genes

MDPI

Article

XGBPRH: Prediction of Binding Hot Spots at Protein–RNA Interfaces Utilizing Extreme Gradient Boosting

Lei Deng [1], Yuanchao Sui [1] and Jingpu Zhang [2,*]

[1] School of Computer Science and Engineering, Central South University, Changsha 410075, China; leideng@csu.edu.cn (L.D.); suiyuanchao@csu.edu.cn (Y.S.)

[2] School of Computer and Data Science, Henan University of Urban Construction, Pingdingshan 467000, China

* Correspondence: zhangjp@csu.edu.cn

Received: 16 January 2019; Accepted: 15 March 2019; Published: 21 March 2019

Abstract: Hot spot residues at protein–RNA complexes are vitally important for investigating the underlying molecular recognition mechanism. Accurately identifying protein–RNA binding hot spots is critical for drug designing and protein engineering. Although some progress has been made by utilizing various available features and a series of machine learning approaches, these methods are still in the infant stage. In this paper, we present a new computational method named XGBPRH, which is based on an eXtreme Gradient Boosting (XGBoost) algorithm and can effectively predict hot spot residues in protein–RNA interfaces utilizing an optimal set of properties. Firstly, we download 47 protein–RNA complexes and calculate a total of 156 sequence, structure, exposure, and network features. Next, we adopt a two-step feature selection algorithm to extract a combination of 6 optimal features from the combination of these 156 features. Compared with the state-of-the-art approaches, XGBPRH achieves better performances with an area under the ROC curve (AUC) score of 0.817 and an F1-score of 0.802 on the independent test set. Meanwhile, we also apply XGBPRH to two case studies. The results demonstrate that the method can effectively identify novel energy hotspots.

Keywords: hot spots; protein–RNA interfaces; XGBoost; two-step feature selection

1. Introduction

The proteins and nucleic acids constitute the two most important types of biological diversity in living organisms, and they each have their structural characteristics and particular fixed function. Protein–RNA interaction site prediction is of great significance and helps us understand how protein function is achieved so that we can better understand and study the various features of cells [1–5]. Among the protein–RNA interface residues, as is known to all, only a small number of hot spots are essential for the binding free energy. Sufficient identification of these hot spots helps to better understand the molecular mechanisms. Moreover, the interactions of protein with small molecule compounds are the basis for drug design, and structure-based drug design has achieved great success in the development of drugs [6,7]. In recent years, the success rate of the discovery of lead compounds by the molecular docking of compound databases and protein structures has significantly improved. The precise localization of hot spots can elucidate the principle of protein–RNA interactions and provide a very significant theoretical support and a basis for target drug preparation. At present, the research of protein–RNA binding and the critical hot spots in protein–RNA interfaces is an important research direction of bioinformatics and cell biology [8,9].

The characteristics for determining protein–protein binding hot spots have been extensively studied. The research proves that the composition at amino acid in hot spot areas differ from that in

non-hot spot areas. For example, Thorn and Bogan [10] found that hot spots are abundant in Arg, Tyr, and Trp because of their conformation and size. Meanwhile, they proved that hot spots are concerned with energetically less essential interfaces, whose O-ring shape seems to occlude bulk water molecules from the hot spots. Furthermore, analysis has demonstrated that Asp and Asn are more common in hot spots than Glu and Gln because of the differences in side-chain conformational entropy. In recent years, a variety of machine learning algorithms have been used to predict protein–protein interaction hot spots with structural and sequence properties [11–16]. However, these protein–protein interaction hot spot prediction methods and features cannot be directly used to predict protein–RNA binding hot spots. So far, only a few methods have been used to predict protein–RNA interaction hotspots. Barik et al. proposed HotSPRing [17] to identify the hot spots with physico-chemical and structural features in protein–RNA complexes using random forest classifiers. Pan et al. proposed a new method named PrabHot (Prediction of protein–RNA binding hot spots) [18], which used an ensemble of conceptually distinct machine learning algorithms to predict the hot spots.

In this paper, we propose XGBPRH, a powerful computational method to identify hot spots in protein–RNA complexes. First, 156 exposure (solvent exposure), network (residue interaction network) [19], structure [20,21], and sequence features are extracted. To remove irrelevant and redundant information, we use an McTWO feature selection algorithm on the 156 features to select six optimal features. Then, the six optimal features are fed into an eXtreme Gradient Boosting (XGBoost) classifier [22] for predicting protein–RNA binding hot spots. We also evaluate the relative importance of the six optimal features. The results show that exposure and network features are crucial for prediction. Furthermore, we compare XGBPRH with two recent methods, namely HotSPRing and PrabHot, using an independent dataset. The experiments demonstrate that XGBPRH gains the highest values of F_1 and area under the ROC curve (AUC), respectively, which are significantly higher than those of the other two methods. The flowchart of XGBPRH is depicted in the following Figure 1.

