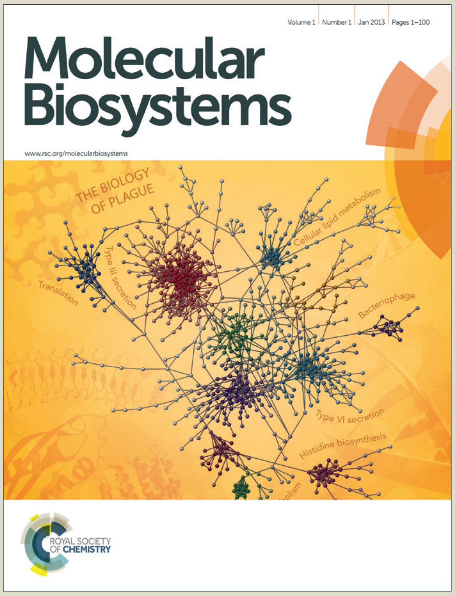

Molecular BioSystems RSCPublishing

ARTICLE

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2014,
Accepted 00th January 2014

DOI: 10.1039/x0xx00000x

www.rsc.org/

A novel feature extraction scheme for prediction of protein-protein interaction sites

Xiuquan Du [a,b,*], Anqi Jing [b], and Xinying Hu [b]

Identifying protein-protein interaction (PPI) sites plays an important and challenge role in some topics of biology. Although many methods have been proposed, this problem is still far away to be solved. Here, a feature selection approach with an 11-sliding window and random forest algorithm is proposed, which is called DX-RF. This method has achieved an accuracy of 88.79%, recall of 82.09%, and precision of 85.76% with top-ranked 34 features on the Hetero test dataset and has got 91.6% accuracy, 89.2% precision, 83.54% recall with top-ranked 25 features set on the Homo test dataset. Compared to other methods, the results indicate DX-RF method has a strong ability to select relevance features to get a higher performance. Moreover, in order to further understand protein interactions, feature analysis in this study are also performed.

1 Introduction

Protein-protein interactions (PPIs) plays a crucial role in many biological processes, including the transduction of signal, pathways of metabolic, regulation of enzymes, translation of gene, and mediation of cell adhesion [1]. Since the PPIs usually form the backbone in these processes, the research about the PPIs has becoming significant. Particularly, identifying protein-protein interaction sites (PPI sites) would bring more insights to the understanding of the structures and functions of proteins, and simplify the identification of drug targets. However, experimental methods of identifying the PPI sites cost high in finance, timely and labour. Therefore, many computational methods have been proposed.

Up to now many methods have been proposed to the prediction of protein-protein interaction sites based on features of protein with the machine learning methods. A large number of properties of protein have been explored in these computational methods to the prediction of interaction sites. These features can be grouped into two classes: sequence-based and structure-based. Ofran et al. analyses that it is possible to predict protein-protein interaction sites from sequence alone, because the amino acid composition of these residues is very different from the rest [2] residues, the most residues in interface are clustered in the consecutive and local sequence segment [3], 70% of prediction in their work was correct. Dhole et al. [4] propose a sequence-based method to predict PPI sites. In their work, they used three attributes of position specific scoring matrix (PSSM), predicted relative solvent accessibility (PRSA) and averaged cumulative hydropath (ACH) with a sliding window. With the help of the L11-logreg classifier, they achieved a Matthews's correlation coefficient (MCC) of 0.175, specificity of 60.3%, recall of 63.8% and accuracy of 60.9%.

What is more, structural information is also one of significant features among interface. Many methods with structural features have achieved successfully in predicting PPI sites [5, 6]. La and Kihara et al. [7] presented a computational method to the prediction of protein binding sites via extracting the MSA of surface patch and computing phylogenetic trees, which achieved an average area under roc curve (AUC) value of 0.624 on training and testing sets. The performance of Wang et al.'s method5 to predict interaction sites is 66.3% sensitivity, 49.7% specificity, 65.4% accuracy and 0.297 MCC.

Some methods with three-dimensional structure information and sequence information have achieved good result [8-10]. Qiu et al. [11] extracted properties based on 3D structure to build a patch-based model and a residue-based model. For the residue-based model, they achieved a specificity rate of 70% and a sensitivity of 0.78; for patch-based model, they achieved a success rate of 0.8. In Li et al.'s work [12], the sequence information, secondary structural information and 3D structural information are extracted to predict protein-protein interaction sites. They achieved accuracy of 0.673 and MCC of 0.348.

[a] Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Anhui, China; E-Mail: dxqllp@163.com

[b] School of Computer Science and Technology, Anhui University, Anhui, China; E-Mails: aqjing0224@gmail.com; lehehehu@163.com;

These authors contributed equally to this work.

They also analysed the features and confirmed 3D structural information which they adopted contributed a lot to the prediction. Minhas et al. [13] presented a method to predict interaction sites with sequence and structure features. They adopted relative accessible surface area, half sphere amino acid composition, protrusion index, PSSM and predicted rASA. In most cases, machine learning methods with features are proposed to predict protein-protein interaction sites. Popular algorithms are Support Vector Machine (SVM) [4, 14, 15], Random Forest (RF) [11, 16-18], neural network (NN) [19] and so on. Chen et al. [14] constructed an integrative profile by developing a support vector machine ensemble. Wang et al. [19] constructed a model using a radial basis function neural network only using evolutionary conservation and spatial sequence profile to predict PPI sites. Their approach has a good performance of the specificity of 67.6%, the sensitivity of 66.6% and accuracy of 68.9%.

Despite many improvements were achieved in recent years, the problem of predicting PPI sites is still far away from being overcome. The performance of classifiers based on machine learning algorithm largely depends on the features which are extracted from protein sequence. So, it is difficult to make a good prediction only based on one characteristic, but it does not mean it must be better if more features were combined together. Therefore, how to select relevance features for predicting protein interaction sites is becoming a problem.

Some researchers also have developed feature selection method to extract the relevance feature subset, for example, Minimum Redundancy Maximal Relevance (mRMR) [12]. However, this feature method need to a large of time to select the best feature subset. For this problem, a novel feature extraction scheme method for predicting PPI sites based on a scoring method (namely DX) is proposed with higher accuracy, which can measure the identifying capability of each feature and is very easy to be computed. According to top-ranked scores, a subset features which are strong identifying capability is selected. Based on this subset, a Random Forest (RF) classifier is built. The experimental results showed that the RF classifier with top-ranked features can get the highest accuracy.

2. Results and Discussion

2.1 Optimization of feature space

In order to evaluate the ability of each feature for predicting PPI sites, first each feature's score is calculated according to Equation (1). 550 features are ranked by the score of each feature from high to low. A feature with a higher DX score implies that it is more important for prediction of PPI sites. The follow procedure is applied to decide the optimal feature set selection: an empty feature set is constructed, and then each feature is added (with score from high to low) into the empty feature set. In order to measure each feature's ability for the prediction, one feature is added every time and the 10-CV experiment using RF algorithm is carried out with the current feature set. In this way, 550 feature sets are constructed and 550 10-CV experiments are performed. In order to eliminate random results, each experiment is performed five times, and then average results are reported. Figure 1 shows the MCC values of 550 10-CV models. The model got the highest MCC when top-ranked 34 features on the Hetero train dataset and top-ranked 25 features on the Homo train dataset, respectively. The cross-validation average performance of this model with top-ranked 34 features, are 89.14% accuracy, 82.5% recall, 86.39% precision. The performance of this model with top-ranked 25 features, are 90.53% accuracy, 82.6% recall, 88.24% precision. Therefore, top-ranked 34 features and top-ranked 25 features are chose as the optimized feature space on the Hetero dataset and Homo dataset, respectively.

2.2 Feature analysis

The 550 features roughly are divided into three types: structure-based features, sequence-based features and physicochemical and biochemical features. In this section, according to the DX score, the top-ranked 34 features and top-ranked 25 features are analyzed. Figure 2 shows the contribution of each feature in the top-ranked 34 features. From Figure 2, we can see that features from HSSP profile, HQI8 and PSAIA contribute mainly to the prediction of PPI sites. Meanwhile some features are not selected in the optimal feature space, such as the RASA, DPX etc. In the optimal features space, there are 14 structure-based features (making up to 41% to the selected), 18 sequence-based features (making up to 53%) and 2 physicochemical and biochemical features (making up to 6%). Figure 3 illustrates that the distribution of the top-ranked 25 features on the Homo dataset. The similar conclusion can be got on the Homo dataset. Figure 4 and Figure 5 illustrates the distribution of features of each residue with the 11-sliding window on the Hetero and Homo dataset, respectively. The two figures all show that the current residue (denotes 0) is more important for predicting protein-protein interaction sites than other residues.

1) Structure-based features analysis

From Figure 6, we can see that 14 structure-based features are selected by the DX with RF algorithm, including 11 features from ACC, 1 total ASA, 1 n-polar ASA and 1 from average CX. Such as RASA, DPX features have not been selected. It shows the accessible surface area is more efficient than other structure-based features on the Hetero dataset. The same results are got on the Homo dataset.

2) Sequence-based features analysis

The distribution of each sequence-based feature is presented in the Figure 7 and Figure 8. From Figure 7, it can be seen that there are 18 sequence features and 5 sequence features contribute to predict protein interaction sites on the Hetero dataset and Homo dataset, respectively. From Figure 8, 5 Entropy, 5 RELENT and 5 sequence variability (VAR) features from Hetero dataset are selected and it contributes more than other sequence-based features. The similar conclusion can be obtained on the Homo dataset. It shows features from HSSP file especially sequence variability and entropy benefit a lot to the prediction but no features from HSSP profile was not selected in the optimal feature space. From Figure 7, we can see that the features number of site -1, site 0 and site 1 are more than other sites.

