Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 528767, 7 pages
http://dx.doi.org/10.1155/2014/528767

Research Article

The Prediction of Protein-Protein Interaction Sites Based on RBF Classifier Improved by SMOTE

Hui Li,[1,2] Dechang Pi,[1] and Chishe Wang[2]

[1] *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Yudao Street 29, Nanjing, Jiangsu 210016, China*
[2] *Department of Information Technology, Jinling Institute of Technology, Nanjing, Jiangsu 210001, China*

Correspondence should be addressed to Hui Li; hui_li.hh@outlook.com

Received 14 April 2014; Accepted 28 May 2014; Published 30 June 2014

Academic Editor: Jingjing Zhou

Protein-protein interaction sites are the basis of biomolecule interactions, which are widely used in drug target identification and new drug discovery. Traditional site predictors of protein-protein interaction mostly based on unbalanced datasets, the classification results tend to negative class, resulting in a lower predictive accuracy for positive class. A method called RBFIS (radial basis function improved by SMOTE) is presented in the paper to address the problem. The intelligent algorithm SMOTE is used to artificially synthesize the imbalanced datasets of negative sample classes. Simultaneously, KNN algorithm is utilized to interpolate values between the minority class samples to generate new samples, making the sample data tend to balance as much as possible. Then, RBF classifier is used to construct the site predictor of protein-protein interaction based on the processed quasi-equilibrium sample sets. The results of experiments indicated that the method had an improvement on recall and f-measure of positive class compared with traditional methods by 12% and 25%. Moreover, many rounds of experiments were performed for different combinations of features. It was observed that the key combination of different multiple features can better efficiently improve the prediction performance. In conclusion, the studies we have performed show that the proposed method is better for dealing with the imbalanced protein interaction sites.

1. Introduction

With the completion of genome projects of human and other species, life science has entered the postgenomic era which focuses on the research of functional genomics. In the postgenomic era, protein as a major embodiment and executive of life activities has made the proteomics an important research field. Moreover, the protein-protein interactions have become an important research focus at the physical and structural level. As one of the major basic matters of life body, the interaction between proteins not only helps to understand life process but also has a great promoting role in the exploration of the mechanism for the treatment of various diseases, development of new drugs, discovery of drug targets, and so on. Currently, the research of protein-protein interaction is mainly based on the microlevel and macrolevel. At the microlevel, it focuses on the study of the binding site [1], that is, the functional sites (interface residues) in protein-protein interaction process. The surface residues whether they are interface residues, which are the residues that interacted, or not are analyzed. At the macrolevel, it focuses on the study of the combining objects, that is, the interactions objects between proteins and the formed interaction network.

The widely used methods for studying protein interaction are biochemical experiments. There are small-scale and high-throughput experiments according to the scale of the experiment. The traditional small-scale experiment methods mainly include plasmon resonance technology (PR), protein affinity chromatography, immunoprecipitation, nuclear magnetic resonance (NMR), X-ray crystallography technology, and two-hybrid technique. Small-scale experiments have advantages of strong purpose and high accuracy but can only detect one or several pairs of interactions between proteins in one time, which brings about time-consuming and labor-intensive result. With the development of new technologies, the high-throughput experimental techniques have emerged,

such as yeast two-hybrid screening techniques, mass spectrometry, and protein chip. The high-throughput techniques can identify more pairs of protein interactions in shorter time, but it easily makes the experimental result false positive and false negative because of the constraints of experimental conditions and data integrity. With the emergence of large numbers of protein data and the development of computing technology, more and more computational methods are used to predict protein-protein interactions, which save much time and effort and make up for the deficiencies of biological experimental methods.

