Published online 11 February 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 8* ***e43***
doi: 10.1093/nar/gkz087

CPPred: coding potential prediction based on the global description of RNA sequence

Xiaoxue Tong and Shiyong Liu*

School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

Received July 18, 2018; Revised January 26, 2019; Editorial Decision January 28, 2019; Accepted February 01, 2019

ABSTRACT

The rapid and accurate approach to distinguish between coding RNAs and ncRNAs has been playing a critical role in analyzing thousands of novel transcripts, which have been generated in recent years by next-generation sequencing technology. Previously developed methods CPAT, CPC2 and PLEK can distinguish coding RNAs and ncRNAs very well, but poorly distinguish between small coding RNAs and small ncRNAs. Herein, we report an approach, CPPred (coding potential prediction), which is based on SVM classifier and multiple sequence features including novel RNA features encoded by the global description. The CPPred can better distinguish not only between coding RNAs and ncRNAs, but also between small coding RNAs and small ncRNAs than the state-of-the-art methods due to the addition of the novel RNA features. A recent study proposes 1335 novel human coding RNAs from a large number of RNA-seq datasets. However, only 119 transcripts are predicted as coding RNAs by the CPPred. In fact, almost all proposed novel coding RNAs are ncRNAs (91.1%), which is consistent with previous reports. Remarkably, we also reveal that the global description of encoding features (T2, C0 and GC) plays an important role in the prediction of coding potential.

INTRODUCTION

Recently, next-generation sequencing technology has generated thousands of novel transcripts (1–4). Many of them are non-coding RNAs (ncRNAs) (5,6). Although they cannot encode proteins, many experiments have also demonstrated that they play important biological roles in various biological processes, such as gene regulation/expression, gene silencing, and RNA modification and processing (7–10). Furthermore, ncRNAs also tend to exhibit striking tissue specificity, functionality conserved (11,12), and have become the key to disease development processes (13–16).

However, there are growing evidence that ncRNAs could contain small open reading frames (sORFs, ≤303 nt) encoding micropeptides (17–25). In 2002, Rohrig *et al.* discover that a long non-coding RNA (lncRNA, >200 nt) with 679 nucleotides is in fact a messenger RNA (mRNA) (25). The RNA is transcribed from a gene called early nodulin 40 (*ENOD40*), whose two open reading frames (ORFs) encode two micropeptides with 12 and 24 amino acids, respectively. In 2007, the mRNA of '*polished rice*' (*pri*) is originally annotated as lncRNA in Drosophila, but it contains sORFs (17,18), which encode four micropeptides with 11, 11, 11 and 32 amino acids, respectively. The *pri* has an essential role as a key transcription factor associating with activating development (17,19). Subsequently, calcium-related sORF encoding peptides are found, which are able to regulate muscle contraction (22–24). Since then, the micropeptides harbored in other ncRNAs are found gradually (20,21). Furthermore, ribosome profiling, mass spectrometry (MS) and proteogenomics have been performed for detection of sORF-encoded peptides recently (26–33).

Nevertheless, the biological significance and function of most ncRNAs remain unclear comparing with coding RNAs. Rapid and accurate coding potential prediction of transcripts is critical for analyzing these data. From a computational perspective, distinguishing between coding RNAs and ncRNAs is a binary classification task and various tools have been developed (34–45). In 2006, Liu *et al.* present an SVM-based tool (43), namely COCN, which could predict coding RNAs from ncRNAs on the basis of a hybrid feature set. However, COCN is slow in calculating abundant datasets. CPC (Coding Potential Calculator) (38) uses support vector machine to differentiate coding RNAs from ncRNAs, which is developed by Kong *et al.* in 2007. Six biologically meaningful features are extracted, such as ORF quality, ORF coverage, ORF integrity, sequence similarity with known proteins. However, the performance of CPC depends on the quality of multiple sequence alignment. The coding potential calculator CPC1 to CPC2 is updated in 2017 (39). The CPC2 (a predictor of coding potential based on ORF length, Fickett score, ORF integrity and isoelectric point) is much faster and more accurate than CPC1, in particular for lncRNAs. Moreover, the model of CPC2 is species-nonspecific. In 2013, Wang *et al.* present a logistic regression model CPAT (42) for differentiating ncRNAs from coding RNAs, which uses four features (ORF

*To whom correspondence should be addressed. Tel: +86 027 87543881; Fax: +86 027 87556576; Email: liushiyong@gmail.com

length, ORF coverage, Fickett score and Hexamer usage bias). They indicate that the length of ORF is the most important feature for coding potential prediction. Overall, those tools mentioned above have been developed for distinguishing ncRNAs from coding RNAs in general.

On the other hand, many tools such as PLEK, iseeRNA, lncRScan-SVM, FEELnc, COME, DeepLNC and LncRNApred can predict lncRNAs from coding RNAs (34,35,37,40,41,44,45). PLEK, IseeRNA and LncRscan-SVM are based on the SVM algorithm. Among them, iseeRNA (35) is developed to predict long intergenic non-coding RNAs (lincRNAs). PLEK (34) is designed by an improved k-mer strategy to identify lncRNAs from coding RNAs in 2014. LncRScan-SVM (45) is established to distinguish coding RNAs and lncRNAs by analyzing gene structure, the conservation of RNA sequence and codon sequence. In 2016, Wucher *et al.* apply ORF coverage, codon usage and the frequency of the nucleotides to presenting FEELnc with Random Forests (40), which could identify lncRNAs even without a training set of non-coding RNAs. Moreover, DeepLNC (37) can be used to identify the lncRNAs from coding RNAs, which uses k-mer frequencies of transcripts and Deep Neural Network. However, it is unclear which features are used. In 2017, Hu *et al.* develop a supervised machine learning tool COME (41), which uses diversified sequence-based and experimental features. In particular, the features are encoded by using decompose-compose method. In 2018, Schneider *et al.* propose a scheme (36), which is based on SVM to differentiate lncRNAs from coding RNAs by using sequence features that are selected by PCA. The sequence features include the relative length of first ORF and frequencies of K-mer. The method is trained and tested on human, mouse and zebrafish data with the accuracy of 98.21%, 98.03% and 96.09%, respectively. They also predict 81.2% of human pseudogenes and 91.7% of mouse pseudogenes. Besides, in 2018, McGillivray *et al.* develop a Bayesian algorithm to predict the function of upstream open reading frames (uORFs) based on 89 features (46). Subsequently, a functional sORF-encoded peptide predictor (FSPP) is built by Li *et al.* to detect the sORF-encoded peptides and their functions (47).

