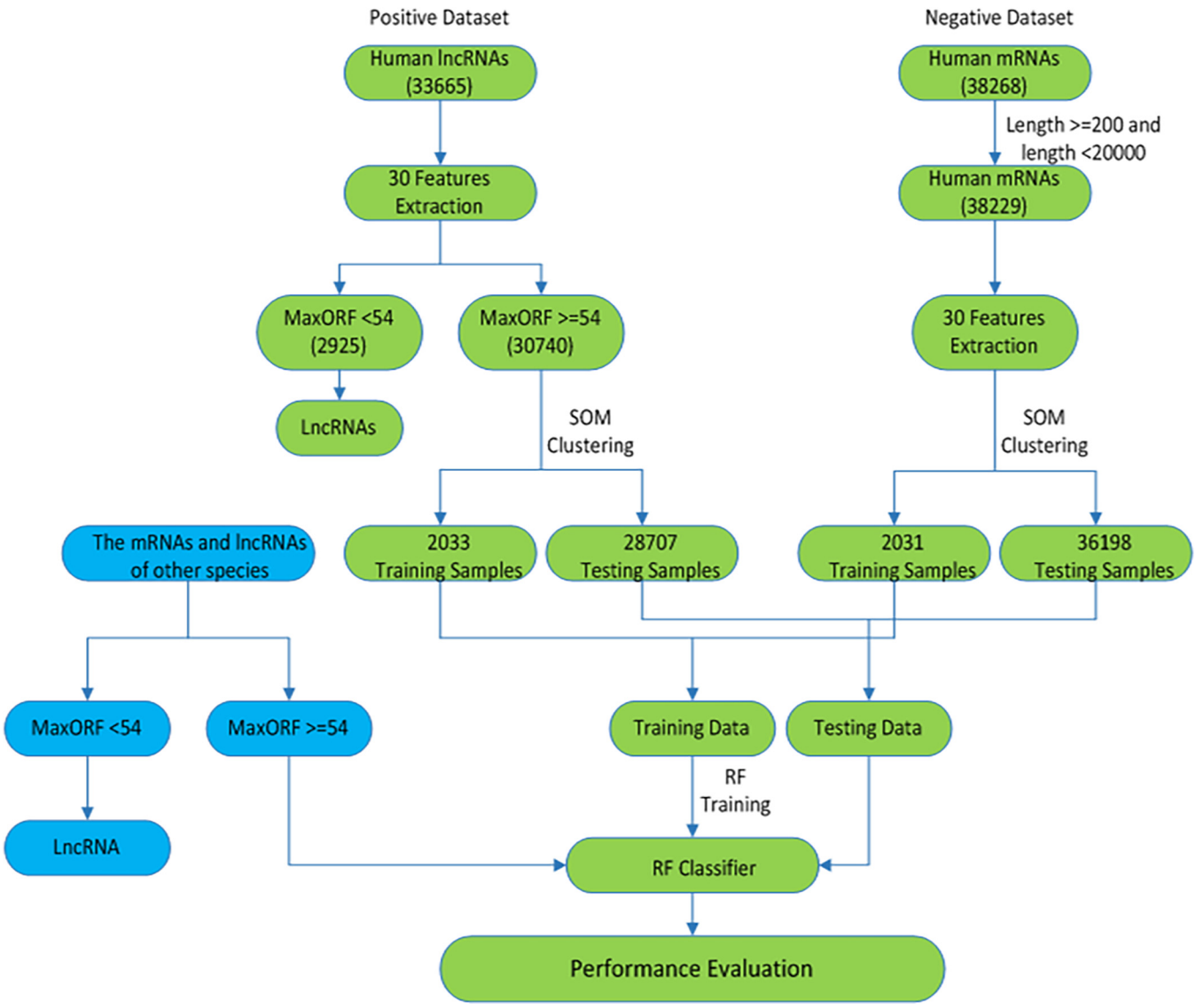


Fig 7. The workflow of lncRNAs identification

doi:10.1371/journal.pone.0154567.g007

Selection of machine learning algorithms

In general, the performance of machine learning algorithms depends on the content of research. Every algorithm has its own advantage. Therefore, we construct three different classifiers by using three algorithms based on the same training dataset to evaluate their performances. The results show that RF algorithm outperforms the two other algorithms for the identification of lncRNAs and mRNAs. To visualize the performance of those three algorithms, we generate ROC curves in Fig 8. The Area Under the Curve (AUC) measures the performance of an algorithm under different thresholds. On average, the AUC of the RF algorithm is about 0.9738. Compared with the AUC of SVM (0.9621) and ANN (0.9649), the robustness of RF model is more obvious. So we use the RF algorithm as the classified model in this work.

