DOI 10.1007/s00232-015-9856-z

Prediction of Protein–Protein Interaction Sites with Machine-Learning-Based Data-Cleaning and Post-Filtering Procedures

Guang-Hui Liu[1,2] · Hong-Bin Shen[3] · Dong-Jun Yu[1]

Received: 7 August 2015 / Accepted: 3 November 2015 / Published online: 12 November 2015

Abstract Accurately predicting protein–protein interaction sites (PPIs) is currently a hot topic because it has been demonstrated to be very useful for understanding disease mechanisms and designing drugs. Machine-learning-based computational approaches have been broadly utilized and demonstrated to be useful for PPI prediction. However, directly applying traditional machine learning algorithms, which often assume that samples in different classes are balanced, often leads to poor performance because of the severe class imbalance that exists in the PPI prediction problem. In this study, we propose a novel method for improving PPI prediction performance by relieving the severity of class imbalance using a data-cleaning procedure and reducing predicted false positives with a post-filtering procedure: First, a machine-learning-based data-cleaning procedure is applied to remove those marginal targets, which may potentially have a negative effect on training a model with a clear classification boundary, from the majority samples to relieve the severity of class imbalance in the original training dataset; then, a prediction model is trained on the cleaned dataset; finally, an effective post-filtering procedure is further used to reduce potential false positive predictions. Stringent cross-validation and independent validation tests on benchmark datasets demonstrated the efficacy of the proposed method, which exhibits highly competitive performance compared with existing state-of-the-art sequence-based PPIs predictors and should supplement existing PPI prediction methods.

Keywords Protein–protein interaction sites · Imbalanced learning · Data cleaning · Random forests · Post-filtering

Introduction

Protein–protein interactions are responsible for carrying out biochemical activities in living systems (Ahmed et al. 2015; Marceau et al. 2013). Previous studies have validated that protein–protein interactions play critical roles in the life cycles of living cells, such as genetic material duplication, regulation of gene expression, cell signal transduction, metabolism, organism growth and reproduction, cell apoptosis, and cell necrosis (Fry 2015; Sharon and Sinz 2015). Therefore, the study of how protein–protein interactions form intermolecular regulatory networks, including genetic regulatory pathways, metabolism, and signal transduction pathway, is of great biological significance (Betel et al. 2007; Hall et al. 2007; Hu et al. 2011; Jia et al. 2015a; Skrabanek et al. 2008). This research will not only help in further understanding various biological processes and mechanisms from a systematic point of view (Gromiha et al. 2009; Yugandhar and Gromiha 2014a, b) but can also help identify new drug targets and pave the way for the development of new drugs (Ako-Adjei et al. 2015; Burgoyne and Jackson 2006; Russell and Aloy 2008).

In preliminary studies, the determination of whether two proteins would interact with each other and which residues were interactive, i.e., protein–protein interaction sites

✉ Dong-Jun Yu
njyudj@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China

[2] School of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210046, China

[3] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai 200240, China

(PPIs), was mostly confined to biological experimental methods (Drewes and Bouwmeester 2003). However, these wet lab methods are difficult to apply to all living organisms because they are both time- and cost-consuming (Edwards et al. 2002; Friedrich et al. 2006). In addition, biological wet lab methods for identifying PPIs tend to involve risks of high false positive and false negative results (Ito et al. 2001; Von Mering et al. 2002). In recent years, researchers have investigated the possibility of utilizing computational approaches to rapidly and accurately predict PPIs on large-scale protein datasets (Ito et al. 2000). Over the past decades, a number of machine learning algorithms, such as neural networks (NNs) (Fariselli et al. 2002), support vector machines (SVMs) (Bradford and Westhead 2005; Wang et al. 2006; Yan et al. 2004), and random forests (RFs) (Jia et al. 2015c; Šikic et al. 2009), have been successfully applied to PPI prediction, and many PPI predictors have emerged (Dhole et al. 2014; Murakami and Mizuguchi 2010a; Ofran and Rost 2007; Porollo and Meller 2007; Singh et al. 2014).

