W12–W16 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 18 May 2017*
doi: 10.1093/nar/gkx428

CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features

Yu-Jian Kang†, De-Chang Yang†, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei and Ge Gao*

State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Center for Bioinformatics, Peking University, Beijing 100871, People's Republic of China

Received March 01, 2017; Revised April 30, 2017; Editorial Decision May 02, 2017; Accepted May 03, 2017

ABSTRACT

With advances in next-generation sequencing technologies, numerous novel transcripts in a large number of organisms have been identified. With the goal of fast, accurate assessment of the coding ability of RNA transcripts, we upgraded the coding potential calculator CPC1 to CPC2. CPC2 runs ∼1000 times faster than CPC1 and exhibits superior accuracy compared with CPC1, especially for long non-coding transcripts. Moreover, the model of CPC2 is species-neutral, making it feasible for ever-growing non-model organism transcriptomes. A mobile-friendly web server, as well as a downloadable standalone package, is freely available at http://cpc2.cbi.pku.edu.cn.

INTRODUCTION

Recent studies have well demonstrated that non-coding RNAs (ncRNAs) are pervasively transcribed from plant to animal genomes (1–4). Increasing evidences indicate that these ncRNAs play critical roles in numbers of important cellular processes, including transcriptional inhibition mediated by microRNAs (5), epigenetic inheritance by Piwi-interacting RNAs (6), cell-cycle regulation (7) or even acting as structural components in ribosomes (8).

With advances in next-generation sequencing technologies, numerous novel transcripts in a large number of diverse organisms, including several non-model ones, have been discovered in rapidly increasing RNA-seq data (9–12). Effective and efficient identification of ncRNAs in the massive dataset is an essential step for following-up function and evolution studies, and demands a fast, accurate and species-neutral assessment tool (13–19).

As a response to the challenge, we updated our Coding Potential Calculator (CPC) algorithm (20) to version 2. Employing a novel discriminative model based on four sequence intrinsic features, CPC2 not only runs ∼1000 times faster than CPC1 but is also more accurate. In addition, CPC2 is species-neutral, making it more useful for the ever-growing non-model organism transcriptomes. CPC2 is available freely at http://cpc2.cbi.pku.edu.cn as both a web server and a downloadable standalone package.

MATERIALS AND METHODS

To identify discriminative features, we first compiled a candidate list of sequence intrinsic features (i.e. features can be derived from transcript sequence directly) based on literature survey (see Supplementary Table S1). A hierarchical feature selection procedure was employed to identify effective features with recursive feature elimination method (random forest functions with 10-fold cross-validation, implemented with the caret R package (21)) adopted in each stage (see Supplementary Figure S1 for details). We identified a final set of four intrinsic features as Fickett TESTCODE score, open reading frame (ORF) length, ORF integrity and isoelectric point (pI). While the Fickett TESTCODE score is derived from the weighted nucleotide frequency of the inputted full length transcript (22), the rest of three features (ORF length, ORF integrity and isoelectric point) are calculated based on the longest putative ORF identified *in silico* (see http://cpc2.cbi.pku.edu.cn/help/feature_selection.php for the full candidate list as well as the script).

We then trained a support vector machine (SVM) model using these four intrinsic features. The LIBSVM (23) package was employed to train an SVM model using the standard radial basis function kernel (RBF kernel) with the training dataset containing 17 984 high-confident human protein-coding transcripts and 10 452 non-coding transcripts (18).