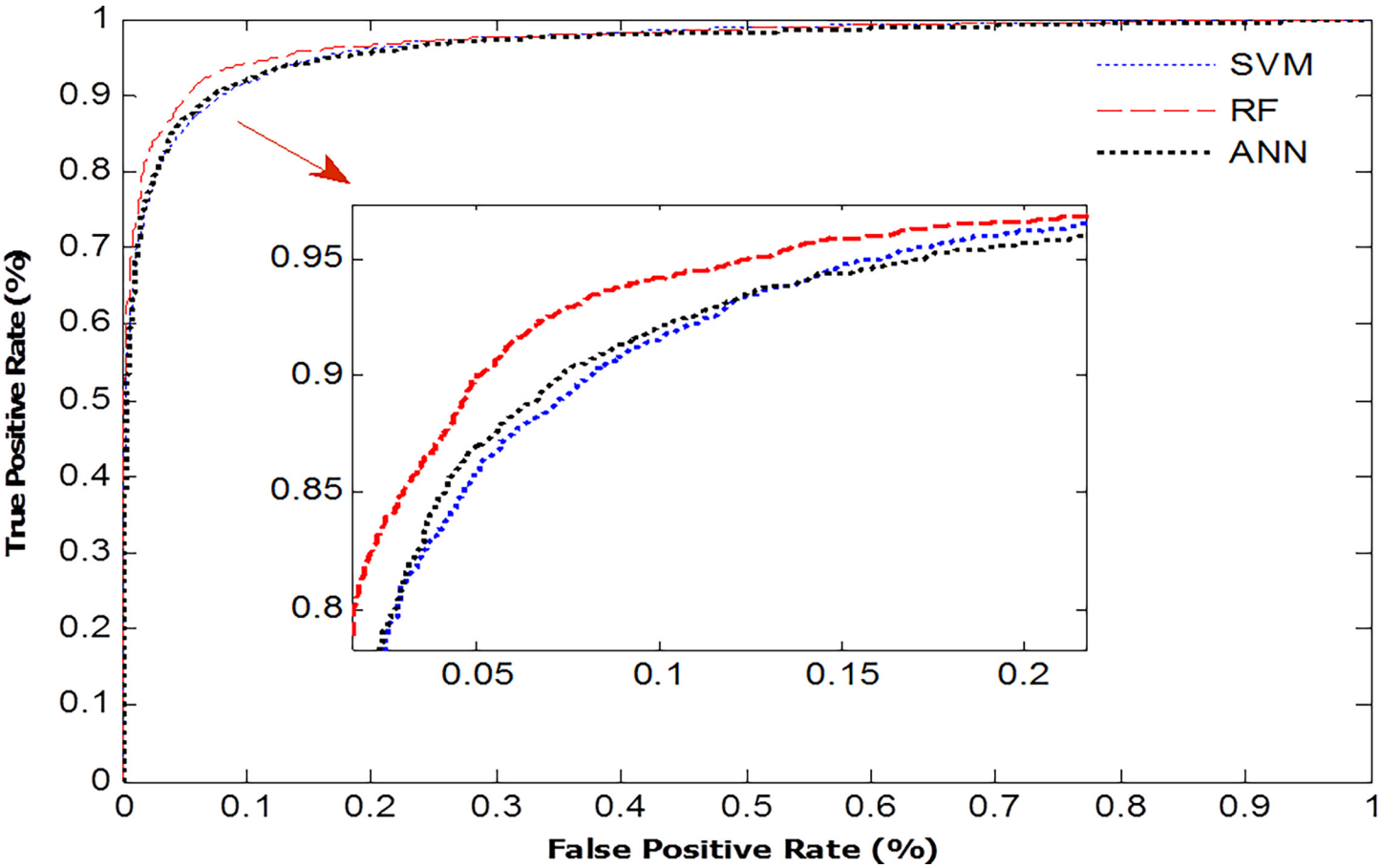


Fig 8. The ROC curves of three different classifiers.

doi:10.1371/journal.pone.0154567.g008

The acquiescent parameters C and g of support vector machine (SVM) are 2 and 1 respectively. In order to improve the accuracy of the identification, the optimal parameters of SVM are 1.97062 and 0.061 by the method of the particle swarm optimization (PSO).

