BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 12 2009, pages 1492–1497
doi:10.1093/bioinformatics/btp257

Sequence analysis

Protein function annotation from sequence: prediction of residues interacting with RNA

R. V. Spriggs[1,†], Y. Murakami[2,†], H. Nakamura[2] and S. Jones[1,*]

[1]Department of Chemistry and Biochemistry, School of Life Sciences, John Maynard-Smith Building, University of Sussex, Falmer, Brighton, BN1 9QG, UK and [2]Research Centre for Structural and Functional Proteomics, Institute for Protein Research, Osaka University, Japan

Received on January 28, 2009; revised on March 24, 2009; accepted on April 8, 2009

Advance Access publication April 23, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: All eukaryotic proteomes are characterized by a significant percentage of proteins of unknown function. Computational function prediction methods are therefore essential as initial steps in the function annotation process. This article describes an annotation method (PiRaNhA) for the prediction of RNA-binding residues (RBRs) from protein sequence information. A series of sequence properties (position specific scoring matrices, interface propensities, predicted accessibility and hydrophobicity) are used to train a support vector machine. This method is then evaluated for its potential to be applied to RNA-binding function prediction at the level of the complete protein.

Results: The 5-fold cross-validation of PiRaNhA on a dataset of 81 RNA-binding proteins achieves a Matthews Correlation Coefficient (MCC) of 0.50 and accuracy of 87.2%. When used to predict RBRs in 42 proteins not used in training, PiRaNhA achieves an MCC of 0.41 and accuracy of 84.5%. Decision values from the PiRaNhA predictions were used in a second SVM to make predictions of RNA-binding function at the protein level, achieving an MCC of 0.53 and accuracy of 76.1%. The PiRaNhA RBR predictions allow experimentalists to perform more targeted experiments for function annotation; and the prediction of RNA-binding function at the protein level shows promise for proteome-wide annotations.

Availability and Implementation: Freely available on the web at www.bioinformatics.sussex.ac.uk/PIRANHA or http://piranha.protein.osaka-u.ac.jp.

Contact: s.jones@sussex.ac.uk.

Supplementary Information: Supplementary data are available at the *Bioinformatics* online.

1 INTRODUCTION

All eukaryotic proteomes are characterized by a significant percentage of proteins of unknown function (PUFs). It has been estimated that PUFs represent 18–40% of a eukaryotic proteome (Friedberg *et al*., 2006; Gollery *et al*., 2007; Horan *et al*., 2008). In order for these proteins to contribute to a wider knowledge of biological systems, function annotations are essential. As more divergent genomes are sequenced the number of PUFs increases, and therefore high-throughput computational methods that predict function from sequence are essential. Such methods can be used as initial screening steps in the function annotation process prior to targeted experimental protocols such as mutagenesis. Function annotation methods need to go beyond the search for, and transfer of annotations from, sequence homologues. One path for function annotation is the analysis of ligands bound to proteins. One functionally significant ligand, bound by 15% of proteins in UniProtKB/Swiss-Prot (UniProt_Consortium, 2008), is RNA. RNA has diverse and essential functions within the cell, including involvement in translation, transcription and catalysis. A common theme of these functions is the interaction with proteins.

Computational function annotation of protein sequences in the field of RNA binding has addressed two problems: (i) predicting RNA-binding residues (RBRs) in protein sequences known to bind RNA and (ii) predicting RNA-binding function for complete PUFs. A number of machine learning techniques have been applied to the first problem with varied levels of success. BindN (Wang and Brown, 2006), RNABindR (Terribilini *et al*., 2007), PPRInt (Kumar *et al*., 2008) and a method by Jeong *et al*. (2004; Jeong and Miyano, 2006) all predict RBRs in RNA-binding protein sequences. BindN (Wang and Brown, 2006) uses support vector machines (SVMs) trained on sidechain pKa, hydrophobicity (H) and molecular mass. RNABindR uses a Naïve Bayes classifier trained on interface amino acid propensities and sequence periodicities of RBR (Terribilini *et al*., 2006, 2007). PPRInt uses an SVM trained using position specific scoring matrices (PSSMs) that capture information on the conservation of RBR (Kumar *et al*., 2008). The method by Jeong *et al*. (2004; Jeong and Miyano, 2006) uses a neural network with a weighted PSSM.

