Published online 17 January 2013

Nucleic Acids Research, 2013, Vol. 41, No. 6 e74
doi:10.1093/nar/gkt006

CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model

Liguo Wang[1,2,3], Hyun Jung Park[2,3], Surendra Dasari[1], Shengqin Wang[4], Jean-Pierre Kocher[1,*] and Wei Li[2,3,*]

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN 55905, USA, [2]Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA, [3]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA and [4]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 210000, China

Received October 10, 2012; Revised December 30, 2012; Accepted January 2, 2013

ABSTRACT

Thousands of novel transcripts have been identified using deep transcriptome sequencing. This discovery of large and 'hidden' transcriptome rejuvenates the demand for methods that can rapidly distinguish between coding and noncoding RNA. Here, we present a novel alignment-free method, Coding Potential Assessment Tool (CPAT), which rapidly recognizes coding and noncoding transcripts from a large pool of candidates. To this end, CPAT uses a logistic regression model built with four sequence features: open reading frame size, open reading frame coverage, Fickett TESTCODE statistic and hexamer usage bias. CPAT software outperformed (sensitivity: 0.96, specificity: 0.97) other state-of-the-art alignment-based software such as Coding-Potential Calculator (sensitivity: 0.99, specificity: 0.74) and Phylo Codon Substitution Frequencies (sensitivity: 0.90, specificity: 0.63). In addition to high accuracy, CPAT is approximately four orders of magnitude faster than Coding-Potential Calculator and Phylo Codon Substitution Frequencies, enabling its users to process thousands of transcripts within seconds. The software accepts input sequences in either FASTA- or BED-formatted data files. We also developed a web interface for CPAT that allows users to submit sequences and receive the prediction results almost instantly.

INTRODUCTION

Although the human genome sequence was released a decade ago, the role of functional noncoding RNAs (ncRNAs) is much less understood compared with their coding counterparts. Several previous studies have demonstrated that the human genome is pervasively transcribed (1–4), but thoroughly cataloging all the RNA species (especially ncRNA) is challenging. Undiscovered ncRNAs might be rare, transient or beyond the detection limits of conventional approaches. Furthermore, ncRNAs also tend to be idiosyncratic to species and tissues (5,6). Nevertheless, advances in RNA-Seq have provided a new method of surveying the whole transcriptome to an unprecedented degree. Recent genome-wide studies revealed tens of thousands of novel transcripts, the majority of which were long noncoding RNAs (lncRNAs, >200 nt) (4–9). Although a few dozen lncRNAs have been characterized to some extent and are reported to have critical roles in diverse cellular and disease development processes (6,10–14), the biogenesis and function of most lncRNAs remain unclear.

Accurate and quantitative assessment of coding potential is the first step toward comprehensive annotation of newly discovered transcripts. Until now, prediction of coding potential heavily relied on sequence alignment, either pairwise homology search for protein evidence such as that used in the Coding-Potential Calculator (CPC) and PORTRAIT methods (15,16) or multiple alignments to calculate the phylogenetic conservation score such as that used in the Phylogenetic Codon Substitution Frequencies (PhyloCSF) and RNAcode methods (17,18). Alignment-based approaches are particularly useful for highly conserved protein-coding genes and, to a lesser extent, short genes encoding housekeeping or regulatory RNAs (e.g. snRNAs, snoRNA, transfer RNA). However, these approaches cannot immediately apply to all the novel transcripts because of several intrinsic limitations. First, most newly discovered