3) Physicochemical and biochemical features analysis

AAFactors are not selected in the top-ranked 34 features on the Hetero dataset, but 2 features from HQI8 are selected in the optimal feature set. It reflects that the clustering method of HQI is more effective in our prediction method than AAFactors. In this study, the two features are hydrophobicity and intrinsic propensities. The names of selected indices defined in HQI8 are BIOV880101 and MIYS990104.

2.3 Comparison with the original feature space

As we introduced above, there are 550 features for a residue before carrying out the DX method. In order to evaluate the capability of the feature selection method, 10-CV experiment based on RF algorithm with the original feature space is performed on the Hetero train dataset and Homo train dataset and this process is carried out five times. From Table 1, it can be seen that the method with DX (89.14% Acc) got higher 12% than the method without DX (76.62% Acc) on the Hetero train dataset. The method with DX (90.53% Acc) got higher 12% than the method without DX (78.18% Acc) on the Homo train dataset. The whole performance exclude recall is improved. Compared to the original feature space, it confirms that DX-RF method has the ability to select 'more useful' features for PPI sites prediction. Therefore, it can be concluded that DX feature selection is necessary before constructing the model.

Table 1. The performance of classifiers with/without feature selection method

Dataset	Feature space	Acc	Rec	Pre	F	MCC
Hetero	Top-ranked 34 features	89.14	82.5	86.39	84.4	0.7613
	Original (without DX)	76.62	93.99	75.63	83.81	0.4693
Homo	Top-ranked 25 features	90.53	82.6	88.24	85.33	0.7845
	Original (without DX)	78.18	95.1	77.37	85.32	0.4833

2.4 Compared to the other methods

In order to further evaluate the performance of DX-RF classifier, four famous machine learning algorithms, Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB) and Decision Tree (DT) are adopted to perform the experiment on the Hetero/Homo test dataset. In this study, SVM model is produced with the original 550 features and the default parameters and then SVM model is applied to predict Hetero/Homo test dataset. The other algorithms are the same as SVM. Table 2 gives that the detail test results. It can be seen that the performance of DX-RF classifier is higher than other classifiers both Hetero and Homo test dataset. It higher (average ~14% Acc) than other classifiers on the Hetero test dataset. It higher (average ~15% Acc) than other classifiers on the Homo test dataset. As we know, duo to the effect of different random decision value, ROC (Receiver Operating Characteristic) can give a reliable performance comparison. Figure 9 shows that the ROC curves comparison of five methods on the Hetero test dataset. Figure 10 shows that the ROC curves comparison of five methods on the Homo test dataset. From Figure 9, DX-RF can get a 0.9391 AUC area, it higher 0.2338, 0.1191, 0.4391, 0.1211 than NB (0.7053), DT (0.82), SVM (0.5), RF (0.818) on the Hetero test dataset, respectively. From and Figure 10, DX-RF can get a 0.9568 AUC area, it higher 0.2458, 0.1156, 0.4568, 0.1307 than NB (0.711), DT (0.8412), SVM (0.5), RF (0.8261) on the Homo test dataset, respectively.

Table 2. The performance comparison of five methods on the Hetero/Homo test dataset

Dataset	Classifiers	Acc	Rec	Pre	F	CC
Hetero	NB	67.76	50.27	55.1	52.57	0.283
	DT	86	79.61	80.74	80.17	0.6936
	SVM	64.45	0	/	/	/
	RF	76.68	44.64	81.35	57.65	0.4711
	DX-RF	**88.79**	**82.09**	**85.76**	**83.88**	**0.7534**
Homo	NB	68.38	55.76	49.97	52.71	0.2916
	DT	88.12	80.95	81.36	81.15	0.7248
	SVM	68.4	0	/	/	/
	RF	78.9	43.54	80.86	56.6	0.4798
	DX-RF	**91.6**	**83.54**	**89.2**	**86.28**	**0.8032**

2.5 Compared to the other feature selection method

Li et al. [12] propose the Minimum Redundancy Maximal Relevance (mRMR) method to select relevance features for protein-protein interaction sites prediction. They also use IFS method and RF algorithm to evaluate the important of each feature. But in their study, mRMR feature selection method need to much time. In spired by Li et al.'s study, considering time, a novel feature selection method (DX) in this study is proposed. In order to compare Li et al.'s method with DX, the experiment is performed again on the Hetero/Homo train dataset to select the optimal features and then RF algorithm is used to construct the model, namely mRMR-RF. After cross-validation, mRMR-RF select top-ranked 31 features and top-ranked 34 features on the Hetero/Homo train dataset, respectively. Then, mRMR-RF model is constructed based on the optimal features set and is applied to predict Hetero/Homo test dataset. Table 3 shows that the performance on the Hetero/Homo test dataset between mRMR-RF and DX-RF. The AUC performance of DX-RF (0.9391) is higher about 2.85% than mRMR-RF (0.9106) on the Hetero test dataset. The AUC performance of DX-RF (0.9568) is higher about 4.03% than mRMR-RF (0.9165) on the Homo test dataset. ROC curves (Figure 11A) and Recall-Precision curves (Figure 11B) are also plotted to compare these two methods objectively. From Figure 11, DX-RF is higher than mRMR-RF both Hetero and Homo test dataset.

Table 3. The performance comparison between mRMR-RF and DX-RF

Dataset	Classifier	FST (s)	Acc	AUC
Hetero	DX-RF	0.641128	**88.79**	**0.9391**
	mRMR-RF	126013	86.05	0.9106

Homo	DX-RF	0.712766	**91.6**	**0.9568**
	mRMR-RF	142572	87.71	0.9165

FST: Feature selection times

2.6 Four experimental examples

Here we give four examples that are predicted by the SVM, RF, DT, NB, mRMR-RF and DX-RF classifiers. The first example (PDB: 1B4U_A) from Hetero test dataset is the crystal structure of an aromatic ring opening dioxygenase LigAB, a protocatechuate 4,5-dioxygenase, under aerobic conditions [20]. DX-RF can predict 46 residues to be interfaced with 78.85% recall and 80.7% precision (Figure 12A). SVM predicts 0 interface residues (Figure 12B). RF predicts 24 interface residues with 92.3% recall and 85.71% precision (Figure 12C). DT predicts 40 interface residues with 75% recall, 75.47% precision (Figure 12D). NB predicts 26 interface residues with 73.08% recall, 65% precision (Figure 12E). mRMR-RF predicts 42 interface residues with 82.69% recall, 82.35% precision (Figure 12F) while the actual interface residues are 52 (Figure 12G).

The second example (PDB: 2GAC_B) from Hetero test dataset is the crystal structures of Flavobacterium glycosylasparaginase. An N-terminal nucleophile hydrolase activated by intramolecular proteolysis [21]. DX-RF can predict 63 residues to be interfaced with 86.21% recall and 94.03% precision (Figure 13A). SVM predicts 0 interface residues (Figure 13B). RF predicts 42 interface residues with 89.66% recall and 93.33% precision (Figure 13C). DT predicts 60 interface residues with 82.76% recall, 92.31% precision (Figure 13D). NB predicts 46 interface residues with 68.97% recall, 83.64% precision (Figure 13E). mRMR-RF predicts 62 interface residues with 82.76% recall, 92.54% precision (Figure 13F) while the actual interface residues are 66 (Figure 13G).

The third example (PDB: 1QQ5_A) from Homo test dataset is the crystal structures of intermediates in the dehalogenation of haloalkanoates by L-2-haloacid dehalogenase [22]. DX-RF can predict 39 residues to be interfaced with 92.86% recall and 84.78% precision (Figure 14A) and SVM, RF, DT, NB, mRMR-RF predicts 0, 22, 41, 20, 37 interface residues (Figure 14B, Figure 14C, Figure 14D, Figure 14E, Figure 14E, Figure 14F), respectively. The actual interface residues are 45 (Figure 14G).

The fourth example (PDB: 2ONE_A) from Homo test dataset is the crystal structure of asymmetric dimer enolase-2-phospho-D-glycerate/enolase-phosphoenolpyruvate [23]. DX-RF can predict 33 residues to be interfaced with 97.6% recall and 89.2% precision (Figure 15A) and SVM, RF, DT, NB, mRMR-RF predicts 0, 16, 34, 18, 33 interface residues (Figure 15B, Figure 15C, Figure 15D, Figure 15E, Figure 15E, Figure 15F), respectively. The actual interface residues are 41 (Figure 15G).

3. Materials and Methods

3.1 Datasets

This study is divided into 3 phases: the feature extraction phase, training phase, and testing phase. We select two sets (hetero complex proteins and homo complex proteins) of non-redundant chains are derived by Koike Asako et al. [24]. For these datasets, first, all sequence pairs were removed by BLAST with 25% similarity cut-off and length of sequence > 100 amino acid residue. Thus, 324 protein sequences of hetero complexes and 674 protein sequences of homo complexes were obtained. Second, according to the definition of Koike Asako et al., hetero complexes were deleted with lower than 20 interfacial residues (IR) and homo complexes were deleted with lower than 30 interfacial residues. Third, hetero complexes and homo complexes were deleted with no HSSP profile. Therefore, 270 hetero complexes and 289 homo complexes were comprised of our dataset. We randomly select 202 chains of all chains as Homo train dataset and 87 chains as Homo test dataset from Homo dataset. We randomly also select 189 chains of all chains as Hetero train dataset and 81 chains as Hetero test dataset from Hetero dataset. Supplement Materials S1 gives the all datasets.

