Predicting Protein-Protein Interaction Sites by Rotation Forests with Evolutionary Information

Xinying Hu[2], Anqi Jing[2], and Xiuquan Du[1,2,*]

[1] Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education, Anhui University, Anhui, China
dxqllp@163.com
[2] The School of Computer Science and Technology, Anhui University, Anhui, China
lehehehu@163.com, aqjing0224@gmail.com

Abstract. In this paper, according to evolutionary information and physicochemical properties, we selected eight features, combined with Rotation Forest (RotF) to predict interaction sites. We built two models on both balanced datasets and imbalanced datasets, named balanced-RotF and unbalanced-RotF, respectively. The values of accuracy, F-Measure, precision, recall and CC of balanced-RotF were 0.8133, 0.8064, 0.8375, 0.7775 and 0.6283 respectively. The values of accuracy, precision and CC of unbalanced-RotF increased by 0.0679, 0.0122 and 0.0361 over balanced-RotF. Precision values of unbalanced-RotF on our four selected testing sets were 0.907, 0.875, 0.878 and, 0.889, respectively. Moreover, experiment only using two physicochemical features showed evolutionary information has effective effects for classification.

Keywords: protein-protein interaction sites, Rotation Forests, evolutionary information, machine learning.

1 Introduction

On the basis of the sequence and structural information of protein, some methods have been proposed [1-6]. Kini and Evans [1] proposed a unique predictive method that detecting the presences of the "proline" because they observed that proline is the most common residue found in the flanking segments of interface residues. Zhu-Hong You.et al[6] have proposed a novel method only using the information of protein sequences, which used the PCA-EELM model to predict protein-protein interactions. Many methods to predict the protein-protein interacting sites are motivated by the different machine learning methods with characteristics of proteins [5-17]. Minhas F.U. et al. [10] presented a novel method called PAIRpred. They selected structure and sequence information of residue pairs, combined Support Vector Machine method, which achieved good and detailed result. Peng Chen and Jinyan Li [11] trained a SVM using an integrative profile by combining the hydrophobic and evolutionary information, where they used a self-organizing map (SOM) technique as input vectors. Based on the Random Forest method, B.L.et al. [12] presented a new method with the

* Corresponding author.

D.-S. Huang et al. (Eds.): ICIC 2014, LNBI 8590, pp. 271–279, 2014.

Minimum Redundancy Maximal Relevance method followed by incremental feature selection. What they took into consideration included the five 3D secondary structures.

2 Methods

2.1 Defining the Protein Interaction Sites

In our article, we adopted the Fariselli's [7] method to define the definition of surface residues and interface residues. If a residue's RASA is at least 16% of its MASA, it is defined to be a surface residue, or it is defined to be a non-surface residue. If a surface residue's difference between ASA and CASA is greater than 1 $Å^2$, than it is defined to be an interface residue, otherwise it is defined to be a non-interface residue.

2.2 Features

In this paper, we adopt eight features to express interaction sites. These features as follows: sequence profiles, entropy, relative Entropy, conservation Weight, accessible surface areas, sequence variability, hydrophobicity [18] and polarity. The first six values of features could be obtained in HSSP database [19].

2.3 Creating Sample Sets

From what we have mentioned above, we use these features to describe a residue. Each residue is made up by 27 values (Sequence profile is 20 values, and the other features is one value.). We used sliding window of size 5. Therefore, there are 27*5 values in each residue's sample. If a residue doesn't have enough neighbors, we substitute the zero for its value.

2.4 Rotation Forests

In our paper, we constructed classifiers using Rotation Forests [20]. Rotation Forests, including many decision trees, is an ensemble learning method for classification. Its output decided by the mode of outputs of decision trees. In our experiment, we used Rotation Forest algorithm in Waikato Environment for Knowledge Analysis (WEKA) [21] to construct classifiers.

2.5 Datasets

In our experiment, we chose the proteins in bos Taurus organism as our training and testing datasets. We downloaded proteins of whose resolution is less than 3.5$Å^2$ in bos Taurus organism from the Protein Data Bank (PDB). Then we gave up those proteins whose length is less than 40 residues and sequence similarity is greater than 30%. Finally, we obtained 292 chains, 65185 residues. According to the definition of surface residues, there were 9291 interface residues and 30899 non-interface residues. Finally, we chose them as our datasets, named Bos. Then interface residues were labeled as "+1", and non-interface residues were labeled as "-1". We have created two training

and testing datasets. One is unbalanced named unbalanced-Bos, containing all surface sites in Bos. The other one is balanced named balanced-Bos.

