Supplementary materials:

Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique

Xiaoying Wang [1,2,3,†], Bin Yu [1,3,4,†,*], Cheng Chen [1,3], Anjun Ma[5,6], Bingqiang Liu [2], Qin Ma [5,6,*]

[1] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

[2] School of Mathematics, Shandong University, Jinan 250100, China

[3] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

[4] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

[5] Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD 57007, USA

[6] Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA

*To whom correspondence should be addressed

†Contributed equally to this work.

Contact: yubin@qust.edu.cn or Qin.Ma@sdstate.edu

Table of Contents

1. Supplementary Method Illustration

Si 1. Naïve Bayes (NB)

The Naïve Bayes algorithm is a method of classification by using the Naïve Bayes principle in probability theory (He *et al.*, 2017; Lin and Chen, 2013). As an expression model combining prior knowledge with new information extraction, it is often used in classification and prediction problems.

The protein sequence information sample is denoted by an n-dimensional feature vector $X = \{x_1,\ x_2,\ \cdots,\ x_n\}$, C means class, Y means interface residue, N means non-interfacial residues. According to the Naïve Bayes theorem, the formula of the classifier is as follows.

$$P(C_Y / X) = P(C_Y / X)P(C_Y) / P(X) \tag{1}$$

Where $P(C_Y)$ means the proportion of the interface residue in the training dataset, $P(X / C_Y)$ means the portion of non-interfacial interface residue with an attribute value $\{x_1,\ x_2,\ \cdots,\ x_n\}$. $P(X)$ indicates the probability of $X = \{x_1,\ x_2,\ \cdots,\ x_n\}$ in the whole dataset. According to the independence of each attribute, the (2) is obtained for classification prediction.

$$P(C_Y / X) = P(C_Y)\prod_{i=1}^{n} P(X_i / C_Y)P(C_Y) / P(X_i) \tag{2}$$

However, Naïve Bayes algorithm requires each attribute to be independent or almost independent, which is often difficult to be realized in real practical experiments. Regarding the feature extraction of protein-protein interaction sites, each feature attribute is not independent. Therefore, Naïve Bayes is often used in the case of small correlation of characteristic attributes. The Naïve Bayes algorithm is complicated to achieve a good classification effect when the correlation is too apparent. Based on this, the application scope of Naïve Bayes is significantly restricted.

Si 2. Support vector machine (SVM)

Support vector machine (SVM) is a machine learning method, which is based on a statistical theory proposed by Vapnik *et al.* (1997). In recent years, SVM has also been widely used in the field of bioinformatics (Chen *et al.*, 2015; Guo *et al.*, 2008; Khan *et al.*, 2017; Kim *et al.*, 2004; Qiu *et al.*, 2018; Song *et al.*, 2017; Wan *et al.*, 2016; Yu *et al.*, 2017; Yu *et al.*, 2017b; Yu *et al.*, 2017c; Zamanighomi *et al.*, 2017; Zhang and Tang, 2016). The characteristic of this method is to improve the generalization ability to learn from the principle of risk minimization. The principle of two classifications using support vector machines is to map the input space sample into a high dimensional feature space through a nonlinear function and to find an optimal hyperplane in the feature space so that the two classes of samples are linearly separable.

The kernel function $K(x_i, x_j)$ is used, instead of an inner product, in the optimal classification plane. Then the optimal function is obtained.

$$F(x) = \operatorname{sgn}((w^*)^T x + b^*) = \operatorname{sgn}(\sum_{i=1}^{n} \alpha_i^* y_i K(x_i, \mathrm{x}_j) + b^*) \tag{3}$$

where, sgn means sign function, b^* implies hyperplane offset, α_i implies lagrange multiplier, $K(x_i, x_j)$ implies kernel function.