3.2 Definition of protein interaction sites

In order to construct the dataset, the surface residue and interface residue need to be defined. There are some different definitions and here we used the Fariselli's method [25]. The ASA (Accessible Surface Area) of each residue can be computed by the DSSP program. If a residue's RASA (Relative Accessible Surface Area) is at least of its (Maximal Accessible Surface Area), it is defined to be a surface site. If a surface site's ASA-CASA (complex accessible surface area) > 1 Å^2, it is defined to be interface residues, otherwise it is defined to be a non-interface residues.

3.3 Features extraction

1) Structure-based features

a) Accessible Surface Area: The accessible surface area (ASA) is the atomic surface area exposed to a solvent. The ASA value (ACC) of each residue was got from HSSP [26] in our work. In addition, Protein Structure and Interaction Analyzer (PSAIA) [27] also is adopted to calculate the ASA value for each residue, including backbone ASA, side-chain ASA, polar ASA and non-polar ASA.

b) Relative Accessible Surface Area: Relative accessible surface area (RASA) was calculated by PSAIA [27]. The following residue attributes are calculated by PSAIA: total RASA, backbone RASA, side-chain RASA, polar RASA and non-polar RASA.

c) Depth index: The residue depth is defined as the minimum distance of a residue from any solvent accessible residue and it has been computed by PSAIA. For residue depth, there are six features were calculated by PSAIA. In this paper, average depth index (DPX) is chose.

d) Protrusion index: The protrusion of a non-hydrogen residue is the ration of the volume of a sphere with a radius of 10.0 A cantered at that atom that is not filled with atoms. Same with the DPX, PSAIA calculates six features for the protrusion and average protrusion index (CX) as a feature for predicting protein-protein interaction sites.

2) Sequence-based features

The sequence profile in HSSP file for each protein chain is composed of L rows and 20 columns. ‘L’ stands for the number of amino acids in a chain and 20 kinds of amino acids index columns. $P_{i,j}$ means the probability of j-th amino acid take the place of i-th residue. We also extracted the other four properties of protein from HSSP [26] file: Entropy, Relative Entropy (RELENT), Conservation Weight (WEIGHT) and Sequence Variability (VAR). Entropy measures the conservation of a residue in the location. Relative Entropy is defined as the standardized entropy which normalized to the scale of 0 to 100. Conservation Weight measures the sequence conservation of a position. Sequence variability contains evolutionary information, on a scale of 0-100 as exported from NAGLIN alignments.

3) Physicochemical and biochemical features

a) High-quality-indices: Since Saha et al. [28] have made a conclusion that physic-chemical features of amino acids play a significant role in identifying the PPI sites, thus properties of amino acids are taken into count as important characteristics in discriminating between interacting sites and non-interacting sites. Recently, 544 physicochemical and biochemical properties of amino acids are released in AAIndex1 database. Based on the statistical analyses, 544 characteristics are divided into eight classes, namely high-quality-indices (HQIs). In this work, HQI8 is used as features, including eight clusters which are composed of electric properties, hydrophobicity, alpha and turn propensities, physicochemical properties, residue propensity, composition, beta propensity and intrinsic propensities. Each cluster is composed of one value and there are 8 values for each amino acid.

b) Amino acid factors (AAFactors): Based on AAindex1, Atchley et al. [29] made statistical analyses on these 544 properties, as well. Different form HQI, they summarized these properties into five patterns, which reflect polarity, secondary structure, molecular volume, codon diversity and electrostatic charge.

Finally, for each residue, 50 features are extracted including 25 features from sequence information (20 from HSSP profile, 1 entropy, 1 relative entropy, 1 conservation weight, 1 sequence variability and 1 ACC), 12 features from structure information (5 features from ASA, 5 features from RASA, 1 feature from DPX, 1 feature from CX), 13 features from Physicochemical and biochemical information (8 features from HQI8 and 5 features from Amino acid factors) and 4 features from HSSP file (entropy, relative entropy, conservation weight and sequence variability). In addition, an 11-size sliding window is chose for each residue. Therefore, 550 features are extracted for each residue.

3.4 Feature selection (DX-score)

It is hard to decide what features we should choose as inputs for learning models because combing more multiple features is not always effective. Some features are not effective for discriminating interface residues from non-interface residues. The purpose of feature selection is to get rid of these lower-capability features, which usually are redundant and irrelevant, and improve the capability of learning models. In our work, we used the DX score to select features. By giving each feature a score, it ranks the importance of the features. The definition of DX is shown below:

$$DX = \frac{(\mathrm{m}_1 - \mathrm{m}_0)^2}{d_1^{\,2} + d_0^{\,2}} \qquad (1)$$

m_1 and m_2 stand for the mean value of the feature in positive training dataset and negative training dataset, respectively. d_1 and d_2 stand for the standard deviation value of the feature in positive training dataset and negative training dataset, respectively. Higher score a feature scored means the feature has the potential capacity to distinguish the PPI sites.

3.5 Classification using random forest with 10-CV method

Random Forest (RF) [30] is an ensemble machine learning method which is typically made up of several individual classification trees. A small random subset of features is selected as inputs for each classification tree. The final prediction of forest is decided by the votes of the predictions of all classification trees. In this work, we use the Random Forest algorithm in Waikato Environment for Knowledge Analysis (WEKA) [31]. In this work, machine learning models were evaluated by a statistical technique using 10-CV methods in following ways: the dataset is partitioned into 10 subsets with equal samples. Each subset is composed of equal number of positive and negative samples. When constructing models, 9 subsets are combined for training a model. The remaining subset is tested on the model. In the 10-CV method, this procedure is repeated 10 times to make sure each subset play the role of the test.

3.6 Measuring methods.

The accuracy (ACC), Recall (Rec), Precision (Pre), F-Measure (F) and Matthews’s correlation coefficient (MCC) are used to evaluate the prediction capability of models

$$Acc = (TP + TN)/(\mathrm{TP+TN+FP+FN}) \qquad (2)$$

$$Rec = TN/(TN + FP) \qquad (3)$$

$$\mathrm{Pr}e = TP/(TP + FP) \qquad (4)$$

$$F = (2 * rec * pre)/(rec + pre) \qquad (5)$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \qquad (6)$$

where TP stands for the true positive residues; TN stands for true negative residues; FP stands for false positive residues; FN stands for false negative residues.

4. Conclusions

In this study, a novel method (DX-RF) is presented to predict PPI sites. Based on the DX feature selection method, an optimized set of features was selected from the original feature set. 10-CV method combined with random forest was employed to construct the top-ranked classifier, which achieved the best performance on the two datasets. In order to look which features are considered as the significant ones, a feature analysis is performed according to their type. In order to confirm the DX algorithm's ability to select the significant features, several methods are used to compare the performance with DX-RF. The results confirmed the DX method have the high ability to select the significant features. Finally, four predicted examples are used to further illustrate the result.

Acknowledgments

This work was partially supported by National Science Foundation grant, (No. 61203290), the start-up Foundation for Doctors of Anhui University (No.33190078) and outstanding young backbone teachers training (No.02303301).

References

1. Skrabanek, L.; Saini, H. K.; Bader, G. D.; Enright, A. J., Molecular biotechnology 2008, 38, 1-17.
2. Ofran, Y.; Rost, B., Journal of molecular biology, 2003, 325, 377-387.
3. Ofran, Y.; Rost, B., FEBS letters, 2003, 544, 236-239.
4. Dhole, K.; Singh, G.; Pai, P. P.; Mondal, S., Journal of Theoretical Biology, 2014.
5. Wang, B.; Chen, P.; Huang, D.-S.; Li, J.-j.; Lok, T.-M.; Lyu, M. R., FEBS letters, 2006, 580, 380-384.
6. Yu, D.; Hu, J.; Yang, J.; Shen, H.; Tang, J., 2013.
7. La, D.; Kihara, D., Proteins: Structure, Function, and Bioinformatics, 2012, 80, 126-141.
8. Konc, J.; Janežič, D., Bioinformatics, 2010, 26, 1160-1168.
9. González, A. J.; Liao, L., BMC bioinformatics, 2010, 11, 537.
10. Chen, C.-T.; Peng, H.-P.; Jian, J.-W.; Tsai, K.-C.; Chang, J.-Y.; Yang, E.-W.; Chen, J.-B.; Ho, S.-Y.; Hsu, W.-L.; Yang, A.-S., PloS one, 2012, 7, e37706.
11. Qiu, Z.; Wang, X., Journal of theoretical biology, 2012, 293, 143-150.
12. Li, B.-Q.; Feng, K.-Y.; Chen, L.; Huang, T.; Cai, Y.-D., PloS one, 2012, 7, e43927.
13. Minhas, A.; ul Amir, F.; Geiss, B. J.; Ben‐Hur, A., Proteins: Structure, Function, and Bioinformatics, 2013.
14. Chen, P.; Li, J., BMC bioinformatics, 2010, 11, 402.
15. Bartoli, L.; L Martelli, P.; Rossi, I.; Fariselli, P.; Casadio, R., Current Protein and Peptide Science, 2010, 11, 601-608.
16. Li, B.-Q.; Hu, L.-L.; Niu, S.; Cai, Y.-D.; Chou, K.-C., Journal of Proteomics, 2012, 75, 1654-1665.
17. Šikić, M.; Tomić, S.; Vlahoviček, K., PLoS computational biology, 2009, 5.
18. Chen, X.-w.; Jeong, J. C., Bioinformatics, 2009, 25, 585-591.
19. Wang, B.; Chen, P.; Wang, P.; Zhao, G.; Zhang, X., Protein and peptide letters, 2010, 17, 1111-1116.
20. Sugimoto, K.; Senda, T.; Aoshima, H.; Masai, E.; Fukuda, M.; Mitsui, Y., Structure,1999, 7, 953-965.
21. Guo, H.-C.; Xu, Q.; Buckley, D.; Guan, C., Journal of Biological Chemistry,1998, 273, 20205-20212.
22. Ridder, I. S.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W., Journal of Biological Chemistry, 1999, 274, 30672-30678.
23. Zhang, E.; Brewer, J. M.; Minor, W.; Carreira, L. A.; Lebioda, L., Biochemistry, 1997, 36, 12526-12534.
24. Koike, A.; Takagi, T., Protein Engineering Design and Selection 2004, 17, 165-173.
25. Fariselli, P.; Pazos, F.; Valencia, A.; Casadio, R., European Journal of Biochemistry, 2002, 269, 1356-1361.
26. Sander, C.; Schneider, R., Proteins: Structure, Function, and Bioinformatics,1991, 9, 56-68.
27. Mihel, J.; Šikić, M.; Tomić, S.; Jeren, B.; Vlahoviček, K., BMC structural biology, 2008, 8, 21.
28. Saha, I.; Maulik, U.; Bandyopadhyay, S.; Plewczynski, D., Amino acids, 2012, 43, 583-594.
29. Atchley, W. R.; Zhao, J.; Fernandes, A. D.; Drüke, T., Proceedings of the National Academy of Sciences of the United States of America, 2005, 102, 6395-6400.
30. Breiman, L., Machine learning, 2001, 45, 5-32.
31. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H., ACM SIGKDD explorations newsletter, 2009, 11, 10-18.