The problem with methods that attempt to solve the first problem: (i) predicting RBR in protein sequences known to bind RNA, is that when applied to non-RNA-binding proteins they still predict RBRs to be present. Hence, such methods cannot be used directly to predict RNA-binding function at the level of the complete protein. Consequently, a number of separate methods for making predictions of RNA binding at the protein level have been developed using SVMs. Fujishima *et al*. (2007) make predictions of RNA-binding function for proteins from *Pyrococcus furiosus*, using amino acid composition and periodicity. Han *et al*. (2004) use amino acid composition combined with van der Waals volume, polarity, charge,

*To whom correspondence should be addressed.

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

surface tension, secondary structure and solvent accessibility as features in an SVM. Finally, in two related papers, Cai and Lin (2003) and Yu *et al*. (2006) use amino acid composition and amino acid properties in the context of sequence length to make predictions at the protein level. However, a recurring problem in the field of machine learning is the ability to compare methods, and this is the case with RNA-binding function prediction at the protein level. No benchmark dataset has been used, and a variety of measures have been applied to assess performance. None of the methods outlined above use either of the machine learning standards, area under a receiver operator characteristic (ROC) curve (AUC) or the Matthews Correlation Coefficient (MCC), to measure performance. The methods achieve accuracies of between 77% and 96%, but the datasets are not available for validation and only one method is available as a server.

In the current article, a method is developed to predict RBRs with a high level of accuracy. The method, named PiRaNhA (capitals denote P-RNA; pronounced 'piranha', as in the fish), uses an SVM with a PSSM and three amino acid properties: interface propensity (IP), predicted solvent accessibility (pA) and hydrophobicity (H), to predict RBR from protein sequence. The method is tested using 5-fold cross-validation on a dataset of 81 RNA-binding proteins, and used to make predictions on an additional 42 RNA-binding proteins not used in training. The method shows a better performance level than all previous methods based on the AUC and MCC values. This article then outlines an evaluation of the use of the PiRaNhA RBR predictions in a second SVM to predict RNA-binding function at the protein level. Hence, this two-stage process has the potential to address both problems in this field: (i) predicting RBR in protein sequences known to bind RNA and (ii) predicting RNA-binding function at the protein level .

PiRaNhA gives experimentalists predictions that will enable them to make fewer, and more targeted, mutations to verify RNA-binding function. The RNA-binding function prediction method (at the protein level) has the potential to be used to scan whole proteomes to identify novel RNA-binding proteins. The significance of an accurate tool for predicting RBR and for predicting RNA-binding function at the protein level is reflected in the central role played by RNA-binding proteins in cellular processes, and, as a consequence, in human diseases such as cancer (Lukong *et al*., 2008).

2 SYSTEM AND METHODS

This article describes the use of a machine learning method to predict RBR within protein sequences, and evaluates the potential to use these predicted RBR to predict RNA-binding function at the protein level. RNA-binding proteins can be described in terms of the common domains they feature, such as the KH and the RRM domains (Chen and Varani, 2005). However, the binding sites themselves are comprised of large numbers of short residue segments, which preclude them from being described (and hence predicted) by generic sequence motifs (unpublished data). SVMs were, therefore, chosen as an effective method of integrating a large number of features across the whole sequence length. SVM have been widely used for prediction in this field (Jeong and Miyano, 2006; Jeong *et al*., 2004; Kumar *et al*., 2008; Terribilini *et al*., 2007; Wang and Brown, 2006; Wang *et al*., 2008), with methods using either PSSMs (Kumar *et al*., 2008) or physicochemical properties (Terribilini *et al*., 2007; Wang and Brown, 2006) as sequence features. PiRaNhA is the first method to combine both of these feature types. Amino acid propensities proved central to our previous work on protein–protein interaction predictions (Murakami and Jones, 2006), and here, too, they prove a key feature in the SVM.