*To whom correspondence should be addressed. Tel: +1 507 538 8315; Fax: +1 507 284 0360; Email: kocher.jeanpierre@mayo.edu
Correspondence may also be addressed to Wei Li. Tel: +1 713 798 7854; Fax: +1 713 798 6822; Email: WL1@bcm.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

transcripts are lncRNAs, which tend to be lineage specific and less conserved (5,6). This greatly limits the discriminatory power of alignment-based methods. For example, only 29 of 550 lncRNAs identified from zebrafish had detectable sequence similarity with putative mammalian orthologs (6), and only 993 of 8195 human lncRNAs have orthologous transcripts in other species (5). Second, considerable fractions of lncRNAs are overlapped with either the sense or antisense strand of protein-coding genes. These lncRNAs cannot be correctly classified by homology searching because they would have significant matches to protein-coding genes (3,8,19). Third, the reliability of alignment-based approaches largely depends on the quality of alignments (20). This is problematic because most widely used multiple-sequence alignment tools use heuristics and do not guarantee optimal alignments. Finally, alignment-based methods are extremely time-consuming. For instance, CPC and PhyloCSF took 2 days to evaluate the coding potential of 14 353 lncRNAs identified by Cabili *et al.* (5). This problem is getting more attention as massive-scale RNA sequencing is increasingly being performed. Consequently, a more accurate, robust and faster method that does not rely on sequence alignment is needed to distinguish ncRNAs, especially lncRNAs, from protein-coding genes.

Here, we present Coding-Potential Assessment Tool (CPAT), an alignment-free program, which uses logistic regression to distinguish between coding and noncoding transcripts on the basis of four sequence features. CPAT is highly accurate (0.967) and extremely efficient (10 000 times faster than CPC and PhyloCSF, and 50 times faster than PORTRAIT). CPAT needs only the sequence or coordinate file as input, and it is straightforward to use. We expanded the availability of CPAT to a larger scientific audience via a web interface, which allows users to submit sequences and receive the prediction results back almost instantaneously (http://lilab.research.bcm.edu/cpat/index.php).

MATERIALS AND METHODS

Coding-potential prediction is essentially a binary decision problem, which makes logistic regression a suitable approach. As an alignment-free method, all selected features (predictor variables) were calculated directly from the sequence. The first feature was the maximum length of the open reading frame (ORF). ORF length is one of the most fundamental features used to distinguish ncRNA from messenger RNA because a long putative ORF is unlikely to be observed by random chance in noncoding sequences. Despite the simplicity, ORF length has high concordance with more sophisticated discrimination methods and remains the primary criterion in almost all coding-potential prediction methods (21). The second feature was ORF coverage defined as the ratio of ORF to transcript lengths. This feature also has good classification power, and it is highly complementary to, and independent of, the ORF length (Supplementary Figures S1 and S3). Some large bona fide ncRNAs may contain putative long ORFs by random chance (5), and thus cannot be classified correctly by ORF length alone. Fortunately, those large ncRNAs usually have much lower ORF coverage than protein-coding RNAs (Figure 1B).

The third feature we used was the Fickett TESTCODE score (termed 'Fickett score' hereafter), which is a simple linguistic feature that distinguishes protein-coding RNA and ncRNA according to the combinational effect of nucleotide composition and codon usage bias (22). Briefly, the Fickett score is obtained by computing four position values and four composition values (nucleotide content) from the DNA sequence. The position value reflects the degree to which each base is favored in one codon position versus another. For example, position value of A (A_{pos}) is calculated as follows:

$$A_1 = \text{Number of As in position } 0, 3, 6\ldots$$
$$A_2 = \text{Number of As in position } 1, 4, 7\ldots$$
$$A_3 = \text{Number of As in position } 2, 5, 8\ldots$$
$$A_{pos} = \frac{MAX(A_1,A_2,A_3)}{MIN(A_1,A_2,A_3)+1}$$

C_{pos}, G_{pos} and T_{pos} are determined in the same way. The percentage composition of each base is also determined. These eight values are then converted into probabilities (p) of coding using a lookup table provided in the original article. Each probability is multiplied by a weight (w) for the respective base, where the value of w reflects the percentage of time each parameter alone successfully predicts coding or noncoding function for the sequences of known function. Finally, the Fickett score is calculated as follows:

$$\text{Fickett Score} = \sum_{i=1}^{8} p_i w_i$$

The Fickett score is independent of the ORF, and when the test region is $\geq$200 nt in length (which includes most lncRNA), this feature alone can achieve 94% sensitivity and 97% specificity, with 'no opinion' on 18% of the sequences (22).