To evaluate the performance of CPC2 across species, we further built an independent testing set for human, mouse, zebrafish, fly, worm and the model plant *Arabidopsis*. We selected protein-coding and non-coding transcripts that met rigorous criteria to obtain a testing set of high quality: for the protein-coding testing set, we obtained all non-predicted mRNAs from the RefSeq database (24) with protein sequences annotated by Swiss-Prot (25) and redundant sequences (i.e identity $\geq$ 0.9) removed using CD-hit with default parameters. Non-coding transcripts were ob-

*To whom correspondence should be addressed. Tel: +86 10 6275 5206; Fax: +86 10 6275 5206; Email: gaog@mail.cbi.pku.edu.cn
†These authors contributed equally to the paper as first authors.

tained from the Ensembl (v87) (26) and EnsemblPlants (v32) (26) databases with transcript status as 'KNOWN'. All sequences in training set were further excluded (Table 1). The full training set and testing set are available for downloading as FASTA files at http://cpc2.cbi.pku.edu.cn/help/data_set.php.

We employed standard performance measurements including sensitivity, specificity and accuracy, with protein-coding calls defined as 'positive' and non-coding calls as 'negative'. The abbreviations in the equations below are as follows: FN, false negative; FP, false positive; TN, true negative; and TP, true positive.

$$\text{Sensitivity} = \frac{TP}{TP + FN}\ ;\ \text{Specificity} = \frac{TN}{TN + FP}$$

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

Back-end of the CPC2 web server is implemented in PHP running on Apache web server. The front-end interface is powered by JavaScript libraries Bootstrap (http://getbootstrap.com/), JQuery (http://jquery.com/), Tablecloth (http://cssglobe.com/lab/tablecloth/) as well as Highcharts (http://www.highcharts.com/).

RESULTS

CPC2 is fast, accurate and species-neutral

Given the large volume of transcriptome data generated by next generation sequencing, the efficiency is becoming vital for a useful tool in the real world. To measure the computational speed, we first randomly selected a sample of 200 sequences that consisted of 100 mRNAs and 100 lncRNAs from the human testing dataset. CPC2 completed its analysis in 1.8 s, whereas CPC1 required >1000-fold time (2815 s) on Intel Xeon E7-8830 2.13GHz CPU in single thread mode. To further evaluate the real world efficiency, we then measured the computational speed on all the coding and non-coding transcripts in Ensembl v87 (26) with gene and transcript status annotated as 'KNOWN'. This dataset consists of 597 996 protein-coding transcripts and 55 277 non-coding transcripts from 69 organisms, which is more similar to the circumstances of users' input. Similar to previous result, CPC2 showed a significant speedup (42 min) than CPC1 (4783 min).

In addition to being efficient, a sensible tool should pose high accuracy in a robust and species-neutral fashion across different organisms. Designed to use rather stringent criteria for non-coding calls, the CPC1 exhibits high sensitivity and relative poor specificity. As many important biological roles of long ncRNAs (lncRNAs) have been revealed by recent studies performed in this decade (7), CPC2 adopted a more balanced calling of protein-coding and non-coding transcripts, which is more suitable for current transcriptome studies. To evaluate the performance across various species, we ran both CPC1 and CPC2 against human, mouse, zebrafish, fly, worm and plant (*Arabidopsis*) testing set. The CPC2 showed better overall accuracy (0.961) than of CPC1 (0.932) with a much more improved specificity (0.970 versus 0.873) and a slightly lower sensitivity (0.952 versus 0.995).

A

		Dataset Size	CPC2	CPC1
Noncoding RNAs	Small ncRNAs	20,649	20,649(100%)	20,457(99.1%)
	Long ncRNAs	22,028	20,754(94.2%)	16,793(76.2%)
Coding RNAs	mRNAs	40,341	38,400(95.2%)	40,150(99.5%)
Overall		**83,018**	**79,803(96.1%)**	**77,400(93.2%)**

B

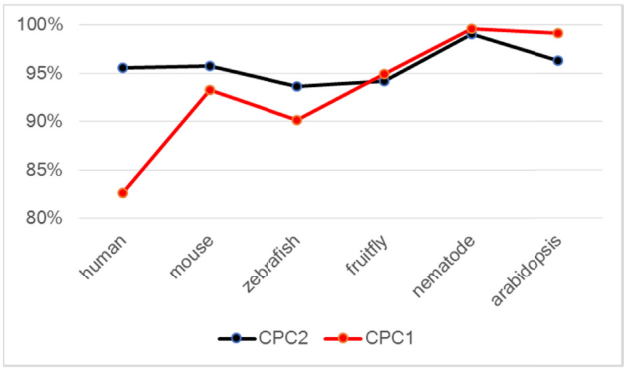

Figure 1. Evaluation on accuracy of CPC1 and CPC2 in six species. (**A**) The overall accuracy (**B**) the detailed accuracy in six organisms. The Long ncRNAs were defined as non-coding RNAs longer than 200 nt.