2.1 Datasets of protein–RNA complexes

Structures of protein–RNA complexes were extracted from the Protein Data Bank (PDB; Berman *et al*., 2000) if they had $\geq$5 protein–RNA contacts, and, if solved by X-ray crystallography, had a resolution $\leq$3.0 Å. No restriction was made on which organism the proteins were derived from. RBRs were defined by calculating all the potential hydrogen bonds and van der Waals interactions occurring between the protein residues and the RNA, using HBPLUS (McDonald and Thornton, 1994). This essentially defined an RBR as any residue $\leq$3.9 Å distance from the RNA (Ellis *et al*., 2007). The protein chains were mapped to UniProtKB/Swiss-Prot (UniProt_Consortium, 2008) accessions using PDBSWS (Martin, 2005), and the type of RNA bound to the protein deduced from the keywords in the UniProtKB/Swiss-Prot record and the details in the PDB header. The resulting dataset was used to create a training set and a testing set:

- RNAset81: a non-redundant set of RNA-binding proteins derived from the 86-protein set used by Kumar *et al*. (2008) as a benchmark dataset [as defined by Jeong and colleagues (Jeong and Miyano, 2006; Jeong *et al*., 2004)]. Upon inspection, the set of 86 proteins provided by Kumar *et al*. was found to contain a number of homologues [using the rules for homology set out by Jeong *et al*.: 70% sequence identity over a 90% overlap on both sequences, and BLASTClust (Altschul *et al*., 1990)]. As a consequence, four sequences were removed. One further protein which had no RBR was also removed. The resultant dataset, denoted RNAset81, was used in training and cross-validation (Supplementary Table S1). The dataset includes, amongst other RNA types, 41 rRNA-binding proteins, 15 tRNA-binding proteins and 7 mRNA-binding proteins.
- RNAtestset42: a non-redundant (defined as proteins that have $\leq$35% sequence identity $\geq$90% of both sequences; using BLASTClust) set of RNA-binding proteins used purely for testing. Members of this set of protein–RNA complexes were deposited in the PDB after 2004 (when Jeong *et al*. defined the benchmark dataset of 86 protein–RNA complexes used by Kumar *et al*.). Any proteins homologous to any protein in RNAset81 were removed. As a second level of homology testing, Pfam (Finn *et al*., 2006) annotations were assigned to proteins in both the post-2004 protein–RNA dataset and RNAset81. Any proteins containing the same PfamA annotations as proteins in RNAset81 were removed from the post-2004 dataset. Multiple proteins with the same PfamA domain annotations were also removed from the post-2004 dataset. This left a set of 42 'new' protein–RNA complexes, denoted RNAtestset42 (Supplementary Table S2). The dataset includes, amongst other RNA types, 24 rRNA-binding proteins, 9 tRNA-binding proteins and 2 mRNA-binding proteins.

2.2 Sequence feature vectors

Four properties of the residues in the protein sequence are used as features in training the SVM.

(a) PSSM: this was created using PSI-BLAST (Altschul *et al*., 1997) with *E*-value 0.001 for three iterations, and the NCBI nr sequence database. The PSSM describes the evolutionary conservation of the residue positions. Feature vectors were constructed from the PSSM by concatenating the rows for each residue.

(b) Interface Propensity (IP): this value describes how likely it is to find a residue of a specific type in an RNA binding site. The propensities were taken from our previous structural analysis of RNA binding sites (Ellis *et al*., 2007).

(c) Predicted accessibility (pA): this was used to indicate the solvent exposure of a residue. SABLE (version 2.0; Wagner *et al*., 2005) was

used to predict the relative solvent accessibility (RSA) of each residue. The pAs are used, rather than measured values from the known structures, because when the method is used with protein sequences of unknown function only pAs will be available. In a comparison between actual and predicted RSA in independent test sets, SABLE achieved overall correlation coefficients of '~0.66' (Adamczak *et al*., 2005).

