Bioinformatics, 34(9), 2018, 1473–1480
doi: 10.1093/bioinformatics/btx822
Advance Access Publication Date: 21 December 2017
Original Paper

Structural bioinformatics

Computational identification of binding energy hot spots in protein–RNA complexes using an ensemble approach

Yuliang Pan[1], Zixiang Wang[1], Weihua Zhan[2] and Lei Deng[1,3,*]

[1]School of Software, Central South University, Changsha 410075, China, [2]School of Electronics and Computer Science, Zhejiang Wanli University, Ningbo 315100, China and [3]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 12, 2017; revised on December 5, 2017; editorial decision on December 16, 2017; accepted on December 19, 2017

Abstract

Motivation: Identifying RNA-binding residues, especially energetically favored hot spots, can provide valuable clues for understanding the mechanisms and functional importance of protein–RNA interactions. Yet, limited availability of experimentally recognized energy hot spots in protein–RNA crystal structures leads to the difficulties in developing empirical identification approaches. Computational prediction of RNA-binding hot spot residues is still in its infant stage.

Results: Here, we describe a computational method, PrabHot (Prediction of protein–RNA binding hot spots), that can effectively detect hot spot residues on protein–RNA binding interfaces using an ensemble of conceptually different machine learning classifiers. Residue interaction network features and new solvent exposure characteristics are combined together and selected for classification with the Boruta algorithm. In particular, two new reference datasets (benchmark and independent) have been generated containing 107 hot spots from 47 known protein–RNA complex structures. In 10-fold cross-validation on the training dataset, PrabHot achieves promising performances with an AUC score of 0.86 and a sensitivity of 0.78, which are significantly better than that of the pioneer RNA-binding hot spot prediction method HotSPRing. We also demonstrate the capability of our proposed method on the independent test dataset and gain a competitive advantage as a result.

Availability and implementation: The PrabHot webserver is freely available at http://denglab.org/PrabHot/.

Contact: leideng@csu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein–RNA interactions play crucial roles in a large number of biological processes (König *et al.*, 2012; Zhang *et al.*, 2017a, b). Among the RNA-binding interface residues, only a small fraction termed hot spots make dominant contributions to the most of the binding free energy. Locating specific RNA-binding hot spot residues is of vital importance for exploring the underlying molecular recognition mechanism, and a good starting point for many applications such as protein engineering and drug design (Cho *et al.*, 2009; Wang *et al.*, 2012). Mutagenesis technologies have been applied to expose the RNA-binding hot spots that dramatically change the binding affinity (Loedige *et al.*, 2014; Yan *et al.*, 2003; Yang *et al.*, 1997). However, the number of experimentally determined hot spots is severely limited since laboratory experiments are costly and time-consuming. Computational approaches provide an alternative pathway to investigate energy hot spots on a large scale.

A series of computational approaches (Cho *et al.*, 2009; Deng *et al.*, 2013, 2014; Petukh *et al.*, 2015) focusing on protein-protein binding hot spots have been developed based on several existing databases (Fischer *et al.*, 2003; Moal and Fernández-Recio, 2012; Thorn and Bogan, 2001) that contain experimentally verified hot spots. Molecular dynamics (MD) simulations, empirical functions and machine learning methods have been suggested to be useful for hot spot prediction (Deng *et al.*, 2013). On the other hand, a lot of efforts have been made to predict RNA-binding sites (Amitai *et al.*, 2004; Castello *et al.*, 2016; Kim *et al.*, 2006; Kumar *et al.*, 2008; Li *et al.*, 2011; Liu *et al.*, 2010; Murakami *et al.*, 2010; Paz *et al.*, 2016; Wang *et al.*, 2010, 2013; Walia *et al.*, 2012, 2014). However, most of these methods are not suitable for discriminating hot spots at the RNA-binding interface. Recently, Barik and coworkers analyzed the degree of evolutionary conservation across the interface residues involved in protein–RNA complexes, and combined Random Forests with structural and physico-chemical features to probe protein–RNA binding hot spots (Barik *et al.*, 2016). Here is some text.

Although these advancements have been made, the problem of predicting RNA-binding hot spots is still in the initial stage. The primary factor is that the experimental determination of hot spot residues in RNA-binding proteins remains elusive because of the high cost and effort. Until now, there is no publicly available database for capturing protein–RNA binding hot spots from mutagenesis experiments, which limits the development of empirical prediction methods. Besides, the relationship between RNA-binding hot spot residues and physic-chemical, evolutionary or structural features has not been clarified. Informative attributes for accurately identifying RNA-binding hot spots have not been thoroughly exploited.

To address this issue, we created a reference dataset of 107 hot spots from 47 protein–RNA crystal structures, which are manually collected from the references. Also, we propose a novel ensemble approach, termed as PrabHot (Prediction of protein–RNA binding hot spots), for predicting hot spots in protein–RNA interfaces. The ensemble vote classifier (EVC) combines conceptually different machine learning classifiers and uses the average predicted probabilities to predict hot spots. We compute an optimal set of 35 features selected from a wide variety of sequence, structure and residue interaction network based features with the Boruta algorithm (Kursa *et al.*, 2010). We further show that PrabHot achieves a significantly improved overall performance on the cross-validation dataset and independent dataset, and is capable of more accurately predicting RNA-binding hot spots compared to other state-of-the-art predictors. The flowchart of PrabHot is shown in Figure 1. A web server of PrabHot is available at http://denglab.org/PrabHot/.