In this paper, we use an artificial neural network (ANN) algorithm called voting based extreme learning machine (V-ELM) as the method of comparison. ELM is a kind of quick training algorithms of generalized SLFNs [76,77]. More and more researchers are interested in this method. The hidden layer parameters of SLFNs do not need to be tuned. ELM provides better generalization performance at a much faster learning speed. Because random parameters of the hidden layer nodes are used and remained unchanged during the training process, some samples may be misclassified, especially for those with position close to the classification boundary. In order to avoid this problem and improve the classification performance of ELM, Gao. et al. [78] proposed a new algorithm called voting based extreme learning machine (V-ELM) by incorporating multiple independent ELMs and making decisions with a majority voting method. We select N = 300 as the number of hidden layer nodes in the V-ELM model.

Random forest is an ensemble learning method by constructing multitude of decision trees. This algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler [79]. Thus "Random Forests" became their trademark. The advantage of a RF algorithm is the robustness provided by random feature selection and the bootstrap aggregating technique [80]. In this paper, we choose N = 300 as the decision trees in our RF model.

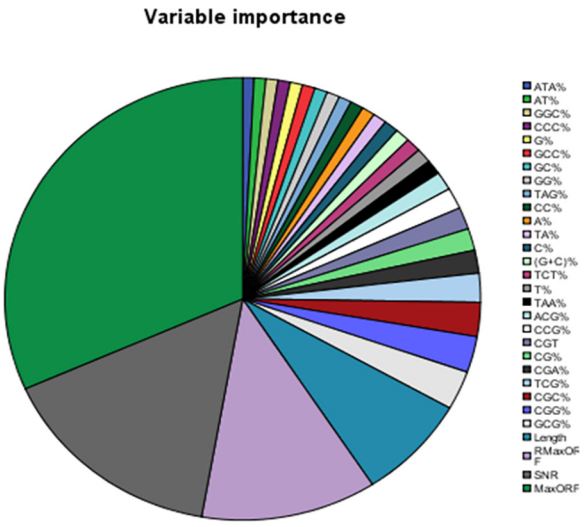


Fig 9. The importance of feature variable.

doi:10.1371/journal.pone.0154567.g009

Importance of each feature variable

In order to determine those features which play an important role in the identification of lncRNAs, we use the pie chart based on permutations to show the importance of each feature variable. The RF model can estimate the importance of a feature based on the increases in prediction error when the out-of-bag (OOB) error for that feature is permuted while other features are unchanged. As shown in Fig 9, the size of the area represents the level of the feature importance. We find that the first four important features are MaxORF, SNR, RMaxORF and Length. This chart shows that newly proposed feature can improve the prediction accuracy of lncRNAs.

Performance evaluation

In this paper, we select 2033 lncRNAs and 2031 mRNAs of human as the training samples by SOM algorithm (S1 and S2 Tables). The remaining 28707 lncRNAs and 36198 mRNAs (S3 and S4 Tables) are used to assess our RF model. As shown in Table 3, the accuracy of lncRNAs and mRNAs are 93.42% (26818/28707) and 92.5% (33483/36198) respectively. Besides, 35851 lncRNAs and 27728 mRNAs (S5 and S6 Tables) of mouse are downloaded from from the database of NONCODE version 3.0.

After removing the lncRNAs of mouse with MaxORF<54, the remaining 35373 lncRNAs and 27728 mRNAs are used to estimate the RF model. Similarly, our RF classifier correctly predicts 95.27% (33699/37373) lncRNAs and 92.7% (25921/27728) mRNAs for the mouse testing dataset.

To further assess the performance of RF model, we download 2113 other species lncRNAs other species from database of NONCODE version 3.0. The last line of Table 3 shows the prediction results of 2113 lncRNAs from other species. The accuracy is 97.78% (2066/2113). These results further indicate the high accuracy of RF classifier for the identification of lncRNAs. What's more, our RF model just needs the training samples of human beings.