The fourth feature we used was hexamer usage bias (termed 'hexamer score' hereafter). This may be the most discriminating feature because of the dependence between adjacent amino acids in proteins (23). The hexamer score can be computed in numerous ways; here, we used a log-likelihood ratio to measure differential hexamer usage between coding and noncoding sequences. For a given DNA sequence, we calculated the probability of the sequence under the model of coding DNA and under the model of noncoding DNA, and then we took the logarithm of the ratio of these probabilities as the score of coding potential. We used F (h_i) (i = 0, 1, …, 4095) and F' (h_i) (i = 0, 1, …, 4095) to represent in-frame hexamer frequency, calculated from coding and noncoding training data sets (described below), respectively. For a given hexamer sequence $S = H_1, H_2, \ldots, H_m$,

$$\text{Hexamer Score} = \frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{F(H_i)}{F'(H_i)}\right)$$

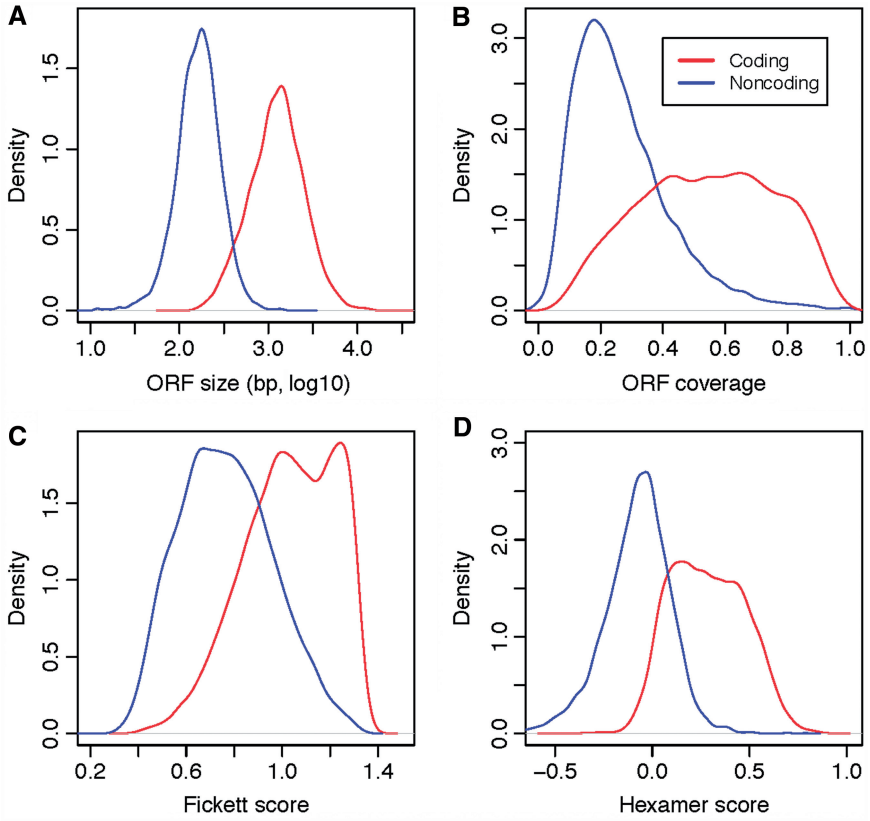

Figure 1. Score distribution between coding (red) and noncoding (blue) transcripts for the four linguistic features selected to build the logistic regression model; training data set containing 10 000 coding and 10 000 noncoding transcripts were used. (**A**) ORF size. (**B**) ORF coverage. (**C**) Fickett score (TESTCODE statistic). (**D**) Hexamer usage bias measured by log-likelihood ratio.