In particular, the CPC2 exhibited superior accuracy (0.942) for long non-coding transcripts, a newly discovered key regulators in several physiological and pathological processes (27–30), than of CPC1 (0.762, Figure 1A). Further comparison with other popular tools (14,17,19) also confirmed CPC2′s superior performance (Supplementary Figure S2).

Even the underlying model in CPC2 was trained based on transcript sequences from human only (the training set used in CPC1 is consist of sequences from multiple organisms), the CPC2 showed a more robust performance across species, with accuracy varied from 0.937 to 0.991 (from 0.826 to 0.997 for CPC1, Figure 1B), which may partly due to the fact that only sequence intrinsic features were employed in CPC2. In particular, while CPC1 shows higher accuracy than CPC2 in *Arabidopsis*, the inter-species variance of accuracy of CPC2 (0.04%) is one order of magnitude lower than CPC1 (0.4%) (Figure 1B), a property that we considered 'species neutral' (also see http://cpc2.cbi.pku.edu.cn/help/species_neutral.php for more details).

The web server of CPC2

For users to access CPC2 conveniently, we established a new web portal at http://cpc2.cbi.pku.edu.cn/. Briefly, the CPC2 web server accepts RNA transcripts as input and outputs its coding probability with detailed supporting features for the coding/non-coding call (Figure 2).

CPC2 web server currently supports both 'interactive mode', in which the nucleotide sequences in FASTA format can be directly copied and pasted into the input box at the home page, and 'batch mode' in which users can upload a local file in either FASTA format or BED/GTF/GFF format. When a new analysis task is submitted, a unique 'Task ID' (TID) will be assigned for tracking the analysis progress and retrieving results later.

Table 1. The independent testing set in human, mouse, zebrafish, fly, worm and the model plant *Arabidopsis thaliana*

Dataset type	Human	Mouse	Zebrafish	Fly	Worm	*Arabidopsis*
Coding	6142	10 638	2344	3680	3551	13 986
Non-coding	12 019	12 251	1528	3556	9470	3853

All testing sets are available for downloading as FASTA file at http://cpc2.cbi.pku.edu.cn/help/data_set.php.

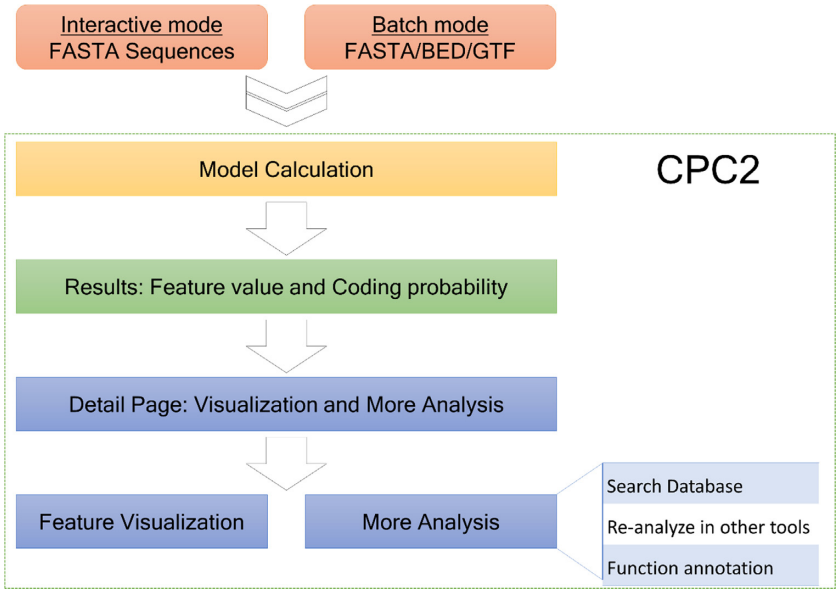

Figure 2. Workflow of the CPC2 web server.