2 Materials and methods

2.1 Datasets

We extracted 63 protein–RNA complexes, which contains experimentally measured binding free energy changes of 350 mutations. Among them, 13 protein–RNA complexes were taken from Barik et al.'s work (Barik *et al.*, 2016), and the other 50 protein–RNA complexes were manually curated from the literature. To remove the redundancy, proteins with sequence similarity > 40% were excluded by using CD-HIT (Li and Godzik, 2006). The interface residues were calculated based on the buried solvent accessible surface area upon complex formation ($\Delta ASA > 1$ Å) and relative solvent accessible surface area (RASA > 5%) by using Naccess (Deng *et al.*, 2009; Hubbard and Thornton, 1993). The interface residues are defined as hot spots if the corresponding binding free energy change ($\Delta\Delta G$) $\geq$ 1.0 kcal/mol, and the remaining residues are defined as non-hot spots. Eventually, we obtained a dataset of 47 protein–RNA complexes containing 107 hot spots and 102 energetically unimportant residues, with a roughly balanced ratio of 1:1. A total of 79 hot spots and 72 non-hot spots from 32 protein–RNA complexes were randomly selected as the benchmark dataset and the rest 15 complexes were used as the independent dataset including 28 hot spots and 30 non-hot spots.

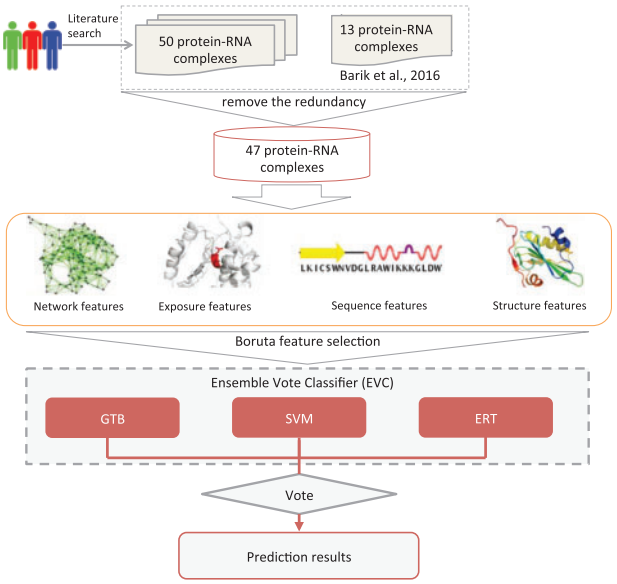

Fig. 1. Flowchart of PrabHot. A reference dataset of 47 protein–RNA crystal structures is generated from literature curation and Barik et al.'s work (Barik *et al.*, 2016). PrabHot calculates four primary sources of information, namely network, exposure, sequence and structure determinants. The optimal features are selected using the Boruta feature selection algorithm. Then three hot spot classifiers, include SVM (Support Vector Machine), GTB (Gradient Tree Boosting) and ERT (Extremely Randomized Trees), are integrated using a voting approach (EVC). Finally, the performance is evaluated on the benchmark dataset and the independent test dataset

2.2 Performance evaluation

To assess the performance, we adopt several widely used measures, including accuracy (ACC), sensitivity (SEN/Recall), specificity (SPE), precision (PRE), F1-score (F1), the Matthew's correlation coefficient (MCC) and the area under the ROC curve (AUC). These measurements are defined as:

$$SEN = TP/(TP + FN), \quad (1)$$

$$SPE = TN/(TN + FP), \quad (2)$$

$$PRE = TP/(TP + FP), \quad (3)$$

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \quad (4)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (5)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (6)$$

2.3 Features extraction

We initially exploit a comprehensive set of 125 features for further selection. The feature set is composed of novel network and exposure determinants, incorporated with sequence and structure features. The details of these characteristics are described below.

2.3.1 Residue interaction network features

Residue interaction network (RIN) has been proved to be useful in some applications of Bioinformatics (Li *et al.*, 2011; Pan *et al.*, 2017). Two residues in a structure will be defined as in contact if the distance between the C_α atoms of them is within 6.5 Å. In this study, we use NAPS (Chakrabarty and Parekh, 2016) to compute a total of 7 topological features which describe the local environment of the target residue in the network, including betweenness, closeness, degree, clustering coefficient, eigenvector centrality, eccentricity and average nearest neighbor degree.

Betweenness is the ratio of all the shortest paths passing through a node and the total number of shortest paths in the network. This is calculated as:

$$C_b(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \sigma_{st}(u)/\sigma_{st}, \tag{7}$$

where V is the set of all nodes and σ_{st} is the number of shortest path between vertices s and t. The term $\sigma_{st}(u)$ indicates the number of shortest path between s and t passing through nodes u.

Closeness is a centrality measure of a vertex and is defined as the average geodesic distance to all other vertices. It can be defined as:

$$C_{cl}(u) = (n-1)/\sum_{v \in V} dist(u, v), \tag{8}$$

where n is the number of nodes in the network and $dist(u, v)$ is the shortest path distance between nodes u and v.