In this study, we use the software package LIBSVM developed by Chang and Lin (2011), which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Si 3. Random Forest (RF)

Random forest (Mayer *et al.*, 2018; You *et al.*, 2015; Yu *et al.*, 2015) is a machine learning algorithm ensembled by multiple independent decision trees. Each decision tree is trained by random bootstrap. When generating decision trees, nodes are randomly divided into a subset of feature space. Because each decision tree is trained based on independent random selection feature subset, there is no need to prune the decision tree to avoid overfitting. Random forests are weighted or averaged by the results of all decision trees. Random forests are widely used because of their high accuracy, high noise immunity, and fast training speed. The specific algorithms are introduced as follows: (1) From the training set, the Bootstrap method is used to select m samples by being put back and generate corresponding decision trees for each sample. (2) Assuming that there are n variables, n_1 variables are randomly selected from each node of each tree, as the split attribute set of the current node and the best variable of classification ability is chosen from the n_1 variables to split the current node. (3) Every tree grows as far as possible without any pruning. (4) The multiple trees generated above constitute a random forest, and the test data is input and classified. The final result of the classification is determined by the most output class of the tree species.

2. Supplementary Tables

Table S1. The influence of different parameters on the evaluation index on Dset186. To choose the best parameter, the parameter *ntree* s selected from 100 to 1000 and the tolerance is 100.

Ntree	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-Measure	MCC
100	71.0	95.7	46.3	64.1	0.767	0.483
200	70.3	96.1	44.5	63.4	0.764	0.474
300	**71.8**	**96.4**	**47.3**	**64.7**	**0.774**	**0.501**
400	71.7	96.1	47.2	64.5	0.772	0.498
500	70.4	96.4	44.4	63.4	0.765	0.477
600	71.5	96.0	46.9	64.4	0.770	0.492
700	70.9	96.3	45.5	63.9	0.768	0.486
800	71.6	96.3	47.1	64.5	0.772	0.497
900	70.9	96.4	45.5	63.9	0.768	0.486
1000	70.8	96.7	44.8	63.7	0.768	0.487

From Table 1, we can see that different values of *ntree* are chosen, and the prediction results are different. For the training dataset Dset186, the maximum prediction accuracy is reached at $ntree$=300 , with a value of 71.8%, 1.5% higher than the Accuracy at $ntree$=200 , and 0.1% higher than that of $ntree$=400 . When $ntree$=300 , the values of MCC, Specificity, Precision, and F-Measure achieve maximum, 0.501, 47.3%, 64.7% and 0.774 respectively. The maximum operating efficiency is reached at $ntree$=300 .

Table S2. The influence of features selection methods. Feature selection has a significant improvement on the classification performance. MDS had the most significant effect on the three data sets, which is a global algorithm that utilizes the similarity between pairs of samples. The purpose is to use this information to construct a suitable low-dimensional space so that the distance between the samples in this space and the similarity between the samples in the high-dimensional space are as consistent as possible. This result showcased that the importance of feature selection in reducing the computational complexity and feature redundancy and improving the prediction performance.

Dataset	Method	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-Measure	MCC
Dset186	Origin	73.1	96.0	54.5	66.8	0.773	0.498
	MDS	**79.1**	**81.7**	**76.6**	**77.7**	**0.784**	**0.584**
	LPP	79.4.	82.2	76.6	77.6	0.799	0.589
	LLE	78.4	81.5	75.2	76.7	0.790	0.567
	FA	78.2	82.7	73.8	75.9	0.792	0.567
	LDA	76.8	79.1	74.5	75.6	0.773	0.537
	NPE	76.3	76.1	76.4	76.4	0.763	0.526
	Autoencoder	58.5	66.1	51.0	57.4	0.614	0.172
Dtestset72	Origin	73.1	90.7	55.5	67.1	0.768	0.493
	MDS	**77.1**	**78.8**	**75.3**	**76.2**	**0.775**	**0.542**
	LPP	76.5	78.2	74.8	75.6	0.769	0.531
	LLE	75.6	75.5	75.7	75.6	0.755	0.512
	FA	75.7	78.5	72.7	74.2	0.764	0.514
	LDA	74.3	72.9	75.6	74.9	0.739	0.485
	NPE	73.5	69.2	77.8	75.7	0.723	0.471
	Autoencoder	55.7	60.0	51.6	55.3	57.4	0.114

