Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural bioinformatics

An ensemble approach to predict binding hotspots in protein-RNA interactions based on SMOTE data balancing and random grouping feature selection strategies

Tong Zhou†, Jie Rong†, Yang Liu, Weikang Gong and Chunhua Li*

Falcuty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China

† Tong Zhou and Jie Rong contribute equally to this work.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The identification of binding hotspots in protein-RNA interactions is crucial for understanding their potential recognition mechanisms and drug design. The experimental methods have many limitations, since they are usually time-consuming and labor-intensive. Thus, developing an effective and efficient theoretical method is urgently needed.

Results: Here we present SREPRHot, a method to predict hotspots, defined as the residues whose mutation to alanine generate a binding free energy change $\geqslant$ 2.0 kcal/mol, while others use a cutoff of 1.0 kcal/mol to obtain balanced datasets. To deal with the dataset imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is utilized to generate minority samples to achieve a dataset balance. Additionally, besides conventional features, we use two types of new features, residue interface propensity previously developed by us, and topological features obtained using node-weighted networks, and propose an effective Random Grouping feature selection strategy combined with a two-step method to determine an optimal feature set. Finally, a stacking ensemble classifier is adopted to build our model. The results show SREPRHot achieves a good performance with SEN, MCC and AUC of 0.900, 0.557 and 0.829 on the independent testing dataset. The comparison study indicates SREPRHot shows a promising performance.

Availability and implementation: The source code is available at https://github.com/ChunhuaLiLab/SREPRHot.

Contact: chunhuali@bjut.edu.cn

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

1 Introduction

Protein-RNA interactions play critical roles in a variety of biological processes by regulating different steps of the gene expression process, from transcription to translation (Keene, 2007). The abnormities in the interactions may lead to multiple diseases such as cancer and neurological disorders (Lukong et al., 2008). It is known that a small fraction of interfacial residues, termed as binding hotspots, contribute to the majority of binding free energy for target RNA (Clackson and Wells, 1995). Thus, the reliable hotspot identification in protein-RNA interactions is crucial for understanding the potential recognition mechanism and for designing drugs. Experimentally, a hotspot residue can be found by evaluating the binding free energy change ($\Delta\Delta G$) upon mutating it to alanine (Krüger et al., 2018). However, these methods are costly and time-consuming, and thus developing an effective and efficient computational method is urgently needed to allow for the hotspot identification on a large scale.

Until now, few methods have been developed for predicting hotspots in protein-RNA interactions, which lags behind protein-protein hotspot

prediction, because of the limited available experimental data. In 2016, Barik et al. proposed HotSPRing, a Random Forest model which uses structural and physicochemical features of interfacial residues to predict the ranges of ΔΔG for RNA binding residue mutations (Barik et al., 2016). The method gives a Matthews correlation coefficient (MCC) of 0.258 where a cutoff threshold of ΔΔG = 1.0 kcal/mol is used. In 2018, Pan et al. developed PrabHot, a better performance tool with MCC being 0.389, which utilizes Boruta (Kursa et al., 2010) feature selection algorithm and a voting machine composed of three different classifiers (Pan et al., 2017). Later in 2019, XGBPRH method was introduced by Deng et al., which adopted McTWO algorithm (Ge et al., 2016) to select out six optimal features to train eXtreme Gradient Boosting (XGBoost) classifier (Chen and Guestrin, 2016), and achieves a MCC improvement to 0.661 (Deng et al., 2019). Despite the advances, the computational prediction of RNA-binding hotspots is still in its infancy.

Besides classifiers, the features used for hotspot prediction are also important. The existing methods mainly use some sequence- and structure-based features. In fact, there are still other features we need to explore to improve the protein-RNA binding hotspot prediction. In a previous work, we extracted a residue-nucleotide pairwise propensity potential from protein-RNA interactions, which shows a good performance in protein-RNA interaction prediction (Wang et al., 2021), discrimination of near-native complex structures (Li et al., 2012; Zhang et al., 2017; Lu et al., 2020) and identification of interfacial residues (Liu et al., 2020). The hotspots are a type of special sites at the interface, and therefore we think the propensity potential could be a good feature for identifying hotspots. In addition, as for the residue topological features from amino acid network (AAN) models, they have been successfully used to explore the functional sites, including catalytic, allosteric and ligand binding residues (Yan et al., 2014). Usually, the topological features are obtained from the traditional unweighted AAN model that ignores the residue node heterogeneity which is critical to the discrimination of the structurally or functionally important residues. In view of this, many weighted AAN models have been developed, among which the node-weighted networks developed by Yan et al. are quite able to characterize the node heterogeneity and have been widely applied in the functional residue prediction (Yan et al., 2018). Thus, we think that the use of residue topological features from the node-weighted networks will probably have a positive role for hotspot prediction.

