# Imbalance Data Processing Strategy for Protein Interaction Sites Prediction

Bing Wang, *Senior Member, IEEE*, Changqing Mei, Yuanyuan Wang, Yuming Zhou, Mu-Tian Cheng, Chun-Hou Zheng, Lei Wang, Jun Zhang, Peng Chen, Yan Xiong

**Abstract**—Protein-protein interactions play essential roles in various biological progresses. Identifying protein interaction sites can facilitate researchers to understand life activities and therefore will be helpful for drug design. However, the number of experimental determined protein interaction sites is far less than that of protein sites in protein-protein interaction or protein complexes. Therefore, the negative and positive samples are usually imbalanced, which is common but bring result bias on the prediction of protein interaction sites by computational approaches. In this work, we presented three imbalance data processing strategies to reconstruct the original dataset, and then extracted protein features from the evolutionary conservation of amino acids to build a predictor for identification of protein interaction sites. On a dataset with 10,430 surface residues but only 2,299 interface residues, the imbalance dataset processing strategies can obviously reduce the prediction bias, and therefore improve the prediction performance of protein interaction sites. The experimental results show that our prediction models can achieve a better prediction performance, such as a prediction accuracy of 0.758, or a high F-measure of 0.737, which demonstrated the effectiveness of our method.

**Index Terms**—- protein interaction sites, imbalanced data, conservative features, prediction performance, prediction bias.

## 1 Introduction

It is well known that the activity of life is mainly involved by proteins, and the protein-protein interaction (PPI) involves various life activities such as metabolism and signal transduction, gene transcription, protein translation, modification and localization, and is also closely related to disease production[1], [2], [3], [4]. But PPI varies from cell to cell, and from time to time within the cell, which poses a challenge for researchers. Experiment techniques, such as yeast two-hybrid systems, affinity purification, and protein fragment complementation assays, had been developed to infer physical interactions among different proteins. However, experimental methods are time-consuming, laborious and prone to the influence of experiment environment, which leads to the residues involved in protein interactions are still mostly unknown [5],[6],[7],[8],[9],[10],[11],[12]. Computational approaches therefore had become an important way to identify protein-protein interaction sites [13],[14],[15],[16],[17],[18], [19], [20].

With the development of machine learning methods, many classical methods have been used to predict protein interaction sites [21],[22],[23],[24],[25]. Fariselli et al. extracted protein interaction related features from the three-dimensional structures of protein complexes, and proposed a neural network based system to protein interaction sites from heterodimers [26]. Wang et al. presented a support vector machine (SVM) based algorithm to identify protein-protein interactions sites on the residues level by incorporating residues spatial sequence profile and evolution rate [27]. Iqbal et al. improved a likelihood ratio-based classifier by combining with a rule induction method for PPI prediction, and the overall performance of the classifier has been significantly improved [28]. Oh et al. applied a two-stage template-based ligand binding site prediction method, by which the template was used to model the protein structure, and then a ligand binding site prediction model was built by the clustering of the structure containing the ligand template [29]. Liu et al. introduced a hybrid feature selection system, where an mRMR filter was followed by a kNNs package to predict protein interaction sites where a PseAA (pseudo amino acid) composition was adopted to encode protein pairs and achieved high quality experimental results [30]. Chen et al. proposed a radial basis function neural networks optimized by the particle swarm optimization algorithm to predict protein interaction sites [1]. Chen and Jeong

- B. Wang is with the School of Electrical and Information Engineering, Anhui University of Technology, Maanshan, Anhui 243002, and Co-Innovation Center for Information Supply & Assurance Technology, Anhui University, Hefei, Anhui 230601, China. Email: wangbing@ustc.edu.
- C. Mei, Y. Wang, Y. Zhou and M.-T. Cheng are with the School of Electrical and Information Engineering, Anhui University of Technology, Maanshan, Anhui 243002, China.
- C.-H. Zheng is with the Computer Science & Technology, Anhui University, Hefei, Anhui 230601, China.
- L. Wang is with the College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, Hunan 410022, China.
- J. Zhang is with the School of Electrical and Antomation, Anhui University, Hefei, Anhui 230601, China.
- P. Chen is with the Institute of Health Sciences, Anhui University, Hefei, Anhui 230601, China. Email: pchen@ahu.edu.cn (Corresponding author)
- Y. Xiong is with the School of Computer Science and Technology, University of Science & Technology, Hefei, Anhui 230026, China. Email: yxiong@ustc.edu.cn (Corresponding author)

developed a random forest-based integrative model for identifying protein interaction sites without any protein structural information [31]. Northey et al. introduced another random forest based predictor, IntPred, which can predict protein-protein interfaces from structural features [32].

