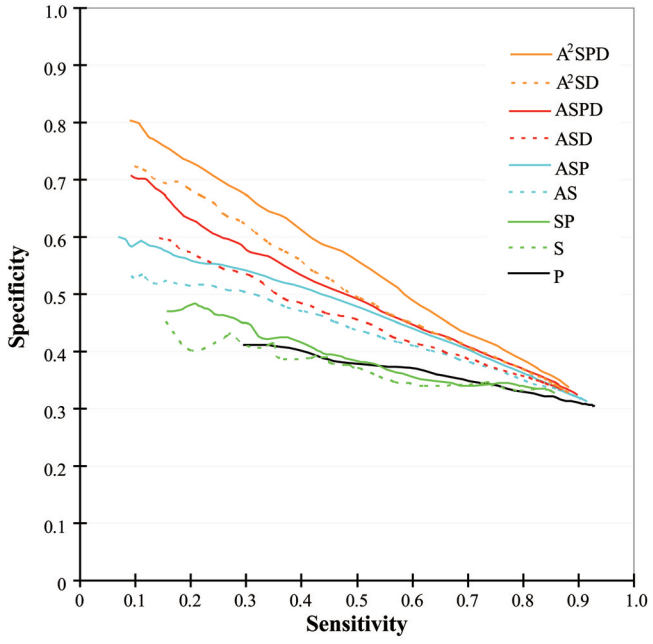

Figure 3. Quality of interface predictions evaluated on 86 jack-knifed datasets. The dotted lines show the evaluation of the prediction with the propensities and the solid lines show that of the prediction with the propensities plus the profile. The prediction using the profile only (**P**) is shown in black solid line. The predictions using the singlet propensities (**S** and **SP**) are shown in green. The predictions using the averaged singlet propensities (**AS** and **ASP**) are shown in cyan. The predictions using the doublet propensities (**ASD** and **ASPD**) are shown in red. The predictions using averaged singlet and doublet propensities (**A^{2}SD** and **A^{2}SPD**) are shown in orange.

of mutually overlapping spheres with 7.0 Å radii. When the singlet score was first weighted by the multiple sequence profile and then averaged over the near neighbor residues (Equation 14), the quality of prediction was further improved (**ASP** in Figure 3). The multiple sequence profile provides the empirical probability of each amino acid type being found at the position under consideration. Therefore, the score weighted by the multiple sequence profile reflects the score for a homologous set of proteins and suppresses possible noise due to random amino acid occurrence. Inclusion of the doublet propensity score further improved our predictive ability (Equation 15) and the predictive ability was further enhanced by the multiple sequence profile weighting (Equation 16) (**ASD** and **ASPD**, respectively in Figure 3).

Residue doublet propensities gave information of residue pair preference shown in Figure 2. The differences between **AS** and **ASD** and between **ASP** and **ASPD** represent the contribution of the doublet propensities. Incorporation of the doublet propensities into the calculations markedly increased the specificity of our predictions and this was especially true when using a high score threshold and the resulting sensitivity was low. Averaging scores over near neighbor residues effectively increased our predictive ability (compare **S** to **AS**) and therefore, we performed an averaging procedure for the doublet information contained in **ASD** and **ASPD**. The singlet information in those scores was already averaged. Therefore, the averaging doublet information in **ASD** and **ASPD** resulted in averaging the singlet information twice (therefore, the names of scores were **A^{2}SD** and **A^{2}SPD**). Thus, residues as far as 14.0 Å away might influence the score. This procedure further enhanced the prediction specificity and sensitivity as seen in Figure 3 (**A^{2}SD** and **A^{2}SPD**).

The quality of all nine prediction methods converged to a point where sensitivity and specificity were ~0.9 and 0.3, respectively. At this point, the threshold for determining the predicted interface was too low and most surface residues were predicted to constitute RNA interfaces in all prediction methods. When all of the surface residues were predicted as RNA interfaces, most of the interface residues were correctly predicted (sensitivity ≈ 1.0) and many non-interface residues (~70% of the surface residues) were also positively predicted (specificity ≈ 0.3).