Additionally, for a relatively small sample size, selecting a subset of significant features is important for building an effective predictor. The commonly used feature selection methods include minimum Redundancy Maximum Relevance (mRMR) (Peng et al., 2005), Random Forest (RF) (Breiman, 2001) and Boruta whose performances are not very ideal for a small sample size. Our strategy to solve this problem is that first the samples are divided into several subsets and the feature selection is performed on all the subsets respectively, and then the commonly selected features are retained as the optimal feature set. Our results demonstrate the effectiveness of the strategy and we call it Random Grouping feature selection strategy in the following.

Another point which needs to be mentioned is the class imbalance problem. To overcome it, most of the existing methods choose ΔΔG = 1.0 kcal/mol as the threshold to define the hotspots, and thus the ratio between the numbers of positive and negative samples is close to 1:1. However, Krüger et al. speculate that there are in fact only about 10% hotspots (ΔΔG ≥ 2.0 kcal/mol) in protein-RNA interfaces (Krüger et al., 2018). Such a high class imbalance will seriously affect the performance of classifier models, inducing an overfitting to the majority class samples (Chawla et al., 2002). Usually, the over-sampling and under-sampling techniques are used to preprocess the imbalanced data, among which the Synthetic Minority Over-sampling Technique (SMOTE) is often used in the field of commercial data mining (Chawla et al., 2002). Different from the naive random over-sampling algorithms that generate minority class samples through a simple random replication, the SMOTE method generates the synthetic samples via some operations in the feature space, which avoids the overfitting problem to some extent (Chawla et al., 2002). Several recent studies have successfully utilized SMOTE to effectively improve the predictions of protein-protein interaction sites (Wang et al., 2019) and drug-target interactions (Redkar et al., 2020).

In this work, we propose an effective method called SREPRHot (a SMOTE and Random grouping strategies-based Ensemble learning model for Protein-RNA binding Hotspot prediction) to predict binding hotspots in protein-RNA interactions, where a threshold of ΔΔG = 2.0 kcal/mol is adopted to define a hotspot. The SMOTE algorithm is introduced to balance the data classes. A subset of optimal features is selected out by our proposed Random Grouping feature selection strategy combined with a two-step method from eight types of candidate features extracted from protein sequences and structures, including the residue-nucleotide pairwise propensity potential and residue topological features from node-weighted AAN models. These features are then utilized to train a stacking ensemble classifier (SCE) to build the hotspot predictor. The framework of SREPRHot method is shown in Figure 1.

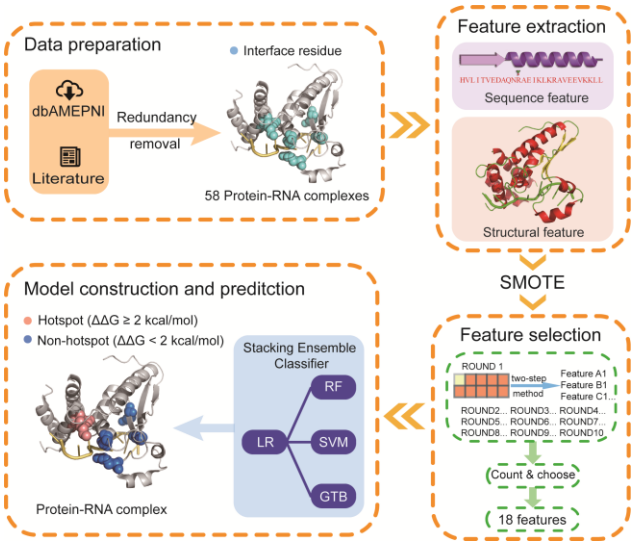

Fig. 1. Framework of SREPRHot for identifying binding hotspots in protein-RNA interactions.