The eccentricity represents that the shortest path distance of the node to the farthest node in the network. It can be calculated as:

$$C_e(u) = max(dist(u, v)). \tag{9}$$

Degree is the number of edges incident to a vertex. This is calculated as:

$$C_d(u) = \sum_{v \in V} A_{uv}, \tag{10}$$

where A_{uv} is the number of contacts between nodes u and v.

The clustering coefficient is a measure of the closeness the neighbors of a vertex. It can be defined as:

$$C_{cc}(u) = \lambda(u)/\gamma(u), \tag{11}$$

where $\lambda(u)$ is the neighbours of u connected by an edge. The formula for $\gamma(u)$ is:

$$\gamma(u) = C_d(u)(C_d(u) - 1)/2. \tag{12}$$

2.3.2 Solvent exposure features

Half-sphere exposure (HSE) has a superior performance with respect to protein stability, conservation among different folds, computational speed and predictability (Hamelryck, 2005). HSE separates a residue's sphere into two half spheres: HSE-up corresponds to the upper sphere in the direction of the chain side of the residue, while HSE-down points to the lower sphere in the direction of the opposite side. HSEpred (Song *et al.*, 2008) is used to compute the features of HSE-up, HSE-down and CN (coordination number). Based on structure, we use hsexpo (Hamelryck, 2005) to calculate the exposure features such as HSEAU (number of C_α atoms in the upper sphere), HSEAD (number of C_α atoms in the lower sphere), HSEBU (the number of C_β atoms in the upper sphere), HSEBD (the number of C_β atoms in the lower half sphere), CN (coordination number), RD (residue depth) and RDa (C_α atom depth).

2.3.3 Features based on protein structure

Structure based features include secondary structure, four-body statistical pseudo-potential, energy scores, solvent accessible area and hydrogen bonds.

1. Secondary structure. We used DSSP (Kabsch and Sander, 1983) to calculate the secondary structure features, including the residue number of first bridge partner, solvent-accessible surface area, peptide backbone torsion angles, C_α atom dihedral and bend angles. Meanwhile, we also used the SPIDER2 (Heffernan *et al.*, 2015) program to predict the secondary structure, solvent accessible surface area and local backbone angles from protein sequences.
2. Four-body statistical pseudo-potential (FBS2P). The four-body statistical pseudo-potential (FBS2P) is based on the Delaunay tessellation of proteins (Liang and Grishin, 2004). Delaunay tessellation is a novel and unique way to define nearest neighbors. The FBS2P is defined as follows:

$$Q^{\alpha}_{ijkl} = \log\left[\frac{f^{\alpha}_{ijkl}}{p^{\alpha}_{ijkl}}\right], \tag{13}$$

where i, j, k and l represent the four amino acids in a protein Delaunay tessellation. The vertices are characterized as the central points of residues. f^{α}_{ijkl} is the observed frequency of the residue form ($ijkl$) in a tetrahedron of type α across a collection of protein structures. p^{α}_{ijkl} denotes the expected random frequency.

1. Energy scores. Seven energy scores, including side-chain energy, residue energy, conservation, interface propensity, two combined scores (combined1 and combined2) and relative solvent accessibility, are calculated with ENDES (Liang *et al.*, 2009).
2. Solvent accessible area (ASA) features. Solvent accessible area (ASA) has been shown to be very useful in identifying hot spot residues from protein-protein complexes (Tuncbag *et al.*, 2009; Xia *et al.*, 2010). Here, Naccess (Hubbard and Thornton, 1993) is used to calculate the absolute and relative solvent accessibilities of the whole residue, entire side chain, main-chain, non-polar side chain and all polar side chain, respectively. Also, we compute the change in the solvent accessible surface area of protein structures in bound and unbound states (ΔASA).
3. Hydrogen bonds. We calculate the number of Hydrogen bonds (Hbond) using HBPLUS (McDonald and Thornton, 1994).

2.3.4 Features based on protein sequence

Based on previous studies, we derived a variety of sequence features.

1. Position-specific scoring matrices (PSSMs). We obtain PSSMs by running PSI-BLAST (Altschul *et al.*, 1997) searching against the NCBI non-redundant database.
2. Local structural entropy (LSE). LSE has been proven to be a useful feature in predicting protein binding hot spots (Chan *et al.*, 2004; Deng *et al.*, 2013). LSE describe the degree of conformational heterogeneity in short protein sequences.
3. Conservation score. The conservation score represents the variability of residues at each position in the protein sequence.

The conservation score calculated based on PSSM and can be defined as follows:

$$Score_i = -\sum_{j=1}^{20} p_{i,j} \log_2 p_{i,j}, \quad (14)$$

where pi, j is the frequency of amino acid j at position i. If a residue has a lower entropy (more conserved), it is given a lower value at a position. In addition, we also use Jensen–Shannon divergence (Capra and Singh, 2007) to calculate the conservation score. Jensen–Shannon divergence is a method of prediction conservation of residues by comparing the distribution of amino acids in multiple sequence alignment (MSA).