Comparison with other methods

In this paper, we compare the LncRNApred with Coding Potential Calculator (CPC). CPC can distinguish coding from noncoding transcripts with high accuracy by using Support Vector

Table 3. The performance of our RF model LncRNApred.

Species	Positive (lncRNAs)	Negative (mRNAs)	Sn	Sp	ACC	MCC
Human	28707	36198	93.42 (26818/28707)	92.5 (33483/36198)	92.96	0.8569
Mouse	35851	27728	95.27 (33699/35851)	93.48 (25921/27728)	94.3	0.8880
Other species	2113	0	97.78 (20668/2113)	0	97.78	0

doi:10.1371/journal.pone.0154567.t003

Machine (SVM) based on six biologically meaningful sequence features. The feature set includes three ORF features (LOG-ODDS SCORE, COVERAGE OF THE PREDICTED ORF, INTEGRITY OF THE PREDICTED ORF) and three sequence alignment features (NUMBER OF HITS, HIT SCORE, FRAME SCORE). In order to compare these two methods, we use the same test dataset which includes 28707 lncRNAs and 36198 mRNAs of human, 35373 lncRNAs and 27728 mRNAs of mouse, 2113 lncRNAs of other species. As shown in Table 4, LncRNApred demonstrates the best performance measured by MCC followed by CPC. While LncRNApred and CPC are applied on human dataset, the values of MCC are 0.8569 and 0.7687 respectively. When LncRNApred and CPC are applied on mouse dataset, the values of MCC are 0.8880 and 0.7520 respectively. Additionally, LncRNApred shows the highest specificity compared to CPC. Although the LncRNApred displays a lower sensitivity, CPC shows a higher false positive rate. A lot of lncRNAs are predicted to be the mRNAs by using CPC.

Web implementation

In this paper, we develop a user-friendly web server named LncRNApred. It is available for free at http://mm20132014.wicp.net:57203/LncRNApred/home.jsp (Fig 10). LncRNApred provides trained RF model based on the training data of human beings. The input of LncRNApred can be a sequence or a fasta file (Fig 10A). The output include sequence ID, Non-coding score, predicted result and the information of features (Fig 10B).

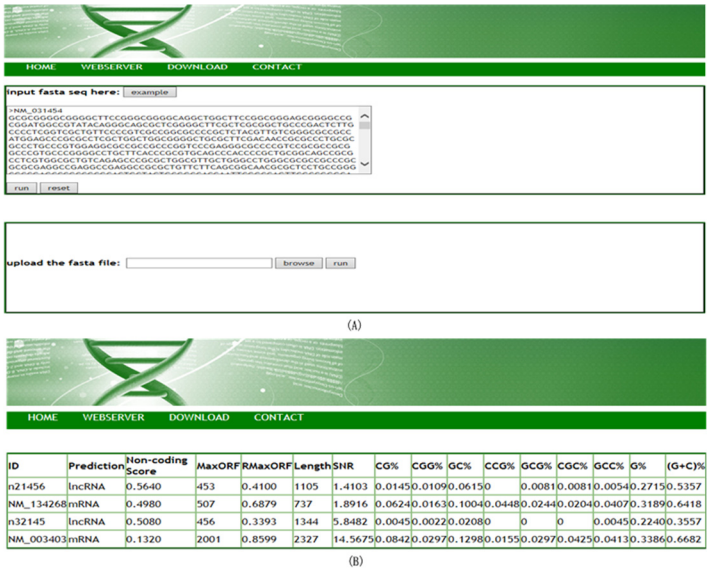
Conclusion

Identification of lncRNAs is the first step to understand the various of regulatory mechanisms. In this paper, we introduce three new features, including MaxORF, RMaxORF and SNR. A new hybrid feature with 89 dimension can be formed by combining 86 sequence features and the above 3 features together. However, not every feature contribute to the classification accuracy. So we optimize the 89 dimensional features using the feature score criterion (FSC). The first 30 features of FSC are selected as the input vector of the classifier. Besides, an RF classifier model is constructed to discover new lncRNAs. Robustness is an advantage of RF model, since it can be used to build the ensemble of trees by randomly selecting features. The accuracy of a RF classifier is highly depends on the selection of training samples. In order to choose representative samples to construct training dataset, we use Self Organizing Feature Map (SOM) to select the training dataset. Finally, we provide a highly reliable and accurate tool called LncRNApred. It can identify the lncRNAs from thousands of assembled transcripts accurately andquickly. Moreover, using

Table 4. The performance of CPC.