The prediction with the averaged singlet and doublet propensities plus the multiple sequence profile (**A^{2}SPD**) achieved the best quality. Its specificity reached as high as 0.8 at the most strict threshold. We used the nine different methods (Equations 10–18) to predict the RNA-binding interface of the known RNA-binding interface for sex-lethal protein (PDB ID: 1B7F) in Figure 4. In this case, the sensitivity of the predictions was fixed close to 0.5. As shown in the figure, the specificities improved in the order of **S**, **P**, **SP**, **AS**, **ASP**, **ASD**, **A^{2}SD**, **ASPD** and **A^{2}SPD**. The improvement was manifested by the reduction in false positive residues, i.e. over-prediction (green residues in Figure 4). The difference between **AS** and **ASD** and between **ASD** and **ASPD** indicated the effects of the doublet propensity and the multiple sequence profile on the predictions, respectively.

Summary of our prediction method

A number of studies have presented structure-based analyses of protein–RNA interactions (27,28,30,37). Those studies showed singlet propensities similar to the present work. In our report, we were able to predict the RNA interface relatively well after incorporating the doublet propensity and the multiple sequence profile into our calculations. The specificity of prediction by **A^{2}SPD** was as high as 80%. With this specificity, the prediction can determine residues that are almost certain to interact with RNA and this could advance wet-lab experiments designed to identify residues constituting the RNA interface. By mutating each of the predicted residues, there is a reasonable probability of experimentally identifying RNA interface residues with minimal cost compared with a random mutagenesis approach.

In our future work, we will try to improve the prediction of RNA interface starting with those residues predicted as RNA interface with high confidence. Our data suggests that the protein–RNA interface area can be described as a set of overlapping spheres and residues near those predicted to comprise the RNA interface are likely involved in protein–RNA interactions. Additional methods capable of delineating interface and non-interface areas around the 'predicted core interface residue' will allow predictions to be made with higher specificity and sensitivity. A prediction improvement of this nature was also suggested by Kloczkowski *et al.* (38) for improving their GOR secondary structure prediction. They suggested that by incorporating the information from a small number of residues predicted with high confidence into the next level of prediction, the overall quality of prediction should improve.