1. Physicochemical features. The eight physicochemical features of an amino acid residue are: average accessible surface area, propensities, atom-based hydrophobic moment, hydrophobicity, polarity, polarizability, hydrophilicity and flexibility parameter for no rigid neighbors. The values of the eight physicochemical features for target residues are obtained from the AAindex database (Kawashima and Kanehisa, 2000).
2. Disordered regions. DISOPRED (Jones and Cozzetto, 2015) and DisEMBL (Linding *et al.*, 2003) are used to predict dynamically disordered regions of each residue in the protein sequence.
3. Blocks substitution matrix. We use BLOSUM62 (Henikoff and Henikoff, 1992) to count the relative frequencies of amino acid and their substitution probabilities.
4. Solvent accessible area (ASA) based on protein sequence. Apart from the calculated solvent accessibility by Naccess from protein structure, we also used the NetSurfP (Petersen *et al.*, 2009), SPIDER2 (Heffernan *et al.*, 2015), ACC and SSpro programs to calculate solvent accessibility from protein sequence, where ACC and SSpro programs are from the SCRATCH package (Cheng *et al.*, 2005).

2.4 Features selection

Feature selection can readily remove redundant and irrelevant features that contribute to further improve the performance of a classifier (Wang *et al.*, 2013). Based on the 125 candidate properties, we use the Boruta algorithm (Kursa *et al.*, 2010) to further select an optimal feature subset. The Boruta algorithm is a wrapper-base feature selection method, which built using random forest (RF) (Breiman, 2001) by default. It aims to find all relevant features useful for prediction, rather than finding minimal-optimal features. To enhance the performance, we replace the Boruta algorithm's default RF with the gradient tree boosting algorithm (GTB) (Friedman, 2002). The evaluation criterion R_c represents the prediction performance of the classifier with different ranking features and is defined as follows:

$$R_c = \frac{1}{n}\sum_{i=1}^{n}\{ACC_i + SEN_i + SPE_i + AUC_i\}, \quad (15)$$

where n is the repeat times of 10-fold cross-validation; ACC_i, SEN_i, SPE_i and AUC_i represent the values of the accuracy, sensitivity, specificity and AUC score of the i-th 10-fold cross-validation, respectively. We select the top-k ranked features with the highest R_c score.

2.5 Ensemble vote classifier

The ensemble vote classifier (EVC) algorithm is based on the combination of different machine learning classifiers for predicting the sample label by using the average predicted probabilities. In this study, we find the optimal combination of the seven classifiers by the result of 10-fold cross-validation on the benchmark dataset. The seven classifiers including support vector machines (SVM) (Chang and Lin, 2011), GTB (Friedman, 2002), RF (Breiman, 2001), decision tree (Breiman *et al.*, 1984), extremely randomized trees (ERT) (Geurts *et al.*, 2006), Bernoulli Naive Bayes (Christopher *et al.*, 2008) and AdaBoost (Freund and Schapire, 1995). We found that the combination of ERT, SVM and GTB achieved the best performance with an AUC score of 0.860.

3 Results

3.1 Selection of optimal features

The 125 candidate characteristics can be grouped into four categories: network (residue interaction network), exposure (solvent exposure), sequence and structure features. We combine the structure and sequence features as Combined1 (sequence + structure), and combine network and exposure features as Combined2 (network + exposure). We compare the predictive performance of different feature categories. As listed in Table 1, the network features obtains the best performance among the four basic feature categories, with the highest SEN, MCC and AUC values of 0.728, 0.409 and 0.779, respectively. We also find that the novel feature combination (Combined2) performs much better than the combination of structure and sequence features (Combined1). As we expected, the combination of all the features (network + exposure + sequence + structure) shows the highest performance. The results suggest that the four categories of features may be complementary and their combination is helpful for predicting RNA-binding hot spots.

Selection of valuable information is an essential step for building accurate hot spot classification models. The traditional Boruta algorithm internal classifier uses Random Forest (RF). To assess the utility of the proposed GTB-based Boruta method, we evaluate the performance by incorporating the EVC model with selected properties that correspond to different ranked features. We compare the Boruta methods (RF-based and GTB-based) and other three widely used feature selection methods: RF, maximum relevance minimum redundancy(mRMR) (Peng *et al.*, 2005) and recursive feature elimination(RFE) (Guyon *et al.*, 2002). We evaluate the performance with 10-fold cross-validation on the benchmark dataset. Table 2 shows the prediction performance of GTB-based Boruta algorithm in comparison with other four existing feature selection methods. The results indicate that the GTB-based Boruta method can substantially boost the prediction performance. We also access the performance of top-k ranked features sorted with the GTB-based Boruta algorithm. As shown in Figure 2, the R_c value is highest when using the top 35 ranked features. As a result, we select the top 35 features as the optimal feature set.

Table 1. Performance comparison of different feature combinations

Feature group	ACC	SEN	SPE	PRE	F1	MCC	AUC
Sequence features	0.648	0.681	0.600	0.714	0.697	0.279	0.718
Structure features	0.672	0.704	0.630	0.656	0.688	0.347	0.740
Exposure features	0.675	0.724	0.621	0.656	0.688	0.346	0.733
Network features	0.703	0.728	0.634	0.705	0.723	0.409	0.779
Combined1	0.701	0.719	0.687	0.700	0.679	0.413	0.768
Combined2	0.720	0.739	0.674	0.706	0.717	0.420	0.802
All features	0.728	0.776	0.719	0.757	0.719	0.480	0.829

Table 2. Prediction performance of GTB-based Boruta algorithm in comparison with other four existing feature selection methods