PDBtestset164	Origin	72.0	96.4	46.2	64.2	0.770	0.492
	MDS	**77.7**	**80.0**	**75.3**	**76.4**	**0.782**	**0.554**
	LPP	77.6	80.0	75.1	76.3	0.781	0.552
	LLE	77.0	80.0	74.0	75.5	0.777	0.541
	FA	76.2	77.7	74.7	75.4	0.765	0.524
	LDA	74.6	76.4	72.8	73.7	0.751	0.492
	NPE	76.1	73.5	78.7	77.5	0.754	0.522
	Autoencoder	63.0	65.4	60.6	0.624	0.639	0.260

Table S3. The prediction of Dset186 after the balancing process with MDS. The protein-protein interaction sites were predicted by different classifiers on the training set Dset186 after SMOTE. It is found that the EL-SMURF method proposed in this study reaches the maximum in Accuracy index. At the same time, the Precision and F-measure indexes of the EL-SMURF method reached the highest respectively. This shows that EL-SMURF has good performance to extract implicit information from training sets. The method has good classification effect in training dataset Dset186.

Methods	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-Measure	MCC
NB	57.6	64.2	51.6	56.8	0.603	0.154
RF	78.9	79.7	73.8	76.1	0.780	0.572
SVM	69.2	73.1	65.3	67.8	0.704	0.385
EL-SMURF	**79.1**	**81.7**	**76.6**	**77.7**	**0.784**	**0.584**

Table S4. The prediction of Dtestset72 after the balancing process with MDS. The protein-protein interaction sites were predicted by different classifiers, NB, RF, SVM, and EL-SMURF, on the independent validation set Dtestset72 after SMOTE. It is found that the Accuracy index of the EL-SMURF method proposed in this study is the largest. At the same time, the Precision, F-measure, and MCC reached the maximum value respectively. This shows that EL-SMURF has a good classification effect in the independent validation set Dtestset72.

Methods	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-Measure	MCC
NB	59.1	66.4	51.8	58.0	0.619	0.184
RF	76.9	78.2	75.0	76.0	0.775	0.542
SVM	69.2	73.5	64.8	67.6	0.774	0.385
EL-SMURF	**77.1**	**78.8**	**75.3**	**76.2**	**0.775**	**0.542**

Table S5. The prediction of PDBtestset164 after the balancing process with MDS. Compared with the different classifiers after SMOTE processing, it is found that the EL-SMURF method proposed in this study has a significant increase in the Accuracy index of the individual classifier. At the same time, the F-measure and MCC also reached the maximum. This shows that EL-SMURF has a good classification effect in the independent validation PDBtestset164.

Methods	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-Measure	MCC
NB	58.2	64.7	51.6	57.3	0.608	0.165
RF	77.1	79.5	74.7	75.9	0.776	0.543
SVM	66.3	74.0	58.6	64.1	0.687	0.330
EL-SMURF	**77.7**	**80.0**	**75.3**	**76.4**	**0.782**	**0.554**

Table S6. Comparison of different methods on Dset186 over leave-one-out cross-validation.

Methods	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-measure	MCC
DC-RF-RUS-PF	65.1	64.3	61.7	28.6	0.373	0.202
SSWRF	67.9	58.1	69.7	32.2	0.386	0.234
CRF	72.7	61.2	67.4	31.8	0.390	0.204
LORIS	60.4	69.8	58.6	28.7	0.384	0.221
PSIVER	67.3	41.6	74.3	30.6	0.353	0.151
EL-SMURF	**79.1**	**81.7**	**76.6**	**77.7**	**0.784**	**0.584**

Table S7. Comparison of different methods on the independent validation dataset Dtestset72.