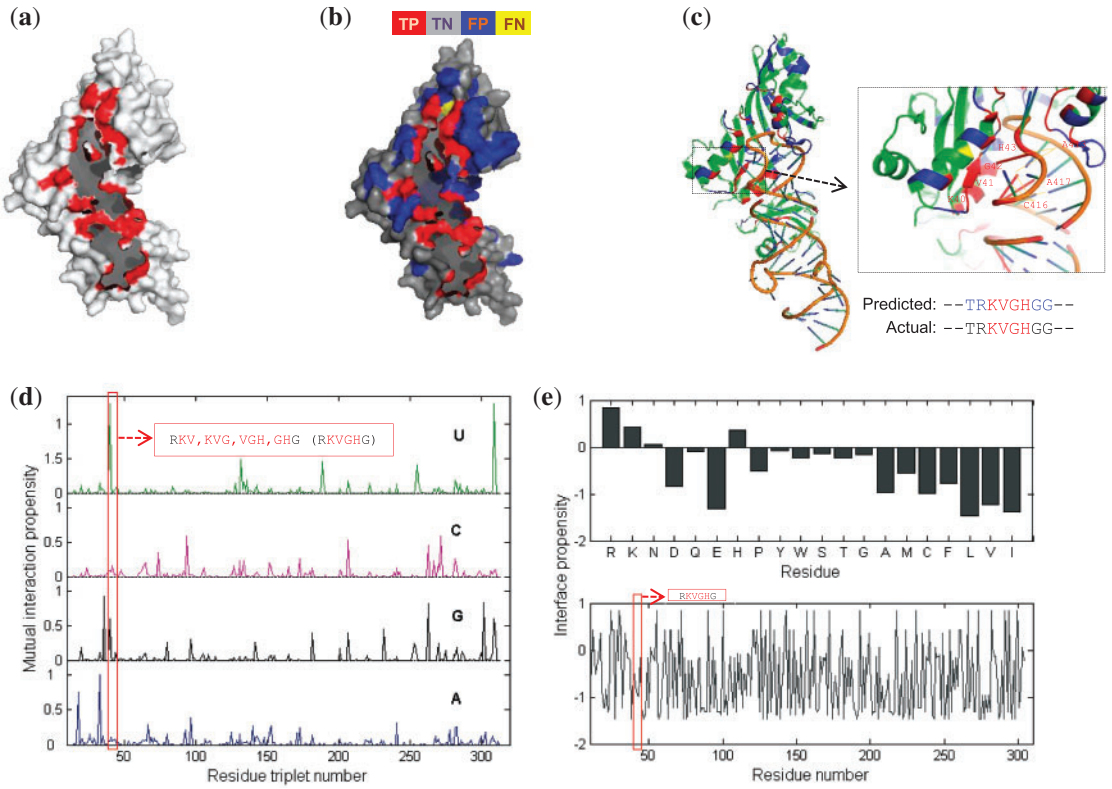

Fig. 2. An example of predicting RNA binding sites. (**a**) Actual interface residues with RNA in protein 1R3E:A. (**b**) Predictions are mapped onto the original structure where different prediction catalogs are represented by different colors. (**c**) Structure of the protein–RNA complex with an example of prediction in the zoomed part. (**d**) Mutual interaction propensity between the triplets and nucleotides in the protein. Triplets are listed by sliding residues through the protein sequence. The box part corresponds to the values of residues in the zoomed part of (c). (**e**) Upper panel shows the interface propensity of each amino acid type in the dataset. It is defined as the proportion of an amino acid in interaction sites divided by the proportion of the residue in the dataset (see more in Supplementary Materials). Lower panel shows the interface propensity of binding with RNA for the residues in the protein. The box part corresponds to the values of the zoomed sites.

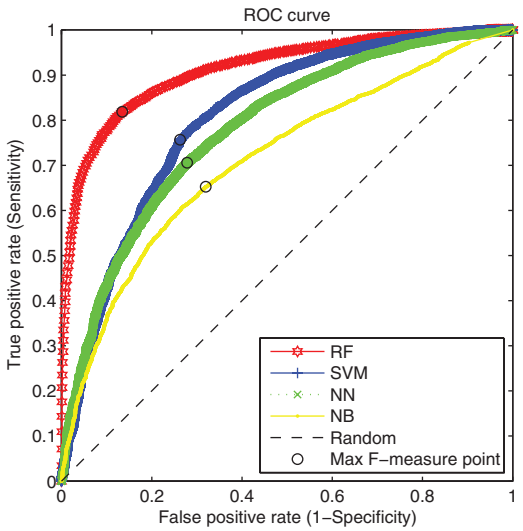

Fig. 3. The ROC performance of several classifiers.

Table 2. Comparison of the prediction performances

Method	SN (%)	SP (%)	ACC (%)	*F*-measure	MCC
Our method	81.9	86.8	82.4	0.843	0.488
RNABindR (opt)	34.2	93.8	87.3	0.501	0.300
RNABindR (sn)	83.7	49.5	62.3	0.623	0.208
RNABindR (sp)	17.1	98.6	89.7	0.292	0.281
BindN (sn)	77.3	52.9	55.5	0.628	0.188
BindN (sp)	51.0	79.6	76.5	0.622	0.225
RNAProB	74.0	65.6	73.1	0.696	0.267
PPRint	78.9	74.1	74.6	0.764	0.355
SVM based	75.7	73.7	75.5	0.747	0.335
NN based	70.6	72.2	70.7	0.714	0.280
NB based	65.3	68.1	65.6	0.667	0.211

residues, we found that the ACC of prediction was declined than that of using all descriptors. For instance, when we deleted the interaction propensity, ACC, *F*-measure and AUC of the prediction became 75.8%, 0.751 and 0.828 comparing to 84.5%, 0.859 and 0.923 with all descriptors, respectively.