Existing machine-learning-based PPI predictors take protein sequential features, structural features, or both as the inputs of prediction models. Because protein 3D structures can provide intuitive and effective clues, they are generally preferred for performing PPI prediction (Agrawal et al. 2014; Cukuroglu et al. 2014; Sudha et al. 2014). For example, Jones and Thornton (1997a, b) carefully analyzed a series of residue patches on the surface of protein 3D structures using six parameters (residue interface propensity, solvation potential, hydrophobicity, planarity, protrusion, and accessible surface area) and proposed a method for calculating the relative combined score of a surface patch for forming protein–protein interactions. Bradford et al. proposed a support vector machine (SVM) predictor based on surface patch analysis (Bradford and Westhead 2005) and then further improved the prediction performance using a Bayesian network (Bradford et al. 2006). Chen et al. (2012) constructed three-dimensional probability density maps of non-covalent interacting atoms on protein surfaces and then applied machine learning algorithms to learn the characteristic patterns of the probability density maps specific to PPIs. However, the number of known protein 3D structures is still considerably smaller than that of sequenced proteins in spite of great efforts made in determining protein structures, which significantly limits the applicability of structure-based PPI prediction (Murakami and Mizuguchi 2010a).

Recently, much attention has been paid to sequence-based PPI prediction, and some progress has been made (Bock and Gough 2001; Zhou and Shan 2001). A number of promising PPI predictors that utilize sequence-derived features and machine-learning algorithms have emerged (Jia et al. 2015a, b; Murakami and Mizuguchi 2010a; Ofran and Rost 2007) (Dhole et al. 2014; Murakami and Mizuguchi 2010a; Singh et al. 2014). Ofran and Rost developed a neural network method called ISIS (Ofran and Rost 2007) for predicting PPIs based on the predicted structural features and evolutionary information calculated from the sub-sequences of nine consecutive residues. Porollo and Meller developed a predictor named SPPIDER (Porollo and Meller 2007) using SVMs and neural networks based on relative solvent accessibility (RSA), and they claimed that the RSA feature possesses a better discrimination capability than that of evolutionary conservation, physicochemical characteristics, structure-derived features, and other previously considered features for performing PPI prediction. Murakami and Mizuguchi used kernel density estimation to construct a naïve Bayesian classifier named PSIVER (Murakami and Mizuguchi 2010a) with position-specific scoring matrices (PSSM) and predicted accessibility (PA) as feature sources. Recently, Dhole et al. implemented SPRINGS (Singh et al. 2014) and LORIS (Dhole et al. 2014) to identify PPIs by applying artificial neural networks and L1-regularized logistic regression, respectively, based on evolutionary conservation, predicted relative solvent accessibility and averaged cumulative hydropathy features.

Overall, significant achievements have been made in the prediction of PPIs. Nevertheless, the performance of PPI prediction is still far from satisfactory, and there is still room for further improvement. In addition, we carefully analyzed the existing machine-learning-based PPIs predictors and observed that the severe class imbalance phenomenon, in which the number of majority samples (non-interacting residues) significantly outnumbers that of minority samples (interacting residues), has not been well considered (Ofran and Rost 2007; Porollo and Meller 2007; Šikić et al. 2009), which may potentially deteriorate the performance of machine-learning-based PPI predictors.

Motivated by all these observations, we proposed a new machine-learning-based method for further improving the performance of sequence-based PPI prediction. The proposed method mainly consists of three steps: First, a novel data-cleaning procedure is developed to relieve the severity of class imbalance by removing those difficult marginal targets from the majority samples in the original training dataset using a machine-leaning model; second, based on the cleaned dataset, a machine-learning-based PPI prediction engine is trained; finally, a post-filtering procedure is applied to the results of the prediction engine to reduce false positive predictions. We performed stringent computer experiments on benchmark datasets with both cross-validation and independent validation tests, and the results demonstrated the feasibility and efficacy of the proposed method.

According to Chou's 5-step rule (Chou 2011), which has been implemented in a series of recent publications (Chen et al. 2014; Jia et al. 2015b; Lin et al. 2014; Liu et al. 2015a;

Xu et al. 2014), to establish a useful sequence-based statistical predictor for a biological system, the following five guidelines should be followed (Chou 2011): (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to perform the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (e) establish a user-friendly web server for the predictor that is accessible to the public. Below, we describe how to address these steps systematically.

Materials and Methods

Benchmark Datasets

In this study, we benchmarked the proposed method on three widely used datasets, denoted Dset186, Dtestset72, and PDBtestset164, to demonstrate the method's feasibility and effectiveness. Among the three datasets, Dset186 was used as a training dataset and the remaining two, i.e., Dtestset72 and PDBtestset164, were used as independent validation datasets. Dset186 was previously constructed by Murakami and Mizuguchi and consists of 186 non-redundant (sequence identity <25 %), heterodimeric, non-transmembrane, and transient protein chains, which have been structurally resolved by X-ray crystallography with a resolution of ≤3.0 Å (Murakami and Mizuguchi 2010a). The interacting residue in the protein chains was defined as a residue that lost absolute solvent accessibility of <1.0 $\mathring{A}^2$ upon complex formation (Singh et al. 2014).

