Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W345–W349***
doi:10.1093/nar/gkm391

CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine

Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei* and Ge Gao*

Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P. R. China

Received January 30, 2007; Revised April 13, 2007; Accepted May 1, 2007

ABSTRACT

Recent transcriptome studies have revealed that a large number of transcripts in mammals and other organisms do not encode proteins but function as noncoding RNAs (ncRNAs) instead. As millions of transcripts are generated by large-scale cDNA and EST sequencing projects every year, there is a need for automatic methods to distinguish protein-coding RNAs from noncoding RNAs accurately and quickly. We developed a support vector machine-based classifier, named Coding Potential Calculator (CPC), to assess the protein-coding potential of a transcript based on six biologically meaningful sequence features. Tenfold cross-validation on the training dataset and further testing on several large datasets showed that CPC can discriminate coding from noncoding transcripts with high accuracy. Furthermore, CPC also runs an order-of-magnitude faster than a previous state-of-the-art tool and has higher accuracy. We developed a user-friendly web-based interface of CPC at http://cpc.cbi.pku.edu.cn. In addition to predicting the coding potential of the input transcripts, the CPC web server also graphically displays detailed sequence features and additional annotations of the transcript that may facilitate users' further investigation.

INTRODUCTION

Recent transcriptome studies have revealed that a large number of transcripts in mammals and other organisms do not encode proteins but function as noncoding RNAs (ncRNAs) instead. *In vivo* experiments have demonstrated important biological roles of noncoding RNAs, including regulation of transcription and translation, RNA modification and epigenetic modification of chromatin structure (1–3). There is immense interest within the biological community to identify and study new noncoding RNAs.

As millions of transcripts are generated by large-scale cDNA and EST sequencing projects every year, there is a need for automatic methods to accurately and quickly distinguish protein-coding RNAs from noncoding RNAs. Since to date no web server and few standalone tools have been designed for this purpose, researchers sometimes used tools developed for other purposes such as cDNA annotation and functionally domain identification (4–12). However these methods showed varied performance on different datasets (12,13). Recently a new algorithm and standalone software named CONC was published that classifies transcripts as 'coding' or 'noncoding' using machine learning methods (13). CONC showed improved performance over previous tools such as ESTScan (6). However, CONC is slow for large datasets and does not have a web-server interface, limiting its usefulness. It works well with high-quality transcripts but may suffer from errors such as frameshifts which are common in ESTs and even occur occasionally in full-length cDNAs (11). Furthermore, CONC only outputs the 'coding'/'noncoding' classification but does not provide an explanation or related information. New tools are desired that are more accurate, run faster, and have a more user-friendly web-based interface.

METHODS

To assess a transcript's coding potential, we extract six features from the transcript's nucleotide sequence. A true protein-coding transcript is more likely to have a long and high-quality Open Reading Frame (ORF) compared with a non-coding transcript. Thus, our first three features assess the extent and quality of the ORF in a transcript. We use the *framefinder* software (14) to identify the

*To whom correspondence should be addressed. Tel: +86-10-6275-5206; Fax: +86-10-62759001; Email: weilp@mail.cbi.pku.edu.cn
Correspondence may also be addressed to Ge Gao. Tel: +86-10-6275-1861; Fax: +86-10-6275-1861; Email: gaog@mail.cbi.pku.edu.cn

longest reading frame in the three forward frames. Known for its error tolerance, *framefinder* can identify most correct ORFs even when the input transcripts contain sequencing errors such as point mutations, indels and truncations (14,15). We extract the LOG-ODDS SCORE and the COVERAGE OF THE PREDICTED ORF as the first two features by parsing the *framefinder* raw output with Perl scripts (available for download from the web site). The LOG-ODDS SCORE is an indicator of the quality of a predicted ORF and the higher the score, the higher the quality. A large COVERAGE OF THE PREDICTED ORF is also an indicator of good ORF quality (14). We add a third binary feature, the INTEGRITY OF THE PREDICTED ORF, that indicates whether an ORF begins with a start codon and ends with an in-frame stop codon.