Although those approaches has made great progress in PPI studies, there are still suffer from a problem that imbalance data distribution of positives and negatives, which is caused by the current definitions of protein interaction sites in bioinformatics. Current definitions of interaction site are mainly based on the distance between the target residue and the residues on the partner protein chains in protein complexes, or the change of accessible surface area (ASA) of the target residue when the protein chain it belongs to was involved into the formation of protein complexes. Both of them are using threshold based discrimination rules to decide which residue is interaction or non-interaction sites, which inevitable bring some mistakes into data collection because it is already found that the residues are contributed differently to protein interface formation, and some residues have relatively low contribution even they are located in the interface. Furthermore, many protein interactions have not yet been known, and which make the number of interaction sites are usually smaller than that of non-interaction sites. Generally, computational methods based PPI studies convert the prediction of protein interaction sites into a problem of classification, i.e., classifying the sites in protein sequences into positives which means the target sites are interaction sites, otherwise, negatives which are non-interaction sites. The data imbalance between positives and negatives within the dataset will cause the incorrect prediction of protein interaction sites. A few works had been aware of this imbalance problem in protein interaction sites prediction. Chen and Jeong constructed many decision trees where each of them was built by randomly selection of the same number of samples for interaction and non-interaction sites, and then converted the imbalance data problem into many balanced data classification problems [31]. Wang et al. implemented a dataset reconstruction strategy by using manifold learning under a hypothesis that the interaction and non-interaction sites have different inherent structure manifolds [33]. Several existing works adopted similar data processing methods as what Chen and Jeong did [31]. However, imbalanced-data problem in protein interaction sites prediction is still in its infancy.

This paper focuses on the extraction of related feature attributes and solves the problem of data imbalance to improve the performance of protein interaction sites prediction. Firstly, five evolutionary conserved features of amino acids based on multiple sequence alignments are extracted, i.e., the spatial sequence of residues, the sequence information entropy, relative entropy of residues, the sequence of residues, conservative weights, and evolutionary rate. Then three methods are proposed to deal with the problem of imbalanced data sets in protein interaction prediction, which are down-sampling method based on nearest neighbor rule (ENNS), sampling method based on nearest neighbor rule and boundary noise factor (ENNB), and Cluster-based sub-sampling method (KCS). Finally, a support vector machines was used to build a predictor to infer the interface residues. The experimental results show that our proposed method can achieve a better prediction performance, such as a prediction accuracy of 0.758.

## 2 Methods

### 2.1 Dataset

Our experimental dataset is derived from datasets used by Ansari and Helms et al. [34]. This dataset contains 170 transient protein interaction pairs and does not include antibody-antigen interactions. In order to improve the data quality, the protein chains with less than 50 residues are dropped out, and for multiple interacting partners with the same chain, only the chain with most interface residues is remained as the representative. Moreover, the protein chains with a sequence similarity of more than 30% was excluded using BLASTCLUST program to remove the possible redundancy in the dataset [35]. To ensure the effectiveness of the evolutionary information, some obsolete protein chains with small family are also discarded. Finally, 91 non-redundant protein chains are used for this study.

The definition of interaction sites in this work is same as what Fariselli et al. did in their work [26]. A residue can be treated as surface residues if amino acid residues with a relative accessible surface area of at least 16% of their maximum accessible surface area after amino acid dehydration condensation when they are involved into the complexes formation. Of the surface residues, a residue having a distance between the two residue alpha carbon atoms in interacting chains of less than 1.2 nm is defined as an interface residue, otherwise, a non-interface residue.

In this way, 91 protein chains can yield 10,430 surface residues, which formed the original residue dataset D in this work, where 2,299 of them are interface residues and 8,131 are non-interface residues. It can be found that D is an obvious imbalance dataset, and positives are only 22% of it. Therefore, an imbalance processing strategy should be considered in prediction of protein interaction sites to void the possible bias in prediction results.

### 2.2 Feature extraction

It is well known that protein interfaces are subject to strong evolutionary constraints for they are crucial for the proper functions of cell can be performed, and there are many previous studies can infer protein interaction sites or interfaces successfully for protein evolutionary information. Therefore, five evolutionary conservation features of amino acids, i.e., residue spatial sequence profile, sequence information entropy, relative entropy, residue sequence weight and residue conservative fraction, are extracted from proteins for vectorization of each residue in the dataset D. The first four features are extracted from the HSSP database, and the fifth one from Consurf Serve.

The spatial sequence profile of amino acid residues indicates the frequency of various amino acids at a given residue position in the primary structure of the protein, which can be achieved from multiple sequence alignment (MSA), and is a frequently used feature in protein structure and/or function-related analysis.

The protein residue sequence entropy can estimate the conservativeness of sequence variability based on Shannon information theory, which is a numerical score can refer to uncertainty of a given protein residues position. The relative entropy is the normalization of the sequence information entropy that can incorporate the observed background frequency from protein sequence databases into the conservation judgment of the target residue. The conserved weight of the residue sequence is a conservative calculation of the position of the protein sequence that is the simplest measure in HSSP database to describe the sequence variation.

The evolutionary rate of residues is a measure which can calculate the conservation of each amino acid position in a statistical point of view, which can be extracted from ConSurf Server. This measure implements the Rate4Site algorithm to calculate the maximum likelihood estimate of the evolution rate from the topology and branch lengths of the phylogenetic tree.

## 2.4 Feature encoding scheme

After the above five features are extracted from HSSP database and ConSurf Server, each amino acid position in protein surface can be represented as a vector of 24 dimensions, of which twenty dimensions are from the spatial sequence profile, and each of other four dimensions from other corresponding feature. However, protein interaction usually works by one or more interfaces, and each interface are formed by the composition of forces of residues within the interface, although the contributions of interface residues for the formation of interface are different. Therefore, the vector of single residue cannot make full advantage of protein interaction information.