2.6 Measuring Method

Accuracy, F-Measure, Recall, Precision, Correlation Coefficient (CC) were calculated to evaluate the performance of our predictors. ROC Area and ROC curve were also used in our article.

$$Accuracy = (TP + TN)/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad Recall = TN/(TN + FP) \quad \Pr ecision = TP/TP + FP$$

$$F - Measure = (2 * recall * precision)/(recall + precision)$$

$$CC = (TP * TN - FP * FN)/\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$$

3 Experimental Procedure and Results

3.1 Experiments on Different Machine Learning Methods

Experimental Procedure: We make experiments on both balanced Bos and unbalanced-Bos using Rotation Forests. We carried out experiments for 10 times and we created 10 models for balanced-Bos. We named the experiment balanced-RotF. For unbalanced-Bos, we carried out experiment for only once with10-cross validation and we named it unbalanced-RotF. Meanwhile we constructed other classifiers by some other machine learning methods in WEKA and LIBSVM [22] software. By comparing those results on different classifiers, we can make an observation which classifier that our sample sets perform on is better. From Figure 1 and Figure 2, we can see the value of accuracy, F-Measure and CC of Rotation Forests are higher than that of other machine learning methods. The values of accuracy, F-Measure, Precision, Recall and CC are 0.8133, 0.8064, 0.8375, 0.7775 and 0.6283，respectively.

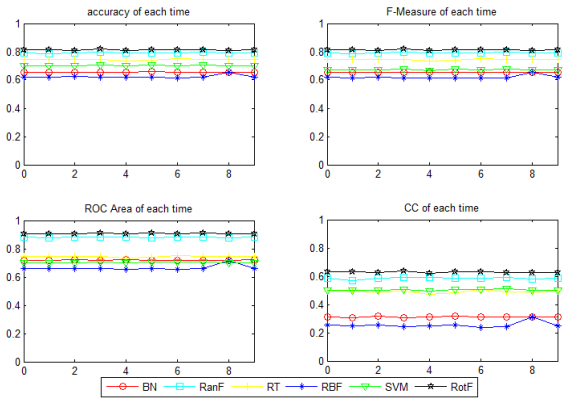

Fig. 1. The accuracy, F-Measure, and CC of different machine learning methods on the balanced-Bos. Figures on the top left corner, the top right corner, the bottom corner, show the accuracy values, F-Measure and CC of ten times of six machine learning methods on the balanced-Bos, respectively.

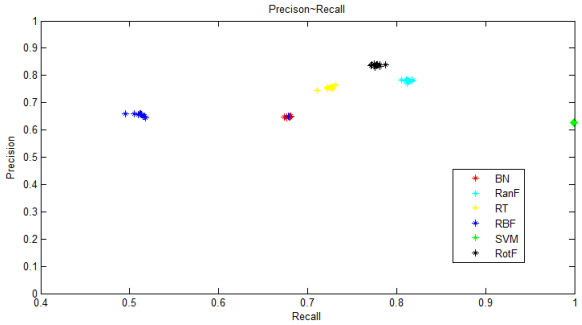

Fig. 2. The performances of Precision and Recall on the balanced-Bos. Ten black points show performances of ten balanced-RotF experiments we carried out. The vertical axis values show the value of recall and the abscissa axis values show the value of precision.

What we expected was it can achieve high accuracy and also get the high recall and precision. Figure 2 shows the values of precision and recall. Obviously, the closer that points can approach to the right top corner, the better performances they have. Black points in Figure 2, standing for Rotation Forest algorithm, we can observe that black points have highest precision. Figure 3 shows performances of different methods on unbalanced-Bos. It is obvious that unbalanced-RotF have better performances than other methods. The values of accuracy, F-Measure, precision, recall and CC were 0.8812, 0.7271, 0.8497, 0.6354 and 0.6644, respectively; expect precision of SVM was higher. From the three figures, we can make a conclusion that Rotation Forests are more suitable for our extracted features. From Table 1, what else we can observe was that the some indicator values on unbalanced-Bos were better than balanced Bos. It shows information that negative samples contained make irreplaceable contributions to the prediction of interaction sites and it should not be abandoned.