From Table 3, we can also identify the importance of these features for the prediction and the contribution of these properties

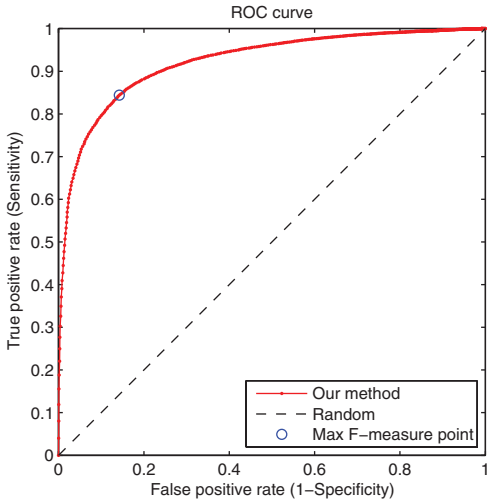

Fig. 1. The ROC curve of predicting performance.

Table 1. The result of 5-fold cross-validation of the RF classifier with different window sizes in the 53 315 residues

Window size	SN (%)	SP (%)	ACC (%)	*F*-measure	AUC
1	72.5	79.5	73.2	0.758	0.844
3	83.1	86.6	83.4	0.848	0.918
5	84.4	85.8	84.5	0.851	0.923
7	84.9	85.0	84.9	0.850	0.922
9	85.0	84.5	84.9	0.850	0.920
11	82.5	86.0	82.8	0.842	0.916
13	83.8	84.0	83.9	0.840	0.915

The performance of *F*-measure and AUC is slightly declined when we increase the window size. The best performance was obtained with the window size of 5 residues. The test provided evidences for the effectiveness of our method to predict the RNA interface residues in proteins. Figures 2a and b show an example of the predicted interface residues with RNA in protein 1R3E:A. Figure 2a presents the actual interface residues of the protein structure in red. Figure 2b uses a different coloring scheme to illustrate the prediction performance on individual residues. Figure 2c gives a part of the predicted versus the actual binding residues (structure is shown in the box panel). Most of the actual interface residues (ACC of 85.57%) are well identified in the protein, and the prediction details are shown in Supplementary Materials.

3.2 Comparison with other methods

In this work, we proposed a novel RF-based method to predict the binding residues in proteins by using the combined features. By considering the two-side information in the interacting partners of amino acids and nucleotides, we show that the proposed interaction propensity feature is able to represent the binding SP of the residues in proteins. The sequence features as well as the structure features of the residues can discriminate the propensity of amino acids to interact with RNA when combined together. The superior performance of the prediction in cross-validation experiments confirms the effectiveness of our method. In this subsection, we further validate our method by comparing our method with other existing methods.

In recent years, several methods have been proposed to predict RNA binding residues in proteins, such as RNABindR (Terribilini *et al*., 2007), BindN (Wang and Brown, 2006), RNAProB (Cheng *et al*., 2008) and PPRint (Kumar *et al*., 2008). For comparison, we tested the prediction performances of these methods. In the 205 protein chains, we constructed a training dataset by randomly selecting 105 protein chains as well as a testing dataset of the rest 100 chains. We trained the RF classifier by using the training dataset and validated the prediction in the testing dataset to compare the performance of the methods. The tradeoff threshold between SN and SP of our method was set to be RF voting score that gave the best performance. Figure 3 shows the ROC curve of the results. The predictions in the testing 100 protein chains by other methods were carried out by using their default parameters. The results are shown in Table 2. RNABindR is a Bayesian classifier for RNA binding sites in proteins. BindN, RNAProB and PPRint are predictors of protein–RNA binding residues based on SVMs. RNABindR was implemented in three different options, i.e. 'optimal prediction (opt)', 'high sensitivity (sn)' and 'high specificity (sp)' prediction. Similarly, BindN was tested with two options of 'expected sensitivity of 80% (sn)' and 'expected specificity of 80% (sp)'. RNABindR and BindN with high SP are obtained at the expense of SN, and vice versa. Clearly, our method consistently outperforms the existing methods in terms of these measures.