In this work, a target residue $i$ is replaced by a local environment in a protein surface because the associations among the neighboring residues can represent the interaction information more effectively. Herein, a sliding window, such as of length 11, centered at the residue $i$ and its 10 nearest neighbors in protein surface is used to encode the feature vector for this residue. These 10 residues can be treated a local interface around the target residue $i$, and each of them is also vectorized as a 24 dimensional feature like what the target residue does. Finally, a 264-dimensional vector is obtained for each residue, and will be used for later predictor construction.

## 2.3 imbalanced-data processing

Obviously, the dataset used in this work is imbalanced, i.e., the positives are only 22% of the samples in whole dataset D. Actually, this kind imbalance is common in current protein interaction sites studies because there is only small part of protein interactions have been confirmed by experiments. Moreover, there are some restrictions on computational methods-based protein interaction studies, such as some kinds of interactions with high tendentious characteristics or the protein chains with higher homology generally should be removed from dataset for avoiding predictive bias. Therefore, if a predictor is built based on conventional machine learning algorithm, it will cause a high bias to the negative samples or inaccurate prediction.

In this work, SVM, a popular and effective machine learning algorithm is used to build the predictor which discriminate the interaction sites from protein surface. To address the problem of data imbalance in our dataset D, three data sampling strategies are developed for prediction performance comparison. The first one is data sampling methods based on nearest neighbor rule (ENNS), the second is nearest neighbor rule and boundary noise factor (ENNB), and the third method is k-means clustering (KCS), have been proposed for data rebalance to improve protein interaction sites prediction.

### 2.3.1 Edited Nearest Neighbor Sampling

To re-balance the number of samples among different classes in the original dataset, random sampling method selects the samples randomly from sub-dataset where the class with bigger sample size. But this kind of sampling approach assigns all of the samples in a given population with a given equal opportunity, and does not consider the data distribution of sample, which will cause some important sample information lost during sampling.

Therefore, Wilson et al. proposed an Edited Nearest Neighbor data sampling method (ENNS) to take advantage of near neighbor information. For each data point in the dataset, this approach screens the class information of the three nearest neighbors, and if the category information of two or more of these three neighbors' classes is inconsistent with the target data point, the target data point will be removed. This nearest neighbor rule can reasonably cut down the samples with suspected class label, and therefore improve the data quality and reduce the computational cost.

Fig.1 shows how the inconsistent samples are removed, where the orange and green balls represent majority and minority class samples, dotted balls represent deleted samples based on ENNS method. It can be found from Fig.1 that some data points near the green ones are removed from the original dataset, and thus the number of samples within the bigger sub-dataset will be reduced and a new dataset with better balance will be obtained.

ENNS data sampling approach can effectively treat the data noise within the bigger sub-dataset. However, the distributions of data with different labels are generally different, and most samples normally have the same labels with their neighbors, so the samples deleted by this method are very limited, and sometimes it cannot rebalance the original dataset effectively.

**Algorithm 1** ENNS

```
Input: The original data set D.
       $X_i$ is the sample in D.

For i=1,2,…,n
    a. Calculate the Euclidean distances between $X_i$
       and other samples in D.
    b. Get the category information of three sam-
```

```
        ples closest to $X_i$.
    c.  If two or more nearest samples' labels are
        different from $X_i$, $X_i$ is removed from D.
END
Output: The balanced data set $D_S$.
```

### 2.3.2 Nearest Neighbor Sampling with Boundary Noise Factor

In many situations, there are some observations that are far from other observations with same labels, which are called outliers in statistics. An outlier can cause serious problems in statistical analyses and therefore mislead data interpretation because outliers within dataset will cause high skewness of data distribution. Thus, outliers with in original dataset should be removed in data analysis to eliminate the influence of these far-away samples.

To address this problem, Yang proposed a parameter, boundary noise factor, based on ENNS method, where outlier detection techniques were applied to handle the sample overlap problem. In the original dataset, each sample was assigned with an isolation degree which can measure the distance between itself and other samples within the dataset [36]. Based on this idea, a boundary noise sample can be found and removed based on boundary noise factor.

For each sample $x_i$ in the dataset $\mathbb{D}$, the nearest neighbor samples $K_D$ can be found using the Euclidean distance between $x_i$ and the other samples within D based on ENNS. Then an over-dimensional sphere $\Theta_D$ can be constructed by taking the biggest distance as radius and $x_i$ as the center. Within $\Theta_D$, all samples with the same class label as $x_i$ are defined as KNS($x_i$), and others as KND($x_i$). Then, the boundary noise factor of any sample data in data set $\mathbb{D}$ is defined as:

$$BNF(x_i) = \alpha\left(\frac{K_D+\delta}{|KNS(x_i)+\delta|}\right) + \beta|KND(x_i)| \tag{1}$$

where $|\text{KNS}(x_i)|$ and $|\text{KND}(x_i)|$ are the size of KNS($x_i$) and KND($x_i$), respectively.