The large and rapidly growing protein sequence databases provide a wealth of information for the identification of protein-coding transcript. We derive another three features from parsing the output of BLASTX (16) search (using the transcript as query, *E*-value cutoff 1*e*-10) against UniProt Reference Clusters (UniRef90) which was developed as a nonredundant protein database with a 90% sequence identity threshold (17). First, a true protein-coding transcript is likely to have more hits with known proteins than a non-coding transcript does. Thus we extract the NUMBER OF HITS as a feature. Second, for a true protein-coding transcript the hits are also likely to have higher quality; i.e. the HSPs (High-scoring Segment Pairs) overall tend to have lower *E*-value. Thus we define feature HIT SCORE as follows:

$$S_i = \underset{j}{\text{mean}}\left\{-\log_{10} E_{ij}\right\}, \quad i \in [0,1,2]$$

$$\text{HIT SCORE} = \underset{i\in\{0,1,2\}}{mean}\{S_i\} = \frac{\sum_{i=0}^{2} Si}{3},$$

where E_{ij} is the *E*-value of the *j*-th HSP in frame *i*, S_i measures the average quality of the HSPs in frame *i* and HIT SCORE is the average of S_i across three frames. The higher the HIT SCORE, the better the overall quality of the hits and the more likely the transcript is protein-coding. Thirdly, for a true protein-coding transcript most of the hits are likely to reside within one frame, whereas for a true non-coding transcript, even if it matches certain known protein sequence segments by chance, these chance hits are likely to scatter in any of the three frames. Thus, we define feature FRAME SCORE to measure the distribution of the HSPs among three reading frames:

$$\text{FRAME SCORE} = \underset{i\in\{0,1,2\}}{variance}\{S_i\} = \frac{\sum_{i=0}^{2}\left(S_i - \bar{S}\right)^2}{2}$$

The higher the FRAME SCORE, the more concentrated the hits are and the more likely the transcript is protein-coding.

We incorporate these six features into a support vector machine (SVM) machine learning classifier (18). Mapping the input features onto a high-dimensional feature space via a proper kernel function, SVM constructs a classification hyper-plane (maximum margin hyper-plane) to separate the transformed data (18). Known for its high accuracy and good performance, SVM is a widely used classification tool in bioinformatics analysis such as microarray-based cancer classification (19,20), prediction of protein function (21,22) and prediction of subcellular localization (23,24). We employed the LIBSVM package (25) to train a SVM model using the standard radial basis function kernel (RBF kernel). The C and gamma parameters were determined by grid-search in the training dataset. We trained the SVM model using the same training data set as CONC used (13), containing 5610 protein-coding cDNAs and 2670 noncoding RNAs.

EVALUATION

We evaluated our method, named Coding Potential Calculator (CPC), by 10-fold cross-validation on the training data sets. The accuracy was 95.77%. For further evaluation we tested CPC on three large datasets including two non-coding RNA datasets from the Rfam 7.0 database (26) and RNAdb databank (27), respectively, and a protein-coding RNA dataset from the EMBL nucleotide databank based on cross-links to the UniProt/SwissProt protein knowledgebase (17,28). We recorded the accuracy and computation time of CPC in Table 1, and compared it with CONC (version 1.01 downloaded from the authors' website http://cubic.bioc.columbia.edu/~liu/conc/ and installed locally). Both CPC and CONC were run in a Linux box with Intel Xeon 3.0G CPU and 4G RAM. Overall, CPC showed better accuracy on all three datasets with an order-of-magnitude faster speed (Table 1). For more stringent evaluations we removed

Table 1. Evaluation of accuracy and CPU time of CPC and CONC on three datasets

Dataset	Dataset type	Dataset size[a]	Accuracy		Time (min)	
			CPC	CONC	CPC	CONC
Rfam	Noncoding	30 770	98.62%	97.12%	3513	46 376
RNADB	Noncoding	3996	91.50%	85.44%	598	7322
Embl cds	Coding	121 914	99.08%	98.70%	69 116	826 210[b]

[a]CONC focuses on sequences with at least 80 nucleotides and assumes shorter sequences unlikely to have coding potential. CPC does not make this assumption and has similar performance on shorter sequences, but to make a direct comparison here we shows results only on sequences with at least 80 nucleotides.