Method	ACC	SEN	SPE	PRE	F1	MCC	AUC
RFE	0.729	0.759	0.722	0.748	0.742	0.478	0.830
RF	0.721	0.743	0.737	0.740	0.726	0.463	0.833
mRMR	0.736	0.778	0.713	0.751	0.756	0.491	0.842
Boruta (RF)	0.742	0.756	0.726	0.791	0.737	0.501	0.848
Boruta (GTB)	0.750	0.784	0.761	0.782	0.754	0.513	0.860

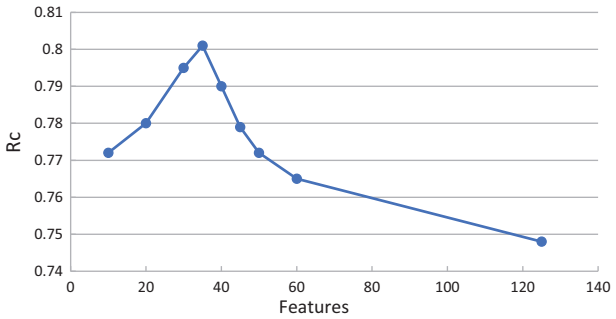

Fig. 2. The R_c values of top-k feature sets obtained by using the GTB-based Boruta algorithm and the EVC model

The relative importance and rankings of the 35 optimal features are displayed in Supplementary Table S1. We found that the residue interaction network (RIN) features and solvent exposure features dominate the top-10 list. Important RIN features include betweenness, eccentricity and closeness. Betweenness ranked second among the 35 features. High betweenness is expected in the case of key residues that act as a bridge in protein structure (del Sol and O'meara, 2005). Eccentricity and closeness are also top-ranked features, where eccentricity represents shortest path distance of the node to the farthest node in the network, and closeness describes the status of a residue located in the entire protein structure where highly central residues have higher closeness values (Amitai *et al.*, 2004). Among the four solvent exposure features, RDa and HSE-up are on the top list. Although some features may not differ significantly between hot spots and other residues, they have a good complementarity and thus collectively contribute to the performance improvement.

We also calculated the numbers of features of each feature type in the candidate full feature set and the selected optimal feature set, respectively, and redrawn the pie chart. As shown in Figure 3, the proportion of novel features (Network + Exposure) has increased significantly in the optimal feature set (from 14% to 29%). All the results suggest that RIN and exposure features are more predictive than traditional sequence and structural features in determining RNA binding hot spot residues.

3.2 Ensemble vote classifier improves predictions

PrabHot uses ensemble vote classifier (EVC) to build the final model with the 35 optimal features. Through the experiment, we find the EVC composed of ERT, SVM and GTB can achieve the best performance. Moreover, we compare EVC with support vector machine (SVM), random forests (RF) and gradient tree boosting (GTB) which are well known to perform relatively well on a variety of tasks. Table 3 shows the prediction performance of EVC and other machine learning methods on the benchmark dataset with 10-fold cross-validation. EVC, GTB, ERT SVM and RF achieve AUC values of 0.860, 0.846, 0.821, 0.804 and 0.771, respectively. Comparing with the other methods, the EVC model can improve the prediction performance in terms of ACC, SEN, F1 and AUC score.

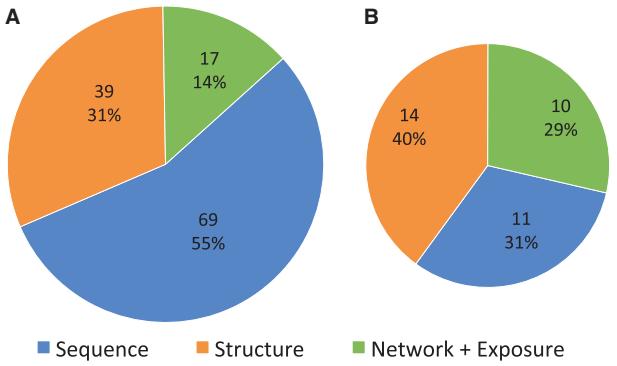

Fig. 3. The proportion of novel features and traditional features in the candidate feature set (**A**) and the optimal feature set (**B**)

Table 3. Prediction performance of EVC classifiers in comparison with four other classifiers on the optimal feature set

Method	ACC	SEN	SPE	PRE	F1	MCC	AUC
RF	0.688	0.702	0.674	0.712	0.697	0.380	0.771
SVM	0.683	0.733	0.679	0.714	0.710	0.410	0.804
ERT	0.735	0.777	0.697	0.741	0.748	0.480	0.821
GTB	0.747	0.763	0.767	0.757	0.743	0.514	0.846
EVC	0.750	0.784	0.761	0.782	0.754	0.513	0.860

3.3 PrabHot outperforms other state-of-the-art approaches

To the best of our knowledge, there only exists one RNA-binding hot spot prediction method HotSPRing (Barik *et al.*, 2016), which was trained on 13 protein–RNA complexes with Random Forest. We calculate P-values with the two-tailed, paired *t*-test (Dietterich, 1998) to compare the performances of our PrabHot method and HotSPRing. We use the 10-fold cross-validation process, and the same random folds are used with each model. We calculate the F1-scores of all the models on the 10-folds and obtain 10 paired F1-score sets. We use MATLAB to compute the P-values. A significance level of 0.05 is used to indicate statistical significance. On the other hand, we repeat the 10-fold cross-validation procedure 50 times to derive a more accurate estimate of model prediction performance. Each time, the 10-fold cross-validation is performed by randomly partitioning into 10 equal-sized subsamples, and the hold-out evaluation is conducted. The cross-validation is repeated 50 times. Then the average performance across all 50 cross-validation trials (PrabHot-50) is computed. Table 4 shows the detailed results. The deviation for PrabHot-50 is calculated based on the average performance across 50 runs, while the deviations for PrabHot and HotSPRing are calculated based on the 10-fold cross-validation. Our PrabHot model shows the best predictive performance (F1 = 0.754, MCC = 0.513 and AUC = 0.86). The F1-score of our PrabHot method is higher than that of HotSPRing (ΔF1 = 0.087), a difference that is statistically significant ($P = 2.16 \times 10^{-3}$). What's more, the average performance across 50 repetitions (PrabHot-50) is also significantly better than that of HotSPRing.