Species	Positive (lncRNAs)	Negative (mRNAs)	Sn	Sp	ACC	MCC
Human	28707	36198	76.35 (21031/28707)	99.2 (36062/36198)	87.7	0.7687
Mouse	35851	27728	75.27 (26986/35851)	99.8 (27647/27728)	82.5	0.7520
Other species	2113	0	93.3 (1971/2113)	0	93.3	0

doi:10.1371/journal.pone.0154567.t004



ID	Prediction	Non-coding Score	MaxORF	RMaxORF	Length	SNR	CG%	CGG%	GC%	CCG%	GCG%	CGC%	GCC%	G%	(G+C)%
n21456	lncRNA	0.5640	453	0.4100	1105	1.4103	0.0145	0.0109	0.0615	0	0.0081	0.0081	0.0054	0.2715	0.5357
NM_134268	mRNA	0.4980	507	0.6879	737	1.8916	0.0624	0.0163	0.1004	0.0448	0.0244	0.0204	0.0407	0.3189	0.6418
n32145	lncRNA	0.5080	456	0.3393	1344	5.8482	0.0045	0.0022	0.0208	0	0	0	0.0045	0.2240	0.3557
NM_003403	mRNA	0.1320	2001	0.8599	2327	14.5675	0.0842	0.0297	0.1298	0.0155	0.0297	0.0425	0.0413	0.3386	0.6682

Fig 10. Screenshots of LncRNApred web server. (A) The input page. Single sequence or a fasta file can be as the input of LncRNApred. (B) The output page. LncRNApred reports sequence ID, Non-coding score, predicted class and the information of features.

doi:10.1371/journal.pone.0154567.g010

LncRNApred, we can also predict protein-coding potential of transcripts. The results indicate that our LncRNApred outperforms CPC. Therefore, we believe that V-ELMpiRNAPred is a valuable tool for the study of lncRNA and protein-coding transcripts.

Supporting Information

S1 Table. The positive training data of LncRNApred. The 2033 human lncRNAs are selected as the positive training data.
(RAR)

S2 Table. The negative training data of LncRNApred. 2031 human mRNAs are selected as the negative training data.
(RAR)

S3 Table. The positive test data (human) of LncRNApred. The 28707 mouse lncRNAs are selected as the positive test data.
(RAR)

S4 Table. The negative test (human) data of LncRNApred. The 36198 human mRNAs are selected as the negative test data.
(RAR)

S5 Table. The positive test data (mouse) of LncRNApred. The 35851 mouse lncRNAs are selected as the positive test data.
(RAR)

S6 Table. The negative test data (mouse) of LncRNApred. The 27728 mouse mRNAs are selected as the negative test data.
(RAR)

Acknowledgments

This work is supported by the National Natural Science Foundation of China (11401311, 31301229) and the Natural Science Foundation of Jiangsu Province (BK20141358).

Author Contributions

Conceived and designed the experiments: CP GLZ. Performed the experiments: ZC YYC JZ TY LYZ. Analyzed the data: CP. Contributed reagents/materials/analysis tools: CP. Wrote the paper: CP GLZ.