(d) Hydrophobicity (H) scores for each residue type were taken from the Kyte and Doolittle (1982) hydropathy scale.

The sequence properties were integrated into a feature vector covering a sub-sequence, with the residue being described in the centre. Sub-sequences, or windows, of various lengths have been tested, covering 5–25 residues. By using windows, the properties of all the residues in the sub-sequence are used to describe the residue in the centre, to enable the sequence environment of the residue to be taken into account; only the interaction status of the centre residue is attached to each feature vector.

3 ALGORITHM

3.1 SVMs

SVMs were trained to distinguish between RNA-binding and non-RBRs in protein sequences. LIBSVM (version 2.8; www.csie.ntu.edu.tw/~cjlin/libsvm; Chang and Lin, 2001) was used to carry out the SVM work, using the Radial Basis Function (RBF) kernel. The RBF kernel is defined as; $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$, where x_i and x_j are two feature vectors, and γ is a training parameter which determines the RBF width. Parameters C (the relationship between margin maximization and training error minimisation) and γ (effective range of distances between points) were optimized.

3.2 Performance measures

Four measures were calculated to assess SVM performance, using counts of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

(a) Sensitivity (SN) measures the proportion of the known RBRs that are correctly predicted and is defined as TP/(TP+FN).

(b) Specificity (SP) measures the proportion of the known non-RBRs that are correctly predicted and is defined as TN/(TN + FP).

(c) Accuracy (ACC) is a measure of what proportion of all the predictions made (RNA binding and non-RNA binding) are correct and is defined as (TP + TN)/(TP + FN + TN + FP).

(d) MCC indicates the size of the correlation between the actual status (RNA binding or not) of the residues and the predicted status of the residues. MCC values range between 1, where all predictions are correct, and –1, where none are correct. MCC is defined as $(TP \times TN) - (FP \times FN) / \sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}$.

(e) $F_{0.5}$-measure (Hripcsak and Rothschild, 2005) combines precision and recall into their harmonic mean, and is defined as $F_\beta = [(1 + \beta^2) \times R \times P]/[(\beta^2 \times P) + R]$, where R = recall, P = precision and β is a weight parameter. In the current work precision is favoured over recall and hence $\beta = 0.5$. Precision is defined as TP/(TP + FP) and recall is defined as TP/(TP + FN).

The 5-fold cross-validation was used to assess the performance of the SVM models. Predictions, as to whether each residue binds RNA or not, are made for the test sub-set at varying SVM decision values. If a residue has an SVM score above the decision value (the threshold value), then the residue is predicted to bind RNA. A range of SVM scores from −4.000 to +4.000 were used as thresholds, in increments of 0.001. The performance measures described above are computed for each threshold, then the average performance measures (over the five iterations) are calculated for each threshold. The averaged sensitivity and specificity values are used to generate a Receiver Operator Characteristic (ROC) curve; plotting SN vs. 1.0-SP at each threshold. To find the best model, the C and γ parameters of the RBF kernel are changed incrementally in turn and the five-fold cross-validation is repeated after each change. The C and γ values that give the largest AUC provide the optimal solution.

Models, trained and tested using the 5-fold cross-validation procedure, were also tested using a dataset of 42 proteins that were not used in training (RNAtestset42). This test, using previously unseen and unrelated sequences, was a true test of how well the model will perform when used to make predictions for a protein of unknown function. The TP, FP, TN and FN values for each protein in the test set are used to calculate overall performance measures.

4 IMPLEMENTATION

4.1 Performance of SVM trained on the RNAset81 dataset

The RNAset81 dataset represents a validated benchmark dataset as defined by Jeong *et al*. (2004; Jeong and Miyano, 2006) and used by Kumar *et al*. (2008), but amended in the current work to remove homologous proteins (Section 2). The dataset contains 2938 RBRs (the positive class) and 16175 non-RBRs (the negative class).