This method applies the nearest neighbor rule and boundary noise factor (ENNB) to sampling, and it can be implemented by several steps. First, the nearest-neighbor rule is used to remove samples labeled majority category that are far from the boundary, and all samples with other labels are retained. We keep samples in which the majority of sample classes in k neighbors are not the same as themselves, and the corresponding sample of the remaining class is removed. Then set an empty set P, and define the data set of the previously removed majority class samples as C; calculate the Euclidean distance between any sample $x_i$ and other samples in the training data set C, and find the closest sample $X_n$, If the category of the nearest neighbor sample $X_n$ is different from $x_i$, the $X_n$ is put into the empty set P until all the samples have calculated the nearest neighbor sample. For the samples in P, the BNF values are calculated according to (4), and they are sorted. The boundary noise samples in P are deleted according to the BNF values, and finally a data set $\Omega^C$ is obtained.

```
Algorithm 2    ENNB
Input: The original data set D.
Set an empty set C, P and $D_B$.
Delete the large class samples away from the boundary from D
by ENNS.
    Put the remaining sample $X_i$ into C.
For i=1,2,…,n
    a.  Calculate the Euclidean distance between $X_i$ and other
        samples in C.
    b.  Find the sample $X_j$ closest to $X_i$.
    c.  If the $X_j$ and $X_i$ categories are different, put $X_j$ in the
        empty set P.
END
For j=1,2,…,n
    a.  For sample $X_j$ in P, calculate BNF value.
    b.  Sort $X_j$.
    c.  Deleting boundary noise samples in P based on BNF
        values, $X_j$ is put into $D_B$.
END
Output: The balanced data set $D_B$DB.
```

### 2.3.3 K-means Clustering Based Sampling Method

K-means clustering based sampling method (KCS) select the representative samples using k-means clustering approach which is a classical method with a high power in finding the similar samples. For the original imbalanced data set $\mathbb{D}$ in this work, the majority are negative sample and minority are positive sample. Therefore the negative dataset is firstly clustered into k clusters, and the ratio $R_i$ of samples of the *i*-th cluster to all negatives can be calculated as:

$$R_i = \frac{N_i}{|ND|}, 2 \leq i \leq k \tag{2}$$

where $N_i$ is the number of the samples in *i*-th cluster, and |ND| is the number of negative data points (NDs).

Then the number of negative samples extracted from each cluster $S_i$ can be calculated as:

$$S_i = |PD| \times R_i, 2 \leq i \leq k \tag{3}$$

where the number of positive data points (PDs) is denoted as |PD|.

The $S_i$ samples which is closest to the cluster center is extracted from the *i*-th cluster, and this processes will be implemented in each cluster. Then a new balanced dataset $D_K$ can be obtained by combine all of the minority class samples, i.e., interface residues, with the samples selected from the majority class, i.e., non-interface residues.

```
Algorithm 3    KCS
Input: The original data set D.
D consists of a majority of samples (negative samples) ND and a
minority of samples (positive samples) PD.
Set an empty set $D_K$.
The ND is clustered into k clusters using k-means clustering, and
the sample of the i-th cluster is $N_i$.
For i=1,2,…,k
    a.  Calculate the ratio of the samples in the i-th cluster to
        all of negative samples.
    b.  Calculate the number of negative samples $S_i$ extracted
```

```
        from each cluster.
    c.  Si samples closest to the cluster center are extracted
        from $N_i$.
END
Combine all minority samples with proportionally extracted
majority samples into $D_K$.
Output: The balanced data set $D_K$.
```

The above is based on k-means clustering extraction, which can obtain a balanced data set. This kind of data set avoids the absence of most class samples containing important information, and can use the training set to obtain a better classification model, and ultimately improve the overall classification performance of the system.

### 2.4 Predictor construction

In this work, the predictor of protein interaction sites is built using support vector machine algorithm, which is a popular machine learning approach for bioinformatics studies. Before the predictor can be trained, each surface residue in the original dataset D is represented by the feature encoding scheme as a 264 dimensional vector. The predictor then are constructed on three smaller and more balanced datasets based on the above data balance methods, i.e., $D_S$ , $D_B$ and $D_K$ from ENNS, ENNB and KCS, respectively. The flowchart of the whole prediction is shown in Fig.2.

### 2.5 Evaluation criteria

To evaluate the performance of the SVM-based predictor in identifying protein interaction sites from other surface residues, four common used indicators, i.e., accuracy, precision, sensitivity, and specificity, are adopted to in this work. Moreover, another two measures, i.e., F-measure and Matthews correlation coefficient (*MCC*) values, are also introduced into predictive performance evaluation. F-measure is a measure which takes into account both recall and precision and therefore is a weighted average indicator of overall performance of predictor. *MCC* is a measure of the quality of binary classification, which is correlation coefficient between the observed and predicted results, which is good indicator for problems with imbalanced data classes.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Specificity = \frac{TN}{FP+TN} \tag{7}$$

$$F-measure = 2 \times \frac{Precision \times Sensitivity}{Precision+Sensitivity} \tag{8}$$

$$MCC = \frac{(TP \times TN-FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{9}$$

where TP, FP, TN and FN represent the number of true positives (correctly predicted interface residues), the number of false positives (incorrectly predicted interface residues), the number of true negatives (correctly predicted non-interface residues) and the number of false negatives (incorrectly predicted non-interface residues), respectively.