As in CPC1, the results will be presented as an intuitive table online which can also be downloaded as a tabular file for further analysis (Figure 3A). In addition, detailed information of each transcript is provided in a separated 'detailed' page, including a summary paragraph, a graphic view of features' distribution in known protein-coding and non-coding transcripts and additional functions (Figure 3B). More analysis such as querying against known databases, re-analyzing in alternative methods and annotating functions can also be run performed for given transcript (Figure 3C and D).

The CPC2 web server implemented a responsive layout, enabling the optimal view for both desktop PCs and mobile devices. A standalone package of CPC2 can also be freely downloaded at http://cpc2.cbi.pku.edu.cn/download.php.

Example

We utilized online CPC2 on a human lncRNA *MEG3* as an example. After inputting its sequence, CPC2 predicted it as a non-coding transcript (Figure 3A). By clicking the 'View' on the last column, more detailed information is shown.

The details page is divided into three parts. A description of *MEG3* summarizing its coding probability and feature values is presented at the top (Figure 3B). In the middle of this page, an interactive visualization of three supporting features including Fickett score, peptide length (synonymous with ORF length) and pI are provided. Taking the graph of peptide length as an example, the black box indicates that *MEG3* has a peptide length of 106 aa and was classified as non-coding. In addition, the position of *MEG3* is noted in the background (Figure 3B). The blue area shows the feature's distribution in non-coding transcripts, whereas the orange one represents protein coding transcripts. Passing the mouse over the distribution curve, the feature value and transcripts frequency of the interval are displayed in a textbox. The static visualizations can be easily downloaded (Figure 3B).

At the bottom, CPC2 also provides additional functions to facilitate the coding/non-coding classification of input sequences (Figure 3C). The first function is querying the transcript against well-annotated databases, including Swiss-Prot (24), RNAdb (31) and lncRNAdb (32) by BLAST (33), to identify more evidence. By placing the mouse over the results, users can view details of predicted ORF and BLAST hits of *MEG3* (Figure 3D). Moreover, the user can also send sequences to alternative tools like CPC1, CPAT and PORTRAIT for re-analysis through the 'Re-analyze' button.

SUMMARY

Employing a novel discriminative model, we upgraded our CPC to version 2. CPC2 runs ∼1000 times faster than CPC1. In addition, the CPC2 model is species-neutral, making it useful for ever-growing non-model organism transcriptomes and even transcriptomes of organisms that are poorly annotated or lack genome assembly. CPC2 is more accurate than CPC1, especially for long non-coding transcripts. In addition, the online CPC2 provides an in-