2 Materials and methods

2.1 Training and testing datasets

We collected the residue mutation thermodynamics data from two sources: the dbAMEPNI database (database of Alanine Mutagenic Effects for Protein-Nucleic Acids Interaction, http://zhulab.ahu.edu.cn/dbAMEPNI) (Liu et al., 2018) and the data gathered in developing the hotspot prediction methods HotSPRing, PrabHot and XGBPRH, which were published in 2016, 2017 and 2019, respectively. Thus, we collected 334 residue mutations from 81 complexes in total. To remove the redundancy, the proteins with sequence similarity > 40% were excluded by using CD-HIT (Li and Godzik, 2006). After that, only the interfacial residues were retained whose absolute

solvent accessibility changes ΔASA (calculated by Naccess (Hubbard and Thornton, 1993)) after binding with target RNAs are > 1.0 Å^2 (Zhang et al., 2020). Finally, we obtained 229 residue mutations across 58 complexes. Among them, the 15 complexes used as the testing dataset by PrabHot and XGBPRH method are considered as our independent testing dataset for easy comparison with them, and the remaining ones are used as the training dataset (Table S1).

Different from the existing methods where the interface residues with ΔΔG ≥ 1.0 kcal/mol are considered as hotspots, our method adopts the criterion of ΔΔG ≥ 2.0 kcal/mol. Thus, there are 35 positive and 136 negative samples in the training dataset, and the corresponding numbers are 10 and 48 in the independent testing dataset.

2.2 Feature extraction

A comprehensive set of 120 features from 8 types was extracted (Table S2). More details on the features are described below.

2.2.1 Physicochemical characteristics of amino acids

Ten physicochemical properties of amino acids (Table S3) are taken from the AAIndex database (Kawashima et al., 2008) and the literatures (Li et al., 2008; Jones and Thornton, 1997a; Voet and Voet, 2004; Ramachandran and Antoniou, 2008), including number of atoms, number of electrostatic charges, number of potential hydrogen bonds, hydrophobicity, hydrophilicity, propensity, isoelectric point, mass, expected number of contacts within a 14 Å sphere and electron-ion interaction potential, which are highly correlated with the interface properties of a protein.

2.2.2 Secondary structural features

SPIDER3 (Heffernan et al., 2015) is applied to compute protein secondary structural features including the main chain torsional angles (φ and ψ), the main chain angles between Cα atoms (θ and τ) and the probabilities of three kinds of secondary structures: alpha-helix, beta-strand and random coil.

2.2.3 Depth index (DPX) and protrusion index (CX)

The geometric shape complementarity at the binding interface is important for protein-RNA interactions. DPX and CX were proposed to characterize the embedded and protruding conditions of an atom surrounded by other non-hydrogen atoms, respectively (Pintar et al., 2003; Pintar et al., 2002). We use PSAIA (Mihel et al., 2008) to calculate the indexes for a protein in bound and unbound states including the means of DPXs and CXs of all atoms of a residue and their standard deviations, and the means of DPXs and CXs of side-chain atoms and their standard deviations. In addition, the differences in the means and standard deviations of all atoms and side-chain atoms of a residue between bound and unbound states are also computed.

2.2.4 Solvent accessible surface area (SASA)

The SASAs of a residue in the complex and monomer are calculated by Naccess for a total of 10 attributes: absolute and relative values for all atoms, total side-chain atoms, main-chain atoms, non-polar atoms and all polar atoms in a residue. Moreover, their changes and the corresponding square roots between the two states are also calculated.

2.2.5 Position-specific scoring matrix (PSSM)

The PSSM gives the probability of occurrence of each kind of amino acid residue at each position, which reflects the evolutionary information of a residue position (Liu et al., 2021). For a protein with N residues, the size of its PSSM matrix is N × 20 and each row encapsulates the evolutionary information for a residue position. The PSSM of a protein is calculated by PSI-BLAST (Altschul et al., 1997) searching against NCBI non-redundant protein sequence database.

2.2.6 Solvent exposure features

Half-sphere exposure (HSE) (Hamelryck, 2005), a kind of solvent exposure measures that describes the contacts between residues and solvent molecules, has been proved to be important for protein structure and function predictions (Sharma et al., 2019). HSE is a two-dimensional measure, where a residue's spatial sphere is divided into two half parts: HSE-up (the upper sphere in the direction of the side chain of a residue) and HSE-down (the lower sphere in the opposite direction). HSEpred (Song et al., 2008) is employed to compute the solvent exposure features HSE-up and HSE-down, and in addition the residue contact number (CN) is also calculated.

2.2.7 Residue interface propensity (IP)

Residue interface propensity is from our previously obtained 20×4 residue-nucleotide pairwise propensity potential that was extracted from 251 protein-RNA interactions (Li et al., 2012), and was later updated (used here, Table S4) based on a larger dataset including 694 interactions (Lu et al., 2020). The propensity of one residue-nucleotide pair is obtained from its observed probability divided by its expected probability of occurring on the interfaces. Here, the interface propensity of a residue type is represented as an average of its paired propensities over the four kinds of nucleotides.