Figures legend

Figure 1 shows that the correlation coefficient of 550 classifiers on the Hetero train dataset and Homo train dataset. Each point is stand for the correlation coefficient (MCC) of each classifier. For example, the point at the (100, 0.679) means when top-ranked 100 features are selected to construct the classifier, the MCC is 0.679.

Figure 2 shows that the distribution of each feature in the top-ranked 34 features on the Hetero dataset.

Figure 3 shows that the distribution of each feature in the top-ranked 25 features on the Homo dataset.

Figure 4 shows the distribution of feature of each residue in the 11-sliding window on the Hetero dataset. For example, -5 denotes the left fifth residue of current residue (denotes 0).

Figure 5 shows that the distribution of feature of each residue in the 11-sliding window on the Homo dataset. For example, -5 denotes the left fifth residue of current residue (denotes 0).

Figure 6 shows that the distribution of structure-based feature with top-ranked 25 features on the Hetero dataset.

Figure 7 shows that the distribution of sequence-based features in the 11-sliding window with top-ranked features.

Figure 8 shows that the number of sequence-based features in the optimal feature space.

Figure 9 shows that the ROC curves comparison of different methods on the Hetero test dataset (AUC LibSVM:0.5;DT:0.82;NB:0.7053;RF:0.818;DXRF:0.9391)

Figure 10 shows that the ROC curves comparison of different methods on the Homo test dataset (AUC LibSVM:0.5;DT:0.8412;NB:0.711;RF:0.8261;DXRF:0.9568)

Figure 11 shows that the ROC curves comparison between DX-RF and mRMR-RF.

Figure 12 shows that predicted the interaction sites on protein (PDB:1B4U_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

Figure 13 shows that predicted the interaction sites on protein (PDB:2GAC_B) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

Figure 14 shows that predicted the interaction sites on protein (PDB:1QQ5_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

Figure 15 shows that predicted the interaction sites on protein (PDB:2ONE_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

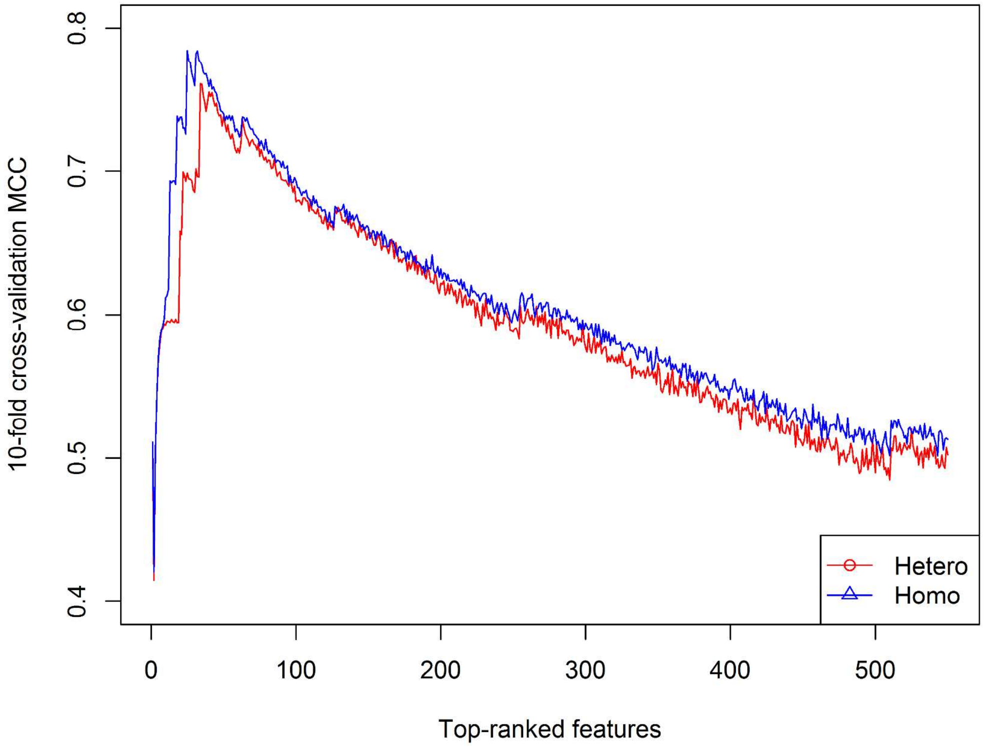

Figure 1. The correlation coefficient of 550 classifiers.

Red indicates the MCC performance on the Hetero train dataset; Blue indicates the MCC performance on the Homo train dataset

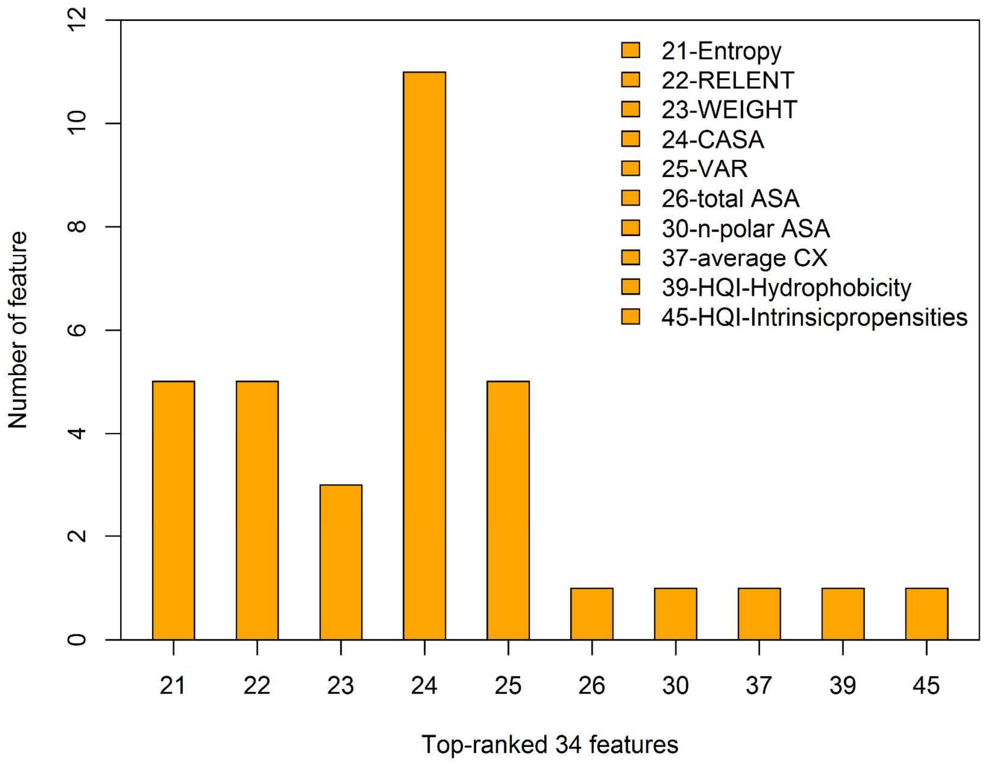

Figure 2. The number of each feature in the top-ranked 34 features on the Hetero dataset.