Hexamer score determines the relative degree of hexamer usage bias in a particular sequence. Positive values indicate a coding sequence, whereas negative values indicate a noncoding sequence.

We build a logistic regression model using these four linguistic features as predictor variables. A χ^2 test was used to evaluate whether our logit model with predictors fits the training data significantly better than the null model, which had only an intercept. We built a high-confidence training data set to measure the prediction performance of our logit model. This data set contained 10 000 protein-coding transcripts selected from the RefSeq database; all transcripts had high-quality protein sequences annotated by the Consensus Coding Sequence project. We also added 10 000 randomly selected noncoding transcripts from the GenCODE database. We evaluate the model with a 10-fold cross-validation and measured its sensitivity, specificity, accuracy, precision and area under the curve (AUC) characteristics. The receiver operating characteristic (ROC) curve and precision–recall (PR) curve were generated using ROCR package (24). We also built a nonparametric two-graph ROC curve for selecting the optimal CPAT score threshold that maximizes the sensitivity and specificity of CPAT while minimizing misclassifications.

We built an independent test data set to compare the performance of CPAT with that of CPC, PhyloCSF and PORTRAIT. This test set composed of 4000 high-quality protein-coding genes (RefSeq annotated) and 4000 lncRNAs from a human lncRNA catalog (5). None of these 8000 genes was included in the training data set for CPAT. Assuming that all 4000 lncRNAs are truly noncoding sequences, we could compute the sensitivity, specificity, accuracy and precision of the algorithms to measure their performance. PhyloCSF could not determine the coding status of 528 (13.2%) noncoding genes. Those 528 genes were equally assigned to the true-negative and false-positive categories. The abbreviations in the equations below are as follows: FN, false negative; FP, false positive; TN, true negative; TP, true positive

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

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

RESULTS

All four selected features were concordantly higher in coding transcripts and lower in noncoding transcripts (Figure 1). We plotted three major features (ORF size, Fickett score and hexamer score) in a three-dimensional space to evaluate their combinatorial effect (Figure 2).

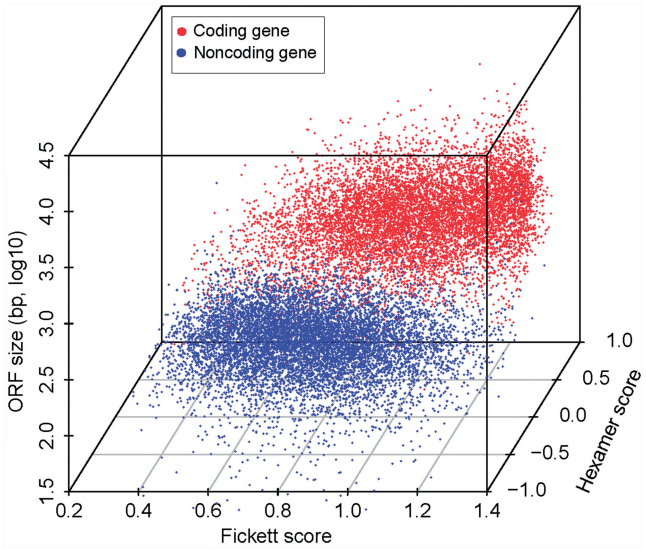

Figure 2. Three-dimensional plot shows combinatorial effects of Fickett score, hexamer score and ORF size on 10 000 coding genes (red dots) and 10 000 noncoding genes (blue dots).