Methods	Acc (%)	Se (%)	Sp (%)	Pr (%)	F-measure	MCC
DC-RF-RUS-PF	64.0	62.7	64.6	32.4	0.336	0.204
SSWRF	64.8	65.4	64.3	26.7	0.224	0.351
CRF	70.6	64.0	64.0	25.6	0.340	0.209
LORIS	61.4	63.1	61.0	23.8	0.324	0.177
SPRINGS	62.4	59.0	63.0	24.1	0.318	0.170
PSIVER	66.1	46.5	69.3	25.0	0.278	0.135
SPPIDER	61.7	45.4	63.7	20.4	0.241	0.081
EL-SMURF	**77.1**	**78.8**	**75.3**	**76.2**	**0.775**	**0.542**

Table S8. Comparison of different methods on the independent validation dataset PDBtestset164.

Methods	Acc (%)	Sp (%)	Se (%)	Pr (%)	F-measure	MCC
DC-RF-RUS-PF	61.1	65.3	52.6	32.4	0.360	0.148
SSWRF	62.1	65.6	52.7	32.3	0.365	0.152
CRF	61.3	64.5	54.3	32.3	0.370	0.113
LORIS	58.8	60.9	53.8	26.3	0.323	0.111
SPRINGS	60.6	64.8	40.7	26.8	0.311	0.108
PSIVER	59.6	63.4	46.4	25.3	0.295	0.078
SPPIDER	71.6	85.1	16.2	23.1	0.129	0.015
EL-SMURF	**77.7**	**80.0**	**75.3**	76**.4**	**0.782**	**0.554**

3. Supplementary Figures

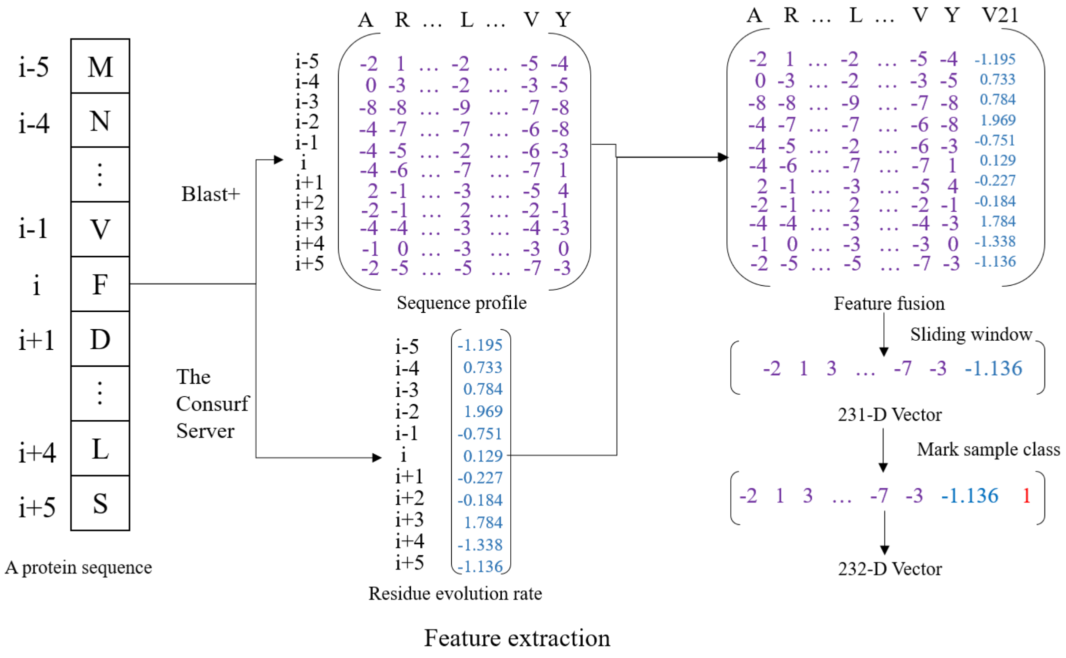

Fig. S1. The procedure of feature extraction.