References

1. Core LJ, Waterfall JJ, Lis JT. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322, 1845–1848. doi: 10.1126/science.1162228 PMID: 19056941
2. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet., 38, 626–635. PMID: 16645617
3. Claverie JM. (2005) Fewer genes, more noncoding RNA. Science, 309, 1529–1530. PMID: 16141064
4. Wilusz JE, Sunwoo H, Spector DL. (2009) Long noncoding RNAs: functional surprises from the RNA world. Genes. Dev., 23, 1494–1504. doi: 10.1101/gad.1800909 PMID: 19571179
5. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science, 316, 1484–1488. PMID: 17510325
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC. (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860–921. PMID: 11237011
7. Hung T, Chang HY. (2010) Long noncoding RNA in genome regulation: Prospects and mechanisms. RNA. Biol., 7, 582–585. PMID: 20930520
8. Birney E, Stamatoyannopoulos JA, Anindya D, Roderic G, Gingeras TR, Margulies EH, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447, 799–816. PMID: 17571346
9. Storz G. (2002) An expanding universe of noncoding RNAs. Science, 296, 1260–1263. PMID: 12016301
10. Costa FF. (2010) Non-coding RNAs: Meet thy masters. BioEssays, 32, 599–608. doi: 10.1002/bies.200900112 PMID: 20544733
11. The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74. doi: 10.1038/nature11247 PMID: 22955616
12. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature, 420, 563–73. PMID: 12466851
13. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. (2005)The transcriptional landscape of the mammalian genome. Science, 309, 1559–1563. PMID: 16141072
14. Johnson JM, Edwards S, Shoemaker D, Schadt EE. (2005)Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends. Genet., 21, 93–102. PMID: 15661355
15. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, 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. doi: 10.1101/gr.132159.111 PMID: 22955988
16. Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, et al. (2012)Long noncoding RNAs are rarely translated in two human cell lines. Genome. Res., 22, 1646–1657. doi: 10.1101/gr.134767.111 PMID: 22955977
17. Esteller M. (2011) Non-coding RNAs in human disease. Nat. Rev. Genet., 12, 861–874. doi: 10.1038/nrg3074 PMID: 22094949
18. Brosnan CA, Voinnet O. (2009) The long and the short of noncoding RNAs. Curr. Opin. Cell. Biol., 21, 416–425. doi: 10.1016/j.ceb.2009.04.001 PMID: 19447594
19. Pauli A, Rinn JL, Schier AF. (2011) Non-coding RNAs as regulators of embryogenesis. Nat. Rev. Genet., 12, 136–149. doi: 10.1038/nrg2904 PMID: 21245830

20. Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. (2011) Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome. Biol. Evol., 3, 1390–1404. doi: 10.1093/gbe/evr116 PMID: 22071789

21. Mercer TR, Dinger ME, Mattuck JS. (2009) Long non-coding RNAs: insights into functions. Nat. Rev. Genet., 10, 155–159. doi: 10.1038/nrg2521 PMID: 19188922

22. Guttman M, Rinn JL. (2012) Modular regulatory principles of large non-coding RNAs. Nature, 482, 339–346. doi: 10.1038/nature10887 PMID: 22337053

23. Wang G, Cui Y, Zhang G, Garen A, Song X. (2009) Regulation of proto-oncogene transcription, cell proliferation, and tumorigenesis in mice by PSF protein and a VL30 noncoding RNA. Proc. Natl. Acad. Sci. U. S. A., 106, 16794–16798. doi: 10.1073/pnas.0909022106 PMID: 19805375

24. Li GB, Zhang HH, Wan XS, Yang XB, Zhu CP, Wang AQ, et al. (2014) Long noncoding RNA plays a key role in metastasis and prognosis of hepatocellular carcinoma. BioMed. Res. Int., 2014.

25. Vucicevic D, Schrewe H, Orom UA. (2014) Molecular mechanisms of long ncRNAs in neurological disorders. Front. Genet., 5.

26. Guay C, Jacovetti C, Nesca V, Motterle A, Tugay K, Regazzi R. (2012) Emerging roles of non-coding RNAs in pancreatic β-cell function and dysfunction. Diabetes. Obes. Metab., 14, 12–21. doi: 10.1111/j.1463-1326.2012.01654.x PMID: 22928560

27. Moskalev EA, Schubert M, Hoheisel JD. (2012) RNA-directed epigenomic reprogramming-an emerging principle of a more targeted cancer therapy? Gene. Chromosome. Canc., 51, 105–110.

28. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. (2013) Long noncoding RNAs and the genetics of cancer. Br. J. Cancer., 108, 2419–2425. doi: 10.1038/bjc.2013.233 PMID: 23660942

29. Shtivelman E, Bishop JM. (1989) The PVT gene frequently amplifies with MYC in tumor cells. Mol. Cell. Biol., 9, 1148–1154. PMID: 2725491

30. Yang F, Yi F, Zheng Z, Ling Z, Ding J, Guo J, et al. (2012) Characterization of a carcinogenesis-associated long non-coding RNA. RNA. Biol., 9, 110–116. doi: 10.4161/rna.9.1.18332 PMID: 22258142

31. Li D, Chen G, Yang J, Fan X, Gong Y, Xu G, et al. (2013) Transcriptome analysis reveals distinct patterns of long noncoding RNAs in heart and plasma of mice with heart failure. PloS. one., 8, e77938. doi: 10.1371/journal.pone.0077938 PMID: 24205036