The calculation methods for the prediction of protein interaction sites are primarily achieved by some machine learning classifiers [2]. The most widely used machine learning methods for the prediction of protein-protein interactions are Bayes, neural networks, and support vector machines (SVM). Bayes is a method of inference analysis based on uncertainty theory, which combines priori knowledge and new evidence collected from the data. Neuvirth et al. [3] proposed an algorithm called ProMate which can successfully predict about 70% of the interface residues of the proteins, and the success rate of the method was equal whether applied on the unbound DB or on the not-intersected bound DB. Wang et al. [4] use naive Bayes classifier to predict the interface residues of protein-protein interaction for a dataset of representative heterologous protein complexes, achieving accuracy rate of 68%, specificity of 40.2%, and sensitivity of 49.9%. Bradford et al. [5] provided a Bayesian network to predict the protein binding sites with an overall success rate of 82% on a dataset of 180 proteins which improved by 36% on a random method. Neural network is also widely used to predict the protein interface residues, which can withstand noise data and classify untrained data [6]. A consensus neural network method called cons-PPISP was presented in [7] which had better predictive power than the best individual models by 3–8 percentage points in accuracy. In [8], radial neural network is used to predict the protein interface residues, reaching accuracy of 68.9%, sensitivity of 66.6%, and specificity of 67.6%. SVM find the best hyperplane to classify the data by mapping data into a high-dimensional space using nonlinear functions. SVM algorithm has also been used to predict interaction interface sites [9–12] and has made different prediction effects on different datasets. Besides, some other machine learning methods have also been used to predict protein-protein interaction interface sites, such as hidden Markov model [13], linear regression [14], score function [15], conditional random fields [16], and random forest [17].

The common points of the previously mentioned methods use one classifier to predict protein interface residues. Recently, in order to further improve the accuracy of prediction, the researchers focused on the combination of some machine classification algorithms, as well as the selection and combination of protein structure eigenvector [18]. Murakami and Mizuguchi [19] described a method called PSIVER which applied the Naïve Bayes classifier with kernel density estimation to predict the protein-protein interaction sites. The method obtained a good prediction effect on a nonredundant set of 186 protein sequences. In [20], a model of RBF network optimized by PSO was offered and achieved good results. Although these algorithms have made some prediction accuracy, most of the results are based on the calculated mean values, and in most cases we are only concerned with the forecast results of positive samples. During the process of protein-protein interaction, the significance of the predicted interaction sites is far greater than the predicted noninteraction sites.

The prediction of protein interaction sites is mostly based on large imbalance datasets. The minority class that is positive is far less than the majority class that is negative. The traditional machine classification methods deal with unbalanced problems in favor of majority class, resulting in a lower accuracy of minority class which is usually the more important classes. For imbalanced datasets, the commonly used method is resampling, which includes oversampling that artificially increases the samples of minority class, and undersampling that artificially reduces the samples of majority class. SMOTE (synthetic minority oversampling technique) [21] is a very popular oversampling method in which the positive class is oversampled in random and has been applied in classification problems combined with classification algorithms [22]. The prediction of protein interaction sites is also a two-class imbalanced problem. In order to improve the accuracy of positive class, SMOTE combined with KNN (K-nearest neighbor) algorithm is used to process the unbalanced training datasets of protein binding sites in the paper. Then, multiple features of protein are selected to construct feature vectors, and the interaction sites are classified by radial basis function (RBF) model. Finally, the results of the experiments show that there are some improvements on the forecast effects of positive class.

The rest of the paper is organized as follows. Section 2 introduces the datasets and our methods. In Section 3, some supporting simulation results and discussion are presented. Finally, conclusions are drawn in Section 4.

2. Datasets and Methods

2.1. Datasets. Here we selected the experimental datasets from the paper of Wang et al. [4], which had strong representation in the prediction of protein binding sites and also had been used by many other researchers. The datasets included 70 heterologous protein complexes as the training data, selected from eight different kinds of protein complexes. The datasets were filtered and then aligned using PSI-BLAST program to exclude the protein chains whose sequence consistency exceeded 25% and sequence length more than 90%. Finally we get the research datasets of 90 protein chains from 57 compounds, as shown in Table 1.

2.2. Surface Residues and Interface Residues. We defined the surface residues from the datasets of protein chains using the common definition way to produce the experimental sample datasets, and then the interface residues from surface residues were defined.

Define 1. If the ratio of ASA (solvent-accessible surface areas) in a single-chain to the largest ASA (as shown in Table 2) of the residue exceeds a certain value (here we choose 0.25),

Table 1: Dataset of protein.