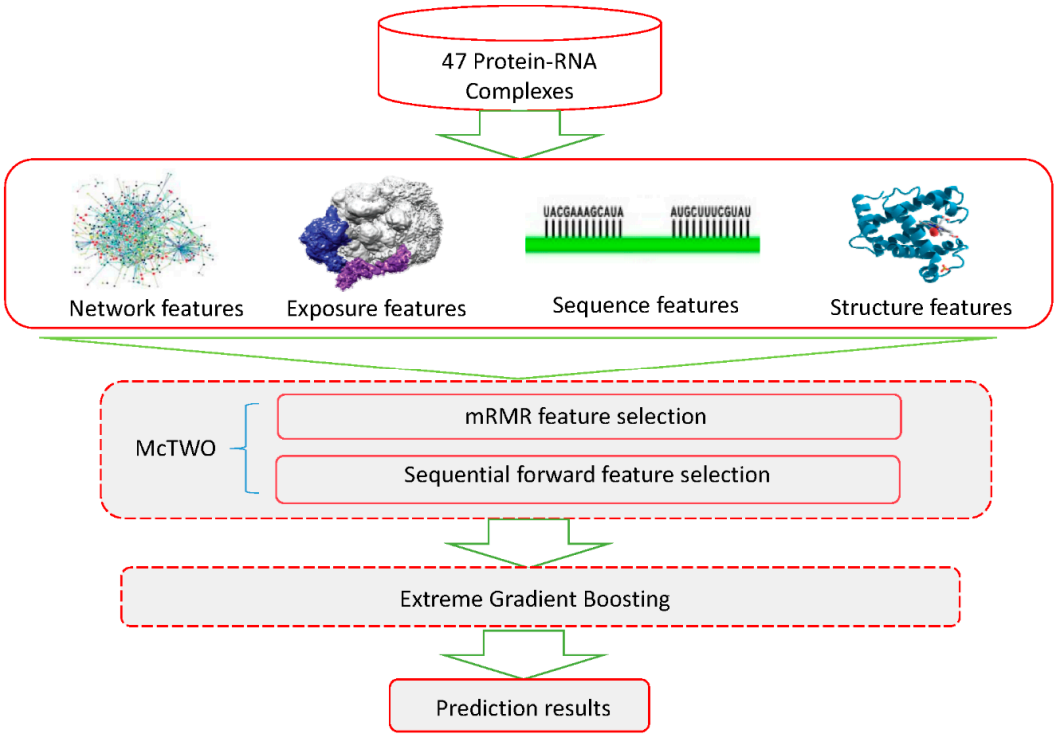

Figure 1. Flowchart of XGBPRH method. The experimental dataset of 47 protein–RNA complexes comes from Pan et al.'s work [18]. We extracted 156 network, exposure, sequence, and structure features. We then adopted the McTWO feature selection algorithm to select the optimal features and used the selected optimal features to train the eXtreme Gradient Boosting (XGBoost) classifier. Finally, we evaluated the performance on the training dataset and independent dataset.

2. Materials and Methods

2.1. Datasets

In this study, the experimental dataset was derived from the works of Barik et al. [17] and Pan et al. [18]. It includes 63 protein–RNA complexes. After removing the redundancy [23] with sequence similarity greater than 40% by using CD-HIT, a dataset of 47 protein–RNA complexes was obtained. Usually, protein–RNA complexes whose corresponding binding free energy change ($\Delta\Delta$ G) $\geq$1.0 kcal/mol are termed as hot spots, and the remaining residues are considered as non-hot spots. Based on this definition, 102 energetically unimportant residues (negative samples) and 107 hot spots (positive samples) were curated from the 47 complexes. Meanwhile, the structural and sequence information of RNAs and proteins in complexes were obtained from the Protein Data Bank (PDB) [24]. The 47 complexes were randomly split into a training benchmark dataset and an independent testing dataset (Table 1). The training dataset has 32 protein–RNA complexes and the independent dataset has 15 complexes. The source code used on this analysis and datasets used are available online at https://github.com/SupermanVip/XGBPRH.

Table 1. The dataset of 47 protein–RNA complexes (Protein DataBank [PDB] codes).

Training dataset	1ASY	1B23	1JBS	1U0B	1URN	1YVP	2BX2	2IX1
	2M8D	2PJP	2Y8W	2ZI0	2ZKO	2ZZN	3EQT	3K5Q
	3L25	3MOJ	3OL6	3VYX	4ERD	4MDX	4NGB	4NKU
	4OOG	4PMW	4QVC	4YVI	5AWH	5DNO	5IP2	5UDZ
Independent testing dataset	1FEU	1WNE	1ZDI	2KXN	2XB2	3AM1	3UZS	3VYY
	4CIO	4GOA	4JVH	4NL3	5EN1	5EV1	5HO4	

2.2. Performance Evaluation

In order to evaluate the performance, we chose the following seven evaluation metrics, which mainly include specificity (SPEC), sensitivity (recall/SENS), F1-score (F1), precision (PRE), the area under the ROC curve (AUC), accuracy (ACC), and the Matthew's correlation coefficient (MCC). These metrics are termed as follows:

$$\mathrm{SPEC} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (1)$$

$$\mathrm{SENS} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (2)$$

$$\mathrm{F1} = 2 \times \mathrm{Recall} \times \mathrm{Precision}/\ (\mathrm{Recall} + \mathrm{Precision}) \quad (3)$$

$$\mathrm{PRE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \quad (4)$$

$$\mathrm{AUC} = \mathrm{P}(\mathrm{P}_{\mathrm{positive}} > \mathrm{P}_{\mathrm{negative}}) \quad (5)$$

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad (6)$$

$$\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})/\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}. \quad (7)$$

2.3. Feature Extraction

Extracting effective features is the key to improving classification performance [25–29]. We initially calculated a combination of 156 features, including exposure features, network features, structural and sequence features. Thirty-one of 156 features were newly curated, and the remaining features were extracted from Pan's work [18]. Details of these features are as follows.

2.3.1. Features Based on Network

According to spatial distance or interaction energy, a residue interaction network (RIN) that captures the inter-residue interactions was obtained. A combination of seven topological features of the

RIN were calculated using the NAPS tool [30]: degree, closeness, eigenvector centrality, betweenness, clustering coefficient, average nearest neighbor degree, and eccentricity.

Degree represents the number of direct neighbors of a node, which is defined as

$$C_d(u) = \sum_{v \in V} A_{uv} \tag{8}$$

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

Closeness is a centrality measure of a node and is termed as the inverse of the shortest path distance of the node to all other nodes in the network.

$$C_{cl}(u) = (n-1) / \sum_{v \in V} dist_{uv} \tag{9}$$

Here, $dist_{uv}$ is the shortest path distance between nodes u and v.