2.2.8 Residue topological features from amino acid network (AAN)

Compared with the traditional unweighted AAN, the node-weighted AAN, which considers residue heterogeneity, can better reflect the residue topological properties (Yan et al., 2018). Here besides the unweighted AAN, the four node-weighted AANs based on residue mass, hydrophobicity, polarity and solvent accessibility respectively are constructed and the corresponding residue topological features including degree, betweenness centrality and closeness centrality are calculated by using the R package "NACEN" (Yan et al., 2018).

2.3 SMOTE dataset balancing algorithm

For the data class imbalance problem, the SMOTE is utilized to generate the minority positive samples to achieve the class balance. First the k-nearest neighbors y (here the default k = 5 adopted) of a sample x in the minority class are found, and then new samples are built by the random interpolation operation according to the following equation:

$$x_{new} = x + (y - x) \times \delta \tag{1}$$

where δ is a random number within the interval of (0,1).

2.4 Feature selection

Here we propose a new Random Grouping strategy combined with a two-step algorithm to select the optimal feature subset. First, the training dataset is randomly divided into 10 equal groups and the feature selection is performed 10 rounds with 9 groups of the 10 used for each round. Then, the selected features in each round are recorded and only the features selected not less than 2 times in the 10 rounds are finally retained as the optimal feature set.

For each round, a two-step method is adopted. First mRMR and Decision Tree (DT) methods (Quinlan, 1979) are combined to sort the importance of the candidate features. Then, Sequential Forward Selection (SFS) (Kohavi and John, 1997) combined with Support Vector Machine (SVM) with default parameters is used to determine the optimal feature combination from the top 60 in the importance list through maximizing the Ec score (Pan et al., 2020) via 10-fold cross-validation repeated 5 times. The Ec score is calculated as:

$$E_c = \frac{1}{R}\sum_{j=1}^{R}\left[\frac{1}{n}\sum_{i=1}^{n}\left(ACC_{ij} + SEN_{ij} + SPE_{ij} + MCC_{ij} + AUC_{ij}\right)\right] \quad (2)$$

where n and R (10 and 5 adopted) are the number of cross-validation folds and the times of the n-fold cross-validation, respectively, and ACC, SEN, SPE, MCC and AUC are the values of accuracy, sensitivity, specificity, Matthew's correlation coefficient and AUC score, respectively. Figure S1 shows the flowchart of our feature selection process.

2.5 Stacking ensemble classifier (SEC)

Stacking (Wolpert, 1992), an ensemble learning strategy that combines multiple base classifiers via a meta-classifier, has been proved to perform better than the single classifiers by many researches. Here we apply three boost classifiers Gradient Tree Boosting (GTB) (Friedman, 2002), Random Forest (RF) (Breiman, 2001) and SVM (Cherkassky, 1997), as base classifiers, and Logistic Regression (LR) (Wright, 1995) as the meta-classifier.

2.6 Performance evaluation

SREPRHot is tuned on the training dataset by a 10-fold cross-validation, and tested on the independent testing set. The evaluation indicators including accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), F1 score (F1) and Matthews correlation coefficient (MCC) are used, which are defined as follows:

$$SEN = \frac{TP}{TP + FN} \quad (3)$$

$$SPE = \frac{TN}{TN + FP} \quad (4)$$

$$PRE = \frac{TP}{TP + FP} \quad (5)$$

$$F_1 = \frac{2 \times SEN \times PRE}{SEN + PRE} \quad (6)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are obtained by comparing the predicted label for each residue with the actual one. We also use the area under a Receiver Operating Characteristic (ROC) curve, named AUC, to measure the performance of the model.

3 Results

3.1 Advantage of the SMOTE algorithm

The use of the criterion of hotspots $\Delta\Delta G \geq 2.0$ kcal/mol leads to a high imbalance between the positive and negative samples, which makes the feature selection and model construction largely dominated by negative samples, therefore disadvantageous for model construction. The SMOTE algorithm was adopted to generate the minority (positive) class samples in the training set to balance the data. In order to explore whether balancing the positive and negative samples to the ratio 1:1 can improve the model performance, we compared the results obtained by the models trained on the balanced data by SMOTE and by Random Repeat Oversampling technique (simple copying operations), and on the initial imbalanced ones, as shown in Table 1.