1A0O_A	1A0O_B	1A2K_A	1A2K_C	1AGR_A	1AGR_E	1AIP_C	1AK4_A	1AK4_C	1AO7_A
1AO7_B	1AO7_D	1ATN_A	1ATN_D	1AVW_B	1BRS_A	1BRS_D	1BTH_I	1CHO_F	1CHO_G
1CSE_E	1CSE_I	1DAN_I	1DAN_T	1DAN_U	1DFJ_E	1DFJ_I	1DHK_A	1DHK_B	1DKG_A
1DKG_D	1EFN_A	1EFN_B	1EFU_A	1EFU_B	1FC2_C	1FC2_D	1FIN_A	1FIN_B	1FLE_I
1FSS_A	1FSS_B	1GG2_B	1GG2_G	1GLA_F	1GLA_G	1GOT_G	1GUA_A	1GUA_B	1HIA_A
1HIA_I	1HWG_A	1HWG_B	1IGC_ A	1JHL_A	1JHL_L	1KB5_B	1MCT_I	1MEL_A	1NCA_N
1NFD_B	1NSN_L	1NSN_S	1OSP_O	1PPF_I	1QFU_A	1QFU_B	1SEB_D	1STF_E	1STF_I
1TGS_I	1TX4_A	1UDI_E	1UDI_I	1VFB_B	1YCS_A	1YCS_B	1YDR_E	2BTF_P	2JEL_P
2PCC_A	2PCC_B	2PTC_I	2SIC_I	2SNI_I	2TRC_P	3SGB_E	4CPA_A	4CPA_I	4HTC_I

Table 2: The nominal maximum area for each type of residue.

Type of residue	A	B	C	D	E	F	G	H	I	K	L	M
ASA (Å^2)	106	160	135	163	194	197	84	184	169	205	164	188
Type of residue	N	P	Q	R	S	T	V	W	X	Y	Z	
ASA (Å^2)	157	136	198	248	130	142	142	227	180	222	196	

Note: amino acid type is expressed by a single letter, B represents D or N, Z represents E or Q, and X represents an unknown amino acid type.

then the residue is defined as a surface residue; otherwise it is defined as nonsurface residues.

Define 2. If the absolute value of the difference of ASA in a protein single-chain (MASA) with ASA in the protein complex (CASA) is greater than 1, the interface is defined as the interface residue; otherwise it is defined as noninterface residues.

By using the definition above, the entire datasets consisted of 16 325 residues, including 8266 surface residues, accounting for 50.6% of the total residues, and including 2476 interface residues, accounting for 15.2% of the total residues and 30% of the surface residues.

2.3. Feature Attributes Selection. Protein has many different biochemical characteristics; how to choose the appropriate biochemical properties is key to predict the protein interaction sites. For the moment, most researchers predicted the interaction sites based on sequence features [19] or structural characteristics [23]; also some others used a combination of characteristics to predict [17, 24, 25]. In order to improve the accuracy of prediction, the paper constructed combined features to form the feature vectors based on key structure and sequence feature attributes.

The feature attributes used in the paper were sequence profiles, ASA, entropy, and weight. Sequence profile is a 20-dimensional vector, for each dimension vector representing related frequency of an amino acid type at each position. Also, sequence profile expresses the evolutionary relationship of protein system effectively, which was extracted from the hssp files of HSSP (homology-derived secondary structure of proteins) database. ASA belongs to the structural characteristic of proteins, which was obtained from DSSP program of protein single-chain. And interface residues have a greater value of ASA than surface residues in general. Entropy is a measure of the sequence change rate on one position, whose value belonged to 0 to 1 normalized by

$$\text{Entropy} = -\frac{1}{\ln d}\sum_{i=1}^{d} f_i \ln f_i, \tag{1}$$

wherein f_i indicates occurrence times of the ith amino acid compared to other 20 kinds of amino acids at the same position of the protein sequence and d represents the kind of amino acids; here we choose 20. Weight is a measure of the sequence conservation on one position, whose value is between 0 and 1.