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

$$C_b(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \sigma_{st}(u) / \sigma_{st} \tag{10}$$

where $\sigma_{st}(u)$ is the total number of shortest paths between nodes s and t passing through node u, and σ_{st} is the number of shortest paths between nodes s and t.

The clustering coefficient is defined as the ratio of numbers of connected neighbors of a node to the total number of connections possible between the neighbors. It is a measure of the closeness of the neighbors of a node.

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

where $\lambda(u)$ is the neighbors of u connected by an edge, and $\gamma(u)$ is defined as follows:

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

The eccentricity indicates the distance from the shortest path of the node to the farthest node in the network:

$$C_e(u) = max(dist(u, v)). \tag{13}$$

2.3.2. Features Based on Solvent Exposure

The solvent exposure measures to what extent a residue is accessible to the solvent (usually water) surrounding the protein. It is crucial for understanding the structure and function of the protein. We used a new 2D exposure measure, a half-sphere exposure (HSE) [31], which divides a residue's sphere into two half spheres: HSE-down and HSE-up. We employed HSEpred [32] to calculate the structure information, including HSE-down, CN (coordination number), and HSE-up. Moreover, the exposure features including HSEAD (number of C_a atoms in the lower sphere), HSEAU (number of C_a atoms in the upper sphere), HSEBD (the number of C_β atoms in the lower half sphere), HSEBU (the number of C_β atoms in the upper sphere), RDa (C_a atom depth), and RD (residue depth) were calculated using the hsexpo program [31].

2.3.3. Features Based on 3D Structure

A protein 3D structure refers to a polypeptide chain that is further coiled and folded on the basis of various secondary structures to form a specific spatial structure. Structure-based features have been widely use to predict protein interaction sites [33].

The solubility and stability of proteins are affected by the surface interacting macromolecules in the form of solvents and small solutes in solution. Consequently, the macromolecular surface is

an important factor for researching the structure and function of molecules. We considered the surface curvature and the molecular surface area, and employed Surface Racer [34] to calculate these two characteristics. We also calculated the total solvent accessible surface area, the framework solvent accessible surface area, the total associated solvent accessible surface area, the backbone relative solvent accessible surface area, the average depth index, the maximum depth index, the average protrusion index, the maximum protrusion index, and the hydrophobicity through PSAIA [35].

2.3.4. Features Based on Protein Structure

The protein structure features were also commonly used, and they are as follows:

1. Solvent accessible area (ASA). ASA represents the relatively accessible surface area, which can be calculated using the Naccess [36] program. These ASA features include values of all atoms (ASA_aaa), relative all atoms (ASA_raa), absolute total side (ASA_ats), and relative total side (ASA_rts). We also computed the ΔASA (the change in the solvent accessible surface area of the protein structure between bound and unbound states).
2. Secondary structure. We calculated seven secondary structure features: the residue number of first bridge partner, the solvent accessible surface area, C_α atom dihedral, peptide backbone torsion angles, and bend angles through DSSP [37] and SPIDER2 [38].
3. Four-body statistical pseudo-potential (FBS2P). The FBS2P score, which is based on the Delaunay tessellation of proteins [39], can be written as the following formula.

$$O^{\alpha}_{ijpq} = \log\left(\frac{f^{\alpha}_{ijpq}}{P^{\alpha}_{ijpq}}\right) \tag{14}$$

   where i, j, p, and q are termed as the four amino acids in a Delaunay tetrahedron of the protein. f^{α}_{ijpq} represents the observed frequency of the residue component ($ijpq$) in a tetrahedron of type a over a set of protein structures, and P^{α}_{ijpq} represents the expected random frequency.
4. Energy scores. We used ENDES [40] to calculate seven energy scores: residue energy (Enrich_re), side-chain energy (Enrich_se), conservation (Enrich_conserv), two combined scores (Enrich_com1 and Enrich_com2), relative solvent accessibility (Enrich_rsa), and interface propensity (Enrich_ip).
5. Hydrogen bonds. The hydrogen bonds were calculated using HBPLUS [41].
6. Helix and sheet. The features of α-helix and β-sheet secondary structure are represented with one-hot encoding [42].

2.3.5. Features Based on Protein Sequence

Besides some common features, we selected a few novel features such as backbone flexibility and side-chain environment. These features can be detailed as follows:

1. Backbone flexibility. The protein is flexible and has a range of motion, especially when looking at intrinsically disordered proteins. The feature is calculated by DynaMine [43].
2. Side-chain environment. The side-chain environment (pKa) represents an effective metric in determining the environmental characteristics of a protein. The value of pKa was acquired from Nelson and Cox, indicating a protein side-chain environmental factor, and has been utilized in previous research.
3. Position-specific scoring matrices (PSSMs). The scoring matrices can be calculated by PSI-BLAST [44].
4. Local structural entropy (LSE). LSE [45] is described as the degree of conformational heterogeneity in short protein sequences.
5. Conservation score. We mainly used Jensen–Shannon divergence [46] to calculate the conservation score, which is calculated as follows:

$$\mathrm{Score_i} = -\sum_{\mathrm{j}=1}^{20} \mathrm{P_{ij}} \log_2 \mathrm{P_{ij}} \tag{15}$$

where P_{ij} is termed as the frequency of amino acid j at position i. The conservation score indicates the variability of residues at each position in the sequence. A value that is small at a position means that the residue is conserved.