Although current computational methods have yielded encouraging results, they are facing certain limitations. For example, they predict poorly on the data of sORFs, which have been studied recently as mentioned above. Inspired by the work of Wang *et al.* (42), we plotted three features (ORF length, Fickett score, Hexamer score) in a three-dimensional space for all RNA transcripts (Human-Training, which contain 33 360 coding RNAs and 24 163 ncRNAs, Figure 1A) and small RNA transcripts (Figure 1B). From Figure 1A, most of the coding RNAs and ncRNAs can be distinguished by using these three features only with slight overlapping. Next, we extracted small coding RNAs with ORF <303 nucleotides in length and small ncRNAs (see the section ‘Datasets’ in ‘Materials and Methods’) from Human-Training dataset and plotted Figure 1B. It can be seen that these three features are incapable of distinguishing between small coding RNAs and small ncRNAs. Herein, we developed a coding potential prediction tool (CPPred), which used SVM to differentiate ncRNAs from coding RNAs on the basis of sequence features, such as ORF length, ORF coverage, ORF integrity, Fickett score, Hexamer score, Isoelectric point (pI) of a predicted peptide, Grand average of hydropathicity (Gravy) of a predicted peptide, estimation of the stability (Instability) of a predicted peptide and global descriptor (CTD) features. The CTD (composition (C), transition (T) and distribution (D)) is originally proposed for predicting protein folding class, which is global protein sequence descriptors established by Dubchak’s work (48). In this work, CTD is used to denote the global transcript sequence descriptors. The CTD features include nucleotide composition, nucleotide transition and nucleotide distribution. It should be noted that the CTD features are firstly proposed by us to distinguish between coding RNAs and ncRNAs in eukaryotes. Distinguishing coding RNAs from ncRNAs based on the CTD features in prokaryote have been explored in 2009 (49); however, overlooked in recent literature. Although the nucleotide composition is as well already represented in the Fickett score, the nucleotide transition and nucleotide distribution are novel features. From Figure 4, the nucleotide transition and nucleotide distribution features (T2, C0 and GC) are important features to classify coding RNAs and ncRNAs. Besides, the mRNA secondary structure of around the start and stop codons has an important potential impact on ribosome pausing (50,51). Thus, RNA secondary structural feature contributes to predicting protein coding potential. Moreover, CTD is built to predict protein folding class primitively (48), so in this work, the CTD features are connected with RNA structural features. The CTD features not only are beneficial to prediction of coding potential (49), but also prove to be important features in RNA-binding protein prediction (52), RNAs functional identity (53) and promoter recognition (54). Subsequently, we trained the model on human dataset, and then tested it on human, mouse, zebrafish, fruit fly and *Saccharomyces cerevisiae*. The data of several popular species are integrated to avoid species specificity. An integrated model is built. The testing results show that CPPred with higher Matthews correlation coefficient (MCC) (55) is particularly more effective on sORF data when compared with other programs. Moreover, to highlight the performance of CPPred on the sORF data, we compared it with sORF finder (56). The sORF finder is developed by Hanada *et al.* and using nucleotide composition to identify sORF. It is proved by Cheng *et al.* to be superior to other tools for predicting sORF (57). Additionally, CPPred is convenient because it only needs FASTA format sequence files as inputs.

MATERIALS AND METHODS

Dataset

Two models are built for distinguishing between coding RNAs and ncRNAs. The first one (Human-Model) is trained by human data, and then tested on human, mouse, fruit fly, zebrafish and *S. cerevisiae*. The second one (Integrated-Model) is built for integrated species, including human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), *S. cerevisiae*, nematode (*Caenorhabditis elegans*) and thale cress (*Arabidopsis thaliana*). As the former, we downloaded

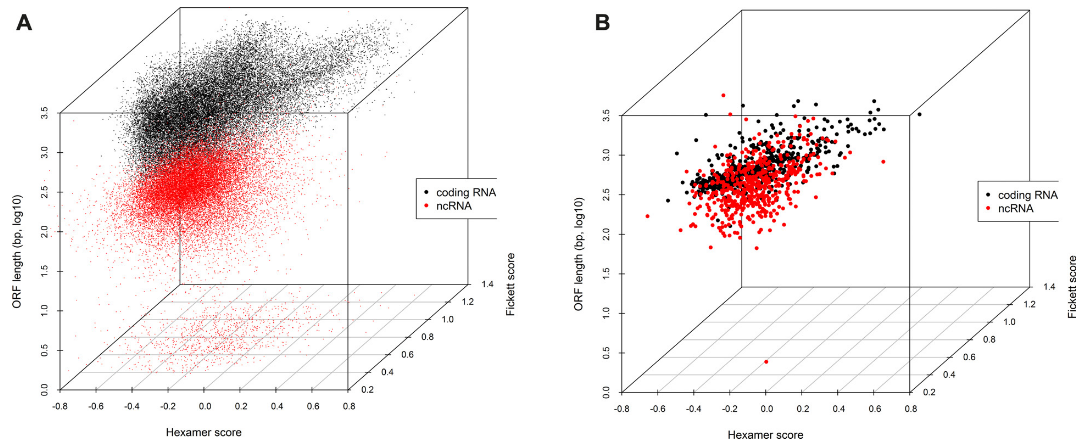

Figure 1. (**A**) Three-dimensional plot of Hexamer score, Fickett score and ORF length on 33360 coding RNAs and 24163 ncRNAs (Human-Training). (**B**) Three-dimensional plot of Hexamer score, Fickett score and ORF length on 508 small coding RNAs and 508 small ncRNAs, which extracted small coding RNAs with ORF <303 nucleotides in length and small ncRNAs from Human-Training.

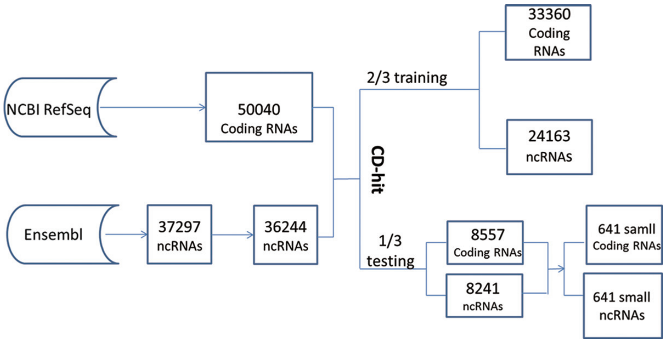

Figure 2. The flowchart of building training set and testing set of human. Human coding RNAs with transcript status 'KNOWN' are downloaded from NCBI RefSeq and human ncRNAs are downloaded from Ensembl. The initial dataset includes 50 040 coding RNAs and 37 297 ncRNAs. For ncRNAs, the data that have no source comments and are not annotated with Havana in the corresponding of gff3 file are removed. After that, the number of coding RNAs and ncRNAs is 50 040 and 36 244, respectively. We randomly selected two-thirds of the data as training set, a collection of 33 360 coding RNAs and 24 163 ncRNAs, which is called Human-Training. Then, the rest of the data are stored as a testing set. At the same time, we reduced redundancy between the testing and training set using CD-hit with sequence identity cutoff ≥80%. Finally, 8557 coding RNAs and 8241 ncRNAs are kept as Human-Testing. Then, the sequences with ORF shorter than 303 nucleotides in length are extracted from coding RNA in Human-Testing. Meanwhile, the same amount of considerable length ncRNAs from Human-Testing are selected randomly. As a result, 641 coding RNAs and 641 ncRNAs are kept as Human-sORF-Testing.