2.4. Hybrid Method. The widely used calculation methods for the sites prediction of protein-protein interaction are some machine learning classifiers. RBF is a commonly used machine learning classifier and is able to deal with nonlinear changes, whose structure is simple, can withstand noise data, and has the classification ability of untrained data. Compared with other neural networks, RBF has the advantages of fast convergence and approximation of any nonlinear function. However, in the process of protein-protein interaction, the number of the interface residues that really interact is little, which leads to the imbalance of positive and negative data, that is, imbalanced two-class problem [22]. In this case, it will make the predictions lead to high precision and low recall [12]. The usual approach is to take the same number of noninterface residues as negative datasets, but this will cause information loss of some samples. Therefore, we retained all the negative sample datasets in the paper and oversampled the positive sample datasets using intelligent SMOTE algorithm. The specific process is as follows. Firstly, analyze the PDB source data of proteins to extract the experimental datasets. Secondly, select the feature attributes to vectorize them to form the sample data, which were balanced using SMOTE algorithm, as the input parameters of RBF model. Then, build the model and train the model by constantly adjusting parameters until it met the requirements. Finally use the model to predict the protein functional sites. Figure 1 shows the specific process.

2.4.1. SMOTE. SMOTE [26] intelligent oversampling algorithm achieved balanced sample data through synthesizing the samples of the new minority class, rather than simply copying the minority class data. The basic principle was

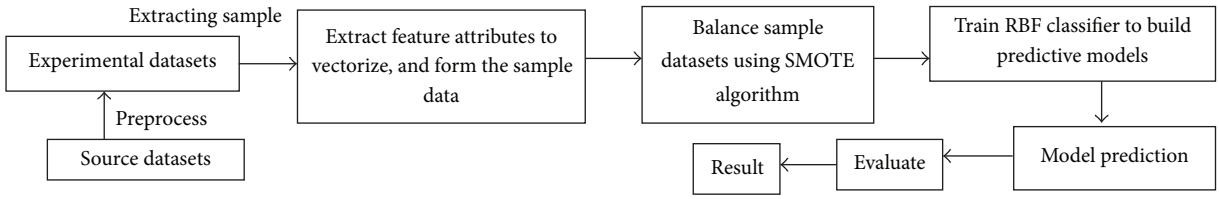

Figure 1: The prediction process of protein interaction sites using RBF improved by SMOTE.

the linear interpolation between the samples of minority class with close proximity and then generation of a new minority class sample. Assume that minority class sample dataset is *Sample*, oversampling rate is N, and the nearest neighbor point is K. The main ideas are as follows.

(1) According to certain rules, calculate K congener nearest neighbor samples for each minority class sample i in the sample datasets of minority class.

(2) Select randomly N samples from each nearest neighbor samples, respectively.

(3) Calculate a new sample according to formula (2) from the minority class sample and each sample of N, that is, the synthetic new sample, and then add to the sample datasets of the minority class:

$$SynthetiSample\,[newindex] = Sample\,[i] + \text{rand} * (NeighborSample\,[i] - Sample\,[i]), \quad (2)$$

wherein *SyntheticSample* is synthetic new samples, *Sample* $[i]$ represents ith sample of sample datasets of minority class, $i = 1, 2, \ldots, T$, T is the number of samples of minority class, rand is a random number between $[0, 1]$, and *NeighborSample* $[i]$ represents a random nearest neighbor sample of sample i.

(4) Continue the process until all of the minority class samples were processed and then ended.

2.4.2. RBF. RBF is able to approximate the performance of nonlinear network, which possesses simple network structure and fast convergence, so it was used to predict the protein interaction sites in the paper. RBF is a feed forward network composed of three layers: the first layer is the input layer, and the number of nodes equals the dimension of the input data; the second layer is the hidden layer, and the number of nodes is determined according to the size of the problem; the third layer is the output layer, and the number of nodes equals the dimension of the output data, wherein the hidden layer is nonlinear and the output layer is linear. The topology is shown in Figure 2.