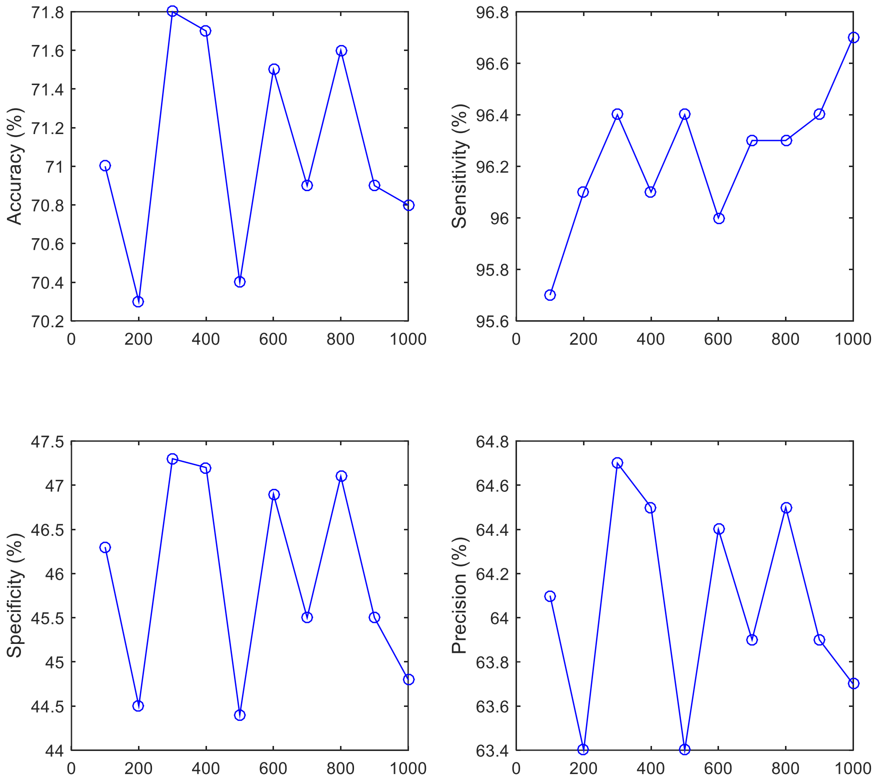

Fig. S2. The influence of different parameters on the evaluation index. When *ntree*=300, the maximum values of the optional ranges on Accuracy, Specificity and Precision are achieved. Considering the efficiency of the program and the prediction accuracy of the prediction model, *ntree*=300 , which is the best parameter value of the random forest classifier under Dset186. Since Dset186 is the training dataset of this study, the *ntree* value of random forests in Dtestset72 and Dtestset164 is set at 300.