## 3 Results and Discussion

The purpose of this work is to identify interaction sites from protein surface, and the predictor will be built using SVM algorithm. The dataset used in this work is composed by 91 non-redundant protein chains, where the original dataset D with 10,430 surface residues, can be obtained. However, the distribution of D is obviously imbalanced because there are 2,299 interface residues which are only 22% of total residues.

Therefore, the main concern will focus on how to reconstruct a more balanced protein surface residues dataset where the number of interface and non-interface residues will be balanced, and then the reconstructed dataset will be used to improve the prediction performance when the SVM predictor is adopted to infer the new potential interface residues.

### 3.1 Dataset reconstruction

In this work, three imbalance data processing strategies, i.e., ENNS, ENNB and KCS, are presented to make the dataset more balanced, to avoid the possible bias in prediction of protein interaction sites.

Although the balance strategies are different among ENNS, ENNB and KCS, the number of negative samples is same as that of positive ones. Therefore, $D_S$ and $D_K$ has same number of 4,598 samples, and half of them are interface residues and another half are non-interface ones. Because the BNF can also remove some boundary noise samples from the positive dataset, $D_B$ has 4164 samples that are smaller than the size of $D_S$ and $D_K$ , but the number of interface and non-interface residues is same.

### 3.2 Prediction performance of three balanced models

We compare ENNS, ENNB and KCS. It can be seen from Table 1 that all three methods can effectively predict protein interaction sites. On the basis of ENNS, ENNB has improved all indicators, with the accuracy rate increased by 4%, the F-measure increased by 3%, and the correlation coefficient increased by 2% to 8%.

It can be seen from the experiment that when k is 36, KCS achieves the best result, and the classification performance is much higher than the other two methods (Acc = 0.758, Spe = 0.842, Sen = 0.676, Pre = 0.811, F measure = 0.737 and MCC = 0.525). The overall recognition rate of KCS is improved by 11% compared with the ENNB. According to the specificity and sensitivity, KCS can improve the recognition rate of minority class samples when the recognition rate of majority class samples does not decrease. 67.6% of protein interaction sites are correctly identified. On the other hand, it also shows that the cluster center of the KCS has the characteristic attributes of the negative sample, and the surrounding negative samples are more representative than others. How to select the value of k is very important for the KCS algorithm-based predictor, and the changes of prediction performance with different k can be found in Fig.3.

Then we further analyze the robustness of the three methods. We saved the five evaluation indicators obtained by cross-validation and compared them with their average values. The results are shown in Fig.4. It can be observed that all three methods have strong robustness. ENNS only fluctuates greatly in specificity, while ENNB and KCS perform well, and their six parameters fluctuate within a range of 0.01.

### 3.3 Prediction performance comparison between balanced and imbalanced dataset

We then compare the above results with the initial imbalanced data set. We do not equalize the data and use SVM to classify directly. The result is shown in Fig.4. We can observe that although the imbalanced data set achieves an accuracy of 0.796, the sensitivity is only 0.015, while the F value and MCC are only 0.030 and 0.058, and the specificity is as high as 0.996. This indicates that in the case where the negative sample far exceeds the positive sample, the experimental result is completely biased toward the former.

From Table 2 we can also see that because the number of negative samples during training is much more than that of positive samples, TN and FN are much more than TP and FP. The negative number in the original label is 4 times the number of positive samples. After classification, regardless of the correctness of the identification, the number of negative samples classified is 160 times the number of positive samples. For the equilibrium dataset, the ratio of negative and positive before and after classification is equivalent, and TP and TN are higher than FP and FN. The prediction effect is far superior to the result of the imbalanced dataset.

### 3.4 Comparison with random sampling method and other approaches

We compare the proposed sampling method with random sampling (RAS). The results are shown in Fig.6. We can intuitively observe that KCS performs better than RAS in all evaluation indicators, with accuracy higher than 0.175 and MCC higher than 0.353. This indicates that although both KCS and RAS are equalizing the sampling process for imbalanced data, KCS can find more representative samples from majority samples, improve the performance of the predictor and improve the classification results.

Li and Kuo[37] also used this dataset to study the prediction of interaction sites between proteins by extracting five different sequence features. In the protein dataset used in this work, It is observed that both the sequence characteristics[37] and the con-served features we extracted can study the prediction of protein interaction sites, but KCS showed better performance, while the accuracy of RAS was 0.27 higher than Li, the number of correctly identified samples is more. This also proves that in this dataset, the evolutionary conservation features we extracted are more representative and can facilitate the prediction of interaction sites.

### 3.5 Visualization of experimental results

We use the molecular visualization tool — Pymol to demonstrate our predictions. Fig.7 shows the results obtained on the protein chain 1 IRA_Y data set under the four sampling methods of RAS, ENNS, ENNB, and KCS. There are 194 balls in the figure that represent the surface residues involved in the prediction. Our methods improves overall predictive performance and successfully predicts most interface residue and non-interface residues and reduces false positives. The ENNB and KCS methods performed well, with only 6.2% of the interface residues in ENNB not predicted, and only 5.2% of the interface residues in KCS were not predicted.