Dtestset72 (Murakami and Mizuguchi 2010b) consists of 72 non-redundant sequences that are non-overlapping with sequences in Dset186. Dtestset72 was constructed based on the protein–protein docking benchmark set version 3.0 (Hwang et al. 2008) using a homology reduction procedure: Any sequences displaying ≥25 % sequence identity over a 90 % overlap with any of the sequences in Dset186 were removed using BLASTClust (Altschul et al. 1997). The obtained Dtestset72 includes rigid body cases (27 protein complexes), medium cases (6 protein complexes), and difficult cases (3 protein complexes). In these cases, each protein complexes consists of two protein chains.

To further explore the prediction performance of PPI prediction models on newly annotated proteins, another independent validation dataset, denoted PDBtestset164, built by Singh et al. (2014) was also used. The PDBtestset164 dataset was obtained using newly annotated proteins from June 2010 to November 2013. The same filter used to obtain Dset186 and Dtestset72 was applied to create PDBtestset164 (Singh et al. 2014). PDBtestset164 consists of non-redundant 164 protein chains extracted from newly deposited proteins (from June 2010 to November 2013) in the Protein Data Bank (PDB) with the same filters that were applied to construct Dset186 and Dtestset72. The software PSAIA (Protein Structure and Interaction Analyzer) (Mihel et al. 2008) was used to identify interacting residues of the protein sequences in PDBtestset64. Because PDBtestset164 consists of new proteins released after the construction of Dset186, we used it as the second independent validation dataset to further evaluate the generalization capability of the proposed method. For details about PDBtestset164, please refer to (Dhole et al. 2014; Singh et al. 2014). Table 1 summarizes the statistics of the three benchmark datasets.

Feature Extraction

To develop a machine-learning-based PPI predictor, a critical step was to represent each residue as a discriminative feature vector. In this study, three feature sources, i.e., position-specific scoring matrix, averaged cumulative hydropathy, and predicted relative solvent accessibility, that have been demonstrated to be useful for PPI prediction were used to construct the discriminative feature for each sample (i.e., a residue in a protein sequence).

Evolutionary information contained in a protein sequence has been demonstrated to be useful for many protein attributes prediction problems, including PPI prediction (Chen and Jeong 2009; Dhole et al. 2014; Murakami and Mizuguchi 2010a; Yan et al. 2003; Yu et al. 2013a). Position-specific scoring matrix (PSSM) obtained by multiple sequence alignment can partially provide the evolutionary information of a protein sequence (Yu et al. 2011). For a given protein sequence, we generated a corresponding $L \times 20$ PSSM using PSI-BLAST (Schäffer et al. 2001) to search the Swiss-Prot database through three iterations, with 0.001 as the E value cutoff for multiple sequence alignment against the sequence (He et al. 2015; Xiao et al. 2015b), where L is the length of the protein sequence. The original PSSM of a protein sequence with L residues generated by PSI-BLAST, denoted as $\mathrm{P}_{\mathrm{original\_pssm}}$, can be formulated as follows:

$$\mathrm{P}_{\mathrm{original\_pssm}} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k,1} & p_{k,2} & \cdots & p_{k,20} \\ \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix}, \quad (1)$$

where $p_{k,j}$ represents the score of the residue k in the protein sequence being mutated to residue type j during the

Table 1 Composition of the training dataset and the two independent validation datasets

Dataset	No. of sequences	(No. of interacting residues, no. of non-interacting residues)
Dset186	186	(5517, 30,702)
Dtestset72	72	(1923, 16,217)
PDBtestset164	164	(6096, 27,585)

evolution process. A positive score indicates that the corresponding mutation occurs more frequently than expected by chance, whereas a negative score indicates the opposite. Note that here we use the numerical code 1, 2, …, 20 to represent the 20 native amino acid types according to the alphabetical order of their single character codes (Zou and Xiao 2015). After obtaining the original PSSM, we further normalized each element of $P_{\text{original\_pssm}}$ to the range (0, 1) with the following logistic function:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

where x is the score in the original PSSM.

A sliding window of size W was then applied to the normalized PSSM to extract feature vectors for each residue of the protein sequence. According to (Dhole et al. 2014), $W = 9$ is a better choice for performing PPI prediction. Therefore, we set $W = 9$ in this study. Accordingly, the dimensionality of the obtained PSSM feature vector for a residue is 9 × 20 = 180-D.