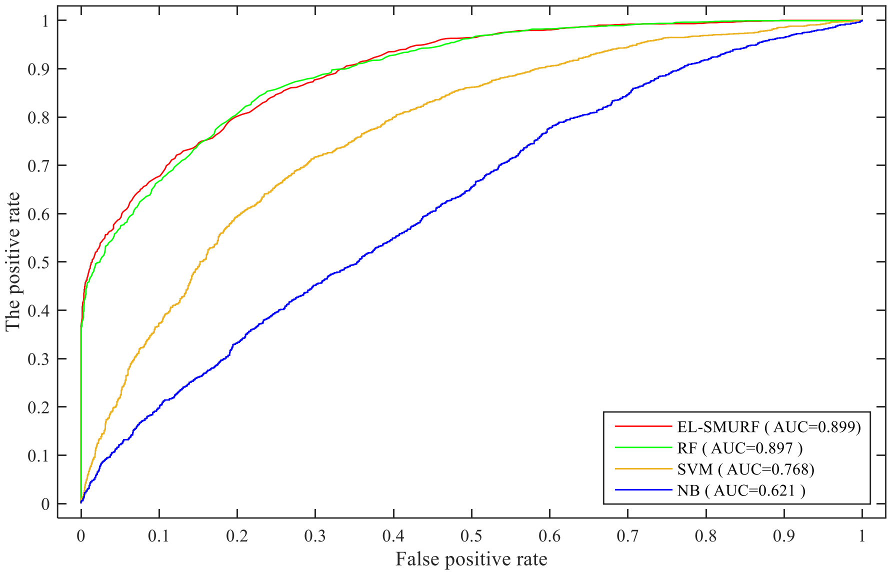

Fig. S3. Roc curves of different classifiers on Dset186. The area under the ROC curve of the EL-SMURF method proposed in this study achieved to 0.899 which is the largest. It is an effective classification prediction method. It can be seen that NB classifier have poor performance in the Dset186 dataset and cannot accurately identify protein-protein interaction sites. The EL-SMURF proposed in this study has a good performance, indicating that it has achieved good classification results in the test data Dset186.

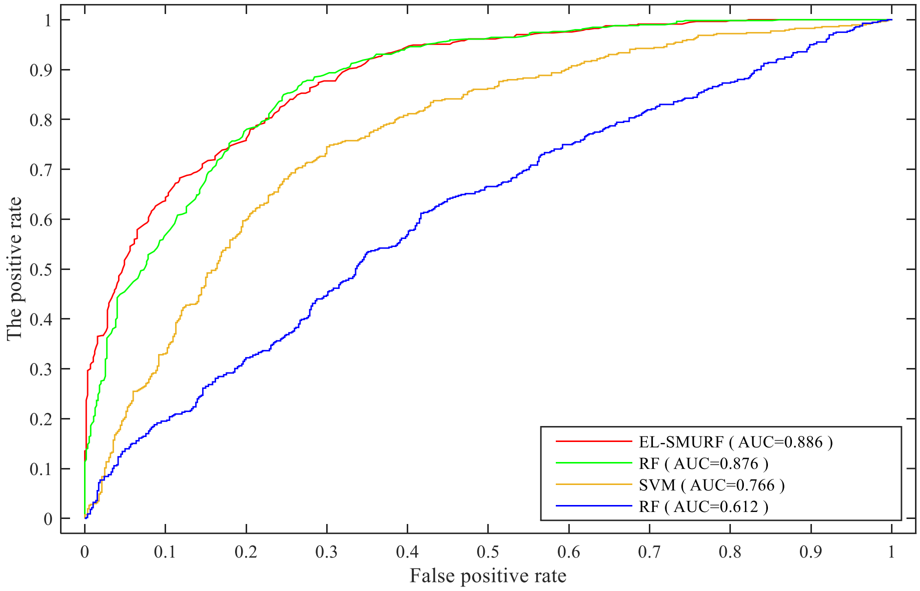

Fig. S4. Roc curves of different classifiers on Dtestset72. The area under the ROC curve of the EL-SMURF method proposed in this study is the largest. It can be seen that SVM classifier and NB classifier have poor performance in Dtestset72 and cannot accurately identify protein-protein interaction sites. The EL-SMURF method proposed in this study has a good classification effect, which shows that it has good generalization performance in the independent test set Dtestset72. The experimental result shows that EL-SMURF is an effective classification and prediction method.