In the radial network, the parameters that need to be trained are the center X_i of radial function in hidden layer, the standard deviation σ of radial function in hidden layer, and the weights w_{ij} between the hidden layer and output layer. Typically, the right weights between input and hidden layers

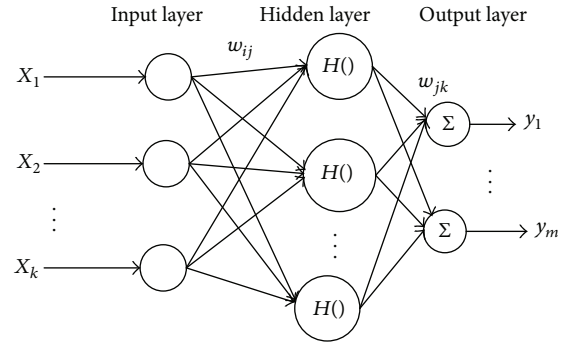

Figure 2: The structure of RBF.

are set to 1. In our work, the radial basis function is defined as follows:

$$\Phi_i(X) = \phi(X, X_i) = \phi(\|X - X_i\|) = \exp\left(-\frac{\|X - X_i\|^2}{2\sigma^2}\right), \quad i = 1, 2, \ldots, N. \quad (3)$$

$\Phi_i(X)$ is the output of the ith node of hidden layer, X is the input vector, X_i is the center of basic functions, the standard deviation is σ, and N is the number of hidden layer nodes. Among them, the center X is calculated using K-means clustering algorithm, and σ is calculated using formula (4). In (4), $d_{\max}$ is the maximum distance between the selected cluster centers:

$$\sigma = \frac{d_{\max}}{\sqrt{2N}}. \quad (4)$$

When the input training sample is X_k, the output result of jth neuron in the network is y_{kj} as shown in

$$y_{kj} = \sum_{i=1}^{N} w_{ij}\phi(X, X_i), \quad j = 1, 2 \ldots J. \quad (5)$$

In (5), w_{ij} is the connection weights between the hidden layer and output layer, J is the number of nodes in the output layer.

2.5. Evaluation Measures. In order to better measure the performance of our classifier, some evaluation indexes shown below were used to evaluate the classifier. TP represents the number of interface residues correctly predicted; TN represents the number of noninterface residues correctly

Table 3: Different values of corresponding evaluation index for different *Percentage* values.

Percentage	Average recall (%)	Average precision (%)	CC (%)	f-ave (%)	AUCC (%)
160%	56.56	55.67	12.13	56.11	58.90
140%	56.71	56.12	12.68	56.41	58.97
120%	58.79	57.83	13.68	57.92	60.50
100%	55.70	56.58	11.97	56.12	58.70
80%	55.21	55.77	12.25	55.21	58.40

predicted; FP represents the number of interface residues mistakenly predicted, which in fact are noninterface residues; FN represents the number of noninterface residues mistakenly predicted, which in fact are interface residues:

$$\text{Recall}^{+} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Precision}^{+} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$\text{Recall}^{-} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \text{Precision}^{-} = \frac{\text{TN}}{\text{TN} + \text{FN}},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{CC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FN})\,(\text{TP} + \text{FP})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}}, \quad (6)$$

$$f\text{-measure} = \frac{2 * (\text{precision} * \text{recall})}{(\text{precision} + \text{recall})},$$

$$\text{Ave-recall} = \frac{(\text{recall}^{+} + \text{recall}^{-})}{2},$$

$$\text{Ave-Precision} = \frac{(\text{Precision}^{+} + \text{Precision}^{-})}{2}.$$

3. Results and Discussion

The LOOCV (leave-one out cross-validation) was used in the study to predict the interface residues of 90 protein chains. For each round of the experiments, we chose a chain as the test data; the other 89 chains as the training data, and the process was done repeatedly 90 times; and the average result from 90-time experiments as the final result of this round. In order to select the best oversampling rate and multiple combined characteristics, we did many rounds of experiments using different values. In order to choose the most appropriate oversampling rate, here we did five rounds of experiments.