human coding RNAs as positives from NCBI RefSeq (58,59) (https://www.ncbi.nlm.nih.gov/nuccore/?term=human[orgn]±AND±src db_refseq_known[prop]±AND±biomol_rna[prop]) and from Ensembl database (60,61) (ftp://ftp.ensembl.org/pub/release-90/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz), human ncRNAs as negatives are obtained as of 26 November 2017 (see Figure 2). The transcript status of coding RNAs is 'KNOWN'. However, for ncRNAs, we removed 1053 ncRNA sequences without source comments, and these sequences are not annotated with Havana in the corresponding gff3 file. Eventually, the total numbers of annotated coding RNAs and ncRNAs are 50 040 and 36 244, respectively. We randomly selected two-thirds (41) of data as a training set, that is, the set including 33 360 coding RNAs and 24 163 ncRNAs, which is called Human-Training. Then, the remaining data are stored as a testing set. To improve the robustness of the assessment of accuracy, we reduced redundancy between the testing and training sets with a threshold of 80% (62–64) by the open-source program cd-hit-est-2d in CD-hit (65), which uses a short word filter to avoid unnecessary alignments and has been widely used as a clustering algorithm, resulting in 8557 coding RNAs and 8241 ncRNAs as Human-testing. Besides, the basic command of cd-hit-est-2d is cd-hit-est-2d –i training_human –i2 testing_human –o testing_human_redundancy –c 0.8 –n 5. Meanwhile, we examined the different threshold (<80%) for removing the redundancy much more stringently between the testing and training sets. The redundancy between the testing and training sets is removed by program cd-hit-2d in CD-hit with threshold of 75%, 70% and 60%, respectively. The performance of CPPred is shown in Supplementary Table S10 on the testing sets. The last row in Supplementary Table S10 is the performance of CPPred in the manuscript. The result shows that removing redundant sequences with ≥80% similarity is enough to obtain independence between test and training sets. Besides, the BLASTCLUST is also tried and used to exclude redundant RNA sequences with the similarity threshold of 75%, 70%, 60%, 50%, 40%, 30%, 20% and 10%, respectively. The result shows the same conclusion. That is because RNA sequences only have four letters, it is easy to get high identity between unrelated sequences. When we run a regular blast search, we often find top hits at very high identities. In general, comparing RNAs at low identity <75% may not be very effective. The threshold of 80% is a stringent cutoff. For example, the sequence redundancy in the datasets of CPC2 (39), lncRScan-SVM (45) and lncADeep (66) is removed by using CD-hit with thresholds of 0.9, 0.8 and 0.95, respectively. The sequence redundancy of CONC's datasets (43) is removed by using NCBI BLASTCLUST with the '-L 0.7' option, and the sequence similarity of CPC's

Table 1. The testing set for mouse, zebrafish, *S. cerevisiae* and fruit fly

	Coding RNAs	ncRNAs
Mouse-Testing	31 102	19 930
Mouse-sORF-Testing	846	1000
Zebrafish-Testing	15 594	10 662
Zebrafish-sORF-Testing	387	500
S.cerevisiae-Testing	6713	413
S.cerevisiae-sORF-Testing	505	413
Fruit-fly-Testing	17 400	4098
Fruit-fly-sORF-Testing	381	381

datasets (38) is removed by using BLASTN with evalue <1e-2. For COME (41), only the overlapping transcripts are eliminated between the training and testing sets. While for the datasets of PhyloCSF (67), CPAT (42), CNCI (68), iSeeRNA (35), PLEK (34), lncRNA-ID (69), lncScore (70), FEElnc (40), longdist (36) DeepLNC (37) and mRNN (71), no redundancy is removed. For the reason that the prediction of sORF is difficult, the sequences with ORF fragments <303 nucleotides in length from coding RNAs in Human-Testing are selected to build an interesting and challenging testing set. Meanwhile, the same amount of comparable length ncRNAs from Human-Testing were filtered out randomly (34,39). As a result, 641 coding RNAs and 641 ncRNAs are kept as Human-sORF-Testing. Based on the same building method of the human testing sets, we constructed Mouse-Testing, Mouse-sORF-Testing, Zebrafish-Testing, Zebrafish-sORF-Testing, *S. cerevisiae*-Testing, *S. cerevisiae*-sORF-Testing, Fruit-fly-Testing and Fruit-fly-sORF-Testing. These datasets are downloaded from Ensembl, as shown in Table 1.

Afterward, several popular species are integrated for the sake of eliminating the problems caused by the specificity of species and the differences between the databases. The data for human, mouse, zebrafish, fruit fly, *S. cerevisiae*, nematode and thale cress are downloaded from NCBI RefSeq including 525 316 coding RNAs and 55 198 ncRNAs. Furthermore, to reduce the computational effort and balance the proportion of human negative-positive data, we randomly selected 52 530 coding RNAs and 27 600 ncRNAs as training set with the same percentage of each species, which is called Integrated-Training. The redundancy of the remaining is removed with sequence identity cutoff ≥80% (62–64). The Integrated-Testing (balance data) is constructed with 13 903 coding RNAs and 13 903 ncRNAs (34,39). As the same building procedure with Human-sORF-Testing, the Integrated-sORF-Testing is obtained, which contains 11 634 small coding RNAs and 11 634 small ncRNAs.

CPPred features

To predict the coding potential of RNA sequences, we extracted features from recently published scientific literature (39,42), and novel CTD features are added.

We used four features proposed by CPAT (42), including ORF length, ORF coverage, Fickett score and Hexamer Score. Similarly, ORF integrity and isoelectric point were derived from CPC2 (39). Next, the Gravy and Instability mentioned by CPC2 are also added. In addition, the algorithm of Hexamer score and Fickett score were discussed in detail (42,72,73). The Fickett score is calculated by considering eight properties of coding sequences. Four of them are composition values with the frequencies of the four nucleotides from RNA sequence. The other four parameters are position value, which reflect the degree to codon preference. Furthermore, the feature of ORF length is used to predict the coding potential (38,39,42,43), but it relies on the full-length transcript (41). In this study, although the ORF length requires annotation of the full-length transcript, we still selected it because of its identifiable power and ease of calculation. The ORF length is the length of the maximum open reading frame, which starts with a start codon and ends with a stop codon (UGA, UAA or UAG). Here, the AUG is selected as the start codon. Although the use of non-AUG has been constantly described (26,29,30,33,74–76), there is no clear consensus on how to choose translation start sites (77).

In particular, we added 30 new features, which were CTD features. In this study, CTD is used to denote the global transcript sequence descriptors. The transcript is a sequence containing four types of nucleotides A, T, G and C. The nucleotide composition (first index C) describes the percent composition of each nucleotide in a transcript sequence, which is contained in Fickett score. The nucleotide transition (second descriptor T) describes the percent frequency with conversion of four nucleotides between adjacent positions. Subsequently, we calculated five relative positions along the transcript sequence of each nucleotide, with the 0 (first one), 25%, 50%, 75% and 100% (last one), to describe the nucleotide distribution (last descriptor D).