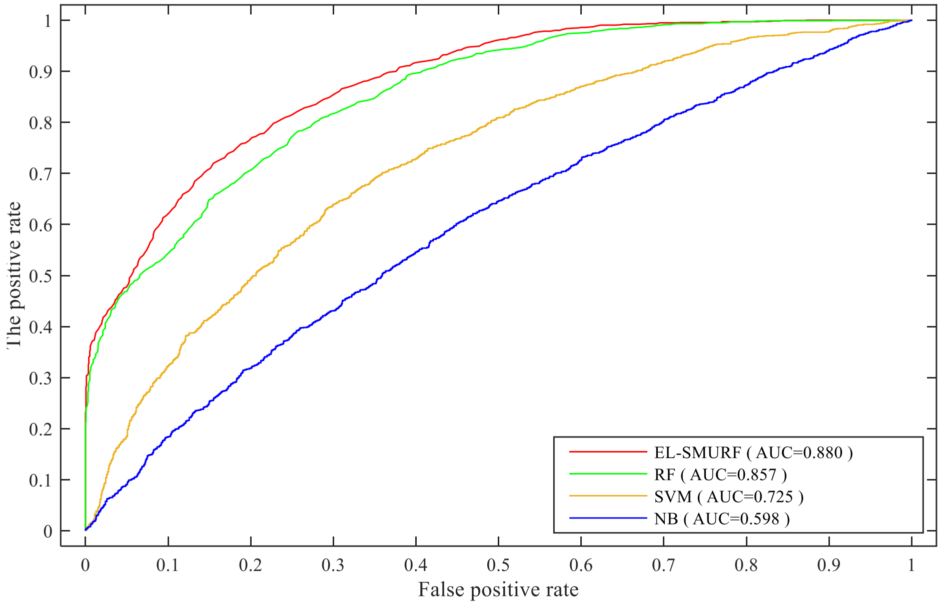

Fig. 5. Roc curves of different classifiers on PDBtestset164. For the independent validation set PDBtestset164, the ROC curve of the EL-SMURF method proposed in this study is 0.880, which has the most extensive coverage area. It is 27.8% higher than NB classifier. So, NB classifier has poor performance in the PDBtestset164 dataset and cannot accurately identify protein interaction sites. The EL-SMURF proposed in this study has a good classification effect, which shows that it has good generalization performance in the independent validation set PDBtestset164. The result indicates that EL-SMURF is a useful classification and prediction method.

4. Supplementary References

Chang, C.C. and Lin, C.J. (2011) LIBSVM: A library for support vector machines, *ACM T. Intel. Syst. Tec.*, **2**, 1-27.

Chen, Z. *et al.* (2015) Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools, and features, *Brief. Bioinform.*, **16**, 640-657.

Guo, Y. *et al.* (2008) Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, *Nucleic Acids Res.*, **36**, 3025-3030.

He, B. *et al.* (2017) NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers, *Bioinformatics*, **33**, 2296-2306.

Khan, M. *et al.* (2017) Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC, *J. Theor. Biol.*, **415**, 13-19.

Kim, J.H. *et al.* (2004) Prediction of phosphorylation sites using SVMs, *Bioinformatics*, **20**, 3179-3184.

Lin, X. and Chen, X.W. (2013) Heterogeneous data integration by tree-augmented naive Bayes for protein-protein interactions prediction, *Proteomics*, **13**, 261-268.

Mayer, J. *et al.* (2018) Sequential feature selection and inference using multi-variate random forests. *Bioinformatics*, **34**, 1336-1344.

Qiu, W. *et al.* (2018) Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, *J. Theor. Biol.*, **450**, 86-103.

Song, Q. *et al.* (2017) Combination of minimum enclosing balls classifier with SVM in coal-rock recognition, *PLoS One*, **12**, e0184834.

Vapnik and Vladimir, N. (1997) The nature of statistical learning theory, *IEEE. T. Neur. Net. Lear.*, **38**, 988-999.

Wan, S. *et al.* (2016) Mem-ADSVM: a two-layer multi-label predictor for identifying multi-functional types of membrane proteins, *J. Theor. Biol.*, **398**, 32-42.

You, Z.H. *et al.* (2015) Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest. *Plos One,* **10**, e0125811.

Yu, B. *et al.* (2017) Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, *Chem. Intell. Lab. Syst.*, **167**, 102-112.

Yu, B. *et al.* (2017b) Accurate prediction of subcellular location of apoptosis proteins combining Chou’s PseAAC and PsePSSM based on wavelet denoising, *Oncotarget*, **8**, 107640-107665.

Yu, B. *et al.* (2017c) Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising, *J. Mole. Graph. Model.*, **76**, 260-273.

Yu, D.J. *et al.* (2015) Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural Information and Random Forest Regression. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, **12**, 611.

Zhang, S.B. and Tang, Q.R. (2016) Protein-protein interaction inference based on semantic similarity of Gene Ontology terms, *J. Theor. Biol.*, **401**, 30-37.

Zamanighomi, M. *et al.* (2017) Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data. *Nucleic Acids Res.,* **45**, 5666-5677.