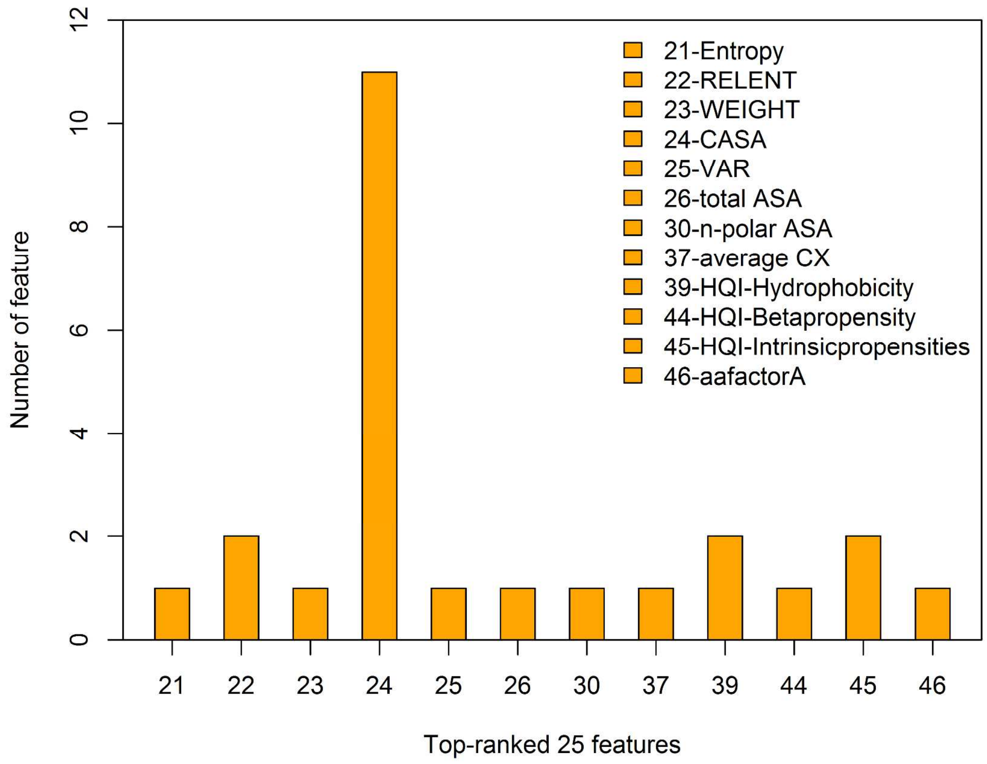

Figure 3. The number of each feature in the top-ranked 25 features on the Homo dataset.

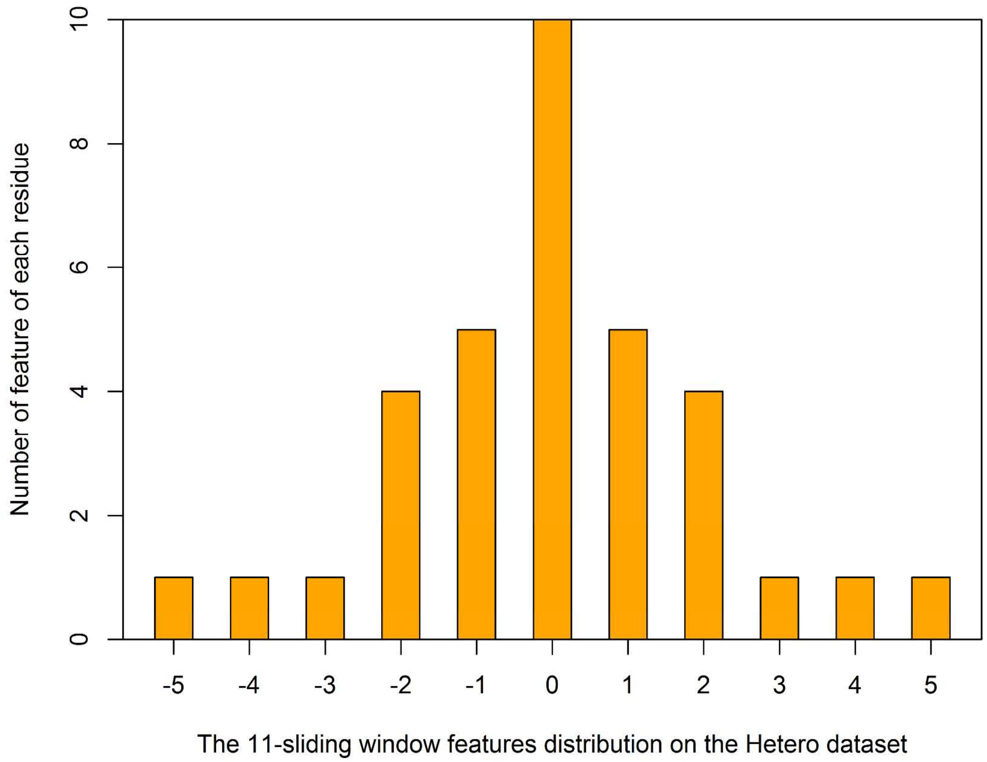

Figure 4. The distribution of feature of each residue in the 11-sliding window on the Hetero dataset.

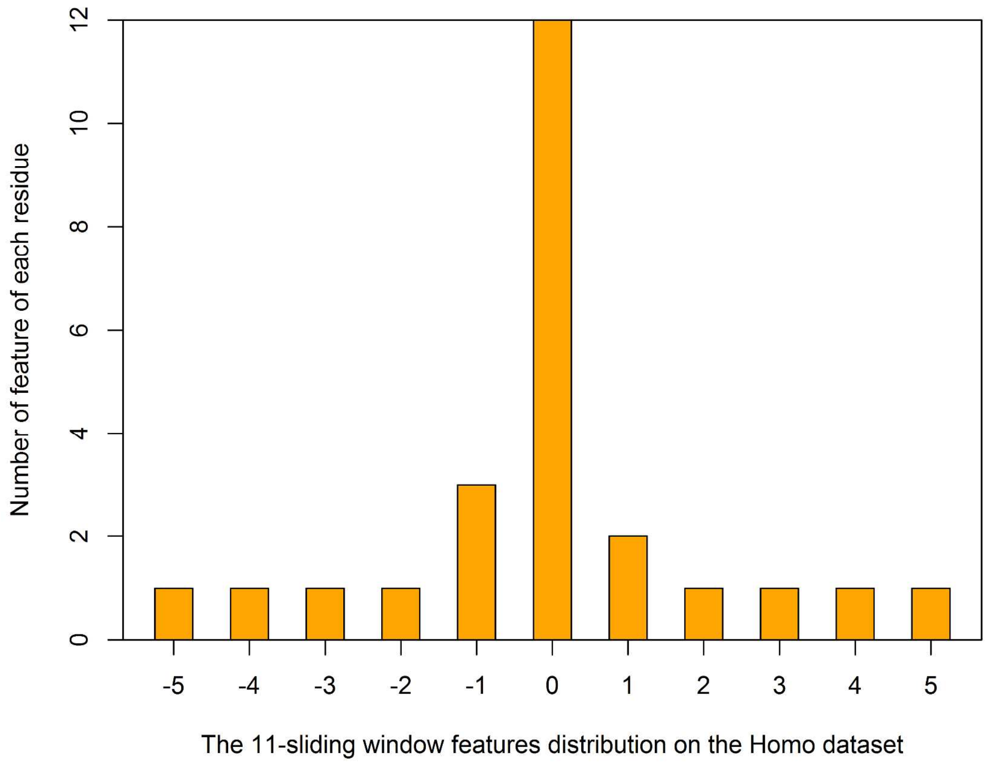

Figure 5. The distribution of feature of each residue in the 11-sliding window on the Homo dataset.

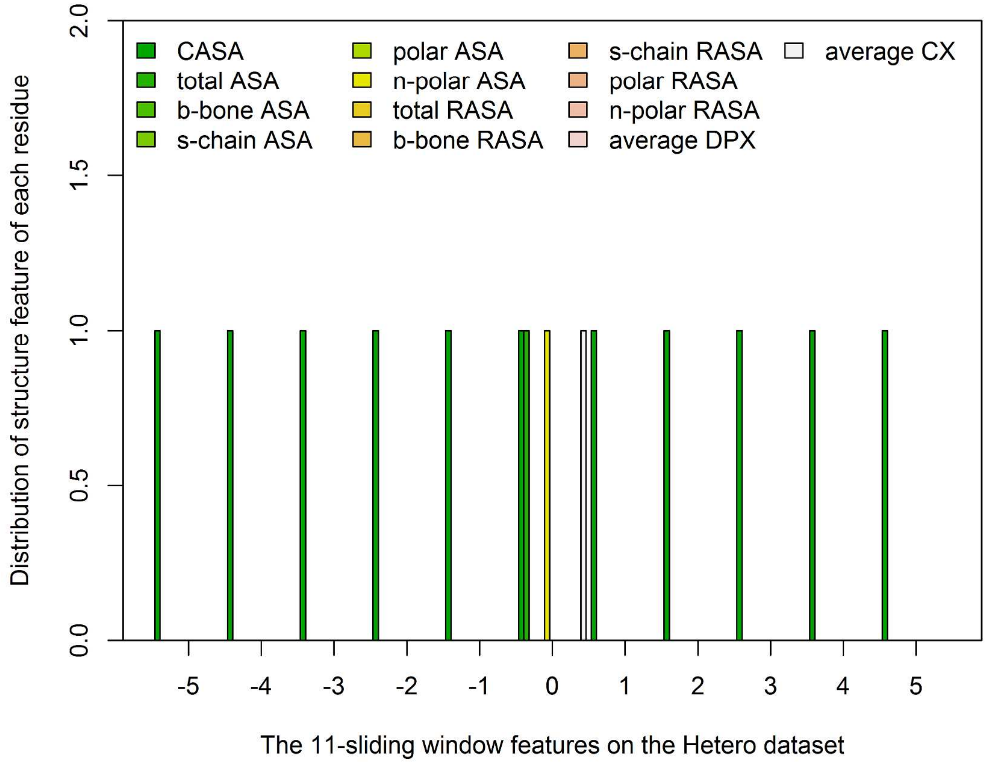

Figure 6. The distribution of structure-based feature with top-ranked 25 features on the Hetero dataset