3.1. Oversampling Rate. The over-sampling rate was expressed by the variable *Percentage*. If the value of *Percentage* is larger, the result tends to in favor of positive cases. When the value was large to a certain extent, the negative samples cannot be predicted or the values of some basic evaluation indexes were abnormal. Here by experiments, the upper limit of *Percentage* is set to 180%. If the value of *Percentage* is smaller, the result tends to in favor of negative cases, which is not beneficial to predict positive class. Here the lower limit of *Percentage* was set to 80%. The prediction test was done many times for different values of 20% increments between the upper limit and the lower limit, and protein sequence spectral and solvent-accessible surface area were took as the feature vectors.

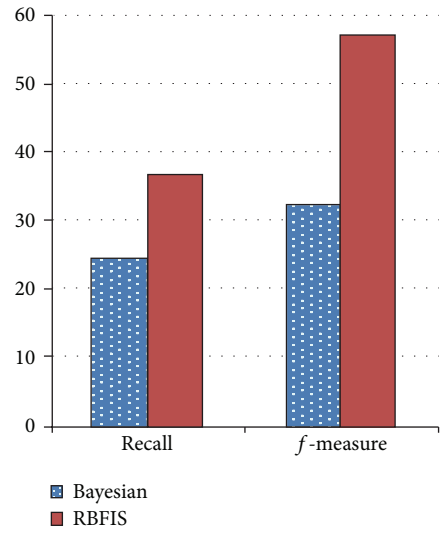

Figure 3: Performance comparison between methods of RBFIS and Bayes.

As can be seen from Table 3, when the value of the oversampling ratio *Percentage* is 120%, the evaluation indexes values of recall, precision, CC, f-measure, and AUCC were better than other situations.

The literature [4] achieved an accuracy of 68.1%, recall of 40.2%, and precision of 49.9%, where recall and precision were the average values of the positive and negative samples. However, the researchers only concerned the accuracy rate of positive sample during the process of sites prediction in most cases. The RBFIS we proposed in the paper for the prediction of the interaction sites was compared with the Bayesian method in [4]. The value of recall and f-measure in our method was improved greatly shown in Figure 3. The values of other evaluation indexes were close to each other, which were not listed here.

As shown in Figure 3, compared with Bayesian method, the recall increased by nearly 12%, and f-measure increased by 25% when using RBFIS to predict interaction sites.

3.2. Multifeature Combination. In order to improve the prediction performance for different combinations of feature vectors, many rounds of experiments were done. When using a combination of features for profile, ASA, entropy, and weight, we get a better evaluation result, as shown in Table 4.

Table 4: Performance index values when using different combinations of features.

Features	Recall$^+$ (%)	Precision$^+$ (%)	CC (%)	AUCC (%)
Profile + ASA	36.70	44.42	12.28	56.2
Profile + ASA + entropy + weight	38.90	44.41	13.76	60.10

Compared with the literature [4] selected combination of features (profile + ASA), recall of positive class increased nearly by 2%, CC increased nearly by 2%, and AUCC increased nearly by 4%, where the *Percentage* value is 120%.

4. Conclusions

With life sciences entering the postgenomic era which is marked by functional genomics, proteins expressed by the gene become a research focus. As the datasets handled by traditional forecasting methods mostly are unbalanced datasets. This makes the results tend to in favour of negative class (majority class), which leads to lower classification accuracy of the positive class (minority class). In this paper, RBF classifier which can approximate nonlinear network was used to construct the interaction sites predictor of protein-protein interaction. Meanwhile, the intelligent algorithm SMOTE was used to artificially synthesize the imbalance data of minority class, and KNN was utilized to interpolate values between the minority class samples to generate new samples in order to achieve data balance. Finally, many rounds of different intersected experiments were done by employing LOOCV method, and the appropriate value of the oversampling ratio was selected to predict the interaction sites, which was greatly improved on the value of f-measure and recall of positive class. In addition, many different experiments were performed for different combinations of features, and the experimental results showed that effective multifeature combinations can better improve the prediction performance. In the future, swarm intelligence algorithms are used to optimize the classification structure to further improve the forecasting results, which is the next research focus.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (NZ2013306), Qing Lan Project, the 333 Project of Jiangsu Province and the Technology Foundation (JSJC2013605C009), Universities Natural Science Research Project of Jiangsu province (13KJD520005), and Modern Educational Technology in Jiangsu Province (2013-R-26144).