3.2 Experiments on Rotation Forests without Evolutionary Information

We make experiments only using hydrophobicity and polarity to confirm whether evolutionary information of proteins makes contributions to predict interaction sites. As the same way, we carried experiments on both balanced-Bos and unbalanced-Bos. From Table 2, we can see the average value of accuracy, F-Measure, precision, recall and CC was 0.6146, 0.6153, 0.6141, 0.6167 and 0.2292 respectively on balanced-Bos. However, the results show all sites including interaction sites and non-interface sites were predicted to be non-interaction sites. Rotation Forests with all eight features performed better on predication of interaction sites. It indicates that evolutionary information has effective effects on classification.

3.3 Experiments on different Independent Testing Datasets

In order to measure performances of RotF-classifiers we created in experiments on different machine learning methods, we built several independent testing sets, which came from escherichia coli (E.coli), bacillus subtilis (B.subtilis), rattus norvegicus

(R.norvegicus) and Yeast bacillus (Y.bacillus). We adopted the models which were produced by RotF-classifiers, including balanced-RotF and unbalanced-RotF. The following Table 3 and Table 4 present the performances of RotF-classifiers on independent testing sets, respectively.

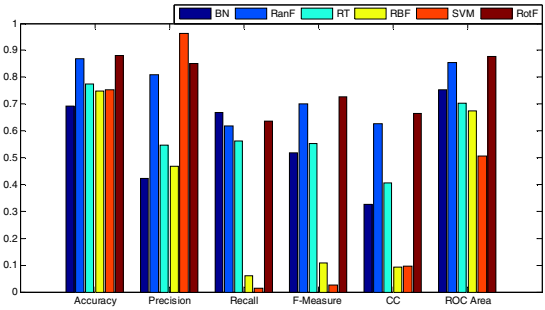

Fig. 3. Performances of unbalanced-Bos using different machine learning methods. It shows accuracy, F-Measure, Precision, Recall, CC and ROC Area using different machine learning methods on the unbalanced-Bos.

Table 1. Performances of Rotation Forests on balanced-Bos and unbalanced-Bos

	Accuracy	F-Measure	Precision	Recall	CC	ROC Area
balanced-Bos	0.8133	0.8064	0.8375	0.7775	0.6283	0.9077
unbalanced-Bos	0.8812	0.7271	0.8497	0.6354	0.6644	0.8790

Table 2. Performances of Rotation Forests only using hydrophobicity and polarity

	Accuracy	F-Measure	Precision	Recall	CC	ROC Area
balanced-Bos	0.6146±0.0030	0.6153±0.0061	0.6141±0.0026	0.6167±0.0123	0.2292±0.0059	0.6576
unbalanced-Bos	0.7509	—[1]	—	0	—	0.5950

[1]: "—" means all residues were predicted to be non-interface sites.

From the Table 3, we can observe that performances on imbalanced samples were good. Values of accuracy of balanced-RotF for these four organisms were 0.8198, 0.8062, 0.8093 and 0.8044, respectively. Values of ROC Area were 0.8441, 0.8278, 0.8030 and 0.8350, respectively. Table 4 shows the performances of unbalanced-RotF models. Values of precision of unbalanced-RotF models for four testing sets were 0.907, 0.875, 0.878 and 0.889, respectively. High precision means more positive samples were predicted correctly. Figure 4 shows the balanced-RotF models of ROC curves on four testing sets and Figure 5 shows the unbalanced-RotF model of ROC curves on four testing sets. We can make a conclusion that our classifiers performed well on independent testing sets, which confirms that our classifiers are suitable for not only the bos Taurus organism, but also for some other organisms. The results confirmed that our Rot-classifiers have extensive adaptation.

Table 3. Performances of Rotation Forests of balanced RotF models on independent testing datasets

	Accuracy	F-Measure	Precision	Recall	CC	ROC Area
E.coli	0.8198	0.7022	0.6976	0.7069	-0.1173	0.8441
variance	0.0101	0.0181	0.0219	0.0149	0.5703	0.0019
B.subtilis	0.8062	0.6665	0.6904	0.6442	0.5308	0.8278
variance	0.0022	0.0037	0.0045	0.0043	0.0052	0.0026
R.norvegicus	0.8093	0.6094	0.6074	0.6118	0.4835	0.8030
variance	0.0036	0.0100	0.0074	0.0208	0.0118	0.0025
Y.bacillus	0.8044	0.6850	0.7089	0.6631	0.5421	0.8350
variance	0.0088	0.0431	0.0435	0.0459	0.0326	0.0024

Table 4. Performances of unbalanced RotF model on independent testing datasets

	Accuracy	F-Measure	Precision	Recall	CC	ROC Area
E.coli	0.849	0.692	0.907	0.560	0.630	0.854
B.subtilis	0.831	0.645	0.875	0.511	0.578	0.839
R.norvegicus	0.856	0.615	0.878	0.473	0.574	0.814
Y.bacillus	0.830	0.674	0.889	0.543	0.598	0.847

According to results, we can also observe that values of accuracy, precision, CC and ROC Area of unbalanced-RotF models were better than balanced samples. The performances of Rot-classifiers in experiments on different machine learning methods also proved it, which means imbalanced samples contained the information of all sites in proteins, including interface sites and non-interface sites. We took the fully use of samples and achieved better results. It meant in our experiment information of non-interface sites should not be abandoned.