32. Schonrock N, Harvey RP, Mattick JS. (2012) Long noncoding RNAs in cardiac development and pathophysiology. Circ. Res., 111, 1349–1362. doi: 10.1161/CIRCRESAHA.112.268953 PMID: 23104877

33. Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, et al. (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev. Cell., 24, 206–214. doi: 10.1016/j.devcel.2012.12.012 PMID: 23369715

34. Kumarswamy R, Bauters C, Volkmann I, Maury F, Fetisch J, Holzmann A, et al. (2014) Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. Circ. Res., 114, 1569–1575. doi: 10.1161/CIRCRESAHA.114.303915 PMID: 24663402

35. Ammosova T, Yedavalli VR, Niu X, Jerebtsova M, Van Eynde A, Beullens M, et al. (2011) Expression of a protein phosphatase 1 inhibitor, cdNIPP1, increases CDK9 threonine 186 phosphorylation and inhibits HIV-1 transcription. J. Biol. Chem., 286, 3798–3804. doi: 10.1074/jbc.M110.196493 PMID: 21098020

36. Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R, et al. (2010) HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. Mol. Cell., 38, 439–451. doi: 10.1016/j.molcel.2010.04.012 PMID: 20471949

37. Muniz L, Egloff S, Ughy B, Jady BE, Kiss T. (2010) Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator Tat. PLoS. Pathog., 6, e1001152. doi: 10.1371/journal.ppat.1001152 PMID: 20976203

38. Eilebrecht S, Brysbaert G, Wegert T, Urlaub H, Benecke BJ, Benecke A. (2011) 7SK small nuclear RNA directly affects HMGA1 function in transcription regulation. Nucleic. Acids. Res., 39, 2057–2072. doi: 10.1093/nar/gkq1153 PMID: 21087998

39. Yoon W, Ma BJ, Fellay J, Huang W, Xia SM, Zhang RJ, et al. (2010) A polymorphism in the HCP5 gene associated with HLA-B* 5701 does not restrict HIV-1 in vitro. AIDS, 24, 155–157. doi: 10.1097/QAD.0b013e32833202f5 PMID: 19935381

40. Catano G, Kulkarni H, He W, Marconi VC, Agan BK, Landrum M, et al. (2008) HIV-1 disease-influencing effects associated with ZNRD1, HCP5 and HLA-C alleles are attributable mainly to either HLA-A10 or HLA-B* 57 alleles. PloS. One., 3, e3636. doi: 10.1371/journal.pone.0003636 PMID: 18982067

41. Zhang Q, Chen CY, Yedavalli VS, Jeang KT. (2013) NEAT1, long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. MBio, 4, e00596–12. doi: 10.1128/mBio.00596-12 PMID: 23362321

42. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic. Acids. Res., 41(DI), D983–D986.

43. Bernhart SH, Hofacker IL. (2009) From consensus structure prediction to RNA gene finding. Brief. Funct. Genomics., 8, 461–471.

44. Rivas E, Eddy SR. (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2, 8. PMID: 11801179

45. Washiet S, Hofacker IL, Stadler PF. (2005) Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U. S. A., 102, 2454–2459. PMID: 15665081

46. Coventry A, Kleitman DJ, Berger B. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A., 101, 12102–12107. PMID: 15304649

47. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput. Biol., 2, e33. PMID: 16628248

48. Tran TT, Zhou F, Marshburn S, Stead M, Kushner SR, Xu Y. (2009) De novo computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics, 25, 2897–2905. doi: 10.1093/bioinformatics/btp537 PMID: 19744996

49. Saetrom P, Sneve R, Kristiansen KI, Snove O Jr, Grunfeld T, Rognes T, et al. (2005) Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. Nucleic. Acids Res., 33, 3263–3270. PMID: 15942029

50. Wang C, Ding C, Meraz RF, Holbrook SR. (2006) PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics, 22, 2590–2596. PMID: 16945945

51. Washietl S, Findeiss S, Muller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA, 17, 578–594. doi: 10.1261/rna.2536111 PMID: 21357752

52. Raasch P, Schmitz U, Patenge N, Vera J, Kreikemeyer B, Wolkenhauer O. (2010) Non-coding RNA detection methods combined to improve usability, reproducibility and precision. BMC Bioinformatics, 11, 491. doi: 10.1186/1471-2105-11-491 PMID: 20920260

53. Salari R, Aksay C, Karakoc E, Unrau PJ, Hajirasouliha I, Sahinalp SC. (2009) smyRNA: A Novel Ab Initio ncRNA Gene Finder. PLoS One., 4, e5433. doi: 10.1371/journal.pone.0005433 PMID: 19415115

54. Bao M, Cervantes Cervantes M, Zhong L, Wang JT. (2012) Searching for non-coding RNAs in genomic sequences using ncRNAscout. Genomics Proteomics Bionformatics, 10, 114–121.

55. Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchatchawalya M. (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. Nucleic. Acids. Res., 41, e21. doi: 10.1093/nar/gks878 PMID: 23012261

56. Liu J, Gough J, Rost B. (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS. Genet., 2, e29. PMID: 16683024

57. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic. Acids.Res., 35, W345–W349. PMID: 17631615

58. Lin MF, Jungreis I, Kellis M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 27, i275–i282. doi: 10.1093/bioinformatics/btr209 PMID: 21685081

59. Sun L, Liu H, Zhang L, Meng J. (2015) lncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine. Plos One., 10(10): e0139654. doi: 10.1371/journal.pone.0139654 PMID: 26437338

60. Sun K, Chen XN, Jiang PY, Song XF, Wang HT, Sun H. (2013) iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics., 14, s7.

61. Wang LG, Park HJ, Dasari S, Wang SQ, Kocher JP, Li W. (2013) ACPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res., 41(6):e74. doi: 10.1093/nar/gkt006 PMID: 23335781

62. Dinger ME, Pang KC, Mercer TR, Mattick JS. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput. Biol., 4, e1000176. doi: 10.1371/journal.pcbi.1000176 PMID: 19043537

63. Guttman M, Rinn JL. (2012) Modular regulatory principles of large non-coding RNAs. Nature, 482, 339–346. doi: 10.1038/nature10887 PMID: 22337053

64. Sun L, Luo HT, Bu DC, Zhao GG, Yu KT, Zhang CH et al. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res., 41(17), e166. doi: 10.1093/nar/gkt646 PMID: 23892401

65. Bu D, Yu K, Sun S, Xie C, Skogerbo G, Miao R, et al. (2012) NONCODE v3.0:integrative annotation of long noncoding RNAs. Nucleic. Acids. Res., 36, D210–D215.

66. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, et al. (2011) The UCSC genome browser database: update 2011. Nucleic. Acids.Res., 39, D876–D882. doi: 10.1093/nar/gkq963 PMID: 20959295

67. Kohonen T. (1990) The self-organizing map. P. IEEE, 78, 1464–1480.

68. Voss RF. (1992) Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett., 68, 3805–3808. PMID: 10045801

69. Yin C, Yau SS. (2007) Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theor. Biol., 247, 687–694.405, 442–451. PMID: 17509616

70. Ficket JW, Tung CS. (1992) Assessment of protein coding measures. Nucleic. Acids. Res., 20, 6441–6450. PMID: 1480466

71. Fickett JW. (1982) Recognition of protein coding regions in DNA sequences. Nucleic. Acids. Res., 10, 5303–5318. PMID: 7145702

72. Tiwari S, Ramachandran S, Bhattacharya A, Bhattachrya S, Ramaswamy R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. CABIOS. 113, 263–270.

73. Yin C, Yau SS-T. (2005) A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. J. Comput. Biol., 9, 1153–1165.

74. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. (1999) Molecular classification of cancer:class discovery and class prediction by gene expression monitoring. Science, 286, 531–537. PMID: 10521349

75. Matthews BW. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta., 405, 442–451. PMID: 1180967

76. Huang GB, Zhu QY, Siew CK. (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. Proc. Int. Joint. Conf. Neural. Netw. 2, 985–990.

77. Huang GB, Zhu QY, Siew CK. (2006) Extreme learning machine: Theory and Applications. Neurocomputing, 70, 489–501.

78. Cao JW, Lin ZP, Huang GB, Liu N. (2012) Voting based extreme learning machine. Inform. Sciences, 185, 66–77.

79. Breiman L. (2001) Random forest. Mach. Learn., 45, 5–32.

80. Breiman L. (1996) Bagging predictors. Mach. Learn., 24, 123–140.