6. Physicochemical feature. The eight physicochemical features can be obtained from the AAindex database [47]. The eight features are as follows: propensities, average accessible surface area, hydrophobicity, atom-based hydrophobic moment, polarity, polarizability, flexibility parameter for no rigid neighbors, and hydrophilicity.
7. Disordered regions. We used the DISOPRED [48] and DisEMBL [49] to predict each residue's disordered regions in the protein sequence.
8. Solvent accessible area (ASA) calculated through the protein sequence. These features can be calculated by NetSurfP [50], SPIDER2 [51], ACC, and SSPro programs [52]
9. Blocks substitution matrix. The substitution probabilities and their relative frequencies of amino acid can be counted by BLOSUM62 [53].

2.4. Feature Selection

Feature selection is vital for the prediction of hot spots in protein–RNA complexes. Feature selection can help us remove irrelevant and redundant features [54–56]. In this paper, we calculated 156 candidate features in all. To select the optimal feature subset, we adopted a new two-step algorithm named McTWO to perform feature selection [57]. First we utilized minimum redundancy maximum relevance (mRMR) [58] to sort the importance of the features. The redundancy and relevance of mRMR was evaluated by mutual information (MI), which is written as follows:

$$\mathrm{I(m,n)} = \iint p(m,n) log \frac{p(m,n)}{p(m)p(n)} dmdn \tag{16}$$

where m and n represent two random variables, and $p(m)$, $p(n)$, and $p(m,n)$ are the probabilistic density functions. By adopting the mRMR algorithm, we obtained 50 optimal features.

Second, we used the XGBoost algorithm to further select features from the top 50 via 10-fold cross-validation. We chose the first three features at random from the 50 optimal features as the original candidate features. We then adopted the method of sequential forward selection (SFS) to add the remaining ones to the three candidate features one by one based on the R_c score. The R_c score is termed as follows:

$$\mathrm{R}_c = \frac{1}{n} \sum_{i=1}^{n} (ACC_i + SENS_i + SPEC_i + AUC_i) \tag{17}$$

where n represents the repeat times of 10-fold cross-validation.

As shown in Figure 2, we sequentially added each feature to the initial feature set and calculated the R_c scores until the 26 features were put into the sets. The R_c score arrives at 3.08 when the number of features is 6. The overall trend of the R_c declines when the number of features continues to increase. In the end, we consider the top 6 features as optimal.

In order to evaluate the effect of the two-step feature selection algorithm, we compared it with four other extensively adopted feature selection approaches, including Boruta [59], recursive feature elimination (RFE) [60], random forest (RF) [61], and mRMR on the training dataset with 10-fold cross validation. The results are displayed in Table 2. The two-step algorithm achieved the highest value of each metric. It is obvious that the performance of the two-step algorithm is better than that of the other four methods.

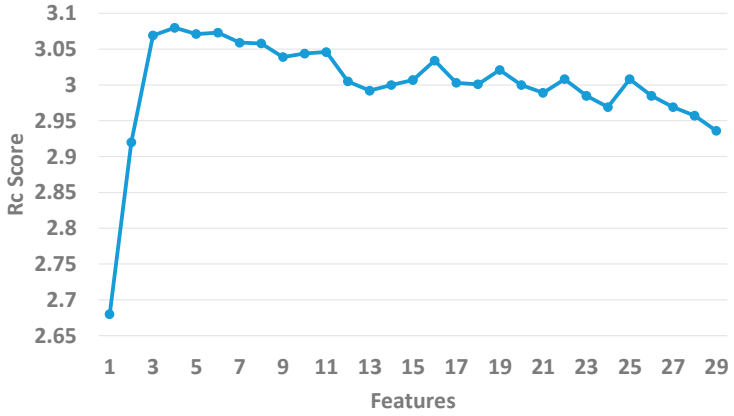

Figure 2. The R_c values of the top 26 features.

Table 2. The performance of the McTWO feature selection algorithm in comparison with four other feature selection algorithms.

Method	ACC	SENS	SPEC	PRE	F1	MCC	AUC
Boruta	0.65	0.603	0.733	0.733	0.634	0.337	0.730
mRMR	0.667	0.661	0.663	0.726	0.662	0.347	0.760
RFE	0.692	0.671	0.702	0.725	0.678	0.366	0.768
RF	0.708	0.698	0.727	0.767	0.711	0.435	0.821
Two-step	0.733	0.732	0.770	0.797	0.743	0.505	0.889

mRMR: Minimum redundancy maximum relevance, RFE: Recursive feature elimination, RF: Random forest, ACC: Accuracy, SENS: Sensitivity, SPEC: Specificity, PRE: Precision, F1: F1-score, MCC: Matthew's correlation coefficient, AUC: Area under the ROC curve.

2.5. Extreme Gradient Boosting Algorithm

The gradient boosting algorithm [62] inherits the advantages of decision trees, and it constructs an ensemble of powerful learners from weak learners. Therefore the extreme gradient boosting algorithm based on the gradient boosting algorithm makes a series of improvements concerning parallelism and predictive accuracy.

In this research, our problem was identifying hot spots and non-hot spots in protein–RNA complexes. This problem can be defined as a binary classification. We used feature vectors F_i (F_i = $\{f_1, f_2, \cdots, f_n\}$, i = 1, 2, $\cdots$, N) as input and used the class label y_i (y_i = $\{-1, +1\}$, i = 1, 2, $\cdots$, N) as the output, where N is the number of rows of the feature vectors, '+1' indicates hot spots, and '-1' represents non-hot spots. The XGBoost algorithm is a combination of classification and regression tree (CART) and a series of the gradient boosting machine [63].

2.6. The XGBPRH Approach

The flowchart of XGBPRH is shown in Figure 1 above. The dataset including 47 protein–RNA complexes was derived from the work of Pan et al as shown in Table 1 above. One hundred fifty-six features were generated from four sources of information: network, exposure, structure, and sequence. Next, we adopted a novel McTWO feature selection algorithm to choose the optimal features. As a result, we obtained a combination of 6 optimal features. Finally, we utilized aXGBoost classifier to predict hot spots and non-hot spots in protein–RNA complexes.