[b]Because the required CPU time is long, the dataset was split and run on 24 nodes in parallel. The reported CPU time was the sum of execution time on individual nodes.

those sequences in the three test datasets that were similar to one or more sequences in the training set (BLASTN E-value cutoff 1*e*-2) and tested CPC on the remaining sequences. We also tested CPC on new entries in the latest UniRef90 release (version 10.1) which were not included in the previous release used to train CPC (version 9.4). In both cases the accuracy of CPC remained high (see section 'More Stringent Evaluation' and Table S1 in Supplementary Data).

We then compared CPC with other prediction algorithms following the same evaluation strategy proposed by Frith *et al.* (12). The results showed that CPC had the highest consistency with expert curation and performed well for the six challenging cases hand-picked by Frith *et al.* (12) (see section 'Comparison with other protein-prediction algorithms following Frith *et al.*' and Table S2 in Supplementary Data). CPC was also able to accurately predict 92% of the 2,849 short peptides with less than 100 amino acids (see section 'Performance on Short Peptides' in Supplementary Data).

WEB SERVER

We developed a user-friendly web interface for CPC (http://cpc.cbi.pku.edu.cn). The CPC web server accepts a set of nucleotide FASTA sequences as input (allowing symbols 'A', 'T', 'G', 'C', and 'U'). The sequences can be pasted directly into the input box or uploaded from a local sequence file. By default, the CPC server runs in 'interactive mode' in that results will be shown in the browser once the computation is finished. For a large set of sequences the user can input an email address to run his/her job in 'batch mode'. The server will send a notice to the user's mailbox upon completion of the job. A unique 'Task ID' (TID) is assigned to each job by the web server. Users can use TID to track the job progress and retrieve the results which are saved on the server for 1 week.

CPC summarizes the main output in a table (Figure 1a). Each row corresponds to one input sequence. The columns show the sequence ID, the coding/noncoding classification, the SVM score (the 'distance' to the SVM classification hyper-plane in the features space), and a 'Details' link (as described later). In general, the farther away the score is from zero, the more reliable the prediction is. As a rule of thumb from our experience, the transcripts with score between −1 and 1 are marked as 'weak noncoding' or 'weak coding'. Results in the summary table can be sorted interactively by sequence id, coding/noncoding classification, and SVM score; they can also be filtered by coding/noncoding classification, and SVM score.

The current version of CPC cannot accurately discriminate transcripts falling entirely within UTR regions from true non-coding transcripts, because neither of them produces amino acid sequences. To handle this limitation, CPC provides the users the option to search database of known UTR sequences, UTRdb (32), using BLAST (see section 'Recognizing Potential UTR regions' and Figure S1 in Supplementary Data).

To 'explain' why a transcript is classified as coding or noncoding, CPC server provides detailed supporting evidence and other related sequence features of the input transcript in an Evidence page (Figure 1b). The Evidence page shows the six features of the transcript, color coded for better visualization. It shows graphically the putative ORF identified by *framefinder* and the BLASTX hits. Mousing over, users can view details of each ORF and BLASTX hits. The Evidence page also provides options for querying the input transcript against well-annotated database, such as the functional domain database Pfam (29), SMART (30) and SuperFamily (31), UTRdb (32) and ncRNA database RNAdb (27). The Evidence page aims to facilitate the user's detailed investigation of the transcript.

We developed the CPC web server on a Java platform using JSP to render the dynamic HTML pages and Apache/Tomcat as the J2EE container. The web site is in compliance with W3C XHTML 1.0 Strict specification and works in both the Microsoft Internet Explorer and Mozilla Firefox browsers. A standalone version of the software is freely available for download on the web site, distributed under GNU GPL. A parallel version with simple distributed computing support is available upon request.