From Table 1, compared with the result from the model trained on the imbalanced data, the corresponding results on the balanced ones have an evident improvement. Furthermore, the model trained on the dataset processed by SMOTE achieves a better prediction than that processed by Random Repeat Oversampling, with SEN, MCC and AUC improved by 26.4%, 12.8% and 8.7% respectively. We argue that the reason for the improvement is that the samples generated by Random Repeat Oversampling technique are only the copies of the original positive ones, not increasing any new information, which may cause a certain degree of overfitting (Chawla et al., 2002).

3.2 Evaluation of different feature selection methods

Our proposed new algorithm, Random Grouping strategy combined with a two-step method (see Materials and Methods), was used to select the optimal feature set. In order to explore the advantages of Random Grouping strategy and the new algorithm, we compared the performances of the four classical methods mRMR, RF, Boruta and SFS (SVM-based) with and without the strategy, and our Random Grouping strategy combined with the two-step method. The results are shown in Table 2.

As shown in Table 2, among the feature selection methods without the Random Grouping strategy, the two-step method reaches the best performance (SEN = 0.768, F1 = 0.591, MCC = 0.525 and AUC = 0.848). Moreover, with the strategy considered, each method's performance has an improvement to some extent, especially in SEN, F1 and MCC scores. Thus, the Random Grouping strategy combined with a two-step algorithm, which we propose to select the optimal features for our model, performs clearly better than the other methods. We speculate the possible reason is that the two-step method considers the complementarity between features and reduces the overfitting (Qiao et al., 2018; Ge et al., 2016), and the Random Grouping strategy reduces the influence of the outlier samples on the feature selection to some extent.

After the dimensional reduction by the Random Grouping strategy combined with the two-step algorithm, we finally obtained the optimal set of 18 features which are shown in Table S5. Among the 18 features, 9 features are sequence-based of four types (physicochemical characteristics of amino acids, PSSM, solvent exposure features and IP) and the other 9 are structure-based of the other three types (DPX and CX, SASA and topological features). It should be pointed out that the residue interface propensity (IP) proposed by us and the two topological features from the node-weighted AAN are selected as the optimal ones, which to our knowledge are used for the first time in protein-RNA hotspot prediction. IP represents the propensity of an amino acid to occur at the interface, while hotspots are a kind of special binding sites, which we think is the possible reason for the helpfulness of IP to the prediction of hotspots at binding interface. As for topological features, some studies have proved that the consideration of the node heterogeneity in network is helpful to the functional residue identification (Yan et al., 2018).

Table 1. Prediction results from models trained on balanced training datasets by SMOTE and Random Repeat Oversampling techniques respectively, and on the initial imbalanced one.

Data	ACC	SEN	SPE	PRE	F1	MCC	AUC
Imbalanced	0.795±0.092	0.492±0.254	0.874±0.076	0.513±0.264	0.490±0.241	0.371±0.294	0.691±0.137
Balanced by Random Repeat Oversampling	0.825±0.078	0.633±0.176	0.874±0.094	0.642±0.215	0.601±0.123	0.515±0.167	0.780±0.096
Balanced by SMOTE	0.833±0.091	0.800±0.227	0.847±0.113	0.602±0.231	0.646±0.171	0.581±0.208	0.848±0.145

Table 2. Prediction results of the models using classical feature selection methods and Random Grouping strategy combined with a two-step method.

Method	ACC	SEN	SPE	PRE	F1	MCC	AUC
mRMR (-)	0.840±0.061	0.625±0.314	0.861±0.102	0.527±0.231	0.554±0.205	0.472±0.201	0.807±0.105
mRMR (+)	0.823±0.089	0.673±0.237	0.853±0.131	0.610±0.225	0.618±0.131	0.523±0.175	0.805±0.136
RF (-)	0.836±0.100	0.580±0.365	0.870±0.110	0.522±0.342	0.534±0.318	0.446±0.340	0.798±0.191
RF (+)	0.857±0.098	0.615±0.242	0.885±0.092	0.584±0.188	0.580±0.180	0.507±0.207	0.830±0.126
Boruta (-)	0.845±0.067	0.683±0.358	0.883±0.084	0.534±0.288	0.586±0.281	0.503±0.296	0.812±0.135
Boruta (+)	0.849±0.099	0.693±0.248	0.889±0.107	0.592±0.237	0.620±0.188	0.541±0.249	0.822±0.132
SFS (-)	0.810±0.086	0.665±0.242	0.837±0.092	0.557±0.188	0.576±0.180	0.482±0.207	0.809±0.128
SFS (+)	0.821±0.081	0.628±0.299	0.868±0.073	0.621±0.198	0.588±0.218	0.501±0.265	0.837±0.204
Two-step (-)	0.831±0.117	0.768±0.292	0.843±0.137	0.522±0.298	0.591±0.279	0.525±0.292	0.848±0.132
Two-step (+)	0.833±0.091	0.800±0.227	0.847±0.113	0.602±0.231	0.646±0.171	0.581±0.208	0.848±0.145