To further evaluate the performance, we compare PrabHot with HotSPRing and four more existing RNA-binding residue prediction

Table 4. Prediction performance of PrabHot in comparison with HotSPRing on the benchmark dataset

Method	SEN	SPE	PRE	F1	MCC	AUC	*P* value
PrabHot	0.784 ± 0.13	0.761 ± 0.12	0.782 ± 0.12	0.754 ± 0.10	0.513 ± 0.11	0.860 ± 0.09	2.16×10^{-3}
PrabHot-50	0.757 ± 0.03	0.748 ± 0.04	0.768 ± 0.03	0.745 ± 0.02	0.502 ± 0.02	0.832 ± 0.02	—
HotSPRing	0.681 ± 0.17	0.552 ± 0.15	0.617 ± 0.16	0.667 ± 0.15	0.280 ± 0.14	0.699 ± 0.13	**

Note: PrabHot-50 represents the average performance across 50 times of 10-fold cross-validation.
**Denotes the reference when calculating the *P*-value.

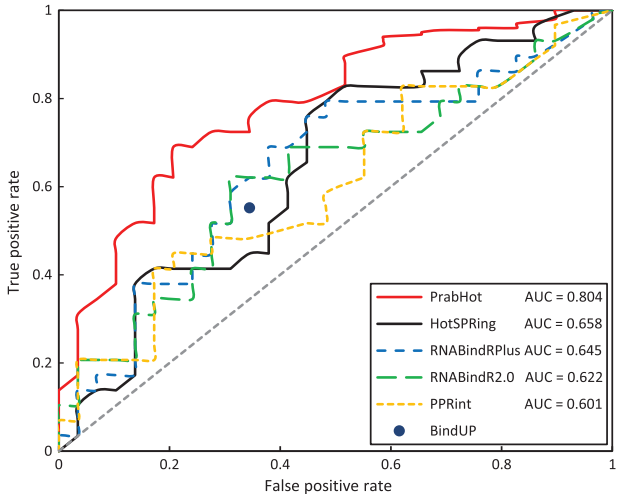

Fig. 4. The ROC curves of different methods on the independent test dataset

methods, including RNABindRPlus (Walia *et al.*, 2014), RNABindR 2.0 (Walia *et al.*, 2012), PPRint (Kumar *et al.*, 2008) and BindUP (Paz *et al.*, 2016), on the independent dataset. The results are presented in Figure 4 and Table 5. Overall, we can see that our PrabHot approach shows a dominant advantage in all the six metrics (SEN, SPE, PRE, F1, MCC and AUC) on the independent test. HotSPRing, which shows the best performance among the existing approaches, correctly predicts RNA-binding hot spots with SEN = 0.655, SPE = 0.552, PRE = 0.604, F1 = 0.633, MCC = 0.258 and AUC = 0.658. When comparing the F1 score with HotSPRing, our PrabHot (0.742) shows greater improvement by 14.6%. Also, the P-value for the difference between PrabHot and HotSPRing is much smaller than 0.05. These observations suggest that our approach has a significant advantage over the pioneer approaches in predicting RNA-binding hot residues. The four RNA-binding site prediction methods have low recognition accuracy in predicting RNA-binding hot spots. A possible reason is that RNA-binding site prediction approaches are designed to discriminate whether a surface residue is an RNA binding site or not, not optimized for predicting RNA binding hot spots.

3.4 Case studies

3.4.1 The MS2 capsid and operator RNA complex

Recombinant MS2 coat protein (PDB ID: 1ZDI, chain A) forms the icosahedral virus shell to protect the viral nucleic acid and acts as a translational repressor by binding with high specificity to a unique site on the RNA (Valegård *et al.*, 1997). Experimentally identified hot spot residues with $\Delta\Delta G \geq 1.0$ kcal/mol in the RNA-binding interface are R49_A, K57_A, K61_A and Y85_A (Supplementary Figure S1A). K43_A and S52_A are found experimentally to be non-hot spots (Hobson and Uhlenbeck, 2006). Prediction results of

Table 5. Prediction performance of PrabHot in comparison with five other prediction tools on the independent dataset

Method	SEN	SPE	PRE	F1	MCC	AUC	*P* value
PrabHot	0.793	0.655	0.697	0.742	0.453	0.804	**
PrabHot-50	0.695	0.690	0.703	0.733	0.389	0.771	—
HotSPRing	0.655	0.552	0.604	0.633	0.258	0.658	1.11×10^{-3}
RNABindRPlus	0.649	0.576	0.611	0.625	0.237	0.645	5.88×10^{-4}
RNABindR 2.0	0.620	0.605	0.622	0.621	0.256	0.622	2.52×10^{-4}
PPRint	0.526	0.538	0.557	0.587	0.134	0.601	8.68×10^{-5}
BindUP	0.552	0.653	0.615	0.582	0.208	—	4.77×10^{-5}

**Denotes the reference when calculating the *P*-value.

PrabHot and HotSPRing are shown in Supplementary Figure S1B and C, respectively. Our method predicts three of the four hot spots correctly: K57_A, K61_A and Y85_A. The three sites have relatively high closeness and betweenness, and low eccentricity compared with other residues in the residue interaction network. Also, the two non-hot spots are also predicted correctly. As a contrast, HotSPRing only correctly predicts a hot spot (K61_A) and a non-hot spot (S52_A).