All possible combinations of the four features (PSSM, IP, pA and H), and window sizes of length 5–25, were used to train the SVM, and the performance of the subsequent models was evaluated (Table 1). Table 1 shows that the best performance is achieved by integrating the PSSM with one or more additional sequence features. Models created from the PSSM with combinations of IP, pA and H all give MCC values ≥ 0.492, larger than for the model with PSSM alone as the feature vector. The additional features do add value to the model as shown by the calculation of the $F_{0.5}$-measure. This measure for all SVM models with additional features (SVM II–VIII) is $\geq$0.56 compared with 0.55 for the model with just the PSSM as the feature vector (SVM I). The SVM model with the largest AUC then MCC value (SVM VIII) was selected as the final model. As models SVM II–VIII gave similar performances in the 5-fold cross-validation, all were tested using the RNAtestset42 dataset (see next section); the PSSM + IP + pA + H model (SVM VIII) ranked amongst those with highest MCC, ACC and SP values, explaining why (in addition to the rankings in Table 1) it was selected as the final model.

The relatively low sensitivity values in Table 1 suggest that combining the predictions from more than one model could have the potential to increase the performance of the method. To test this hypothesis, the overlap between the specific predictions made by each model was calculated (Supplementary Tables S3a–e). The results showed that creating a union between models did not improve the method, as the overlap between the models was high. Hence, the

highest performing model remained the SVM that combined all four features.

The performance measures for the best SVM model from Table 1 are compared with the published performance measures for five other RBR prediction methods (Table 2). This comparison shows that our method (SVM VIII, named PiRaNhA) outperforms all previously published methods in terms of accuracy, AUC and MCC values. The increase in MCC seen with PiRaNhA compared with PPRInt is greater than the improvement they achieved over the next best-published method at that time, namely the NN-based method.

4.2 Predicting binding residues in the RNAtestset42 dataset

A true test of any prediction tool is to make predictions for proteins not related to those used in training. In the current work, predictions were made for proteins in RNAtestset42 using the best model trained on RNAset81 (PiRaNhA) and, for comparison, the PPRInt (Kumar *et al*., 2008), RNABindR (Terribilini *et al*., 2007) and BindN (Wang and Brown, 2006) servers (Table 3). The predictions were carried out, for all methods, using default parameters (and using 'optimal predictions' from RNABindR); the default threshold for PiRaNhA is –0.44 (the SVM score that gave the best performance in the cross-validation). In the absence of AUC values (AUC values cannot be calculated as single-threshold values are used for each model), the results are evaluated based on MCC value and the $F_{0.5}$-measure. The MCC score combines all four of the TP, FP, TN and FN counts into one score, and is considered to be the best evaluation of the overall performance of a method (Baldi *et al*., 2000).

Based on MCC and $F_{0.5}$-measure values, PiRaNhA outperforms the RNABindR, PPRInt and BindN servers. Each method displays a different balance of sensitivity and specificity, with RNABindR being ranked second to PiRaNhA based on MCC and the $F_{0.5}$-measure, with a higher specificity and accuracy, but a considerably lower sensitivity.

4.3 Predicting RNA-binding function at the protein level

To address the problem of predicting RNA-binding function at the protein level, statistical features of the SVM decision values from PiRaNhA were used to train a second SVM classifier. For this purpose, a non-redundant dataset of 268 proteins [134 RNA-binding proteins (positives) and 134 non-RNA-binding proteins (negatives)] was selected. The positive set was created from UniProtKB/Swiss-Prot using strict keywords (RNA-, rRNA- and tRNA-binding). The sequences retrieved were then clustered by Pfam annotation, and one protein retained from each cluster. Any sequences with Pfam domains in common with proteins in RNAset81 (the set used for training PiRaNhA), were discarded. The negative set was created in the same way, using UniProtKB/Swiss-Prot entries that did not feature any keywords that could imply an RNA-binding function. The complete negative set contained 4849 proteins, and a random set of 134 was selected to create equal sized positive and negative sets for training.