## 4 Conclusion

This paper presents an imbalance data processing strategy for improving protein interaction site prediction. Firstly, 91 protein chains have been selected from the datasets used by Ansari and Helms et al.[34]. Five conservative features have been extracted from the HSSP database and Consurf sever and merged them. Three sampling methods of ENNS, ENNB and KCS have been proposed, and SVM classifier has constructed for experiments. The experimental results show that the ENNB and KCS methods achieve excellent classification results, which yielded the accuracy of 0.644 and 0.758 when ten-fold cross-validation had been adopted. KCS method can achieve overall best performance, i.e., the highest precision of 0.811 and the highest F-measure of 0.737. This work shows that the imbalance data processing strategy can improve the prediction of protein-protein interaction sites, which is important for understanding of cell activities and drug design.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61472282, 61672035, 61872004, 61873221, and 61672447), Anhui Province Funds for Excellent Youth Scholars in Colleges (gxyqZD2016068), the fund of Co-Innovation Center for Information Supply & Assurance Technology in AHU (ADXXBZ201705), Anhui Scientific Research Foundation for Returned Scholars, and the Natural Science Foundation of Hunan Province (Nos.2018JJ4058 and 2017JJ5036).

## References

[1] Y. Chen, J. Xu, B. Yang *et al.*, "A novel method for prediction of protein interaction sites based on integrated RBF neural networks," *Computers in Biology & Medicine,* vol. 42, no. 4, pp. 402-407, 2012.

[2] W. Bing, C. Peng, H. De-Shuang *et al.*, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," *Febs Letters,* vol. 580, no. 2, pp. 380-384, 2006.

[3] B. Wang, P. Chen, J. Zhang *et al.*, "Inferring protein-protein interactions using a hybrid genetic algorithm/support vector machine method," *Protein Pept Lett,* vol. 17, no. 9, pp. 1079-84, Sep, 2010.

[4] B. Wang, H. S. Wong, and D. S. Huang, "Inferring protein-protein interacting sites using residue conservation and evolutionary information," *Protein Pept Lett,* vol. 13, no. 10, pp. 999-1005, 2006.

[5] P. J. Wei, D. Zhang, H. T. Li *et al.*, "DriverFinder: A Gene Length-Based Network Method to Identify Cancer Driver Genes," *Complexity,* vol. 2017, no. 99, pp. 1-10, 2017.

[6] P. J. Wei, D. Zhang, J. Xia *et al.*, "LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network," *Bmc Bioinformatics,* vol. 17, no. Suppl 17, pp. 467, 2016.

[7] X. Zhang, Y. Tian, R. Cheng *et al.*, "A Decision Variable Clustering Based Evolutionary Algorithm for Large-scale Many-objective Optimization," *IEEE Transactions on Evolutionary Computation,* vol. 22, no. 1, pp. 97-112, 2018.

[8] Y. Tian, R. Cheng, X. Zhang *et al.*, "An Indicator Based Multi-Objective Evolutionary Algorithm with Reference Point Adaptation for Better Versatility," *IEEE Transactions on Evolutionary Computation,* vol. 22, no. 4, pp. 609-622, 2018.

[9] K. Yan, H. L. Cheng, Z. Ji *et al.*, "Accelerating smooth molecular surface calculation," *J Math Biol,* vol. 76, no. 3, pp. 779-793, Feb, 2018.

[10] Z. Ji, B. Wang, K. Yan *et al.*, "A linear programming computational framework integrates phosphor-proteomics and prior knowledge to predict drug efficacy," *BMC Syst Biol,* vol. 11, no. Suppl 7, pp. 127, Dec 21, 2017.

[11] K. Yan, B. Wang, H. Cheng *et al.*, "Molecular Skin Surface-Based Transformation Visualization between Biological Macromolecules," *J Healthc Eng,* vol. 2017, pp. 4818604, 2017.

[12] P. Ping, L. Wang, L. Kuang *et al.*, "A Novel Method for LncRNA-Disease Association Prediction Based on an lncRNA-disease Association Network," *IEEE/ACM Trans Comput Biol Bioinform*, Apr 16, 2018.

[13] P. Chen, S. S. Hu, J. Zhang *et al.*, "A Sequence-Based Dynamic Ensemble Learning System for Protein Ligand-Binding Site Prediction," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 13, no. 5, pp. 901-912, Sep-Oct, 2016.

[14] S. G. Ge, J. Xia, W. Sha *et al.*, "Cancer Subtype Discovery Based on Integrative Model of Multigenomic Data," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 14, no. 5, pp. 1115-1121, Sep-Oct, 2017.

[15] S. S. Hu, P. Chen, B. Wang *et al.*, "Protein binding hot spots prediction from sequence only by a new ensemble learning method," *Amino Acids,* vol. 49, no. 10, pp. 1773-1785, Oct, 2017.

[16] Z. W. Ji, B. Wang, K. Yan *et al.*, "A linear programming computational framework integrates phosphor-proteomics and prior knowledge to predict drug efficacy," *BMC Syst Biol,* vol. 11, Dec 21, 2017.