Since some prediction scores of the compared methods are not available, their ROC curves cannot be drawn. To remove the possible biases in the comparison, we also compared the underlying machine learning algorithms in these predictors. We implemented several different algorithms, i.e. support vector machine (SVM), naive Bayes (NB) and neural network (NN), using the same procedure as our RF-based method. Figure 3 shows the ROC curves of different classifiers in the testing dataset. The performance details are given in Table 2. In Figure 3, AUC of RF-, SVM-, NN- and NB-based predictors are 0.912, 0.801, 0.782 and 0.713, respectively. Our RF-based method clearly outperformed other classifiers. As to the different datasets, we also tested our methods in several benchmarks. In RNA binding protein datasets RB86 (Kumar *et al*., 2008; Terribilini *et al*., 2006), RB107 (Kumar *et al*., 2008; Wang and Brown, 2006), RB109 (Terribilini *et al*., 2007) and RB149 (Terribilini *et al*., 2007), our method can achieve the AUC of 0.884, 0.885, 0.884 and 0.858, respectively, which also demonstrate the effectiveness of the proposed method. We also carried out more comparison with RNAProB (Cheng *et al*., 2008). The details of these results can be found in the Supplementary Materials.

3.3 Evaluation of feature importance

We combined various features of the residues in addition to the mutual interaction propensity to represent the specific interaction properties of protein residues with RNA nucleotides. Seven descriptors were contained in a hybrid feature vector. To verify their effects on the prediction of binding sites, we tested the performance of the selected features of these descriptors. Table 3 presents the results of prediction performance of the 5-fold cross-validation by subtracting one of the descriptors individually in the scoring scheme. After subtracting each descriptor in describing these

interacting with the RNA. By comparing predicted interactions with the ones observed in RNP complexes, we established true and false positives and negatives. We used these values to create receiver operating characteristic (ROC) curves by plotting the false positive rate (1 – specificity, FPR; Eq. (1) against the true positive rate (sensitivity, TPR; Eq. (2) for each method. We have also estimated the areas under the ROC curve (AUC) using the composite trapezoidal rule and calculated the Matthews Correlation Coefficient (MCC; Eq. (3)) for each method.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{1}$$

Eq. (1). True positive rate (TPR; sensitivity). TP – number of correctly predicted interacting residues, FN – number of incorrectly predicted non-interacting residues.

$$\mathrm{FPR} = 1 - \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{2}$$

Eq. (2). False positive rate (FPR; 1 – specificity). TN – number of correctly predicted non-interacting residues, FP – number of incorrectly predicted interacting residues.

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{3}$$

Eq. (3). Matthews Correlation Coefficient (MCC). TP – number of correctly predicted interacting residues, TN – number of correctly predicted non-interacting residues, FP – number of incorrectly predicted interacting residues, FN – number of incorrectly predicted non-interacting residues.

6. Results of the RNA binding site prediction benchmark

The results of our benchmark presented in Table 3 and Fig. 1 provide an overview of the performance of 10 third-party tools for the prediction of protein–RNA interactions and one ad hoc created meta-predictor (described in a separate section below). We were not able to perform the ROC analysis in case of the sequence-based method PRBR, and all three structure-based methods (DRNA, KYG and OPRA). The reason was that the output of those methods did not include scores for individual residues describing their RNA-binding propensity, which is a compulsory requirement for such an analysis. For the remaining methods, we performed the ROC analysis within a range of observed scores describing the predicted RNA-binding propensity. In case of RNABindR, the output for each protein sequence contained three

Table 3
Results of a benchmark of 10 methods predicting protein–RNA interactions – seven sequence-based methods listed in Table 1.