Coding and noncoding transcripts in our training data set were grouped into two distinct clusters, indicating good concordance between features. The χ^2 test P value was $<.001$ ($\chi^2 = 23\,548.44$; degrees of freedom = 4), indicating that the logit model as a whole fits significantly better than the null model. Ten-fold cross-validation showed that CPAT could achieve very high accuracy, with an AUC of 0.9927 (Figure 3A). We also provide the PR curve because the ROC curve can be misleading when the test data are largely skewed (Figure 3B). We use nonparametric two-graph ROC curves to determine an optimal CPAT score threshold that maximizes the discriminatory power (Figure 3C and D). According to Figure 3D, a score threshold of 0.364 gave the highest sensitivity and specificity (0.966 for both) for human data.

We compared the performance of CPAT with that of CPC, PhyloCSF and PORTRAIT (protein-independent support vector machine model) using an independent test data set composed of 4000 coding genes and 4000 noncoding genes. A multiple alignment of 45 vertebrate genomes, including that of human, was downloaded from the UCSC (University of California, Santa Cruz) Genome Browser and was used as the input alignment for PhyloCSF. In general, CPAT (sensitivity: 0.96, specificity: 0.97) had greater classification power compared with all other programs (Figure 4; Supplementary Tables S1 and S2). Although CPC had the highest sensitivity (0.99), it suffered from poor specificity (0.74). One possible explanation is that a significant proportion of ncRNAs has a certain degree of sequence similarity to protein-coding genes. PhyloCSF had the least sensitivity (0.90) and the lowest specificity (0.63). Part of the reason for these outcomes is that nonconserved transcripts cannot be processed by PhyloCSF. If we consider those 528 nonconserved transcripts as noncoding, the specificity increased from 0.63 to 0.69, and the sensitivity remained unchanged. PORTRAIT had relatively balanced sensitivity (0.96) and specificity (0.87). CPAT achieved highest overall accuracy (0.97) when compared with CPC (0.87), PhyloCSF (0.76) and PORTRAIT (0.92). CPAT's excellent discriminatory power was further demonstrated by the greatest separation between the score distributions of coding and noncoding sequences (Figure 5). Unlike CPC, PhyloCSF and PORTRAIT, choosing a smaller CPAT score threshold to increase the sensitivity will not sacrifice too much specificity.

One could argue that PhyloCSF underperformed in this study because we used whole transcripts for testing rather than consecutive protein-coding exons and intergenic regions as used in its original article (17). To address this issue, we compiled another single-exon test data set consisting of 184 protein-coding and 278 noncoding transcripts. The test results with this data set indicated that CPAT (sensitivity: 0.962, specificity: 0.842) still outperformed PhyloCSF (sensitivity: 0.832, specificity: 0.588, Supplementary Figure S2). However, when tested on PhyloCSF's original data set in *Lin et al.* (25), PhyloCSF (sensitivity 0.91, specificity 0.99) has better performance than CPAT (sensitivity 0.50, specificity 0.98). This is reasonable because lncRNAs in our test data set are poorly conserved, whereas lncRNAs in Lin *et al.* test data set are highly conserved because they are taken from multiple-sequence alignments of three closely related *Drosophila* species. Hence, we argue that PhyloCSF works better if the transcripts are highly conserved, which are rare to find in lncRNAs (5,6). This also highlights the Achilles' heel of the alignment-based methods for detecting lncRNAs. In contrast, the dramatic decrease in CPAT's sensitivity is due to the lack of ORF information in Lin *et al.* test data set, which is largely composed of individual exons, and not exon-length complete transcripts. This, however, will not limit the application scope of CPAT because most full-length transcripts can be constructed at the current sequencing depth (8).

We measured the computational speed of CPAT, CPC and PhyloCSF on a sample of 200 sequences randomly selected from the test data set. CPAT took 0.67 s to process the data, and it was four orders of magnitudes faster than both CPC [11 945 s (3.3 h)] and PhyloCSF [11 737 s (3.3 h)]. Furthermore, computational time for the PhyloCSF did not include the time spent preparing multiple-alignment files for analysis. PORTRAIT was significantly faster than CPC and PhyloCSF, and therefore all 8000 test genes were used to evaluate its speed: CPAT took 23.83 s to process the test set, and it was 48 times faster than PORTRAIT [1146.30 s (19 min)].