DISCUSSION

With the rapidly increasing amount of data generated by large-scale transcriptome sequencing and intensifying attention on the study of noncoding RNAs, methods that can discriminate noncoding RNAs from protein-coding ones with high reliability and fast speed are important. Integrating multiple sequence features with biological significance, CPC is shown to have good accuracy in both cross-validation and several test datasets. It also runs an order-of-magnitude faster than the previous state-of-the-art tool, and thus is more suitable for high-throughput analysis. CPC uses far fewer features than CONC does (6 versus 180) but achieved comparable, even better, performance in the evaluation. The results demonstrated that the sequence features used by CPC have powerful discriminating power and may reflect the intrinsic properties of coding transcript. Using fewer, sequence-based features also significantly reduced computing cost, thus removing a hurdle for a web server to be developed. Additional information such as potential functional domains and similarity to known UTR regions or ncRNA is useful to users. This and other supplementary information is available in the Evidence pages of CPC, making the results of CPC more easily interpretable and biology-meaningful.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENT

This work was supported by the China National High-tech (863) Program (2006AA02Z334, 2006AA02Z314),

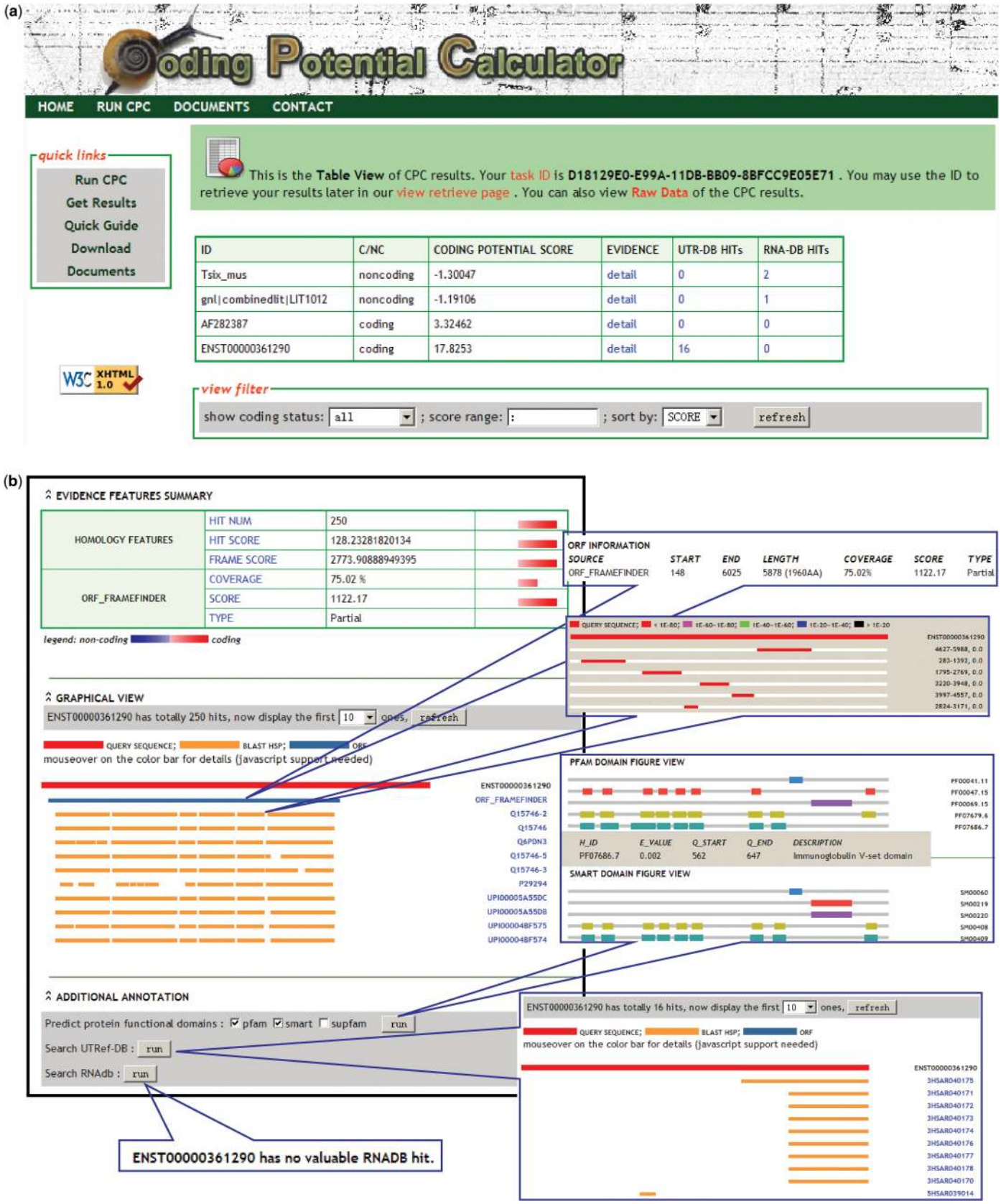

Figure 1. Screenshots of CPC output. (**a**) Results are summarized in a 'Table View'. (**b**) Sequence features and additional annotations of an input transcript are shown in an Evidence page. Users can mouse over or click to see more details.