3.4 Compared to Other Methods

For comparison purpose, we make experiments with our sample sets using other methods. We make experiments on following websites: SPIDDER [23] (http://sppider.cchmc.org/): It uses solvent accessibility, based on artificial neural network method. InterProSurf [24] (http://curie.utmb.edu/usercomplex.html): It was based on solvent accessible surface area, a propensity scale for interface residues and a clustering algorithm to classify interaction sites.

In order to make a convenient comparison to the method of SPIDDER, we redefined interface residues according to the definition of SPIDDER and InterProSurf to calculate accuracy, F-Measure, precision, recall and CC. Table 5 shows performances on the two methods. The results of SPIDDER achieved high accuracy: 0.723, but the values of precision and recall are low. The results of InterProSurf were similar to the SPIDDER. Its precision was higher than SPIDDER, but similarity, the value of recall was unsatisfactory, only 0.4894. (There were some problems with our experiments on InterProSurf. There were 85 chains didn't obtain predicted results, so the results were obtained after deleting these 85 chains).

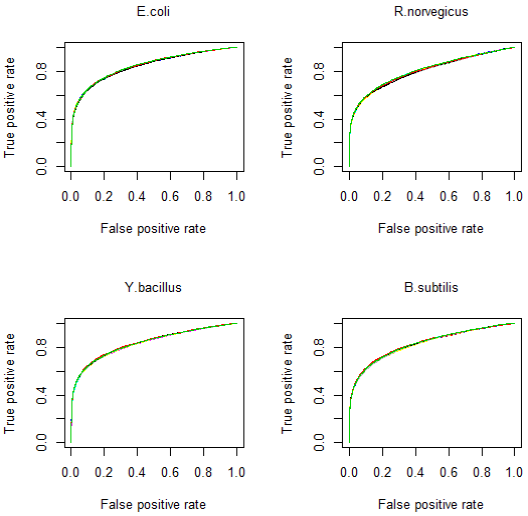

Fig. 4. Balanced ROC curves of four testing sets. It shows ROC curves of four testing sets, which produced by balanced-RotF classifiers from the experiment on Rotation Forests. E.coli stands for the Escherichia coli, which consists of ten curves, where each curve stands for the experimental result of each model from experiment on Rotation Forests. The same as E.coli, B.subtilis, R.norvegicus and Y.bacillus stand for the bacillus subtilis, rattus norvegicus and yeast bacillus, respectively.

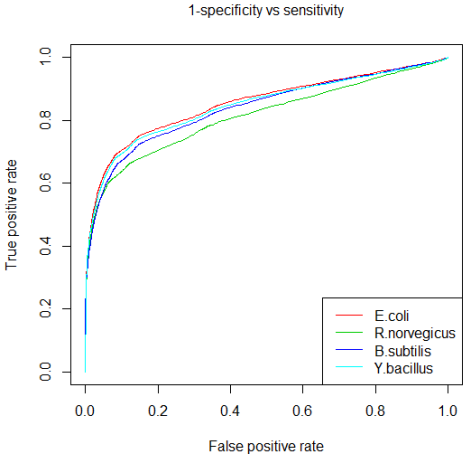

Fig. 5. Unbalanced ROC curves of four testing sets. It shows the ROC curves from four testing sets, which produced by unbalanced-RotF classifier from the experiment on Rotation Forests. There are four curves in the figure, where different color curves stand for different sample sets.

Table 5. Performances of SPIDDER and InterProSurf

	Accuracy	F-Measure	Precision	Recall	CC
SPIDDER	0.723	0.450	0.510	0.402	0.272
InterProSurf	0.851	0.608	0.802	0.489	0.226

4 Conclusion

In our paper, a new method was proposed to predict the interaction sites. At first, we extracted eight features, combined with the sliding window and it contained five amino acids. We created two classifiers named balanced-RotF and unbalanced-RotF, which showed good performances with high accuracy, F-Measure and CC, especially the Recall and Precision. Meanwhile, we make experiments on different machine learning methods: RanF, SVM, RT, BN and RBF. The results show that features that we selected are more suitable for the Rotation Forests method. What is more, we confirmed evolutionary information make contribution to prediction. Moreover, our models were tested on independent datasets, which achieved good results, as well, which proved that our models have extensive adaptation. For comparison, we made our datasets test on other methods. Performances show our results were better than theirs.