3. Results

3.1. Assessment of Feature Importance

To evaluate the relative importance of the six optimal features, we calculated the average F-score of each feature on the training dataset using XGBoost with 10-fold cross-validation. The results are

summarized in Figure 3 and Table 3 over 50 trials. It is obvious that the RDa (C_a atom depth) feature achieves the highest F-score of 0.693. Closeness and eccentricity follow, with values of 0.679 and 0.675, respectively. This indicates that solvent exposure features and network features are vital for discriminating hot spots and non-hot spots. In our six optimal features, there are two network features (closeness and eccentricity), two exposure features (RDa and HSEBD), and two structure features (Enrich_conserv and ASA_rts).

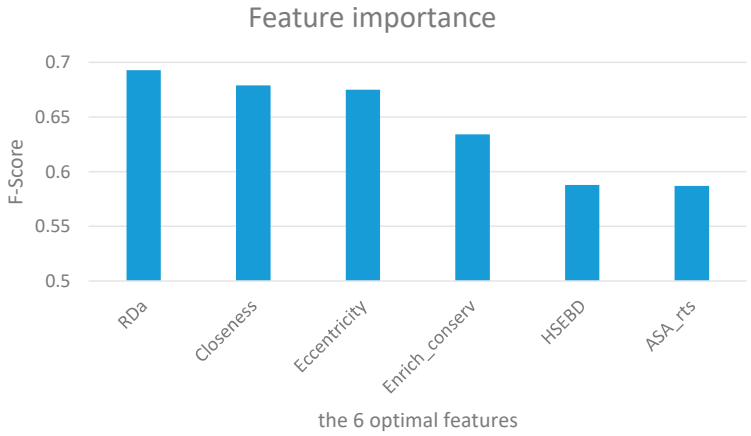

Figure 3. Ranking of feature importance for the six optimal features in terms of F-score.

Table 3. The F-score of the six optimal features using XGBoost with 10-fold cross-validation over 50 trials.

Rank	Feature Name	Symbol	F-Score
1	C_α atom depth	RDa	0.693
2	Closeness	Closeness	0.679
3	Eccentricity	Eccentricity	0.675
4	Enrich conservation	Enrich_conserv	0.634
5	The number of C_α atoms in the lower half sphere	HSEBD	0.588
6	ASA (relative total_side)	ASA_rts	0.587

3.2. Comparison of Different Machine Learing Methods

XGBPRH employs XGBoost as the classifier to determine the hot spots in protein–RNA interfaces with the six optimal features. In order to demonstrate the effectiveness of XGBoost, we used support vector machines (SVMs) [64], random forest (RF), and gradient tree boosting (GTB) to build different models and compared them with XGBPRH. Comparisons were performed with 10-fold cross validation over 50 trials according to the six optimal features. As shown in Table 4, in terms of almost all metrics, XGBoost has the best performance on the training dataset (ACC = 0.744, SENS = 0.740, SPEC = 0.755, precision = 0.785, F1-score = 0.744, MCC = 0.494, AUC = 0.822) except that the score of specificity is lower than that of the RF.

Table 4. Performance comparison of different machine learning methods.

Method	ACC	SENS	SPEC	PRE	F1	MCC	AUC
RF	0.710	0.650	0.781	0.779	0.690	0.430	0.783
SVM	0.741	0.738	0.741	0.775	0.741	0.480	0.802
GTB	0.740	0.728	0.755	0.784	0.739	0.481	0.810
XGBoost	0.744	0.740	0.755	0.785	0.744	0.494	0.822

SVM: support vector machines, GTB: Gradient Tree Boosting.

3.3. Performance Evaluation

As of now, there are two other hot spots prediction methods: PrabHot and HotSPRing. In order to evaluate the performance of our XGBPRH, we compared it with these. We calculated the best results and 50 repetitions' average performance (XGBPRH-50) on the independent test dataset, respectively. As shown in Table 5 and Figure 4, the predictive performance (F1 = 0.870, MCC = 0.661, and AUC = 0.868) significantly outperforms HotSPRing and PrabHot. Moreover, the average performance over 50 trials is superior to that of PrabHot. The results prove that that our method has the best performance in predicting protein–RNA hot spot residues.

Table 5. Prediction performance of XGBPRH in comparison with PrabHot and HotSPRing on the independent dataset.

Method	SENS	SPEC	PRE	F1	MCC	AUC
XGBPRH	0.909	0.733	0.833	0.870	0.661	0.868
XGBPRH-50	0.880	0.537	0.739	0.802	0.454	0.817
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

PrabHot: Prediction of protein–RNA binding hot spots, "-50": 50 repetitions' average performance of the proposed method.

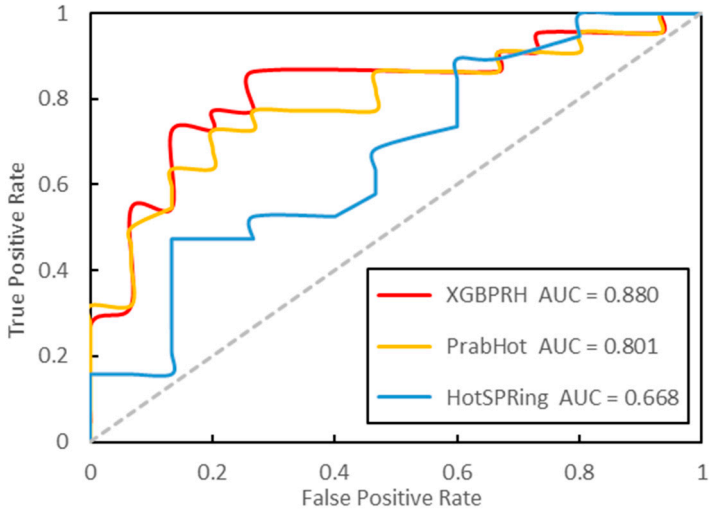

Figure 4. The ROC curves (receiver operating characteristic curve) of the three approaches on the independent test dataset.