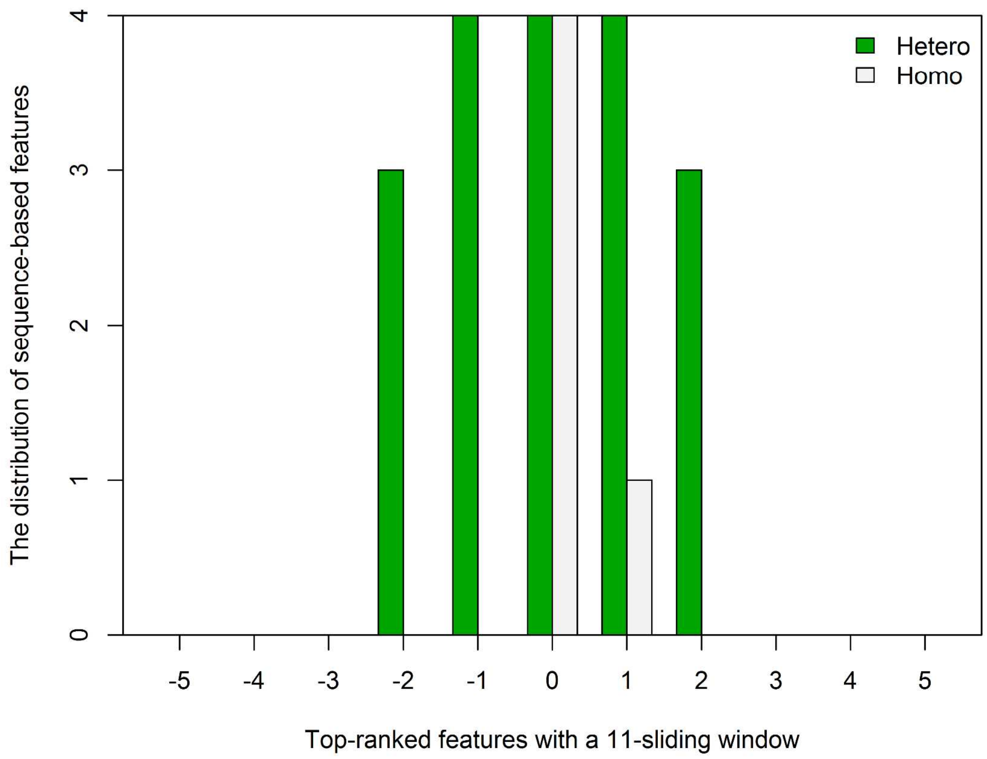

Figure 7. The distribution of sequence-based features in the 11-sliding window with top-ranked features.

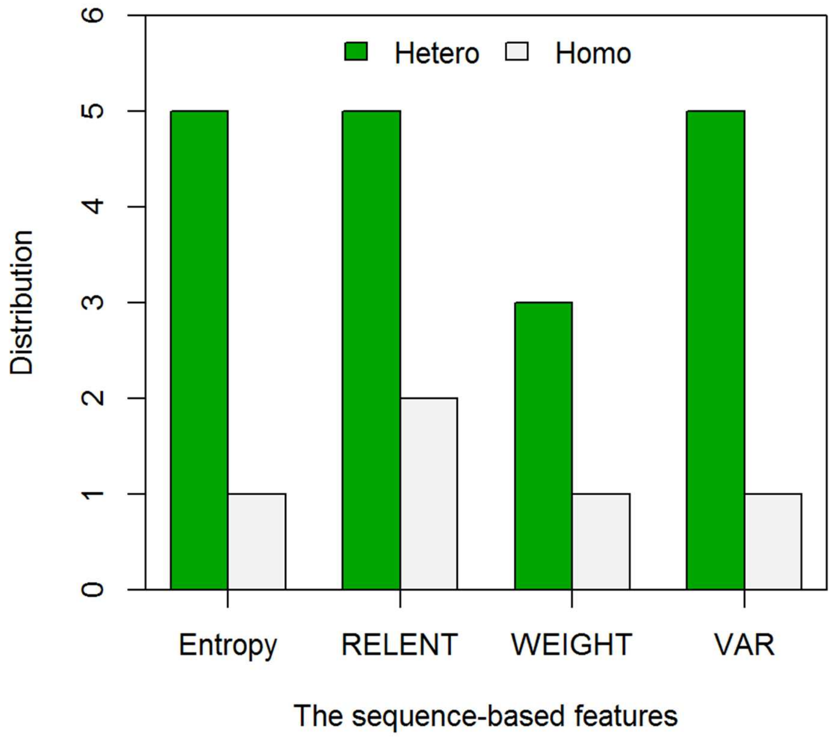

Figure 8. The number of sequence-based features in the optimal feature space.

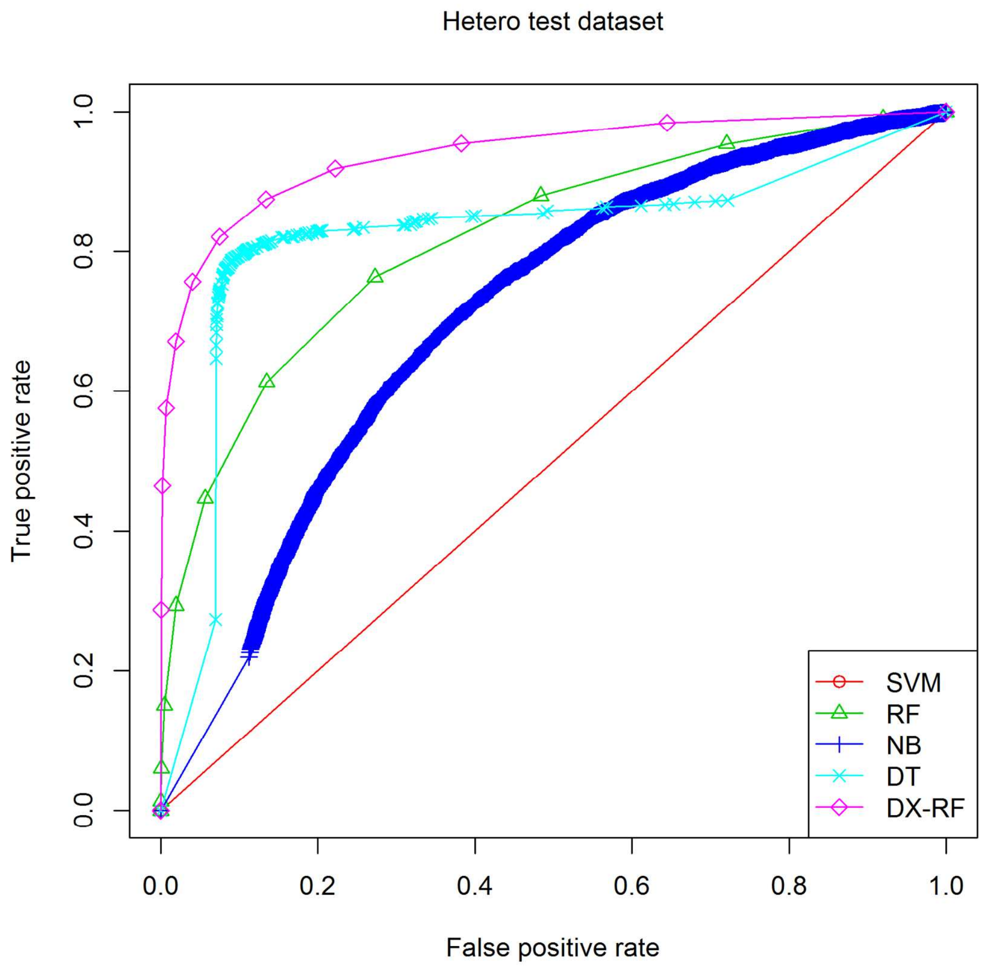

Figure 9. The ROC curves comparison of different methods on the Hetero test dataset (AUC LibSVM:0.5;DT:0.82;NB:0.7053;RF:0.818;DXRF:0.9391)

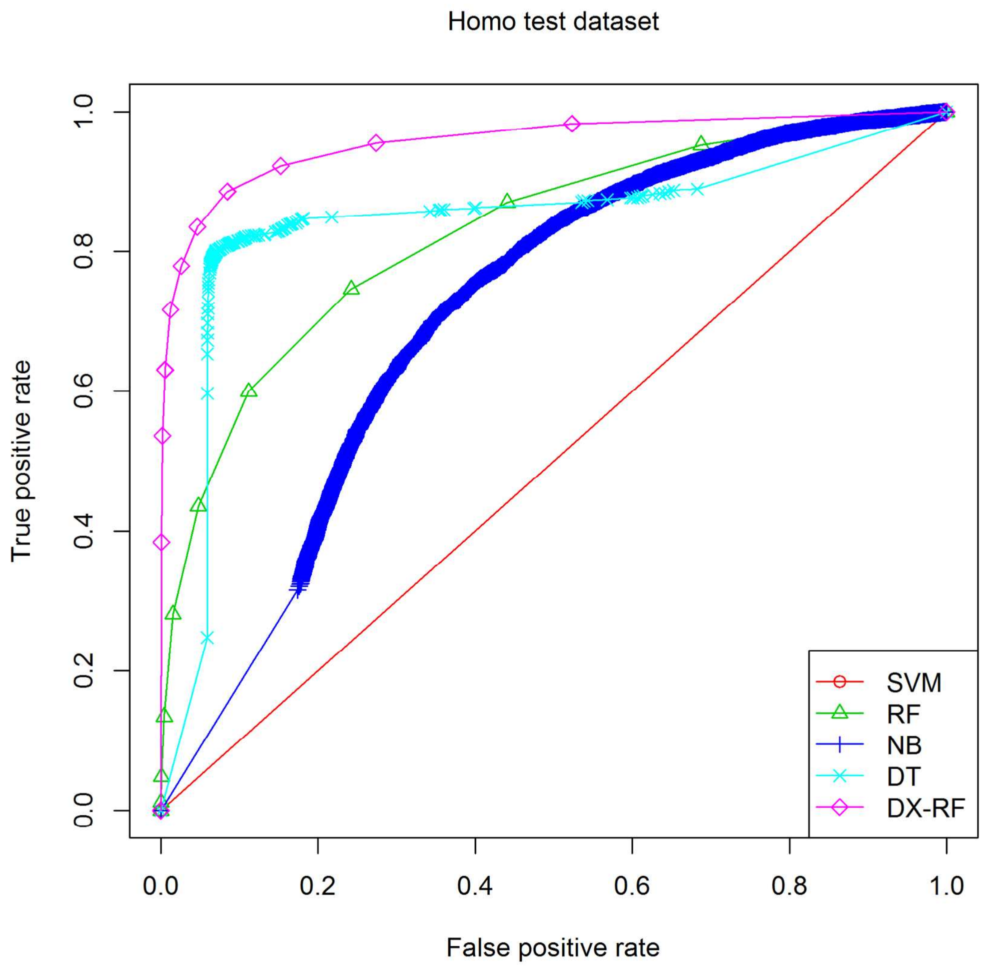

Figure 10. The ROC curves comparison of different methods on the Homo test dataset (AUC LibSVM:0.5;DT:0.8412;NB:0.711;RF:0.8261;DXRF:0.9568)