5 Funding

This project was supported by the National Natural Science Foundation of China (Grant No.61203290), Startup Foundation for Doctors of Anhui University (No.33190078) and outstanding young backbone teachers training (No.02303301).

References

1. Kini, R.M., Evans, H.J.: Prediction of potential protein-protein interaction sites from amino acid sequence identification of a fibrin polymerization site. FEBS Lett. 385(1-2), 81–86 (1996)
2. Tuncbag, N., Keskin, O., Nussinov, R., Gursoy, A.: Fast and accurate modeling of protein–protein interactions by combining template-interface-based docking with flexible refinement. Proteins: Structure, Function, and Bioinformatics 80(4), 1239–1249 (2012)
3. Zhang, S.W., Hao, L.Y., Zhang, T.H.: Prediction of protein–protein interaction with pairwise kernel Support Vector Machine. International Journal of Molecular Sciences 15(2), 3220–3233 (2014)
4. Konc, J., Janezic, D.: Protein-protein binding-sites prediction by protein surface structure conservation. J. Chem. Inf. Model. 47(3), 940–944 (2007)
5. Yuehui, C., Jingru, X., Bin, Y., Yaou, Z., Wenxing, H.: A novel method for prediction for protein interaction sites based on integrated RBF neural networks. Computers in Biology and Medicine 42(4), 402–407 (2012)
6. You, Z.H., Lei, Y.K., Zhu, L., Xia, J., Wang, B.: Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinformatics 14(suppl. 8), S10 (2013)

7. Fariselli, P., Pazos, F., Valencia, A., Casadia, R.: Prediction of protein-protein interaction sites in heterocomplexes with neural networks. Eur. J. Biochem. 269, 1356–1361 (2002)
8. Sriwastava, B.K., Basu, S., Maulik, U., Plewczynski, D.: PPIcons: Identification of protein-protein interaction sites in selected organisms. J. Mol. Model. 19(9), 4059–4070 (2013)
9. Chen, C.T., Peng, H.P., Jian, J.W.: Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. PLOS One 7(6), e37706 (2012)
10. Minhas, F.U., Geiss, B.J., Ben-hur, A.: PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. Proteins: Structure, Function, and Bioinformatics (2013)
11. Chen, P., Li, J.: Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. BMC Bioinformatics 11, 402–416 (2010)
12. Li, B.Q., Feng, K.Y., Chen, L., Huang, T., Cai, Y.D.: Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. PLOS One 7(8), e43927 (2012)
13. Li, C.X., Drena, D., Vasant, H.: HomPPI: a class of sequence homology based protein-protein interface prediction methods. BMC Bioinformatics 12, 244–267 (2011)
14. Jordan, R.A., EI-Manzalawy, Y., Dobbs, D., Honavar, V.: Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics 13, 41–44 (2012)
15. Xu, B., Wei, X., Deng, L., Guan, J., Zhou, S.: A semi-supervised boosting SVM for predicting hot spots at protein-protein interfaces. BMC Syst. Biol. 6(suppl. 2) (2012)
16. Sriwastava, B.K., Basu, S., Maulik, U., Plewczynski, D.: PPIcons: identification of protein-protein interaction sites in selected organisms. J. Mol. Model. 19(9), 4059–4070 (2013)
17. Hwang, H., Vreven, T., Weng, Z.: Binding interface prediction by combining protein-protein docking results. Proteins 10, 1002 (2013)
18. Gallet, X., Charloteaux, B., Thomas, A., Brasseur, R.: A fast method to predict protein interaction sites from sequences. J. Mol. Biol. 302(4), 917–926 (2000)
19. Chris, S., Reinhard, S.: Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Bioinforma. 9(1), 56–68 (1991)
20. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
22. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(27), 1–27 (2011)
23. Porollo, A., Meller, J.: Prediction-based fingerprints of protein-protein interactions. Proteins Struct. Funct. Bioinforma 66(3), 630–645 (2007)
24. Neqi, S.S., Schein, C.H., Oezquen, N., Power, T.D., Braun, W.: InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23(24), 3397–3399 (2007)