DISCUSSION

A number of linguistic features characterizing coding RNA sequences have been developed over the past 30 years. These include maximum ORF size, dinucleotide usage, codon usage bias, hexamer usage bias, nucleotide composition bias between codon positions and imperfect periodicity in base occurrences (23,26). Among these features, we selected ORF features (size and coverage) because of their discriminatory power and ease of calculation (21). In-frame hexamer score was selected because it

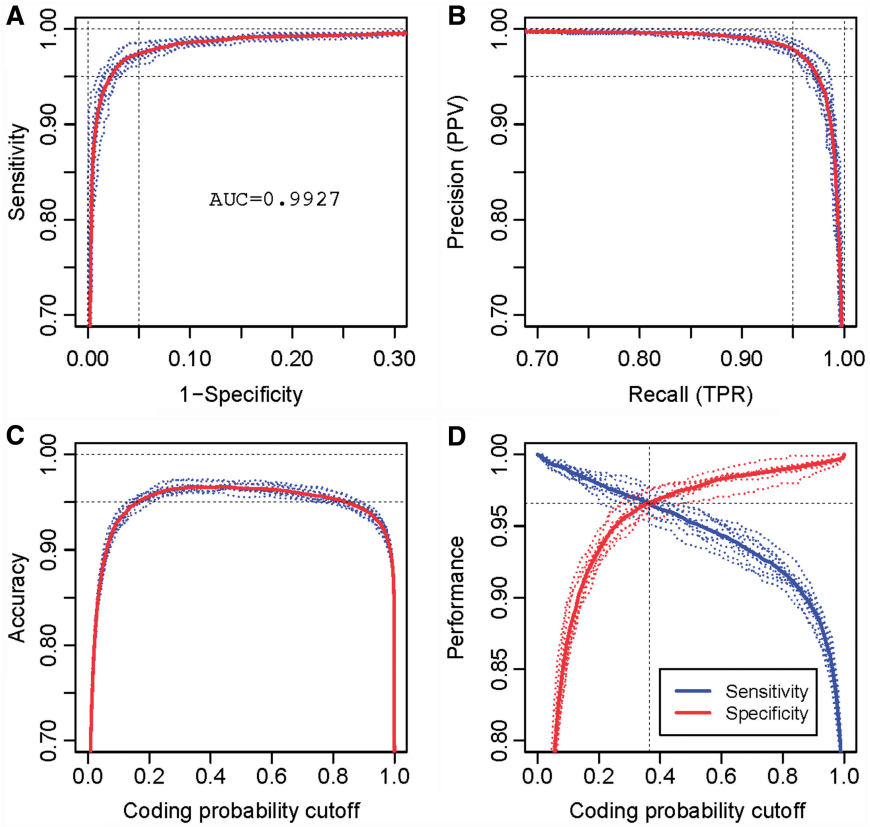

Figure 3. Performance evaluation using 10-fold cross-validation. Dashed curves represent the 10-fold cross-validation; solid curves represent the averaged curve from 10 validation runs. (**A**) ROC curve. (**B**) PR curve. PPV = positive predictive value, TPR = true positive rate. (**C**) Accuracy versus cutoff value. (**D**) Two-graph ROC curve is used to determine the optimum cutoff value.