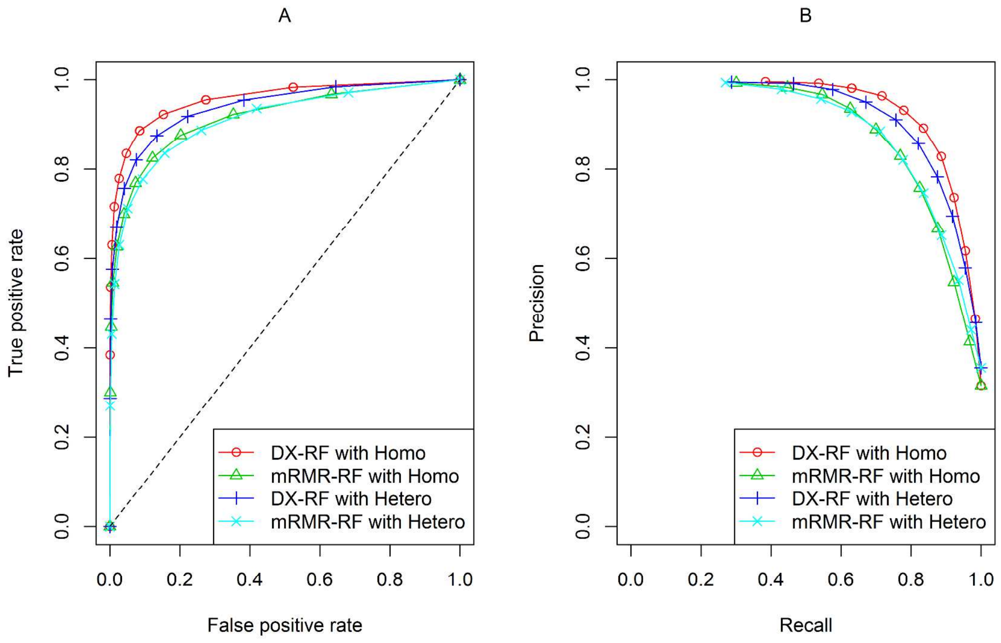

Figure 11. The ROC curves comparison between DX-RF and mRMR-RF.

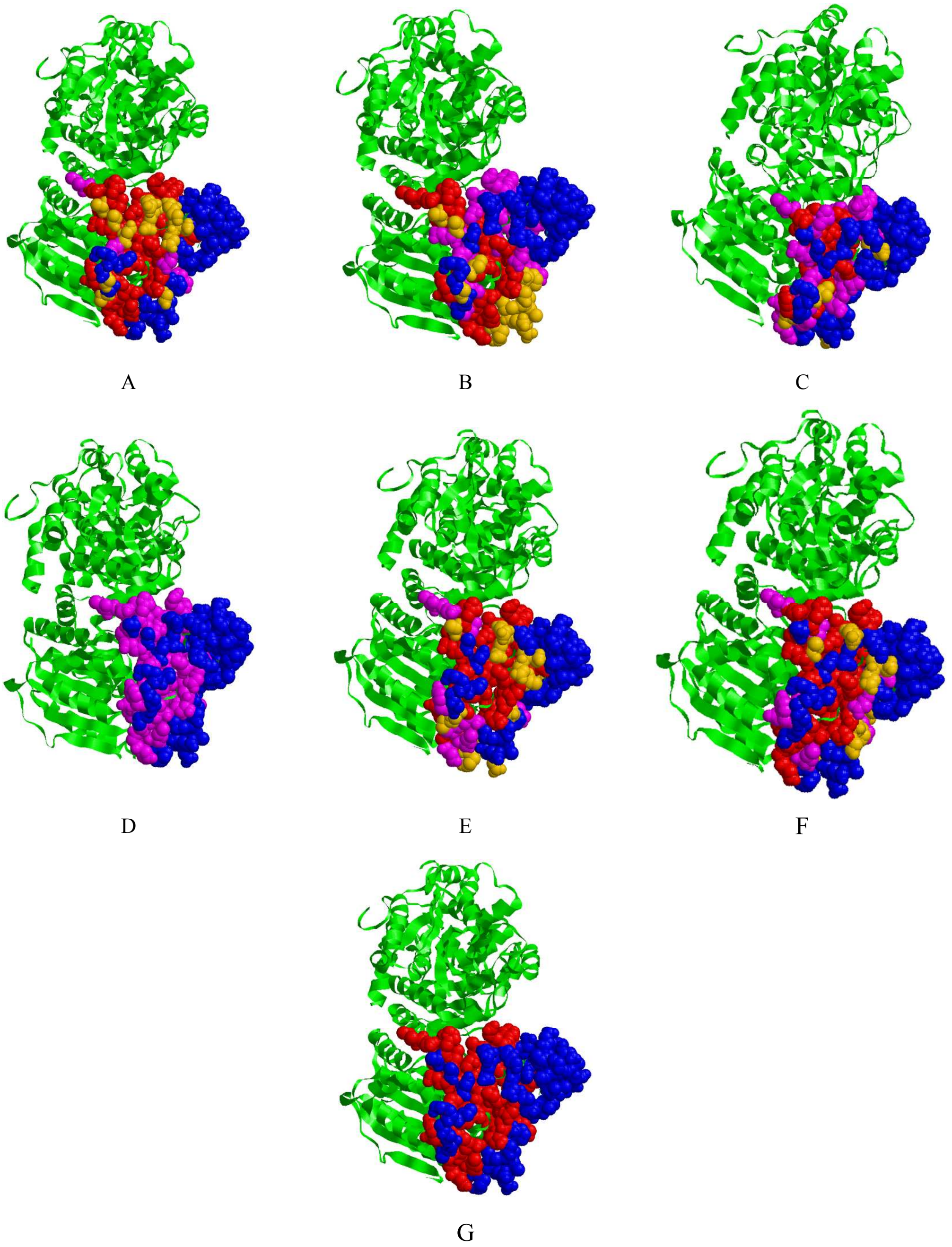

Figure 12. Predicted the interaction sites on protein (PDB:1B4U_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