In XGBPRH, the computing time depends on the number of residues of the protein in the protein–RNA complex. Large proteins usually require more computation time than that of smaller proteins. We compared the computing time of XGBPRH with that of the PrabHot web server. The results indicate that most predictions can be finished in 5–30 min using XGBPRH. For example, a protein of 490 residues (PDB ID: 1FEU, chain A) required a calculation time of about 25 min, almost the same as PrabHot's calculation time.

3.4. Case Study

3.4.1. Structure of the Star Domain of Quaking Protein in Complex with RNA

The complex (PDB ID: 4JVH, chain A) [65] has six hot spots (K120_A, K190_A, N97_A, Q193_A, R130_A, and R124_A). As shown in Figure 5, we chose a planned combination of colors to show the results: a helix is labeled in red, a sheet labeled in green, and a loop colored in blue. We used purple to

label the true positives. It is obvious to see that our XGBPRH method correctly identified all hot spots (K120_A, K190_A, N97_A, Q193_A, R130_A, and R124_A).

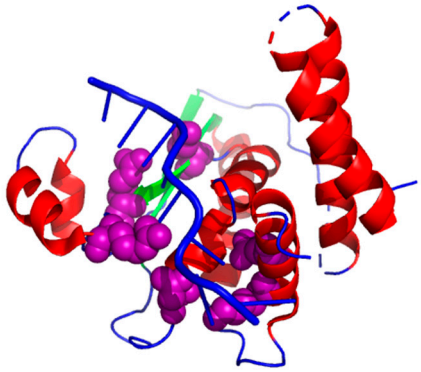

Figure 5. The prediction results on 4JVH using XGBPRH method. True positives colored in purple.

3.4.2. The TL5 and *Escherichia coli* 5S RNA Complex

Thermus thermophilus TL5 (PDB ID: 1FEU, chain A) [66] belongs to the so-called CTC family of bacterial proteins. TL5 [67] binds to the RNA with the help of its N-terminal domain. The complexes have three non-hot spots (K14_A, R20_A, and S16_A) and four hot spots (D87_E, H85_A, R10_A, and R19_A). As shown in Figure 6, our XGBPRH method correctly identified four hot spots (H85_A, R10_A, D87_E, and R19_A) and two non-hot spots (R20_A and K14_A).

Figure 6. The prediction results on 1FEU using XGBPRH. True positives are labeled in purple, true negatives are labeled in yellow, and false negatives are labeled in orange.

4. Discussion

Effective prediction of protein–RNA interaction energy hotspots is of great significance in protein engineering and drug design. In this study, we combined 156 exposure, network, structural, and sequence features. To eliminate the redundant information, we utilized the McTWO feature selection algorithm combined with XGBoost to choose the most useful features, which is the difference between XGBPRH and PraHot. We demonstrated the prediction performance on the independent test dataset. The results show that XGBPRH has superior prediction accuracy. Although our method has achieved good results, there is still room for improvement. First, the protein–RNA interaction hotspot data set is still relatively small, and it is necessary to continue adding experimental data to expand the data set. Semi-supervised learning methods can also be used to improve the prediction performance using a large number of unlabeled data. Secondly, no single feature can fully identify hot spots from the protein–RNA binding interfaces. There is a need to find more effective features or feature combinations to further improve the prediction accuracy.

Author Contributions: L.D., Y.S., and J.Z. conceived this work. L.D. and Y.S. designed and performed the experiments. Y.S. and J.Z. collected the data and analyzed the experiment's results. Y.S. wrote the manuscript, and L.D. and J.Z. revised and approved the paper.

Funding: This project was funded in part by the Natural Science Foundation of Hunan Province [grant number 2017JJ3412 and 2018zzts621] and the National Natural Science Foundation of China [grant number 61672541].