Table 1. The 5-fold cross-validation of all possible combinations of sequence features using RNAset81

SVM	Feature vector	Window size	SN (%)	SP (%)	ACC (%)	MCC	AUC
VIII	PSSM + IP + pA + H	23	56.3	92.8	87.2	0.499	0.860
III	PSSM + pA	25	52.7	93.9	87.6	0.497	0.860
VI	PSSM + IP + H	25	58.4	92.3	87.1	0.506	0.859
VII	PSSM + pA + H	23	58.5	92.1	87.0	0.504	0.859
V	PSSM + IP + pA	23	56.9	92.6	87.1	0.500	0.858
II	PSSM + IP	25	53.9	93.7	87.5	0.499	0.859
IV	PSSM + H	23	58.3	91.7	86.6	0.492	0.855
I	PSSM	23	60.1	90.8	86.1	0.490	0.855

Models are ordered by AUC value. AUC shown to three decimal places.

Table 3. RBR predictions for the RNAtestset42 dataset using PiRaNhA and using RNABindR, PPRInt and BindN

Method	SN (%)	SP (%)	ACC (%)	MCC	$F_{0.5}$
PiRaNhA	53.0	90.0	84.5	0.41	0.49
RNABindR (Terribilini *et al*., 2007)	37.4	93.8	85.5	0.36	0.48
PPRInt (Kumar *et al*., 2008)	70.4	73.9	73.4	0.34	0.36
BindN (Wang and Brown, 2006)	55.0	80.2	76.4	0.29	0.35

Methods are ordered by MCC value.

Table 2. Comparison of reported performance measures for five other prediction methods and the best model from the current work based on the RNAset81 dataset

Method	RBR definition (Å)	Cross-validation	SN (%)	SP (%)	ACC (%)	AUC	MCC
PiRaNhA (SVM VIII)	3.9 (HBPLUS)	5-fold	56.3	92.8	87.2	0.86	0.50
PPRInt (Kumar *et al*., 2008)	6	5-fold	53.1	89.6	81.2	–	0.45
PRINTR (Wang *et al*., 2008)	ENTANGLE (Allers and Shamoo, 2001)	7-fold	55.9	–	87.1	0.83	0.43
NN-based (Jeong and Miyano, 2006; Jeong *et al*., 2004)	6	10-fold	–	–	–	0.77	0.41
RNABindR (Terribilini *et al*., 2007)	5	leave-1-out	33	95	83	–	0.36
BindN (Wang and Brown, 2006)	3.5	5-fold	66.3	69.8	69.3	0.73	0.27

Methods are ordered by MCC value.

The PiRaNhA server was used to make predictions for all the residues in the 268 proteins. A series of statistical features of the SVM decision values from these RBR predictions were then evaluated as features in a second SVM to predict RNA-binding function at the protein level. The features tested for this second stage SVM included the minimum, median, mean, variance, skew, range and kurtosis (measure of the 'peakedness' of a distribution) of the decision values from PiRaNhA. The SVM which used minimum, maximum, mean, range, third quantile, skew and kurtosis in the feature vector achieved the highest MCC score (0.53), and an accuracy of 76.1%, in a 5-fold cross-validation (Table 4).

5 DISCUSSION

This article presents an accurate method (PiRaNhA) for the prediction of RBR, using a machine learning approach that shows an improvement in performance over previous methods. The method is based on an SVM that is the first to integrate PSSMs, amino acid interface proponsities, predicted residue accessibility and hydrophobicity in the feature vector. The method has been assessed using 5-fold cross-validation on a dataset of 81 RNA-binding protein sequences (RNAset81) and by prediction on a set of 42 RNA-binding protein sequences not related to those used in training (RNAtestset42). The highest performing model uses all the features over a window of 23 residues. This model achieves an MCC of 0.50, an AUC of 0.86 and an accuracy of 87.2% in a 5-fold cross-validation. In a true test of performance, predictions for the RNAtestset42 dataset were made using PiRaNhA and three previously published prediction servers (PPRInt, RNABindR and BindN). This test showed that PiRaNhA is the best performing method, with an MCC of 0.41 and an $F_{0.5}$-measure of 0.49.