Method	MCC	AUC
Meta-predictor**	0.460	0.835
PiRaNhA	0.435	0.822
BindN+	0.397	0.821
KYG*	0.382	N/A
DRNA*	0.382	N/A
PPRInt	0.339	0.779
RNABindR	0.317	0.708
OPRA*	0.296	N/A
BindN	0.297	0.733
PRBR	0.294	N/A
NAPS	0.215	0.679

Methods were sorted in descending order according to MCC. N/A – not available, MCC – Matthews Correlation Coefficient, AUC – area under curve.
* Three structure-based from Table 2 (KYG, OPRA and DRNA).
** An ad hoc meta-predictor developed during this study based on top three sequence-based methods according to our benchmark (PiRaNhA, PPRInt and BindN+).

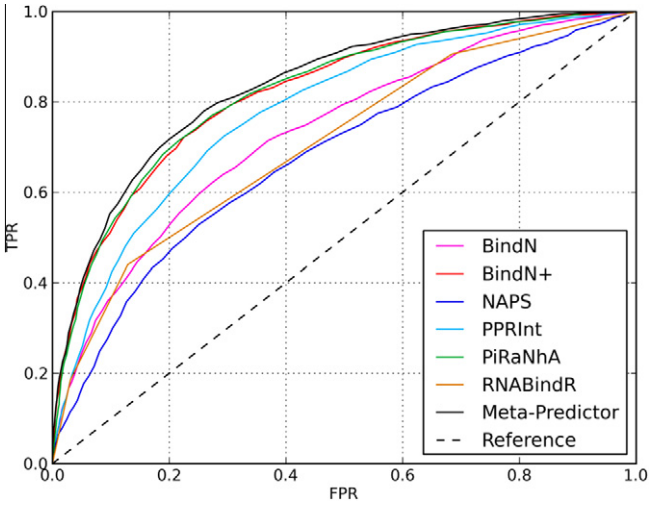

Fig.1. ROC curves of six methods for prediction of RNA-binding residues from sequence (Table 1) and a meta-predictor created during this study using the best three sequence-based methods (PiRaNhA, PPRInt and BindN+).

predictions – optimal, high specificity, and high sensitivity. Because of the fact that no additional scores describing RNA-binding propensities were provided, we calculated scores based on the congruency of predictions. Therefore, as seen in Fig. 1, the RNABindR ROC adopts an unusual shape. In case of all other methods, the ROC analysis was performed based on the scores assigned to protein residues.

According to our benchmark, methods KYG and DRNA were the best among the structure-based methods tested, with MCC values reaching 0.382 in both cases. This value may be overestimated, as the predictions were tested for protein structures taken from protein–RNA complexes (i.e. correspond to the RNA-bound conformation), while in real life the predictions are made for proteins with known structures, but unknown mode of RNA binding. A test of predictions made for unbound variants may be done in the future, when the number of structure pairs with and without RNA grows. Among the sequence-based methods, the ranking was topped by PiRaNhA, for which the MCC reached 0.435 and the AUC was estimated to be 0.822. The next two best-scored methods were BindN+ (MCC: 0.397, AUC: 0.821) and PPRInt (MCC: 0.339, AUC: 0.779).

7. Meta-predictor for protein–RNA interactions

Following the benchmark of primary predictors of RNA-interacting residues, we developed an 'ad hoc' meta-predictor based on three sequence-based primary predictors that ranked highest in our tests (PiRaNhA, PPRInt and BindN+). The meta-predictor works as follows: first, for a query protein sequence, predictions are collected from the three above-mentioned primary predictors. Then, a new meta-score for each residue is calculated as a weighted mean of three scores using the AUC values from the benchmark as weights. As the output, the meta-predictor returns a set of scores for all residues of a given protein sequence query. A threshold to discriminate between RNA-binding and non-binding residues was defined according to the point on the meta-predictor's ROC curve closest to the values of FPR = 0.0 and TPR = 1.0 (upper left corner). Once the threshold value was selected, we were able to calculate the MCC value for the meta-predictions. Our meta-predictor outperformed PiRaNhA only by 1.6% according to AUC (0.835 vs. 0.822) and by 5.7%, according to MCC (0.460 vs. 0.435), which suggests that the predictions of the currently best methods are strongly correlated with each other and combining