For example, the RNA sequence is ACTTGCAGCC CCCCGCCTGTCCCGAG CCGCGCGGGCGCCAGC TCAGTTTGTCCGCGGCGG, which contains 5 adenines (As), 9 thymines (Ts), 20 guanines (Gs) and 26 cytidines (Cs). The features of first descriptor C are 5/60 = 0.083, 9/60 = 0.15, 20/60 = 0.33 and 26/60 = 0.43, respectively. We use A, T, G and C to represent the four features. For the second descriptor T, there is zero transition between A and T, five transitions between A and G, four transitions between A and C, five transitions between T and G, six transitions between T and C and twenty transitions between G and C. Therefore, the frequencies of these transitions are 0/59 = 0.00, 5/59 = 0.085, 4/59 = 0.068, 5/59 = 0.085, 6/59 = 0.10 and 20/59 = 0.34. We use AT, AG, AC, TG, TC and GC to represent the six features. The first, 25%, 50%, 75% and 100% of As are located within 1, 1, 25, 40 and 45 residues, respectively. The D descriptors for As are 1/60 = 0.017, 1/60 = 0.017, 25/60 = 0.42, 40/60 = 0.67 and 45/60 = 0.75. Likewise, the D descriptors for Ts, Gs and Cs are 0.05, 0.067, 0.72, 0.80, 0.85, 0.083, 0.40, 0.57, 0.83, 1.00, 0.033, 0.22, 0.38, 0.65 and 0.97, respectively. We use A0, A1, A2, A3, A4, T0, T1, T2, T3, T4, G0, G1, G2, G3, G4, C0, C1, C2, C3 and C4 to represent the 20 features.

Feature selection

The 38 features as mentioned may include redundant features, so a feature selection process is used to filter out redundant features for coding potential prediction. In this work, mRMR-IFS method is used to select the best subset of features (78–81). The mRMR program is developed by Peng *et al.* (82), which selects good features based on mutual information with the minimal redundancy, maximal relevance criteria. First, the mRMR program is used to rank the 38 features in the training set. Then, Incremental Feature Selection (IFS) (78) is used to increase the features one by one in descending based on the mRMR ordering. For each additional feature, a new subset of feature is generated. Therefore, a total of 38 feature subsets are generated for 38 sorting features.

According to the obtained 38 feature subsets, we select the corresponding feature sets from the whole training set. Through 10-fold cross-validation on the training set, the best subset of features is selected and served as the final model.

SVM classifier

We used the Libsvm (83) (Libsvm-3.22) for predicting coding RNAs and ncRNAs. The radial basis function is selected as the kernel function, and the MCC value is used as the function to optimize the parameters (C and γ). Here, for the Human-Training, the optimal values of $C = 1024.0$, $\gamma = 0.5$ are obtained by grid search method with the selecting top 37 features. For the Integrated-Training, the optimal values of $C = 90.5096679919$, $\gamma = 1.0$ are gained with 38 features. Moreover, for the Human-Training with CTD features and non-CTD features, the optimal values of (C, γ) are (1024.0, 2.0) and (16384.0, 1.0), respectively.

Performance evaluation of CPPred

The CPPred is evaluated by the widely used standard performance metric, which are sensitivity (SN), specificity (SP), accuracy (ACC), precision (PRE), F-score, AUC and MCC (55). These evaluation indexes are defined as follows:

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

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

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

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

$$F - \text{score} = \frac{2 * PRE * SN}{PRE + SN}$$

$$\text{Matthews Correlation Coefficient } (MCC) = \frac{TP * TN - FP * FN}{\sqrt{(TP+FN)*(TP+FP)*(TN+FP)*(TN+FN)}}$$

where TP stands for true positive, which is the number of positive samples identified correctly, FN, TN, FP represent false negative, true negative and false positive, which denote the number of positive samples identified incorrectly, negative samples identified correctly, negative samples identified incorrectly, respectively. The MCC is an overall measurement of performance and another objective assessment index. AUC is the area under the receiver operating characteristic curve.

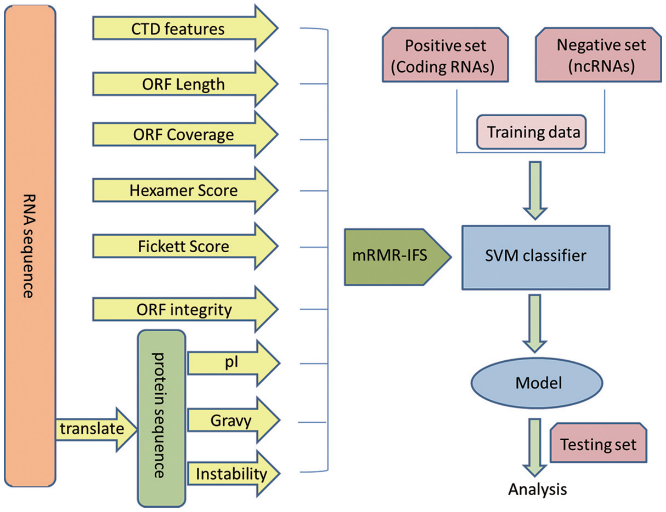

Figure 3. Pipeline of the CPPred. Multiple features are extracted from RNA or protein sequences. Herein, the CTD features include nucleotide composition, nucleotide transition and nucleotide distribution. The ORF coverage is defined as the ratio of ORF to the length of a transcript. The ORF length, Hexamer score and Fickett score are discussed in the 'CPPred features' section. The integrity of the ORF is defined as whether the ORF starts with a start codon (AUG) and ends with a stop codon (UGA, UAA or UAG). PI, Gravy and Instability are calculated by the ProtParam. After that, using mRMR-IFS, the best feature subset is selected and used as input to the SVM classifier. Eventually, we got the final model, which is tested and evaluated by the testing sets.

RESULTS AND DISCUSSION

Pipeline of CPPred

In CPPred, an SVM model is applied to calculating coding potential of a transcript by using features derived from RNA and protein sequences (see Figure 3), which is designed for distinguishing between coding RNAs and ncRNAs. First, we constructed a training dataset, which contained coding RNAs and ncRNAs. Then, 38 features are calculated for each RNA or protein sequence. The uninformative features are reduced by using mRMR-IFS and the best feature subset is picked out. Based on the feature subset, the SVM classifier is used to obtain a model on the training set. Finally, the CPPred is analyzed and evaluated on testing sets.

Feature selection by the mRMR-IFS method

Here, mRMR-IFS (78) method is chosen for feature selection. For each feature subset, the corresponding features are selected and 10-fold cross-validation is performed on the training set. In Figure 4, for Human-Training, the best predictive performance is obtained by using top 37 features with the highest MCC value of 0.953 (SN, SP, ACC and AUC are 97.81% 97.57%, 99.68% and 0.977, respectively). While for the Integrated-Training, the best predictive performance is achieved by applying all the features with MCC value of 0.941 (SN, SP, ACC and AUC are 97.62% 96.75%, 97.32% and 0.995, respectively). Therefore, in order to predict coding potential of transcripts, we selected top 37 features and all features as the optimal feature sets to create the final models.