PiRaNhA uses a relatively large window size of 23 residues. The use of a window enables the sequence environment of the residue in question to be taken into account. It has been observed for many types of protein–ligand interaction that residues not bound directly to the ligand, but in close proximity to those that do, make a contribution to binding. This is known to occur in protein–protein interactions where inner and outer rings of residues have been identified, each making different contributions to the energy of association (Bogan and Thorn, 1998). SVM with larger sequence windows prove more effective as they use information from a larger sequence environment, reflecting the fact that residues beyond those directly bound to the RNA contribute to the association.

The RNA-binding function predictions at the level of the whole protein show a level of accuracy within the range observed for previous methods (Table 4). However, as stated previously, benchmark datasets, and AUC and MCC performance measures have not been used in this field previously, and hence detailed comparisons between the current and previous methods cannot be made. The current method uses a 2-layer feature vector, and it is the high accuracy of the PiRaNhA residue predictions that make this method effective.

Whilst the two-stage SVM method requires further testing, it does show potential to accurately predict *if* a protein binds RNA, as well as *where*. Such a tool will enable proteomes to be scanned for novel RNA-binding proteins, the identification of which will lead to a wider understanding of disease pathways and potentially highlight new drug targets.

PiRaNhA for RBR predictions is freely available at piranha.protein.osaka-u.ac.jp and www.bioinformatics.sussex.ac.uk/PIRANHA.

Funding: Medical Research Council (grant number 70760 to R.V.S.); Strategic International Cooperative Program, Japan Science and Technology Agency (to Y.M. and H.N).

Conflict of Interest: none declared.

Table 4. Performance measures for the five-fold cross-validation of the 2-stage SVM to predict RNA-binding function at the protein level for a non-redundant dataset of 268 proteins.

Feature vector	SN (%)	SP (%)	ACC (%)	AUC	MCC
Minimum, maximum, mean, range, third quantile, skew, kurtosis of decision values from PiRaNhA	80.0	72.3	76.1	0.80	0.53

REFERENCES

Adamczak,R. *et al*. (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Struct. Funct. Bioinf*., **59**, 467–475.

Allers,J. and Shamoo,Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol*., **311**, 75–86.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baldi,P. *et al*. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol*., **280**, 1–9.

Cai,Y.D. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *BBA-Proteins and Proteomics*, **1648**, 127–133.

Chen,Y. and Varani,G. (2005) Protein families and RNA recognition. *FEBS J.*, **272**, 2088–2097.

Ellis,J.J. *et al*. (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins: Struct. Funct. Bioinf*., **66**, 903–911.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Fujishima,K. *et al*. (2007) Proteome-wide prediction of novel DNA/RNA-binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon *Pyrococcus furiosus*. *DNA Res.*, **14**, 91–102.

Han,L.Y. *et al*. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355–368.

Hripcsak,G. and Rothschild,A.S. (2005) Agreement, the F-measure, and reliability in information retrieval. *J. Am. Med. Inform. Assoc.*, **12**, 296–298.

Kumar,M. *et al*. (2008) Prediction of RNA-binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lukong,K.E. *et al*. (2008) RNA-binding proteins in human genetic disease. *Trends Genet.*, **24**, 416–425.

Martin,A.C. (2005) Mapping PDB chains to UniProtKB entries. *Bioinformatics*, **21**, 4297–4301.

McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.

Murakami,Y. and Jones,S. (2006) SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, **22**, 1794–1795.

Terribilini,M. *et al*. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.

UniProt_Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Wagner,M. *et al*. (2005). Linear regression models for solvent accessibility prediction in proteins. *J. Comput. Biol.*, **12**, 355–369.

Wang,L.J. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.

Wang,Y. *et al*. (2008) PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*

Yu,X.J. *et al*. (2006) Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.*, **240**, 175–184.