Acknowledgments: Gratitude is expressed to L.D. for his guidance as tutor. We also acknowledge the Experimental Center of School of Software of Central South University for providing Graphics Processing Unit (GPU).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, Z.; Zhao, X.; Chen, L. Identifying responsive functional modules from protein-protein interaction network. *Mol. Cells* **2009**, *27*, 271–277. [CrossRef] [PubMed]
2. Zhang, C.; Sun, B.; Tang, W.; Sun, P.; Ma, Z. Prediction of conformational B-cell epitope binding with individual antibodies using phage display peptides. *Int. J. Clin. Exp. Med.* **2016**, *9*, 2748–2757.
3. Shen, C.; Ding, Y.; Tang, J.; Jiang, L.; Guo, F. LPI-KTASLP: Prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* **2019**, *7*, 13486–13496. [CrossRef]
4. Zou, Q.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene subsequence embedding for prediction of mammalian N6-Methyladenosine sites from mRNA. *RNA* **2019**, *25*, 205–218. [CrossRef] [PubMed]
5. Zhang, J.; Zhang, Z.; Wang, Z.; Liu, Y.; Deng, L. Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics* **2018**, *34*, 1750–1757. [CrossRef] [PubMed]
6. Cho, K.I.; Kim, D.; Lee, D. A feature-based approach to modeling protein–protein interaction hot spots. *Nucleic Acids Res.* **2009**, *37*, 2672–2687. [CrossRef] [PubMed]
7. Chen, L.; Chu, C.; Zhang, Y.H.; Zheng, M.Y.; Zhu, L.C.; Kong, X.Y.; Huang, T. Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* **2017**, *12*, 526–534. [CrossRef]
8. Deng, L.; Guan, J.; Dong, Q.; Zhou, S. Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinform.* **2009**, *10*, 426. [CrossRef]
9. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2017**, *384*, 135–144. [CrossRef]
10. Xia, J.-F.; Zhao, X.-M.; Song, J.; Huang, D.-S. APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinform.* **2010**, *11*, 174. [CrossRef] [PubMed]
11. Deng, L.; Zhang, Q.C.; Chen, Z.; Meng, Y.; Guan, J.; Zhou, S. PredHS: A web server for predicting protein–protein interaction hot spots by using structural neighborhood properties. *Nucleic Acids Res.* **2014**, *42*, W290–W295. [CrossRef] [PubMed]
12. Deng, L.; Guan, J.-H.; Dong, Q.-W.; Zhou, S.-G. SemiHS: an iterative semi-supervised approach for predicting proteinprotein interaction hot spots. *Protein Pept. Lett.* **2011**, *18*, 896–905. [PubMed]
13. Ozdemir, E.S.; Gursoy, A.; Keskin, O. Analysis of single amino acid variations in singlet hot spots of protein–protein interfaces. *Bioinformatics* **2018**, *34*, i795–i801. [CrossRef] [PubMed]
14. Wang, H.; Liu, C.; Deng, L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep.* **2018**, *8*, 14285. [CrossRef] [PubMed]
15. Geng, C.; Vangone, A.; Folkers, G.E.; Xue, L.C.; Bonvin, A.M. iSEE: Interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 110–119. [CrossRef] [PubMed]
16. Moreira, I.S.; Koukos, P.I.; Melo, R.; Almeida, J.G.; Preto, A.J.; Schaarschmidt, J.; Trellet, M.; Gümüş, Z.H.; Costa, J.; Bonvin, A.M. SpotOn: High accuracy identification of protein-protein interface hot-spots. *Sci. Rep.* **2017**, *7*, 8007. [CrossRef] [PubMed]
17. Barik, A.; Nithin, C.; Karampudi, N.B.; Mukherjee, S.; Bahadur, R.P. Probing binding hot spots at protein-RNA recognition sites. *Nucleic Acids Res.* **2015**, *44*, e9. [CrossRef] [PubMed]
18. Pan, Y.; Wang, Z.; Zhan, W.; Deng, L. Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* **2017**, *34*, 1473–1480. [CrossRef] [PubMed]

19. Ding, Y.; Tang, J.; Guo, F. Identification of residue-residue contacts using a novel coevolution- based method. *Curr. Proteom.* **2016**, *13*, 122–129. [CrossRef]
20. Tang, Y.; Liu, D.; Wang, Z.; Wen, T.; Lei, D. A boosting approach for prediction of protein-RNA binding residues. *BMC Bioinform.* **2017**, *18*, 465. [CrossRef] [PubMed]
21. Ding, Y.; Tang, J.; Guo, F. Identification of protein–ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Modeling* **2017**, *57*, 3149–3161. [CrossRef] [PubMed]
22. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Acm sigkdd International Conference on Knowledge Discovery & Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
23. Zou, Q.; Lin, G.; Jiang, X.; Liu, X.; Zeng, X. Sequence clustering in bioinformatics: An empirical study. *Brief. Bioinform.* **2019**. [CrossRef] [PubMed]
24. Rose, P.W.; Beran, B.; Bi, C.; Bluhm, W.F.; Dimitropoulos, D.; Goodsell, D.S.; Prlic, A.; Quesada, M.; Quinn, G.B.; Westbrook, J.D.; et al. The RCSB Protein Data Bank: Redesigned web site and web services. *Nucleic Acids Res.* **2011**, *39*, D392–D401. [CrossRef] [PubMed]
25. Sharma, R.; Raicar, G.; Tsunoda, T.; Patil, A.; Sharma, A. OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics* **2018**, *34*, 1850–1858. [CrossRef] [PubMed]
26. Sharma, R.; Sharma, A.; Patil, A.; Tsunoda, T. Discovering MoRFs by trisecting intrinsically disordered protein sequence into terminals and middle regions. *BMC Bioinform.* **2019**, *19*, 378. [CrossRef] [PubMed]
27. Sharma, R.; Sharma, A.; Raicar, G.; Tsunoda, T.; Patil, A. OPAL+: Length-specific MoRF prediction in intrinsically disordered protein sequences. *Proteomics* **2018**, e1800058. [CrossRef]
28. Zheng, N.; Wang, K.; Zhan, W.; Deng, L. Targeting virus-host protein interactions: Feature extraction and machine learning approaches. *Curr. Drug Metab.* **2018**. [CrossRef]
29. Liu, S.; Liu, C.; Deng, L. Machine learning approaches for protein–protein interaction hot spot prediction: Progress and comparative assessment. *Molecules* **2018**, *23*, 2535. [CrossRef]
30. Chakrabarty, B.; Parekh, N. NAPS: Network analysis of protein structures. *Nucleic Acids Res* **2016**, *44*, W375–W382. [CrossRef]
31. Hamelryck, T. An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins Struct. Funct. Bioinform.* **2005**, *59*, 38–48. [CrossRef]
32. Song, J.; Tan, H.; Takemoto, K.; Akutsu, T. HSEpred: Predict half-sphere exposure from protein sequences. *Bioinformatics* **2008**, *24*, 1489–1497. [CrossRef] [PubMed]
33. Šikić, M.; Tomić, S.; Vlahoviček, K. Prediction of protein–protein interaction sites in sequences and 3D structures by Random Forests. *PLoS Comput. Biol.* **2009**, *5*, e1000278. [CrossRef] [PubMed]
34. Lee, B.; Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400. [CrossRef]
35. Mihel, J.; Šikić, M.; Tomić, S.; Jeren, B.; Vlahoviček, K. PSAIA—Protein structure and interaction analyzer. *BMC Struct. Biol.* **2008**, *8*, 21. [CrossRef]
36. Hubbard, S.J. *NACCESS: Program for Calculating Accessibilities*; Department of Biochemistry and Molecular Biology, University College of London: London, UK, 1992.
37. Kabsch, W.; Sander, C. Dictionary of protein secondary structure. *Biopolymers* **1983**, *22*, 2577–2637. [CrossRef] [PubMed]
38. Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **2015**, *5*, 11476. [CrossRef] [PubMed]
39. Liang, S.; Grishin, N.V. Effective scoring function for protein sequence design. *Proteins* **2010**, *54*, 271–281. [CrossRef]
40. Liang, S.; Meroueh, S.O.; Wang, G.; Qiu, C.; Zhou, Y. Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins* **2009**, *75*, 397–403. [CrossRef]
41. Mcdonald, I.K.; Thornton, J.M.J. Satisfying hydrogen bonding potential in proteins. *Mol. Biol.* **1994**, *238*, 777–793. [CrossRef]
42. Northey, T.; Barešic, A.; Martin, A.C. IntPred: A structure-based predictor of protein-protein interaction sites. *Bioinformatics* **2017**, *34*, 223–229. [CrossRef]