(+) and (-): with and without Random Grouping strategy. mRMR: minimum Redundancy Maximum Relevance. RF: Random Forest. SFS: Sequential Forward Selection.

3.3 Comparison between different machine learning methods

We needed to select an appropriate machine learning method to build our model. To this aim, we compared the performances of six classic classifiers using 10-fold cross-validation on the training dataset, with the results shown in Table S6. Compared with the classifiers k-Nearest Neighbor (kNN) (Cover, 1968), Adaptive Boosting (Adaboost) (Freund and Schapire, 1995) and eXtreme Gradient Boosting (XGBoost), Gradient Tree Boosting (GTB), RF and SVM achieve the best performances in PRE, F1 and MCC scores. In view of this, we adopted the three classifiers GTB, RF and SVM as the first-layer classifiers of our Stacking Ensemble Classifier (SEC), and the Logistic Regression (LR) as the second layer to output the final result, which can reduce the risk of overfitting to some extent. As a result, generally SEC far outperforms the other classifiers with ACC, PRE, F1, MCC and AUC of 0.833, 0.602, 0.646, 0.581 and 0.848, respectively. Thus, the SEC is used as the machine learning classifier of SREPRHot because of its superior performance.

3.4 Performance comparison of SREPRHot with other approaches

We carried out the hotspot prediction using our method SREPRHot on the training and independent testing datasets, respectively, with the results shown in Table S7. To precisely estimate SREPRHot, we repeated 10-fold cross validation on the training dataset 50 times, obtaining ACC, SEN, F1, MCC and AUC values of 0.818±0.016, 0.814±0.036, 0.638±0.022, 0.565±0.023 and 0.859±0.019, respectively. The results indicate the performances of our model are relatively stable and robust.

Table 3. Comparison of SREPRHot with existing methods on independent testing dataset.

Method	SEN	SPE	PRE	F1	MCC	AUC
XGBPRH (1.0 kcal/mol)	0.909	0.733	0.833	0.870	0.661	0.868
PrabHot 1.0 kcal/mol)	0.793	0.655	0.697	0.742	0.453	0.817
HotSPRing (1.0 kcal/mol)	0.655	0.552	0.604	0.633	0.258	0.658
SREPRHot (2.0 kcal/mol)	0.900	0.792	0.474	0.621	0.557	0.829

Additionally, we compared the performance of SREPRHot on the independent testing dataset with the existing methods PrabHot, XGBPRH and HotSPRing, and the results shown in Table 3. It should be pointed out that the former two were developed to predict the hotspots with a threshold of $\Delta\Delta G = 1.0$ kcal/mol, and the latter was proposed to predict the range of $\Delta\Delta G$ for a residue mutation. Deng et al. (Deng et al., 2019), the developer of XGBPRH, in order to compare the performance of XGBPRH with HotSPRing, adopted a threshold of 1.0 kcal/mol to define hotspots for the results from HotSPRing. The results corresponding to HotSPRing in Table 3 are from the literature (Deng et al., 2019) as HotSPRing is currently unavailable. From Table 3, generally XGBPRH achieves the best performance with SEN, MCC and AUC of 0.909, 0.661 and 0.868, respectively. Considering that our method uses a stricter criterion of $\Delta\Delta G \geq 2.0$ kcal/mol, SREPRHot achieves a good perfor-

mance with SEN, MCC and AUC reaching 0.900, 0.557 and 0.829, respectively. The comparison indicates our approach shows a promising performance, and can be a complement to the methods with the threshold of 1.0 kcal/mol used.