973 Program (2006CB0D0804, 2003CB715906) and China Postdoctoral Science Foundation. Funding to pay the Open Access publication charges for this article was provided by China Ministry of Education 111 Project (B06001).

Conflict of interest statement. None declared.

REFERENCE

1. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet*., **2**, 919–929.
2. Mattick,J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet*., **5**, 316–323.
3. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*., **15**, R17–R29.
4. Furuno,M., Kasukawa,T., Saito,R., Adachi,J., Suzuki,H., Baldarelli,R., Hayashizaki,Y. and Okazaki,Y. (2003) CDS annotation in full-length cDNA sequence. *Genome Res*., **13**, 1478–1487.
5. Hatzigeorgiou,A.G., Fiziev,P. and Reczko,M. (2001) DIANA-EST: a statistical analysis. *Bioinformatics*, **17**, 913–919.
6. Lottaz,C., Iseli,C., Jongeneel,C.V. and Bucher,P. (2003) Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, **19**(Suppl. 2), II103–II112.
7. Shafer,P., Lin,D.M. and Yona,G. (2006) EST2Prot: mapping EST sequences to proteins. *BMC Genomics*, **7**, 41.
8. Carninci,P., Waki,K., Shiraki,T., Konno,H., Shibata,K., Itoh,M., Aizawa,K., Arakawa,T., Ishii,Y. *et al*. (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res*., **13**, 1273–1289.
9. Okazaki,Y. and Hume,D.A. (2003) A Guide to the Mammalian Genome. *Genome Res*., **13**, 1267–1272.
10. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
11. Maeda,N., Kasukawa,T., Oyama,R., Gough,J., Frith,M., Engstrom,P.G., Lenhard,B., Aturaliya,R.N., Batalov,S. *et al*. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet*., **2**, e62.
12. Frith,M.C., Bailey,T.L., Kasukawa,T., Mignone,F. and Kummerfeld,S.K. (2006) Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol*., **3**, 40–48.
13. Liu,J., Gough,J. and Rost,B. (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet*., **2**, e29.
14. Slater,G.S.C. (2000) *Algorithms for the Analysis of Expressed Sequence Tags,* University of Cambridge, Cambridge.
15. Nagaraj,S.H., Gasser,R.B. and Ranganathan,S. (2006) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform*, **8**, 6–21.
16. 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.
17. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.
18. Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman, 340 Pine St, 6th floor San Fransisco, CA 94104, USA.
19. Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
20. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
21. Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.
22. Borgwardt,K.M., Ong,C.S., Schonauer,S., Vishwanathan,S.V., Smola,A.J. and Kriegel,H.P. (2005) Protein function prediction via graph kernels. *Bioinformatics*, **21**(Suppl. 1), i47–i56.
23. Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.
24. Lei,Z. and Dai,Y. (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **6**, 291.
25. Chang,C.C. and Lin,C.J. (2001), *Software available at* http://www.csie.ntu.edu.tw/cjlin/libsvm, Vol. 80, pp. 604–611.
26. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121–D124.
27. Pang,K.C., Stephen,S., Engstrom,P.G., Tajul-Arifin,K., Chen,W., Wahlestedt,C., Lenhard,B., Hayashizaki,Y. and Mattick,J.S. (2005) RNAdb–a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res*., **33**, D125–D130.
28. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.
29. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
30. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*., **34**, D257–D260.
31. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*., **32**, D235–D239.
32. Mignone,F., Grillo,G., Licciulli,F., Iacono,M., Liuni,S., Kersey,P.J., Duarte,J., Saccone,C. and Pesole,G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*., **33**, D141–D146.