[17] Q. Liu, P. Chen, B. Wang *et al.*, "Hot spot prediction in protein-protein interactions by an ensemble system," *BMC Syst Biol,* vol. 12, no. Suppl 9, pp. 132, Dec 31, 2018.

[18] Q. Liu, P. Chen, B. Wang *et al.*, "dbMPIKT: a database of kinetic and thermodynamic mutant protein interactions," *BMC Bioinformatics,* vol. 19, no. 1, pp. 455, Nov 27, 2018.

[19] B. Wang, P. Chen, P. Wang *et al.*, "Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes," *Protein Pept Lett,* vol. 17, no. 9, pp. 1111-6, Sep, 2010.

[20] M. Zhu, X. Song, P. Chen *et al.*, "dbHDPLS: A database of human disease-related protein-ligand structures," *Comput Biol Chem,* vol. 78, pp. 353-358, Feb, 2019.

[21] B. Wang, D. S. Huang, and C. J. Jiang, "A New Strategy for Protein Interface Identification Using Manifold Learning Method," *IEEE Trans Nanobioscience,* vol. 13, no. 2, pp. 118-123, Jun, 2014.

[22] S. Xia, P. Chen, J. Zhang *et al.*, "Utilization of rotation-invariant uniform LBP histogram distribution and statistics of connected regions in automatic image annotation based on multi-label learning," *Neurocomputing,* vol. 228, no. C, pp. 11-18, 2016.

[23] P. Chen, C. Liu, L. Burge *et al.*, "DomSVR: domain boundary prediction with support vector regression from sequence information alone," *Amino Acids,* vol. 39, no. 3, pp. 713-26, Aug, 2010.

[24] J. W. Yu, P. Y. Ping, L. Wang *et al.*, "A Novel Probability Model for LncRNA-Disease Association Prediction Based on the Naive Bayesian Classifier," *Genes,* vol. 9, no. 7, Jul, 2018.

[25] J. Jiang, N. Wang, P. Chen *et al.*, "Prediction of Protein Hotspots from Whole Protein Sequences by a Random Projection Ensemble System," *Int J Mol Sci,* vol. 18, no. 7, Jul 18, 2017.

[26] P. Fariselli, F. Pazos, A. Valencia *et al.*, "Prediction of proteinâ□"protein interaction sites in heterocomplexes with neural networks ," *Febs Journal,* vol. 269, no. 5, pp. 1356-1361, 2010.

[27] B. Wang, P. Chen, D. S. Huang *et al.*, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," *FEBS Lett,* vol. 580, no. 2, pp. 380-4, Jan 23, 2006.

[28] M. Iqbal, A. A. Freitas, and C. G. Johnson, *A Hybrid Rule-Induction/Likelihood-Ratio Based Approach for Predicting Protein-Protein Interactions*, 2009.

[29] M. Oh, and K. J. Joo, "Protein-binding site prediction based on three-dimensional protein modeling," *Proteins Structure Function & Bioinformatics,* vol. 77 Suppl 9, no. S9, pp. 152, 2009.

[30] L. Liu, Y. Cai, W. Lu *et al.*, “Prediction of protein–protein interactions based on PseAA composition and hybrid feature selection,” *Biochemical & Biophysical Research Communications,* vol. 380, no. 2, pp. 318-322, 2009.

[31] X. W. Chen, and J. C. Jeong, “Sequence-based prediction of protein interaction sites with an integrative method,” *Bioinformatics,* vol. 25, no. 5, pp. 585-591, 2009.

[32] T. Northey, A. Baresic, and A. C. R. Martin, “IntPred: a structure-based predictor of protein-protein interaction sites,” *Bioinformatics*, Sep 18, 2017.

[33] B. Wang, D. S. Huang, and C. Jiang, “A new strategy for protein interface identification using manifold learning method,” *IEEE Trans Nanobioscience,* vol. 13, no. 2, pp. 118-123, 2014.

[34] S. Ansari, and V. Helms, “Statistical analysis of predominantly transient protein–protein interfaces,” *Proteins Structure Function & Bioinformatics,* vol. 61, no. 2, pp. 344-355, 2010.

[35] A. SF, G. W, M. W *et al.*, “Basic local alignment search tool,” *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-10, 1990.

[36] M. M. Breunig, "LOF: identifying density-based local outliers." pp. 93-104.

[37] T. H. Kuo, and K. B. Li, “Predicting Protein–Protein Interaction Sites Using Sequence Descriptors and Site Propensity of Neighboring Amino Acids,” *International Journal of Molecular Sciences,* vol. 17, no. 11, pp. 1788, 2016.

## FIGURE CAPTIONS

Fig. 1. Schematic diagram of ENNS, where green circles denote positive samples, yellow circles are negative samples, and white circles in the right subplot are the samples will be removed.

Fig. 2. The flowchart of the proposed method.

Fig.3. The changes of prediction performance with different values of k.

Fig.4. The different measures of prediction performance in 5 repetitions.

Fig.5. Classification performance evaluation of three sampling methods on datasets.

Fig.6. Compared with RAS methods and the former's evaluation performance.

Fig.7. Visualization of RAS and three sampling methods. A) RAS, B) ENNS, C) ENNB, D) KCS. Herein, Green balls, red balls, yellow balls and blue balls represent TP, TN, FP and FN predictions, respectively.