3.5 Case Study

As a case study, Figure 2 shows the prediction results by SREPRHot on two protein-RNA complexes. The first is bacteriophage MS2 coat protein-RNA complex (PDB ID: 1ZDI) (Valegârd et al., 1997). Alanine scanning experiment gives three non-hotspots (K43A, R49A and S52A) and three hotspots (K57A, K61A and Y85A) with $\Delta\Delta G \geq 2.0$ kcal/mol. As shown in Figure 2A, SREPRHot identifies four non-hotspots (K43A, R49A, S52A and K61A) among which three are correctly identified. The two identified hotspots (K57A and Y85A) are all correct predictions. For the second case, which is the structure of STAR domain of Quaking protein in complex with target RNA (PDB ID: 4JVH) (Teplova et al., 2013), the experiment gives four non-hotspots (N97A, K120A, R124A and R130A) and two hotspots (K190A and Q193A). Impressively, SREPRHot correctly identifies all the non-hotspots and hotspots, as shown in Figure 2B.

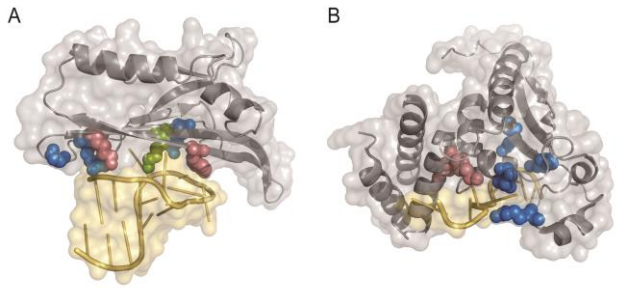

Fig. 2. Prediction results of SREPRHot on 1ZDI (A) and 4JVH (B). The gray surface denotes protein chain while the yellow surface represents RNA chain. True positives are labeled in red, true negatives marked in blue, and false negatives colored in green.

4 Conclusion

The effective prediction of binding hotspots in protein-RNA interactions is essential for understanding their specific recognition and interaction mechanisms. In this paper, a new method SREPRHot is proposed for identifying the binding hotspots, which takes the 18 features of predicted protein residues as input and gives their classification results as output. In order to deal with the data class imbalance problem caused by adopting a stricter criterion of hotspots with $\Delta\Delta G \geq 2.0$ kcal/mol, not 1.0 kcal/mol often used by the existing methods, SMOTE algorithm is utilized to generate the minority (positive) class samples to reach a data class balance. Besides conventional sequence and structural features, the two new feature types, residue interface propensity developed by us and topological features from the node-weighted AAN are extracted as candidate features. From them, our proposed Random Grouping feature selection strategy combined with a two-step method is utilized to pick out an optimal feature set. Finally, a stacking ensemble model is adopted, which combines three well-performing classifiers GTB, RF and SVM via logistic regression to construct the classification method. Compared with the existing methods, SREPRHot achieves a promising performance. We believe that our method is a new beginning in predicting binding hot spots, and in addition the strategies proposed to preprocess the data and select optimal features can also be used as a reference for future prediction works.

One thing that needs to be pointed out is that in SREPRHot performance, the protein interfacial residues need to be known. We can use the currently proposed RNA-binding residue predictors to obtain the information which include aPRBind (Liu et al., 2020), DRNApred (Yan and Kurgan, 2017), NucBind (Su et al., 2019) and NCBRPred (Zhang et al., 2021). In addition, as for features, many tools including BioSeq-Analysis2.0 (Liu et al., 2019), BioSeq-BLM (Li et al., 2021) and DescribePROT (Zhao et al., 2021) have been developed to generate sequence- and structure-based features which can be tried to construct a powerful predictor for hot spot identification in the future.

Funding

This work was supported by the National Natural Science Foundation of China [31971180, 11474013].

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.

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

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

Chawla,N.V. et al. (2002) SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res., 16, 321-357.

Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 785-794.

Cherkassky,V. (1997) The nature of statistical learning theory. IEEE Trans. Neural Netw., 8, 1564.

Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. Science, 267, 383-386.

Cover,T.M. (1968) Rates of convergence for nearest neighbor procedures. In Proceedings of the Hawaii International Conference on Systems Sciences, 413-415.

Deng,L. et al. (2019) XGBPRH: prediction of binding hot spots at protein-RNA interfaces utilizing extreme gradient boosting. Genes, 10, 242.

Freund,Y. and Schapire,R.E. (1995) A desicion-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci., 55, 119-139.

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

Ge,R. et al. (2016) McTwo: A two-step feature selection algorithm based on maximal information coefficient. BMC Bioinformatics, 17, 142.

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.

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

Jones,S. and Thornton,J.M. (1997a) Analysis of protein-protein interaction sites using surface patches. J. Mol. Biol., 272, 121-132.

Kawashima,S. et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res., 36, D202-D205.

Keene,J. (2007) RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet., 8, 533-543.