43. Cilia, E.; Pancsa, R.; Tompa, P.; Lenaerts, T.; Vranken, W.F. From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* **2013**, *4*, 2741. [CrossRef] [PubMed]
44. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
45. Chan, C.H.; Liang, H.K.; Hsiao, N.W.; Ko, M.T.; Lyu, P.C.; Hwang, J.K. Relationship between local structural entropy and protein thermostability. *Proteins* **2004**, *57*, 684–691. [CrossRef] [PubMed]
46. Capra, J.A.; Singh, M. *Predicting Functionally Important residues from Sequence Conservation*; Oxford University Press: Oxford, UK, 2007; pp. 1875–1882.
47. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2012**, *36*, D202–D205. [CrossRef] [PubMed]
48. Jones, D.T.; Cozzetto, D. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **2015**, *31*, 857–863. [CrossRef]
49. Linding, R.; Jensen, L.J.; Diella, F.; Bork, P.; Gibson, T.J.; Russell, R.B. Protein disorder prediction: Implications for structural proteomics. *Structure* **2003**, *11*, 1453–1459. [CrossRef]
50. Petersen, B.; Petersen, T.N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009**, *9*, 51. [CrossRef]
51. Yang, Y.; Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Zhou, Y. *SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks*; Springer: New York, NY, USA, 2017; p. 55.
52. Cheng, J.; Randall, A.Z.; Sweredoski, M.J.; Baldi, P. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33*, 72–76. [CrossRef]
53. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef]
54. Yu, L.; Sun, X.; Tian, S.W.; Shi, X.Y.; Yan, Y.L. Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr. Bioinform.* **2018**, *13*, 253–259. [CrossRef]
55. Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: Predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* **2016**, *10*, 114. [CrossRef]
56. Zou, Q.; Zeng, J.; Cao, L.; Ji, R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* **2016**, *173*, 346–354. [CrossRef]
57. Ge, R.; Zhou, M.; Luo, Y.; Meng, Q.; Mai, G.; Ma, D.; Wang, G.; Zhou, F. McTwo: A two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinform.* **2016**, *17*, 142. [CrossRef]
58. Mundra, P.A.; Rajapakse, J.C. SVM-RFE with MRMR filter for gene selection. *IEEE Trans. Nanobiosci.* **2010**, *9*, 31–37. [CrossRef]
59. Kursa, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A System for Feature Selection. *Fundam. Inform.* **2010**, *101*, 271–285.
60. Granitto, P.M.; Furlanello, C.; Biasioli, F.; Gasperi, F.J.C.; Systems, I.L. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 83–90. [CrossRef]
61. Yaqub, M.; Javaid, M.K.; Cooper, C.; Noble, J.A. Improving the Classification Accuracy of the Classic RF Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation. In Proceedings of the International Conference on Machine Learning in Medical Imaging, Toronto, ON, Canada, 18 September 2011; pp. 184–192.
62. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
63. Babajide Mustapha, I.; Saeed, F. Bioactive molecule prediction using extreme gradient boosting. *Molecules* **2016**, *21*, 983. [CrossRef]
64. Guo, H.; Liu, B.; Cai, D.; Lu, T. Predicting protein–protein interaction sites using modified support vector machine. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 393–398. [CrossRef]
65. Teplova, M.; Hafner, M.; Teplov, D.; Essig, K.; Tuschl, T.; Patel, D.J. Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. *Genes Dev.* **2013**, *27*, 928–940. [CrossRef]

66. Fedorov, R.; Meshcheryakov, V.; Gongadze, G.; Fomenkova, N.; Nevskaya, N.; Selmer, M.; Laurberg, M.; Kristensen, O.; Al-Karadaghi, S.; Liljas, A.; et al. Structure of ribosomal protein TL5 complexed with RNA provides new insights into the CTC family of stress proteins. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2001**, *57*, 968–976. [CrossRef]
67. Gongadze, G.M.; Korepanov, A.P.; Stolboushkina, E.A.; Zelinskaya, N.V.; Korobeinikova, A.V.; Ruzanov, M.V.; Eliseev, B.D.; Nikonov, O.S.; Nikonov, S.V.; Garber, M.B. 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.* **2005**, *280*, 16151–16156. [CrossRef] [PubMed]