3.4.2 The TL5 and *Escherichia coli* 5 S RNA complex

Thermus thermophilus TL5 (PDB ID: 1FEU, chain A) belong to the so-called CTC family of bacterial proteins. TL5 binds to the RNA through its N-terminal domain (Fedorov *et al.*, 2001). As shown in Supplementary Figure S2A, four hot spots (R10_A, R19_A, H85_A and D87_E) and three non-hot spots (K14_A, S16_A and R20_A) have experimentally been determined in TL5 (Gongadze *et al.*, 2005). In these seven alanine mutated residues, our PrabHot method correctly identified three residues (R10_A, H85_A and D87_E) as hot spots and the rest as non-hot spots (Supplementary Figure S2B). R19_A and non-hot residues are located at the edges of the binding site and have high relative solvent accessibilities, while the hot spot residues (R10_A, H85_A and D87_E) have more densely connected edges and higher closeness and betweenness in the residue interaction network. Supplementary Figure S2C shows the prediction results of HotSPRing, only a hot spot (R10_A) and two non-hot spots (K14_A and R20_A) are correctly predicted. A possible reason is that HotSPRing uses evolutionary conservation to predict hot spots and R10_A is highly conserved. However, there is no obvious correlation between evolutionary conservation and hot spots for the binding association. Most of the RNA-binding hot spots are not identified by HotSPRing.

4 Conclusion

Accurate prediction of binding energy hot spot residues in protein–RNA complexes is essential for understanding the underlying molecular recognition mechanism. In this work, we described a

computational identification method, the PrabHot method, to predict RNA-binding hot spots. The essential component of this work was the generation of high-quality datasets with a large number of manually curated hot spots. In developing the approach, we were particularly interested in using residue interaction network and new solvent exposure information in conjunction with other types of features. We also utilized the GTB-based Boruta algorithm to select an optimal feature set, which was proved to be able to improve the prediction accuracy and reduce the risk of overfitting. Also, the ensemble vote classifier combines three equally well-performing models to balance out their weaknesses. Experiments results showed that our method significantly outperformed the other state-of-the-art approach on both benchmark and independent test dataset. We believe that PrabHot can be a useful tool for accurately identifying RNA-binding hot spots with the increasing availability of experimentally determined binding free energy changes of mutations.

Funding

This work was supported by National Natural Science Foundation of China [grant number 61672541]; Natural Science Foundation of Hunan Province [grant number. 2017JJ3287]; Natural Science Foundation of Zhejiang [grant number LY13F020038]; and Shanghai Key Laboratory of Intelligent Information Processing [grant number IIPL-2014-002].

Conflict of Interest: none declared.

References

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.

Amitai,G. *et al*. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol*., **344**, 1135.

Barik,A. *et al*. (2016) Probing binding hot spots at protein–RNA recognition sites. *Nucleic Acids Res*., **44**, e9–e9.

Breiman,L. (2001) Random forests. *Mach. Learn*., **45**, 5–32.

Breiman,L. *et al*. (1984) *Classification and Regression Trees*. CRC press, Boca Raton, FL.

Capra,J.A., and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.

Castello,A. *et al*. (2016) Comprehensive identification of RNA-binding proteins by RNA interactome capture. *Methods Mol. Biol*., **1358**, 131.

Chakrabarty,B., and Parekh,N. (2016) Naps: network analysis of protein structures. *Nucleic Acids Res*., **44**, W375–W382.

Chan,C.H. *et al*. (2004) Relationship between local structural entropy and protein thermostabilty. *Proteins*, **57**, 684–691.

Chang,C.C., and Lin,C.J. (2011) Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol*., **2**, 27.

Cheng,J. *et al*. (2005) Scratch: a protein structure and structural feature prediction server. *Nucleic Acids Res*., **33**, W72–W76.

Cho,K.i. *et al*. (2009) A feature-based approach to modeling protein–protein interaction hot spots. *Nucleic Acids Res*., **37**, 2672–2687.

Christopher,D.M. *et al*. (2008) Introduction to information retrieval. *Introd. Inform. Retriev*., **151**, 177.

del Sol,A., and O'meara,P. (2005) Small-world network approach to identify key residues in protein–protein interaction. *Proteins*, **58**, 672–682.

Deng,L. *et al*. (2009) Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics*, **10**, 426.

Deng,L. *et al*. (2013) Boosting prediction performance of protein–protein interaction hot spots by using structural neighborhood properties. *J. Comput. Biol*., **20**, 878–891.