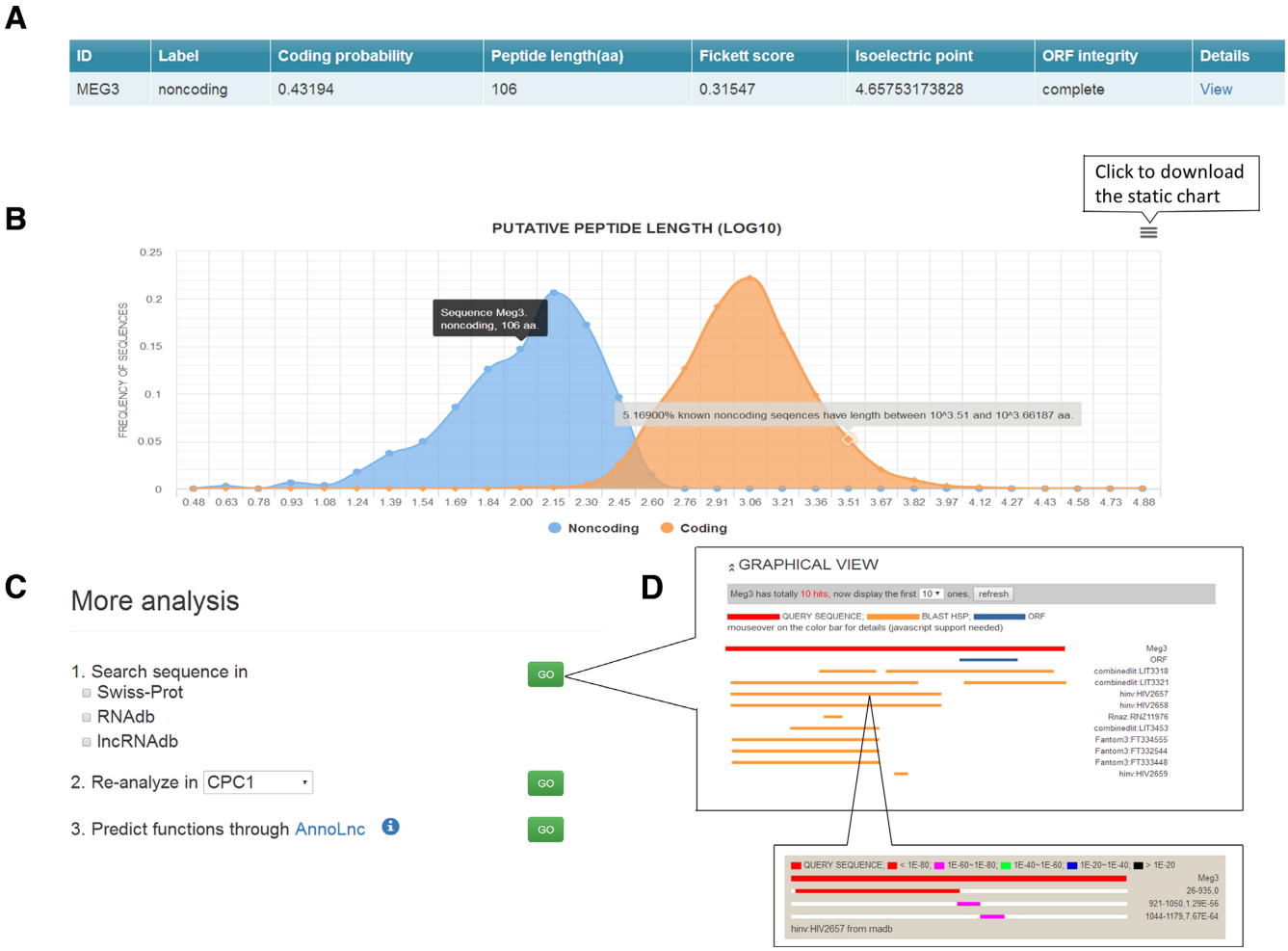

Figure 3. Screenshot of the CPC2 web server. (**A**) Summary tabular output with coding probability; (**B**) graphical view of features' distribution in the 'Details' page; (**C**) more analysis for querying against known databases, re-analyzing in alternative methods and annotating functions; (**D**) the rendered BLAST output including both ORF position and BLAST hits in queried databases.

formative graphic view of results and more integrated functions. The web server is mobile-friendly and more accessible on mobile devices such as the iPad.

Independent of external resources, CPC2 adopted four sequence intrinsic features that are easily comprehensible and biologically meaningful. At the DNA level, the Fickett score captures the position of each base favored in the sequence (18). At the RNA level, ORF length and integrity are powerful because the protein-coding transcript is more likely to have a long and high-quality ORF. Moreover, based on the assumption that the hypothetical peptide identified in a non-coding transcript should have different chemical properties than these real ones encoded by *bona fide* coding sequences, we also added several peptide level features into the candidate list, and eventually adopted pI in the final SVM model.