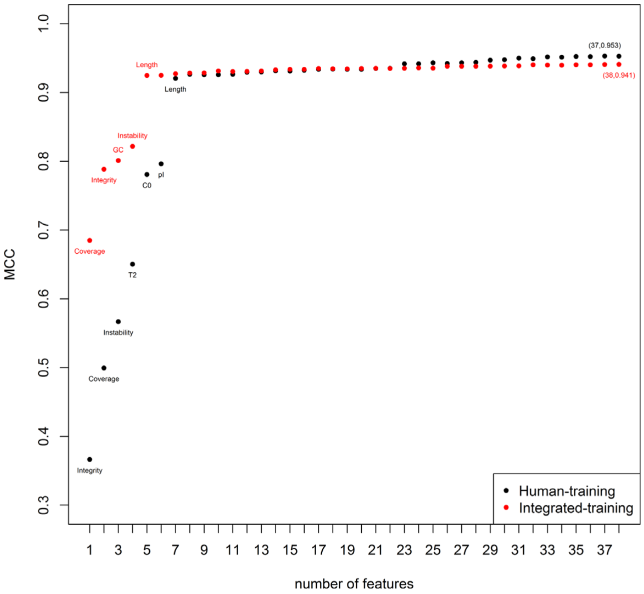

Figure 4. mRMR-IFS feature selection. The mRMR-IFS scatter plot of the feature subsets are drawn by R tool, which corresponds to the two training sets that are Human-Training and Integrated-Training, respectively. Wherein the *x*-coordinate is the number of features in the feature subset, and the *y*-coordinate represents the MCC of the corresponding 10-fold cross-validation.

In addition, the value of MCC dramatically increased to 0.920 on the Human-Training in Figure 4 after adding the seventh feature and then the MCC tended to be stable. The top seven features include ORF integrity, ORF Coverage, Instability, T2 (CTD feature), C0 (CTD feature), isoelectric point and ORF length (Supplementary Table S7). In particular, the seventh feature is the length of ORF. At the same time, for the Integrated-Training, the MCC value (MCC = 0.925) increased suddenly after the fifth feature was added. The top five features contain ORF coverage, ORF integrity, GC (CTD feature), Instability and ORF length (Supplementary Table S8). Moreover, the fifth feature is also the length of ORF. In conclusion, the length of the ORF is a more crucial feature for distinguishing between coding RNAs and ncRNAs, which is consistent with the findings by Wang *et al.* (42).

The top seven features of Human-Training and the top five features of Integrated-Training are important to differentiate coding RNAs from ncRNAs. Among these important features, ORF coverage and ORF length are from CPAT (42). The isoelectric point, length and integrity of the ORF are from CPC2 (39). Of the remaining important features, T2, C0 and GC are derived from CTD features, which indicate the importance of CTD features for distinguishing between coding RNAs and ncRNA. Besides, in Figure 4, it also reveals that the ORF length, ORF coverage, ORF integrity and Instability are shared among species, while pI, T2 and C0 are human-specific.

Performance of CPPred (Human-Model)

To evaluate our method CPPred, for Human-Model, we compared CPPred with CPAT (42), CPC2 (39) and PLEK (34) on testing sets of human, mouse, zebrafish, *S. cerevisiae* and fruit fly. Moreover, for sORF testing sets of human, mouse, zebrafish, *S. cerevisiae* and fruit fly, the method sORF finder (56) has also been added for comparison. The results of Human-Testing are shown in Table 2. The accuracy of CPPred, CPAT, CPC2 and PLEK are 96.23%, 94.33%, 93.07% and 96.73%. The AUCs of them are 0.992, 0.984, 0.982 and 0.993. The MCCs of them are 0.925, 0.886, 0.862 and 0.935. For Human-sORF-Testing, which is a challenging set including 641 small coding RNAs and 641 small ncRNAs, the results are listed in Table 3. Significantly, CPPred is much better than the other three methods (CPAT, CPC2 and sORF finder) with MCC of 0.654 versus 0.472, 0.125 and 0.262. Overall, from the two testing sets, our method performs better than CPAT and CPC2, however, slightly worse than PLEK, probably due to the high level of redundancy between the two human testing sets and PLEK's training set (84). Herein, we downloaded the human training set of PLEK from https://sourceforge.net/projects/plek/. Subsequently, we used the CD-HIT tool with a value of 0.8 as the threshold of sequence identity to analyze the redundancy of between human testing set of CPPred and human training set of PLEK. The results show that 3944 (3944/8557 = 46.1%) coding RNAs and 5411 (5411/8241 = 65.7%) ncRNAs are redundant sequences between the human testing set of CPPred and human training set of PLEK, respectively. Besides, the predictive performance of PLEK dropped behind the other methods (CPPred, CPAT, CPC2 and sORF finder) when testing with mouse, zebrafish, *S. cerevisiae* and fruit fly (Tables 4 and 5; Supplementary Tables S1–6). Moreover, from Table 4 the CPPred outperformed CPAT, CPC2 and PLEK with MCC of 0.926 versus 0.923, 0.909 and 0.796 on Mouse-Testing. Besides, CPPred predicted better than CPAT, CPC2, PLEK and sORF finder with MCC of 0.518 versus 0.392, 0.140, 0.402 and 0.182 on Mouse-sORF-Testing in Table 5.

In addition, for the testing sets of zebrafish and *S. cerevisiae*, CPPred also achieved the best performance. The MCCs of CPPred in the Zebrafish-Testing, Zebrafish-sORF-Testing, *S. cerevisiae*-Testing and *S. cerevisiae*-sORF-Testing are 0.9, 0.387, 0.9 and 0.440, respectively (Supplementary Tables S1–4). However, the MCCs of CPPred in Fruit-fly-Testing and Fruit-fly-sORF-Testing are 0.837 and 0.225 (Supplementary Tables S5 and 6), which are worse than the MCCs of CPAT (MCCs are 0.916 and 0.565), respectively. The reason may be that the model of CPAT is trained on the fruit fly dataset, while the model of CPPred is trained on the human dataset.

It is noteworthy that, in most cases, the CPPred in the testing of sORF is much better than CPAT, CPC2, PLEK and sORF finder.