References

[1] H.-X. Zhou and S. Qin, "Interaction-site prediction for protein complexes: a critical assessment," *Bioinformatics*, vol. 23, no. 17, pp. 2203–2209, 2007.

[2] B. Wang, W. Sun, J. Zhang, and P. Chen, "Current status of machine learning-based methods for identifying protein-protein interaction sites," *Current Bioinformatics*, vol. 8, no. 2, pp. 177–182, 2013.

[3] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: a structure based prediction program to identify the location of protein-protein binding sites," *Journal of Molecular Biology*, vol. 338, no. 1, pp. 181–199, 2004.

[4] C. Wang, J. Cheng, S. Su et al., "Identification of interface residues involved in protein-protein interactions using Naïve Bayes classifier," *Advanced Data Mining and Applications*, vol. 3, no. 3, pp. 293–302, 2009.

[5] J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, "Insights into protein-protein interfaces using a Bayesian network prediction method," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 365–386, 2006.

[6] A. Porollo and J. Meller, "Prediction-based fingerprints of protein-protein interactions," *Proteins: Structure, Function and Genetics*, vol. 66, no. 3, pp. 630–645, 2007.

[7] H. Chen and H. Zhou, "Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data," *Proteins*, vol. 61, no. 1, pp. 21–35, 2005.

[8] B. Wang, P. Chen, P. Wang, G. Zhao, and X. Zhang, "Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes," *Protein and Peptide Letters*, vol. 17, no. 9, pp. 1111–1116, 2010.

[9] J. Chung, W. Wang, and P. E. Bourne, "Exploiting sequence and structure homologs to identify protein-protein binding sites," *Proteins: Structure, Function and Genetics*, vol. 62, no. 3, pp. 630–640, 2006.

[10] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.

[11] Q. Dong, X. Wang, L. Lin, and Y. Guan, "Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins," *BMC Bioinformatics*, vol. 8, article 147, 2007.

[12] A. Koike and T. Takagi, "Prediction of protein-protein interaction sites using support vector machines," *Protein Engineering, Design and Selection*, vol. 17, no. 2, pp. 165–173, 2004.

[13] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, "Modelling interaction sites in protein domains with interaction profile hidden Markov models," *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.

[14] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, and R. Abagyan, "PIER: protein interface recognition for structural proteomics," *Proteins: Structure, Function and Genetics*, vol. 67, no. 2, pp. 400–417, 2007.

[15] N. J. Burgoyne and R. M. Jackson, "Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces," *Bioinformatics*, vol. 22, no. 11, pp. 1335–1342, 2006.

[16] M.-H. Li, L. Lin, X.-L. Wang, and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, no. 5, pp. 597–604, 2007.

[17] M. Šikić, S. Tomić, and K. Vlahoviček, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLOS Computational Biology*, vol. 5, no. 1, pp. 1–5, 2009.

[18] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress, "Progress and challenges in predicting protein-protein interaction sites," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 233–246, 2009.

[19] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, Article ID btq302, pp. 1841–1848, 2010.

[20] Y. Chen, J. Xu, B. Yang, Y. Zhao, and W. He, "A novel method for prediction of protein interaction sites based on integrated RBF neural networks," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 402–407, 2012.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011.

[23] B. Nisius, F. Sha, and H. Gohlke, "Structure-based computational analysis of protein binding sites for function and druggability prediction," *Journal of Biotechnology*, vol. 159, no. 3, pp. 123–134, 2012.

[24] M. M. Gromiha, N. Saranya, S. Selvaraj, B. Jayaram, and K. Fukui, "Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes," *Proteome Science*, vol. 9, supplement 1, article S13, 2011.

[25] L. Wang, Z. Liu, X. Zhang, and L. Chen, "Prediction of hot spots in protein interfaces using a random forest model with hybrid features," *Protein Engineering, Design and Selection*, vol. 25, no. 3, pp. 119–126, 2012.

[26] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RSB∗: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory," *Knowledge and Information Systems*, vol. 33, no. 2, pp. 245–265, 2012.