Kohavi,R. and John,G. (1997) Wrappers for feature subset selection. Artifcial Intelligence, 97, 273-324.

Krüger,D.M. et al. (2018) Protein-RNA interactions: structural characteristics and hotspot amino acids. RNA, 24, 1457-1465.

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

Li,C.H. et al. (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins, 80, 14-24.

Li,H.L. et al. (2021) BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. Nucleic Acids Res., 49, e129.

Li,N. et al. (2008) Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC bioinformatics, 9, 553.

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.

Liu,B. et al. (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res., 47, e127.

Liu,L. et al. (2018) dbAMEPNI: a database of alanine mutagenic effects for protein-nucleic acid interactions. Database, 2018, bay034.

Liu,Y. et al. (2020) aPRBind: protein-RNA interface prediction by combining sequence and I-TASSER model-based structural features learned with convolutional neural networks. Bioinformatics, 37, 937-942.

Liu,Y. et al. (2021) SNB-PSSM: A spatial neighbor-based PSSM used for protein-RNA binding site prediction. J. Mol. Recognit., 34, e2887.

Lu,L. et al. (2020) Preferences of sequence and structure for protein-RNA interfaces and its application in scoring potential construction for docking. Prog. Biochem. Biophys., 47, 634-644.

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

Mihel,J. et al. (2008) PSAIA protein structure and interaction analyzer. BMC Struct. Biol., 8, 21.

Pan,Y. et al. (2017) Computational identification of binding energy hot spots inprotein-RNA complexes using an ensemble approach. Bioinformatics, 34, 1473-1480.

Pan,Y. et al. (2020) Computationally identifying hot spots in protein-DNA binding interfaces using an ensemble approach. BMC Bioinformatics, 21, 384.

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

Pintar,A. et al. (2002) CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics, 7, 980-984.

Pintar,A. et al. (2003) DPX: for the analysis of the protein core. Bioinformatics, 19, 313-314.

Qiao,Y. et al. (2018) Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics, 19, 14.

Quinlan,J.R. (1979) Discovering rules from large collections of examples: a case study. Expert Systems in the Micro-electronic Age, Edinburgh University Press, Edinburgh, UK.

Ramachandran,P. and Antoniou,A. (2008) Identification of hot-spot locations in proteins using digital filters. IEEE Journal of Selected Topics in Signal Processing, 2, 378-389.

Redkar,S. et al. (2020) A Machine Learning Approach for Drug-Target Interaction Prediction using Wrapper Feature Selection and Class Balancing. Mol. Inform., 39, e1900062.

Sharma,R. et al. (2019) Discovering MoRFs by trisecting intrinsically disordered protein sequence into terminals and middle regions. BMC Bioinformatics, 19, 378.

Song,J. et al. (2008) HSEpred: Predict half-sphere exposure from protein sequences. Bioinformatics, 24, 1489-1497.

Su,H. et al. (2019) Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. Bioinformatics, 35, 930-936.

Teplova,M. et al. (2013) Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. Genes Dev., 27, 928-940.

Valegârd,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.

Voet,D. and Voet,J.G. (2004) Biochemistry. J. Wiley & Sons, Hoboken, NJ.

Wang,J. et al. (2021) EDLMFC: an ensemble deep learning framework with multi-scale features combination for ncRNA-protein interaction prediction. BMC Bioinformatics, 22, 133.

Wang,X. et al. (2019) Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. Bioinformatics, 35, 2395-2402.

Wolpert,D.H. (1992) Stacked generalization. Neural Netw, 5, 241-259.

Wright,R.E. (1995) Logistic regression. Reading & Understanding Multivariate Stats., 68, 497-507.

Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res., 45, e84.

Yan,W. et al. (2014) The construction of an amino acid network for understanding protein structure and function. Amino Acids, 46, 1419-1439.

Yan,W. et al. (2018) Node-weighted amino acid network strategy for characterization and identification of protein functional residues. J. Chem. Inf. Model., 58, 2024-2032.

Zhang,J. et al. (2021) NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. Brief Bioinform., 22, bbaa397.

Zhang,S. et al. (2020) A feature-based approach to predict hot spots in protein-DNA binding interfaces. Brief Bioinform, 21, 1038-1046.

Zhang,Z. et al. (2017) A combinatorial scoring function for protein-RNA docking. Proteins, 85, 741-752.

Zhao,B. et al. (2021) DescribePROT: database of amino acid-level protein structure and function predictions. Nucleic Acids Res., 49, D298-D308.