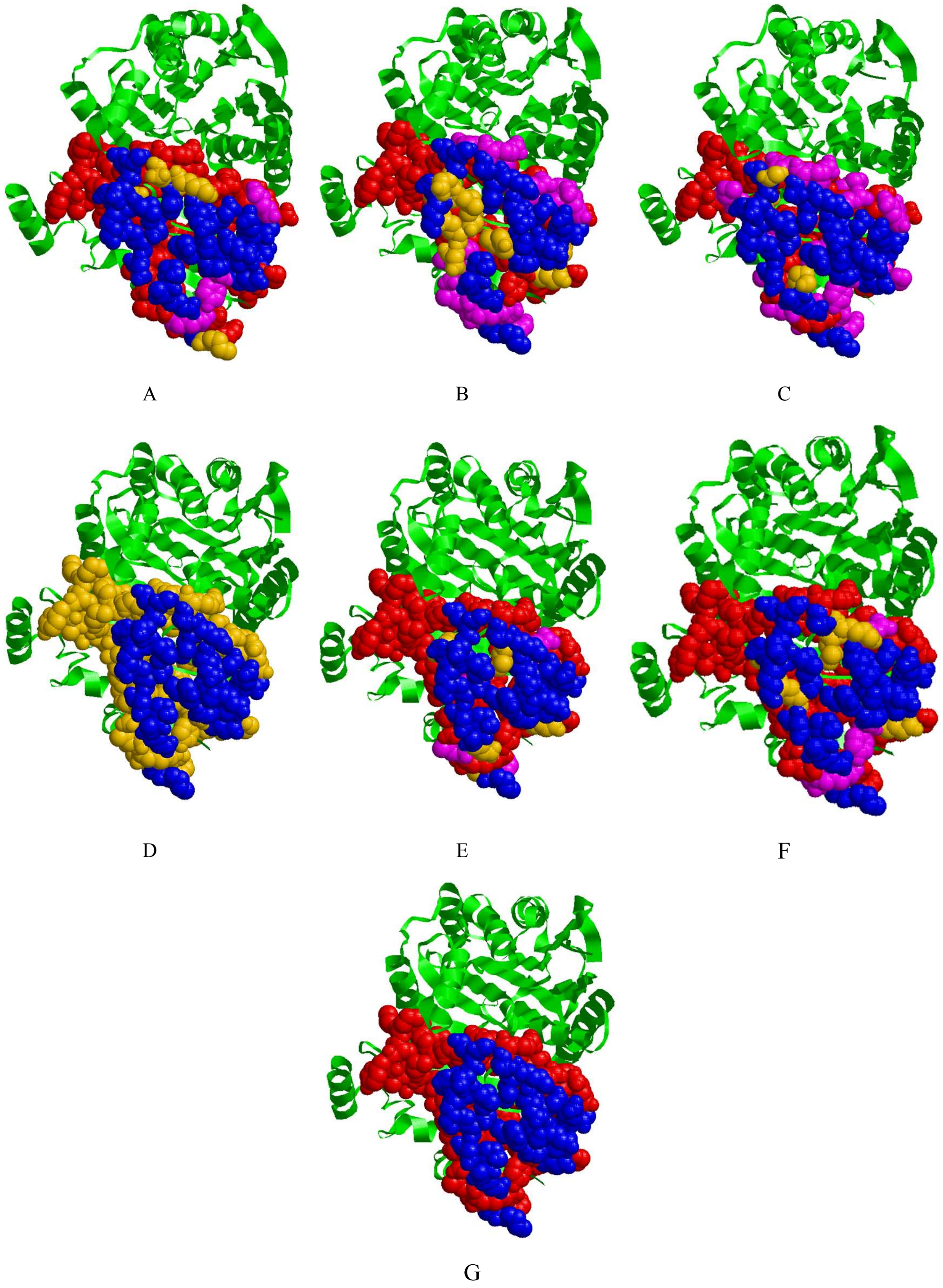

Figure 13. Predicted the interaction sites on protein (PDB:2GAC_B) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

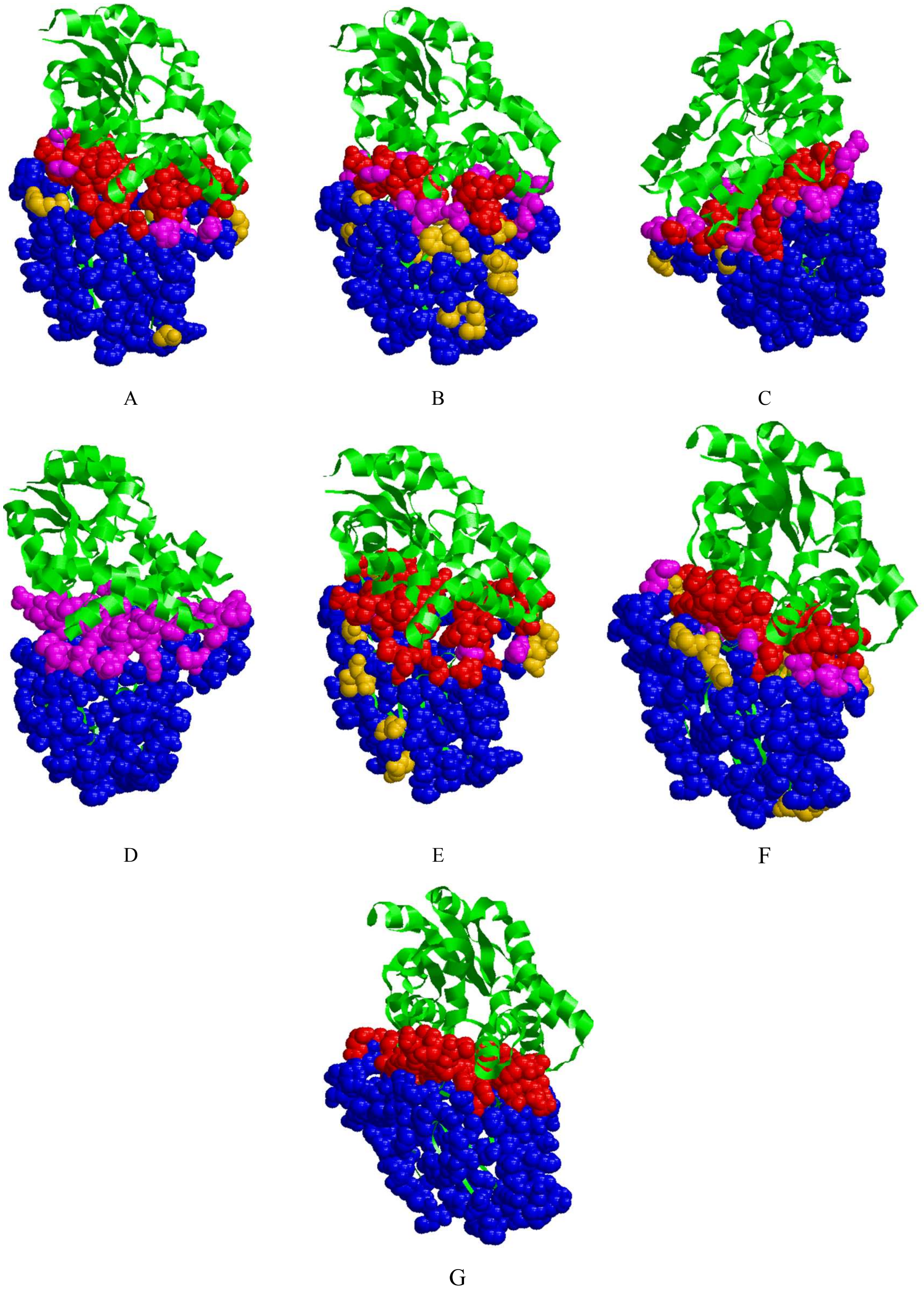

Figure 14. Predicted the interaction sites on protein (PDB:1QQ5_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.

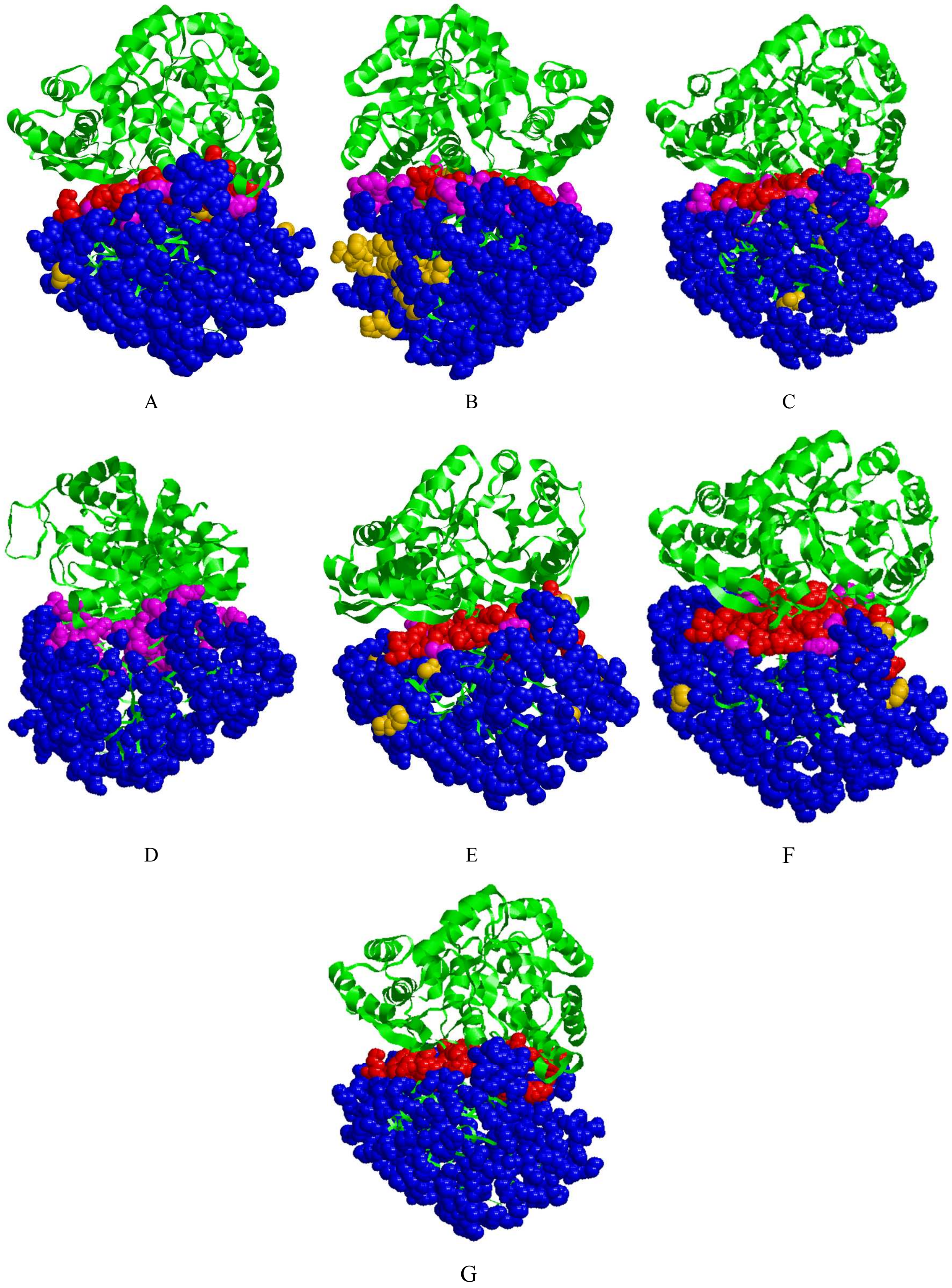

Figure 15. Predicted the interaction sites on protein (PDB:2ONE_A) identified by (A) DX-RF, (B) NB, (C) RF, (D) SVM, (E) DT, (F) mRMR-RF and (G) is the actual interface residues. Red denotes true positive residues, pink denotes false negative residues, gold denotes false positive residues, and blue denotes true negative residues.