Performance of CPPred (Integrated-Model)

Due to the specific differences among the species, we tested Integrated-Model in the Integrated-Testing and Integrated-sORF-Testing, and the results are presented in Tables 6 and 7, respectively. CPPred performed better than CPAT (42), CPC2 (39), PLEK (34) and sORF finder (56) (see MCC, AUC and ACC in Tables 6 and 7). Noteworthy, the model of CPC2 is species-neutral (39), and the performance of CPC2

Table 2. Comparison of CPPred (Human-Model) and CPAT, CPC2, PLEK on Human-Testing

Method	SP (%)	SN (%)	PRE (%)	ACC (%)	*F*-score	AUC	MCC
CPPred	97.04	95.44	97.10	96.23	0.963	0.992	0.925
CPAT	94.07	94.58	94.30	94.33	0.944	0.984	0.887
CPC2	95.30	90.92	95.26	93.07	0.930	0.982	0.862
PLEK	98.10	95.42	98.11	96.73	0.967	0.993	0.935

Table 3. Comparison of CPPred (Human-Model) and CPAT, CPC2, PLEK, sORF finder on Human-sORF-Testing

Method	SP (%)	SN (%)	PRE (%)	ACC (%)	*F*-score	AUC	MCC
CPPred	97.97	63.34	96.90	80.66	0.766	0.928	0.654
CPAT	95.63	45.09	91.17	70.36	0.603	0.850	0.472
CPC2	95.48	11.23	71.29	53.35	0.194	0.799	0.125
PLEK	97.19	77.85	96.52	87.52	0.862	0.953	0.765
sORF finder	29.33	91.11	56.32	60.22	0.696	0.592	0.262

Table 4. Comparison of CPPred (Human-Model) and CPAT, CPC2, PLEK on Mouse-Testing

Method	SP(%)	SN(%)	PRE(%)	ACC(%)	F-score	AUC	MCC
CPPred	97.70	95.57	98.48	96.40	0.970	0.993	0.926
CPAT	96.65	96.10	97.81	96.32	0.970	0.993	0.923
CPC2	95.86	95.86	97.30	95.61	0.964	0.991	0.909
PLEK	93.43	87.61	95.41	89.88	0.913	0.969	0.796

Table 5. Comparison of CPPred (Human-Model) and CPAT, CPC2, PLEK, sORF finder on Mouse-sORF-Testing

Method	SP(%)	SN(%)	PRE(%)	ACC(%)	F-score	AUC	MCC
CPPred	97.00	46.81	92.96	74.00	0.623	0.906	0.518
CPAT	96.20	33.69	88.24	67.55	0.488	0.848	0.392
CPC2	95.10	12.77	68.79	57.37	0.215	0.789	0.140
PLEK	90.80	44.21	80.26	69.45	0.570	0.782	0.402
sORF finder	21.30	91.84	49.68	53.63	0.645	0.538	0.182

is worse than CPPred with MCC 0.869 versus 0.919 for Integrated-Testing and 0.502 versus 0.765 for Integrated-sORF-Testing.

OCTD features

To highlight the importance of CTD features, we built a model (OCTD-Model) by using only 30 CTD features (OCTD). The OCTD-Model is trained by Human-Training and then tested on Human-Testing and Human-sORF-Testing. As compared with OCTD-Model, we obtained NCTD-Model on Human-Training by using non-CTD features (i.e. ORF length, ORF coverage, Fickett score, Hexamer Score, Gravy, pI, ORF integrity and Instability) and then also tested on Human-Testing and Human-sORF-Testing. The results are shown as Supplementary Table S9. For Human-sORF-Testing, the results of OCTD-Model are much better than NCTD-Model with the accuracy of 79% versus 56% and MCC of 0.587 versus 0.214. But as Human-Testing, the results of OCTD-Model are slightly worse than NCTD-Model with the accuracy of 89% versus 93%, MCC of 0.783 versus 0.873. The above results show that CTD features are important in predicting the RNA coding potential, especially for sORF data. Overall, introducing the CTD features can improve the performance of CPPred on sORF data significantly.

The ability of CPPred to estimate novel coding RNAs

To further evaluate our method, we tested its capacity of predicting new coding RNAs and compared it with other methods (CPAT, CPC2, PLEK and sORF finder) on human and mouse. With time elapsing, some new coding RNAs were annotated. We collected the mRNA transcripts as new coding RNAs from 27 November 2017 to 3 April 3 2018. We obtained 74 novel coding RNAs of human, 3278 novel coding RNAs of mouse from the RefSeq database. Among them, there are five human coding RNAs only containing sORF (5 sORF-RNAs), 95 sORF-RNAs of mouse. Moreover, 1178 novel coding genes, which included 1335 coding transcripts, are extracted by Pertea *et al.* in May 2018 in bioRxiv (BioRxiv: https://doi.org/10.1101/332825). Subsequently, Jungreis *et al.* reported that nearly all the novel protein-coding predictions from Pertea *et al.* are false positives in July 2018 in bioRxiv (BioRxiv: https://doi.org/10.1101/360602). Our method CPPred and other methods are used to predict the coding potential of the novel coding RNAs, which are compared according to the number of coding RNAs predicted correctly. As can be seen from Table 8, CPPred correctly predicted 67 in 74 novel human coding RNAs and 3099 out of 3278 novel coding RNAs of mouse. Interestingly, for the 1335 recently annotated coding RNAs from the work of Pertea *et al.*, only 119 transcripts are predicted as coding RNAs by the CPPred. In fact, almost all novel coding RNAs are ncRNAs (91.1%), and the conclusion is consistent with the view of Jungreis *et al.* On the other hand, although the novel coding RNAs recently predicted by Pertea *et al.* have many false positives, there are still some transcripts (119 coding RNAs) that require attention and further experimental validation (Supplementary File S1.xlsx). Moreover, 4 out of 5 novel sORF-RNAs of human are correctly predicted and 35 among 95 novel sORF-RNAs of mouse are also predicted successfully. The performance of CPPred, CPAT, CPC2 and PLEK is similar on coding RNAs. For sORF-RNAs, CPPred outperformed CPAT and CPC2. However, our method CPPred did not perform as well as PLEK for the new mouse sORF-RNAs. PLEK showed a higher false positive rate in Mouse-

Table 6. Comparison of CPPred (Integrated-Model) and CPAT, CPC2, PLEK on Integrated-Testing

Method	SP (%)	SN (%)	PRE (%)	ACC (%)	*F*-score	AUC	MCC
CPPred	94.93	96.91	95.03	95.92	0. 960	0.990	0.919
CPAT	93.86	92.66	93.75	93.26	0.932	0.980	0.865
CPC2	95.54	91.27	95.34	93.40	0.933	0.979	0.869
PLEK	90.17	66.32	87.09	78.24	0.753	0.872	0.582

Table 7. Comparison of CPPred (Integrated-Model) and CPAT, CPC2, PLEK, sORF finder on Integrated-sORF-Testing

Method	SP (%)	SN (%)	PRE (%)	ACC (%)	*F*-score	AUC	MCC
CPPred	94.92	80.83	94.09	87.88	0. 870	0.955	0.765
CPAT	94.49	68.97	92.56	81.77	0.790	0.925	0.657
CPC2	96.20	47.65	92.62	71.93	0.629	0.891	0.502
PLEK	91.40	34.50	80.04	62.95	0. 482	0.737	0.315
sORF finder	55.35	67.09	60.04	61.22	0.634	0.560	0.226

Table 8. CPPred (Human-Model), CPAT, CPC2, PLEK and sORF finder are tested on the novel coding RNAs of human and mouse

Data type	Organism	Number of new data	CPPred	CPAT	CPC2	PLEK	sORF finder
All coding RNAs	Human	74	67	69	67	69	N/A
	Mouse	3278	3099	3155	3095	2961	N/A
	Human (Pertea *et al.*, 2018)	1335	119	177	248	105	N/A
Small coding RNAs	Human	5	4	0	0	1	5
	Mouse	95	35	32	8	43	95

sORF-Testing (Table 5), which may be the main reason for a higher accuracy rate predicted by PLEK than CPPred. Moreover, CPPred performed worse than sORF finder. The reason may be that the sORF finder has a higher false positive rate due to base on a single feature (Tables 3, 5 and 7; Supplementary Tables S2 and S6).