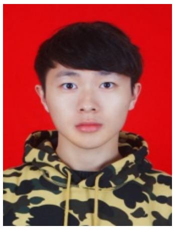
## TABLE TITLES

TABLE 1
Prediction Performance Of ENNB And KCS

TABLE 2
Classification  results of Three Sampling Methods and imbalance dataset

## BRIEF AUTHOR BIOGRAPHIES

**Bing Wang** (SM'14) received the B.S. and M.S degree from Hefei University of Technology, Hefei, China in 1998 and 2004 respectively. He received the Ph.D degree from University of Science and Technology of China, Hefei, China in 2006. He worked as a senior research associate in City University of Hong Kong, 2006-2007, and a postdoctoral fellow in University of Louisville and Van-Vanderbilt University, USA, from 2008 to 2012. Currently, Dr. Wang is serving as a full professor in the School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan, China. He has more than 100 publications, and more than 1,000 citations. His research interests mainly focus on machine learning, data mining, computational biology and cheminformatics.



**Chang-Qing Mei** is a graduate student in the School of Electrical and Information Engineering, Anhui University of Technology. His current research interests include machine learning, data mining, digital image processing and deep learning.

**Yuan-Yuan Wang** received the M.Sc degree from Anhui University of Technology in 2018. Her current work interests include big data, data mining and data management.

**Yuming Zhou**(M'18) is a Full Professor at the Anhui University of Technology. He received the B.S., M.S., and Ph.D. from Huazhong University of Science and Technology, Wuhan, China, in 1993, 2004, and 2007, respectively. Since 2007, he has been with the College of Physics and electronics, Hunan University, where he is a lecturer. In 2010, he moved to the School of Electrical and Information, Anhui University of Technology. He was a visiting scholar with the Auburn University, USA, from July 2013 and June 2014. His research interests mainly focus on model simulation.

**Mu-Tian Cheng** received the B.S. from Anqing Normal University, Anqing, China in 2003. He received the Ph.D degree from Wuhan University, Wuhan, China in 2008. He visited Oklahoma State University and Texas A&M University from 2016 to 2017. Currently, Dr. Cheng is serving as a full professor in the School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan, China. His research interests mainly focus on Plasmonics, Nano photonics and Quantum machine learning.

**Chun-Hou Zheng** received the B.Sc degree in Physics Education in 1995 and the M.Sc. degree in Control Theory & Control Engineering in 2001 from QuFu Normal University, and the Ph.D degree in Pattern Recognition & Intelligent System in 2006, from University of Science and Technology of China. From Feb. 2007 to Jun. 2009 he worked as a Postdoctoral Fellow in the Hefei Institutes of Physical Sceience, Chinese Academy of Sciences. From Jul. 2009 to Jul. 2010 he worked as a Postdoctoral Fellow in the Dept. of Computing, The Hong Kong Polytechnic University. He is currently a Professor in the School of Computer Science and Technology, Anhui University, China. His research interests include Pattern Recognition and Bioinformatics.

**Lei Wang** received the B.S. and M.S degree from Xiangtan University, China in 1994 and 1997 respectively. He received the Ph.D. degree from Hunan University, China in 2005. From 2005 to 2007, he was a postdoctoral research fellow at Tsinghua University, China. From 2007 to 2009, he was a visiting scholar at Lakehead University, Canada and Duke University, USA, separately. Currently, Dr. Wang is serving as a full professor in the College of Computer Engineering & Applied Mathematics, Changsha University, China. His research interests mainly focus on network security and bioinformatics.

**Jun Zhang** received his Bachelor degree from HeFei University of Technology, Master degree from Institute of Intelligent Machine, Chinese Academy of Sciences and Ph.D degree from University of Science and Technology of China. He served in University of Louisville, USA (2009-2011, as Postdoc fellow). He focuses on deep learning with application to bioinformatics, Cheminformatics and computer vision etc. He has published more than 40 papers in international conferences and journals. Currently, he is an associate professor in School of Electrical Engineering and Automation, Anhui University, 230601 Hefei, China.

**Peng Chen** specializes in machine learning and data mining with applications to bioinformatics, drug discovery, computer vision, etc. He has published more than 60 high quality referred papers in international conferences and journals. He is an Professor in the Institute of Physical Science and Information Technology, Anhui University, 230601 Hefei, China. He received his Bachelor degree from Electronic Engineering Institute, PLA, Master degree from Kunming University of Science and Technology, and Ph.D degree from University of Science and Technology of China. Prior to joining Anhui University, he served in City University of Hong Kong (2006, as senior research associate), Howard University, USA (2008-2009, as Postdoc Fellow), Nanyang Technological University, Singapore (2009-2010, as Research fellow), and King Abdullah University of Science and Technology (KAUST), Saudi Arabia (2012-2014, as Postdoc Fellow). From 2011 to 2013, he was an Associate Professor in Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, China.

**Yan Xiong** receives the B.S., M.S., and Ph.D. degrees from the University of Science and Technology of China in 1983, 1986, and 1990, respectively.

He is a professor with the School of Computer Science and Technology, University of Science and Technology of China. His main research interests include distributed processing, mobile computing, computer network, and information security.