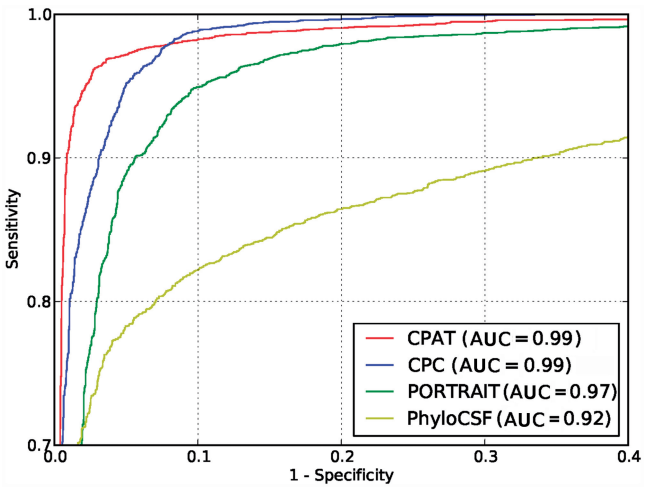

Figure 4. Performance comparison between CPAT, CPC, PhyloCSF and PORTRAIT using ROC curves.

has the highest prediction accuracy (average of sensitivity and specificity) as evaluated by Fickett and Tung in 1992 (23). Fickett score was selected because it simultaneously captures the compositional bias and position asymmetry, which are orthogonal to the ORF features. Supplementary Figure S3 shows the performance of these individual features as well as the combined feature set. The combined feature set has very high sensitivity and specificity (>0.966), leaving very little room for further improvement.

Annotation of genomes has always been a challenging task for biologists, and these efforts have been accelerated by deep transcriptome sequencing. Distinguishing between protein-coding and noncoding sequences is the first and arguably the most crucial step in genome annotation. Most novel transcripts are less conserved and species-specific ncRNAs. Detecting the coding-potential of these transcripts via alignment-based software is intractable. We developed CPAT, a highly accurate alignment-free method, which uses a logistic regression model to discriminate between coding and noncoding transcripts using pure linguistic features. Compared with other tools, CPAT is more robust, markedly faster and more convenient to use. Taken together, CPAT is able to accurately assess the coding potential of tens of thousands of transcripts in real-time, and will be a valuable tool for the rapidly growing RNA-seq community.

AVAILABILITY AND IMPLEMENTATION

Source code was implemented in C and Python and is freely available at: http://code.google.com/p/cpat/. The web server was implemented in PHP, MYSQL and