Researchers have found that protein–protein interaction interfaces are generally hydrophobic patches on the surfaces of proteins (Chothia and Janin 1975; Jones and Thornton 1995, 1997a). Hence, the hydropathy index should be beneficial to the identification of protein–protein interaction sites, which has been demonstrated by related studies (Dhole et al. 2014; Singh et al. 2014). In this study, the hydropathy index proposed by Kyte and Doolittle (1982) was used. The hydropathy index of a residue represents the hydrophobic or hydrophilic properties of a residue's side chain. The hydropathy index of residue is an indicator of hydrophilic and hydrophobic properties (Gallet et al. 2000). We explored the hydropathy indices of a residue and its neighborhood to extract the residues averaged cumulative hydropathy (ACH) feature. More specifically, for a target residue, its five ACH indices corresponding to five windows of different sizes (i.e., sizes of 1, 3, 5, 7, and 9) centered on the residue were calculated using the Python codes provided by Dhole et al. (2014); the five ACH indices were then concatenated to form a 5-D feature vector for the target residue.

We extracted the predicted relative solvent accessibility (PRSA) features of residues with the SANN web server developed by Joo et al. (2012). The SANN web server is freely available at http://lee.kias.re.kr/~newton/sann/. For a given protein sequence, SANN predicts the discrete states (two or three states) and a continuous value of solvent accessibility for each residue in the sequence (Dhole et al. 2014). In this study, we used the predicted continuous value of solvent accessibility to encode each residue into a 1-D PRSA feature.

Finally, a residue can be encoded into a 186-D feature vector by serially combining its 180-D PSSM feature, 5-D ACH feature and 1-D PRSA feature.

Data-Cleaning Procedure

The purpose of data cleaning is to remove marginal targets that are difficult to classify from the majority samples to relieve the severity of class imbalance of the original training dataset. Figure 1 illustrates the workflow of the proposed machine-learning-based data-cleaning (DC) procedure.

Let $S = \{s_1, s_2, \cdots, s_i, \cdots, s_N\}$ be the set of N protein sequences in the original dataset, where s_i is the ith sequence. We use '+' and '−' to represent the class labels of minority and majority residues (i.e., interactive and non-interactive residues), respectively.

The proposed DC procedure removes potential marginal targets from one protein sequence each time. For the i-th sequence $s_i \in S$, the proposed DC procedure first trains a prediction engine, denoted M_i, based on the samples from $S - \{s_i\}$ with a machine learning algorithm; then, the trained prediction engine M_i is used to predict the class labels of residues of sequence s_i; each majority residue (labeled '−') of s_i will then be screened to determine whether it is a marginal target (a majority residue is considered marginal if it is predicted to belong to the minority class by the prediction engine M_i); the cleaned sequence s_i^{c} is obtained by removing all the marginal targets from s_i; this procedure will be repeated N times until all of the N sequences in S have been cleaned. The cleaned dataset is denoted $S^{\mathrm{c}} = \{s_1^{\mathrm{c}}, s_2^{\mathrm{c}}, \cdots, s_i^{\mathrm{c}}, \cdots, s_N^{\mathrm{c}}\}$.

In essence, any machine learning algorithm can be used to construct the prediction engine for data cleaning. In this study, the random forest (RF) (Breiman 2001) algorithm was used as an example to implement the proposed DC procedure.

We explain the rationality of the proposed DC procedure as follows. Clearly, for each sequence $s_i \in S$, the prediction