CONCLUSION

In this work, based on SVM classifier algorithm, we developed a tool CPPred to predict coding potential using multiple features, which are extracted from CPAT (42) and CPC2 (39), and CTD features are added particularly. Here, we used CTD features to predict coding potential for the first time in eukaryotes. Moreover, we found that the features of T2, C0 and GC (CTD features) play a key role in predicting coding potential.

Our method CPPred is trained on Human-Training and Integrated-Training to obtain the Human-Model and the Integrated-Model. As the former, CPPred is tested on the dataset of human, mouse, zebrafish, *S. cerevisiae* and fruit fly, obtaining AUC from 0.72 to 0.99. Besides, our CPPred is compared with other methods CPAT, CPC2, PLEK and sORF finder. The CPPred outperforms CPAT, CPC2, PLEK and sORF finder on the testing sets of mouse, zebrafish and *S. cerevisiae*. However, CPPred does not perform as well as PLEK on human testing sets, which may be due to the high level of redundancy between the testing set of human and PLEK's training set (84) (Tables 4 and 5). We analyzed the testing set of human and PLEK's training set, and found 46.1% coding RNAs and 65.7% ncRNAs redundant sequences between the human testing set of CPPred and human training set of PLEK, respectively. For fruit fly, CPPred performs worse than CPAT, which may be due to the fact that CPAT is trained on fruit fly while CPPred is trained on human dataset. Thus, the second model is built by Integrated-Training. From Tables 6 and 7, we compared CPPred with other tools (CPAT, CPC2, PLEK and sORF finder) and found some improvement in MCCs by >5% and >11% on Integrated-Testing and Integrates-sORF-Testing, respectively. Moreover, the CTD features are particularly important for predicting the coding potential of sORF datasets (Supplementary Table S9). Overall, the results demonstrate that CPPred performs well on long RNA datasets and much better than other tools on sORF datasets.

DATA AVAILABILITY

Source code was implemented in Python and is freely available at http://www.rnabinding.com/CPPred.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the National Supercomputer Center in Guangzhou for the support of computing resources.

FUNDING

National Natural Science Foundation of China [31100522]; National High Technology Research and Development Program of China [2012AA020402]; the Fundamental Research Funds for the Central Universities [2016YXMS017]; Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) [U1501501]. Funding for open access charge: Fundamental Research Funds for the Central Universities [2016YXMS017].

Conflict of interest statement. None declared.

REFERENCES

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

2. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
3. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
4. Junttila,S. and Rudd,S. (2012) Characterization of a transcriptome from a non-model organism, Cladonia rangiferina, the grey reindeer lichen, using high-throughput next generation sequencing and EST sequence data. *BMC Genomics*, **13**, 575.
5. Wang,Y., Li,Y., Wang,Q., Lv,Y., Wang,S., Chen,X., Yu,X., Jiang,W. and Li,X. (2014) Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. *Gene*, **533**, 94–99.
6. Pauli,A., Valen,E., Lin,M.F., Garber,M., Vastenhouw,N.L., Levin,J.Z., Fan,L., Sandelin,A., Rinn,J.L., Regev,A. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, **22**, 577–591.
7. Hannon,G.J. (2002) RNA interference. *Nature*, **418**, 244–251.
8. Machado-Lima,A., Del,P.H. and Durham,A.M. (2008) Computational methods in noncoding RNA research. *J. Math Biol.*, **56**, 15–49.
9. Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat Rev Genet*, **15**, 423–437.
10. Jamalkandi,S.A. and Masoudi-Nejad,A. (2009) Reconstruction of Arabidopsis thaliana fully integrated small RNA pathway. *Funct. Integr. Genomics*, **9**, 419–432.
11. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, **25**, 1915–1927.
12. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.
13. Hung,T. and Chang,H.Y. (2010) Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol.*, **7**, 582–585.
14. Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
15. Cheetham,S.W., Gruhl,F., Mattick,J.S. and Dinger,M.E. (2013) Long noncoding RNAs and the genetics of cancer. *Br J Cancer*, **108**, 2419–2425.
16. Batista,P.J. and Chang,H.Y. (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell*, **152**, 1298–1307.
17. Kondo,T., Hashimoto,Y., Kato,K., Inagaki,S., Hayashi,S. and Kageyama,Y. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol.*, **9**, 660–665.
18. Galindo,M.I., Pueyo,J.I., Fouix,S., Bishop,S.A. and Couso,J.P. (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.*, **5**, e106.
19. Kondo,T., Plaza,S., Zanet,J., Benrabah,E., Valenti,P., Hashimoto,Y., Kobayashi,S., Payre,F. and Kageyama,Y. (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science*, **329**, 336–339.
20. Pauli,A., Norris,M.L., Valen,E., Chew,G.L., Gagnon,J.A., Zimmerman,S., Mitchell,A., Ma,J., Dubrulle,J., Reyon,D. *et al.* (2014) Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science*, **343**, 1248636.
21. Chng,S.C., Ho,L., Tian,J. and Reversade,B. (2013) ELABELA: a hormone essential for heart development signals via the apelin receptor. *Dev. Cell*, **27**, 672–680.
22. Magny,E.G., Pueyo,J.I., Pearl,F.M., Cespedes,M.A., Niven,J.E., Bishop,S.A. and Couso,J.P. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, **341**, 1116–1120.
23. Anderson,D.M., Anderson,K.M., Chang,C.L., Makarewich,C.A., Nelson,B.R., McAnally,J.R., Kasaragod,P., Shelton,J.M., Liou,J., Bassel-Duby,R. *et al.* (2015) A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, **160**, 595–606.
24. Nelson,B.R., Makarewich,C.A., Anderson,D.M., Winders,B.R., Troupes,C.D., Wu,F., Reese,A.L., McAnally,J.R., Chen,X., Kavalali,E.T. *et al.* (2016) A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, **351**, 271–275.
25. Rohrig,H., Schmidt,J., Miklashevichs,E., Schell,J. and John,M. (2002) Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 1915–1920.
26. Zhu,Y., Orre,L.M., Johansson,H.J., Huss,M., Boekel,J., Vesterlund,M., Fernandez-Woodbridge,A., Branca,R. and Lehtio,J. (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.*, **9**, 903.
27. Nesvizhskii,A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*, **11**, 1114–1125.
28. Omasits,U., Varadarajan,A.R., Schmid,M., Goetze,S., Melidis,D., Bourqui,M., Nikolayeva,O., Quebatte,M., Patrignani,A., Dehio,C. *et al.* (2017) An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res.*, **27**, 2083–2095.
29. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
30. Ma,J., Ward,C.C., Jungreis,I., Slavoff,S.A., Schwaid,A.G., Neveu,J., Budnik,B.A., Kellis,M. and Saghatelian,A. (2014) Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.*, **13**, 1757–1765.
31. Olexiouk,V., Van Criekinge,W. and Menschaert,G. (2018) An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.*, **46**, D497–D502.
32. Crappe,J., Van Criekinge,W., Trooskens,G., Hayakawa,E., Luyten,W., Baggerman,G. and Menschaert,G. (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, **14**, 648.
33. Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.
34. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, **15**, 311.
35. Sun,K., Chen,X., Jiang,P., Song,X., Wang,H. and Sun,H. (2013) iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics*, **14**(Suppl. 2), S7.
36. Schneider,H.W., Raiol,T., Brigido,M.M., Walter,M. and Stadler,P.F. (2017) A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts. *BMC Genomics*, **18**, 804.
37. Tripathi,R., Patel,S., Kumari,V., Chakraborty,P. and Varadwaj,P.K. (2016) DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Net. Model. Anal. Health Inform. Bioinform.*, **5**, 21.
38. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
39. Kang,Y.J., Yang,D.C., Kong,L., Hou,M., Meng,Y.Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16
40. Wucher,V., Legeai,F., Hedan,B., Rizk,G., Lagoutte,L., Leeb,T., Jagannathan,V., Cadieu,E., David,A., Lohi,H. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*, **45**, e57.
41. Hu,L., Xu,Z., Hu,B. and Lu,Z.J. (2017) COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.*, **45**, e2.
42. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.P. and Li,W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
43. Liu,J., Gough,J. and Rost,B. (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLos Genet.*, **2**, e29.
44. Pian,C., Zhang,G., Chen,Z., Chen,Y., Zhang,J., Yang,T. and Zhang,L. (2016) LncRNApred: classification of long non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. *PLoS One*, **11**, e154567.