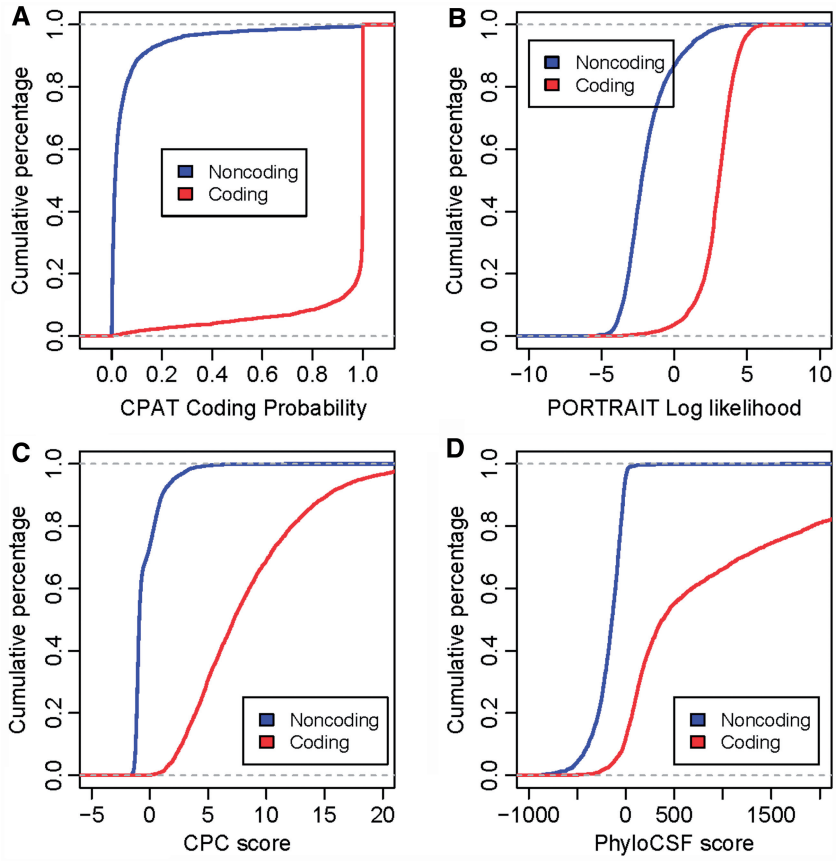

Figure 5. Cumulative curves of coding-potential assessment score for (**A**) CPAT, (**B**) PORTRAIT, (**C**) CPC and (**D**) PhyloCSF.

Apache, with support for all major browsers: http://lilab.research.bcm.edu/cpat/index.php.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figures 1–3.

ACKNOWLEDGEMENTS

The authors thank Chen Wang (Mayo Clinic) and two anonymous reviewers for their valuable suggestions. We also thank Mayo's section of scientific publication for their copy-editing services.

FUNDING

Department of Defense Prostate Cancer Program [PC094421 to W.L.]; the Cancer Prevention and Research Institute of Texas [RP110471-C3 to W.L.]; the Center for Individualized Medicine (CIM) at Mayo Clinic (to J.P.K.). Funding for open access charge: Cancer Prevention and Research Institute of Texas [RP110471-C3 to W.L.].

Conflict of interest statement. None declared.

REFERENCES

1. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science (New York, NY)*, **306**, 2242–2246.
2. Kapranov,P., St Laurent,G., Raz,T., Ozsolak,F., Reynolds,C.P., Sorensen,P.H.B., Reaman,G., Milos,P., Arceci,R.J., Thompson,J.F. *et al.* (2010) The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.*, **8**, 149.
3. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermüller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, NY)*, **316**, 1484–1488.
4. Mercer,T.R., Gerhardt,D.J., Dinger,M.E., Crawford,J., Trapnell,C., Jeddeloh,J.A., Mattick,J.S. and Rinn,J.L. (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.*, **30**, 99–104.
5. 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.
6. 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.
7. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

8. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
9. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.-C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
10. Chu,C., Qu,K., Zhong,F.L., Artandi,S.E. and Chang,H.Y. (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell*, **44**, 667–678.
11. Ørom,U.A., Derrien,T., Beringer,M., Gumireddy,K., Gardini,A., Bussotti,G., Lai,F., Zytnicki,M., Notredame,C., Huang,Q. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46–58.
12. Huarte,M., Guttman,M., Feldser,D., Garber,M., Koziol,M.J., Kenzelmann-Broz,D., Khalil,A.M., Zuk,O., Amit,I., Rabani,M. *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, **142**, 409–419.
13. Pauli,A., Rinn,J.L. and Schier,A.F. (2011) Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.*, **12**, 136–149.
14. Hung,T. and Chang,H.Y. (2010) Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol.*, **7**, 582–585.
15. Arrial,R.T., Togawa,R.C. and Brigido,M.deM. (2009) Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics*, **10**, 239.
16. 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.
17. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics (Oxford, England)*, **27**, i275–i282.
18. Washietl,S., Findeiss,S., Müller,S.A., Kalkhof,S., Bergen,von,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.
19. The FANTOM Consortium. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
20. Schloss,P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.*, **6**, e1000844.
21. Frith,M.C., Bailey,T.L., Kasukawa,T., Mignone,F., Kummerfeld,S.K., Madera,M., Sunkara,S., Furuno,M., Bult,C.J., Quackenbush,J. *et al.* (2006) Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol.*, **3**, 40–48.
22. Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
23. Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
24. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, **21**, 3940–3941.
25. Lin,M.F., Deoras,A.N., Rasmussen,M.D. and Kellis,M. (2008) Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput. Biol.*, **4**, e1000067.
26. Dinger,M.E., Pang,K.C., Mercer,T.R. and Mattick,J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.