Deng,L. *et al*. (2014) Predhs: a web server for predicting protein–protein interaction hot spots by using structural neighborhood properties. *Nucleic Acids Res*., **42**, W290–W295.

Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*., **10**, 1895–1923.

Fedorov,R. *et al*. (2001) Structure of ribosomal protein tl5 complexed with RNA provides new insights into the ctc family of stress proteins. *Acta Crystallograph. D*, **57**, 968–976.

Fischer,T. *et al*. (2003) The binding interface database (bid): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **19**, 1453–1454.

Freund,Y., and Schapire,R.E. (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, Springer, pp. 23–37.

Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal*., **38**, 367–378.

Geurts,P. *et al*. (2006) Extremely randomized trees. *Mach. Learn*., **63**, 3–42.

Gongadze,G.M. *et al*. (2005) The crucial role of conserved intermolecular h-bonds inaccessible to the solvent in formation and stabilization of the tl5·5 srRNA complex. *J. Biol. Chem*., **280**, 16151–16156.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn*., **46**, 389–422.

Hamelryck,T. (2005) An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *Proteins*, **59**, 38–48.

Heffernan,R. *et al*. (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep*., **5**, 11476.

Henikoff,S., and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci*., **89**, 10915–10919.

Hobson,D., and Uhlenbeck,O.C. (2006) Alanine scanning of ms2 coat protein reveals protein–phosphate contacts involved in thermodynamic hot spots. *J. Mol. Biol*., **356**, 613–624.

Hubbard,S.J., and Thornton,J.M. (1993) Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London, 2 (1).

Jones,D.T., and Cozzetto,D. (2015) Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.

Kabsch,W., and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kawashima,S., and Kanehisa,M. (2000) Aaindex: amino acid index database. *Nucleic Acids Res*., **28**, 374–374.

Kim,O.T.P. *et al*. (2006) Amino acid residue doublet propensity in the protein–RNA interface and its application to RNA interface prediction. *Nucleic Acids Res*., **34**, 6450–6460.

König,J. *et al*. (2012) Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet*., **13**, 77–83.

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

Kursa,M.B. *et al*. (2010) Feature selection with the boruta package. *J. Stat. Softw*., **36**, 1–13.

Li,W., and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li,Y. *et al*. (2011) Predicting disease-associated substitution of a single amino acid by analyzing residue interactions. *BMC Bioinformatics*, **12**, 14.

Liang,S., and Grishin,N.V. (2004) Effective scoring function for protein sequence design. *Proteins*, **54**, 271–281.

Liang,S. *et al*. (2009) Consensus scoring for enriching near-native structures from protein–protein docking decoys. *Proteins*, **75**, 397–403.

Linding,R. *et al*. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Liu,Z.P. *et al*. (2010) Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616.

Loedige,I. *et al*. (2014) The nhl domain of brat is an RNA-binding domain that directly contacts the hunchback mRNA for regulation. *Genes Dev*., **28**, 749–764.

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

Moal,I.H., and Fernández-Recio,J. (2012) Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.

Murakami,Y. *et al.* (2010) Piranha: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res.*, **38**, 412–416.

Pan,Y. *et al.* (2017) Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. *PLoS One*, **12**,

Paz,I. *et al.* (2016) Bindup: a web server for non-homology-based prediction of dna and RNA binding proteins. *Nucleic Acids Res.*, **44**, W568.

Peng,H. *et al.* (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Anal. Mach. Intell.*, **27**, 1226–1238.

Petersen,B. *et al.* (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.

Petukh,M. *et al.* (2015) Predicting binding free energy change caused by point mutations with knowledge-modified mm/pbsa method. *PLoS Comput. Biol.*, **11**, e1004276.

Song,J. *et al.* (2008) Hsepred: predict half-sphere exposure from protein sequences. *Bioinformatics*, **24**, 1489–1497.

Thorn,K.S., and Bogan,A.A. (2001) Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.

Tuncbag,N. *et al.* (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

Valegård,K. *et al.* (1997) The three-dimensional structures of two complexes between recombinant ms2 capsids and RNA operator fragments reveal sequence-specific protein–RNA interactions. *J. Mol. Biol.*, **270**, 724–738.

Walia,R.R. *et al.* (2012) Protein–RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics*, **13**, 89.

Walia,R.R. *et al.* (2014) RNAbindrplus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One*, **9**, e97725.

Wang,L. *et al.* (2010) Bindn+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**, S3.

Wang,L. *et al.* (2012) Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng. Des. Sel.*, **25**, 119–126.

Wang,Y. *et al.* (2013) De novo prediction of RNA-protein interactions from sequence information. *Mol. Biosyst.*, **9**, 133.

Xia,J.F. *et al.* (2010) Apis: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **11**, (1), 174.

Yan,K.S. *et al.* (2003) Structure and conserved RNA binding of the paz domain. *Nature*, **426**, 469–474.

Yang,M. *et al.* (1997) Alanine-scanning mutagenesis of bacillus subtilis trp RNA-binding attenuation protein (trap) reveals residues involved in tryptophan binding and RNA binding. *J. Mol. Biol.*, **270**, 696–710.

Zhang,J. *et al.* (2017a) Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi: 10.1109/TCBB.2017.2701379.

Zhang,Z. *et al.* (2017b) Katzlgo: large-scale prediction of lncRNA functions by using the katz measure based on multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, doi: 10.1109/TCBB.2017.2704587.