45. Sun,L., Liu,H., Zhang,L. and Meng,J. (2015) lncRScan-SVM: a tool for predicting long Non-Coding rnas using support vector machine. *PLoS One*, **10**, e139654.
46. McGillivray,P., Ault,R., Pawashe,M., Kitchen,R., Balasubramanian,S. and Gerstein,M. (2018) A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.*, **46**, 3326–3338.
47. Li,H., Xiao,L., Zhang,L., Wu,J., Wei,B., Sun,N. and Zhao,Y. (2018) FSPP: A tool for Genome-Wide prediction of smORF-Encoded peptides and their functions. *Front. Genet.*, **9**, 96.
48. Dubchak,I., Muchnik,I., Holbrook,S.R. and Kim,S.H. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 8700–8704.
49. Han,G., Yu,Z., Anh,V. and Chan,R.H. (2009), Distinguishing coding from non-coding sequence in a prokaryote complete genome based on the global descriptor. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China*. pp. 42–46.
50. Vandivier,L.E., Anderson,S.J., Foley,S.W. and Gregory,B.D. (2016) The conservation and function of RNA secondary structure in plants. *Annu. Rev. Plant. Biol.*, **67**, 463–488.
51. Mortimer,S.A., Kidwell,M.A. and Doudna,J.A. (2014) Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.*, **15**, 469–479.
52. Zhang,X. and Liu,S. (2017) RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics*, **33**, 854–862.
53. Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res.*, **29**, 3928–3938.
54. Yang,J.Y., Zhou,Y., Yu,Z.G., Anh,V. and Zhou,L.Q. (2008) Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinformatics*, **9**, 113.
55. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
56. Hanada,K., Akiyama,K., Sakurai,T., Toyoda,T., Shinozaki,K. and Shiu,S.H. (2010) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics*, **26**, 399–400.
57. Cheng,H., Chan,W.S., Li,Z., Wang,D., Liu,S. and Zhou,Y. (2011) Small open reading frames: current prediction techniques and future prospect. *Curr. Protein Pept. Sci.*, **12**, 503–507.
58. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
59. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
60. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
61. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
62. Lertampaiporn,S., Thammarongtham,C., Nukoolkit,C., Kaewkamnerdpong,B. and Ruengjitchatchawalya,M. (2014) Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. *Nucleic Acids Res.*, **42**, e93.
63. Liu,B., Fang,L., Liu,F., Wang,X., Chen,J. and Chou,K.C. (2015) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One*, **10**, e121501.
64. Sun,L., Liu,H., Zhang,L. and Meng,J. (2015) lncRScan-SVM: A tool for predicting long Non-Coding RNAs using support vector machine. *PLoS One*, **10**, e139654.
65. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
66. Yang,C., Yang,L., Zhou,M., Xie,H., Zhang,C., Wang,M.D. and Zhu,H. (2018) LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, **34**, 3825–3834.
67. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–i282.
68. Sun,L., Luo,H., Bu,D., Zhao,G., Yu,K., Zhang,C., Liu,Y., Chen,R. and Zhao,Y. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.*, **41**, e166.
69. Achawanantakun,R., Chen,J., Sun,Y. and Zhang,Y. (2015) LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, **31**, 3897–3905.
70. Zhao,J., Song,X. and Wang,K. (2016) lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. *Sci. Rep.*, **6**, 34838.
71. Hill,S.T., Kuintzle,R., Teegarden,A., Merrill,E.R., Danaee,P. and Hendrix,D.A. (2018) A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.*, **46**, 8105–8113.
72. Fickett,J.W. and Tung,C.S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
73. Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
74. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.
75. Lee,S., Liu,B., Lee,S., Huang,S.X., Shen,B. and Qian,S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2424–E2432.
76. Gao,X., Wan,J., Liu,B., Ma,M., Shen,B. and Qian,S.B. (2015) Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods*, **12**, 147–153.
77. Mackowiak,S.D., Zauber,H., Bielow,C., Thiel,D., Kutz,K., Calviello,L., Mastrobuoni,G., Rajewsky,N., Kempa,S., Selbach,M. *et al.* (2015) Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.*, **16**, 179.
78. He,Z., Zhang,J., Shi,X.H., Hu,L.L., Kong,X., Cai,Y.D. and Chou,K.C. (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.
79. Li,B.Q., Hu,L.L., Niu,S., Cai,Y.D. and Chou,K.C. (2012) Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. *J. Proteomics*, **75**, 1654–1665.
80. Li,B.Q., Feng,K.Y., Chen,L., Huang,T. and Cai,Y.D. (2012) Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PLoS One*, **7**, e43927.
81. Ma,X. and Sun,X. (2014) Sequence-based predictor of ATP-binding residues using random forest and mRMR-IFS feature selection. *J. Theor. Biol.*, **360**, 59–66.
82. Peng,H., Long,F. and Ding,C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern. Anal. Mach. Intell.*, **27**, 1226–1238.
83. Chang,C. and Lin,C. (2011) LIBSVM. *ACM T Intel. Syst. Tec.*, **2**, 1–27.
84. Achawanantakun,R., Chen,J., Sun,Y. and Zhang,Y. (2015) LncRNA-ID: long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, **31**, 3897–3905.