Since the first release of CPC1 at 2007, number of statistic-based tools have been developed to distinguish non-coding and protein-coding transcripts based on multiple lines of evidences. Many of them show high levels of accuracy (13–20). We hereby argue that the community should, in the coming years, shift from continuous improvement of discriminative performance to biological insights revealed by their statistical models which might further shed light onto the ultimate discriminative mechanism used by the Mother Nature.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Drs Cheng Li, Zemin Zhang and Jian Lu at Peking University for their helpful comments and suggestions during the study.

FUNDING

National Key Research and Development Program [2016YFC0901603]; China 863 Program [2015AA020108]; State Key Laboratory of Protein and Plant Gene Research; National Program for Support of Top-notch Young Professionals (to G.G.) (in part). Part of the analysis was

performed on the Computing Platform of the Center for Life Sciences of Peking University. Funding for open access charge: National Key Research and Development Program [2016YFC0901603].
Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
2. Fu,X.D. (2014) Non-coding RNA: a new frontier in regulatory biology. *Natl. Sci. Rev.*, **1**, 190–204.
3. He,S., Liu,C., Skogerbo,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
4. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.*, **15**, R17–R29.
5. Ambros,V. (2001) microRNAs: tiny regulators with great potential. *Cell*, **107**, 823–826.
6. Brennecke,J., Malone,C.D., Aravin,A.A., Sachidanandam,R., Stark,A. and Hannon,G.J. (2008) An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, **322**, 1387–1392.
7. Rinn,J.L. and Chang,H.Y. (2012) Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, **81**, 145–166.
8. Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.*, **31**, 442–443.
9. Cahais,V., Gayral,P., Tsagkogeorga,G., Melo-Ferreira,J., Ballenghien,M., Weinert,L., Chiari,Y., Belkhir,K., Ranwez,V. and Galtier,N. (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Mol. Ecol. Resour.*, **12**, 834–845.
10. Ellegren,H. and Galtier,N. (2016) Determinants of genetic diversity. *Nat. Rev. Genet.*, **17**, 422–433.
11. 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**, 575–584.
12. Schunter,C., Vollmer,S.V., Macpherson,E. and Pascual,M. (2014) Transcriptome analyses and differential gene expression in a non-model fish species with alternative mating tactics. *BMC Genomics*, **15**, 167–179.
13. Arrial,R.T., Togawa,R.C. and Brigido,M.M. (2009) Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics*, **10**, 239–247.
14. 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.
15. 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**, 311–320.
16. 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.
17. 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.
18. 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.
19. Washietl,S., Findeiss,S., Muller,S.A., Kalkhof,S., von Bergen,M., Hofacker,I.L., Stadler,P.F. and Goldman,N. (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578–594.
20. 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.
21. Kuhn,M. (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software, **28**, https://www.jstatsoft.org/article/view/v028i05.
22. Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
23. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM : a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.
24. 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.
25. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: how to Use the Entry View. *Methods Mol. Biol.*, **1374**, 23–54.
26. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
27. Kitagawa,M., Kitagawa,K., Kotake,Y., Niida,H. and Ohhata,T. (2013) Cell cycle regulation by long non-coding RNAs. *Cell Mol. Life Sci.*, **70**, 4785–4794.
28. Lee,J.T. and Bartolomei,M.S. (2013) X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell*, **152**, 1308–1323.
29. Ng,S.Y., Johnson,R. and Stanton,L.W. (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.*, **31**, 522–533.
30. Park,J.Y., Lee,J.E., Park,J.B., Yoo,H., Lee,S.H. and Kim,J.H. (2014) Roles of long non-coding RNAs on tumorigenesis and glioma development. *Brain Tumor Res. Treat.*, **2**, 1–6.
31. Pang,K.C., Stephen,S., Dinger,M.E., Engstrom,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
32. Quek,X.C., Thomson,D.W., Maag,J.L., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
33. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.