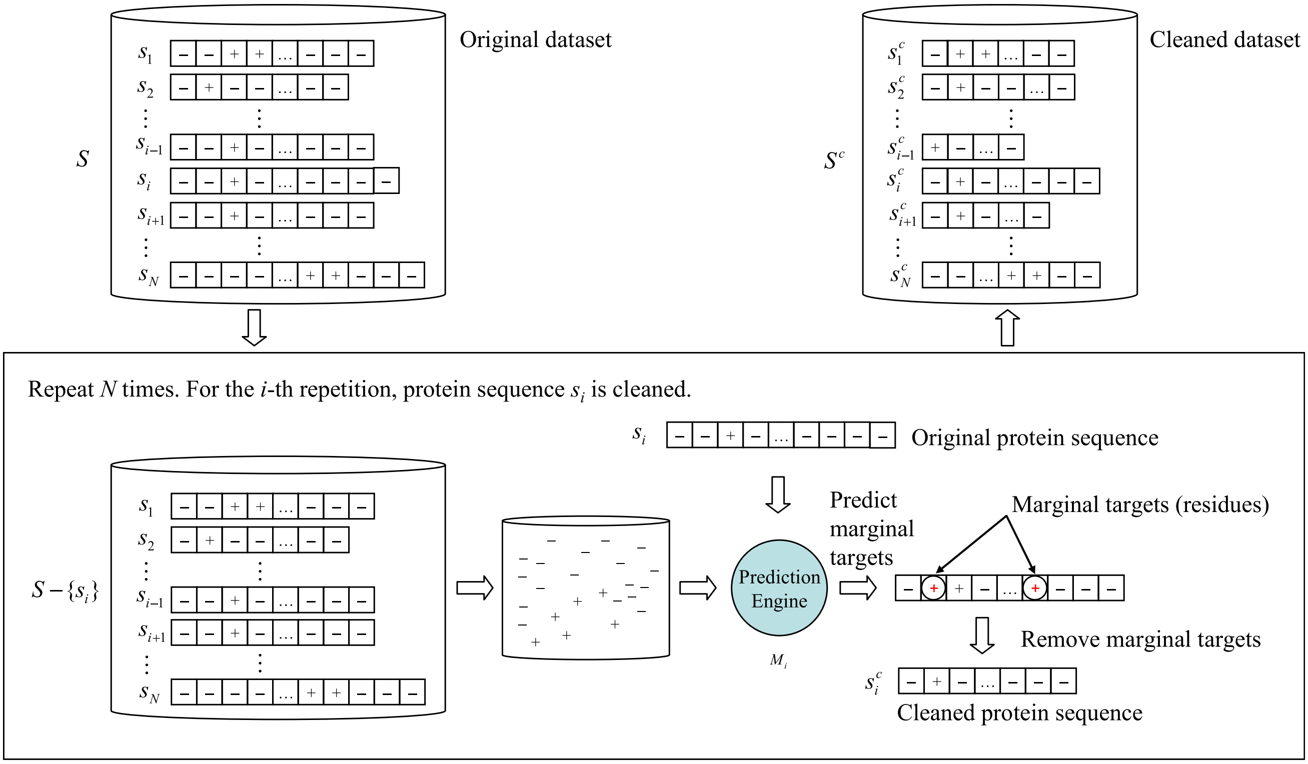

Fig. 1 Workflow of the proposed machine-learning-based data-cleaning procedure

engine is trained on an imbalanced dataset $S - \{s_i\}$, where the number of majority samples is significantly larger than that of minority samples. In other words, the trained model M_i will be biased towards the majority class. Consequently, it is reasonable to consider a majority residue in s_i as a marginal target if it is still being predicted as minority under the majority-prone model M_i.

We processed the original dataset Dset186 with the proposed DC procedure. It was observed that 13,450 majority residues were cleaned. The ratio between the number of majority samples and that of minority samples was decreased from 5.56 to 3.13, demonstrating that the severity of data imbalance was reduced.

To better understand the proposed DC procedure, we took a sequence (PDB ID: 1AY7_A) from Dset186 as an example to vividly illustrate the effect after data cleaning, as shown in Fig. 2. The images in Fig. 2 were generated using PyMOL (DeLano 2002).

Figure 2a, b shows images of the 3-D structures of 1AY7_A in sphere style and cartoon style, respectively, before the DC procedure; on the other hand, Fig. 2c, d shows images of the 3-D structures of 1AY7_A in sphere style and cartoon style, respectively, after the DC procedure. In Fig. 2, interactive and non-interactive residues are highlighted in yellow and cyan, respectively. Note that residues highlighted in red are also non-interactive ones located inside the inner compartment of a protein. These non-interactive residues are also called non-surface residues. In this study, non-surface residue were determined by calculating their relative solvent accessibility (RSA) using NACCESS (Hubbard and Thornton 1993). A residue was considered to occur on the surface if its RSA was <5 % (Jones and Thornton 1997a, b).

From Fig. 2, two observations can be made: (1) These non-surface residues are spatially located inside the inner compartment of a protein and thus cannot interact with other proteins. On the other hand, these non-surface residues (highlighted in red) are locate close to those interactive residues (highlighted in yellow) and thus could have a negative effect on training a machine-learning prediction model with a clear classification boundary. (2) Parts of the non-interactive residues (highlighted in cyan) are also located close to those interactive residues (highlighted in yellow) and thus could have the same negative effect as the non-surface residues.

In this study, we referred to these non-interactive and non-surface residues, which are spatially located close to interactive residues and have a negative effect on training a machine-learning prediction model with a clear classification boundary, as marginal residues. As previously mentioned, on the one hand, removing these (or parts of) marginal residues can reduce the severity of data imbalance in the original training dataset; on the other hand, removing the residues can help construct a much more compact

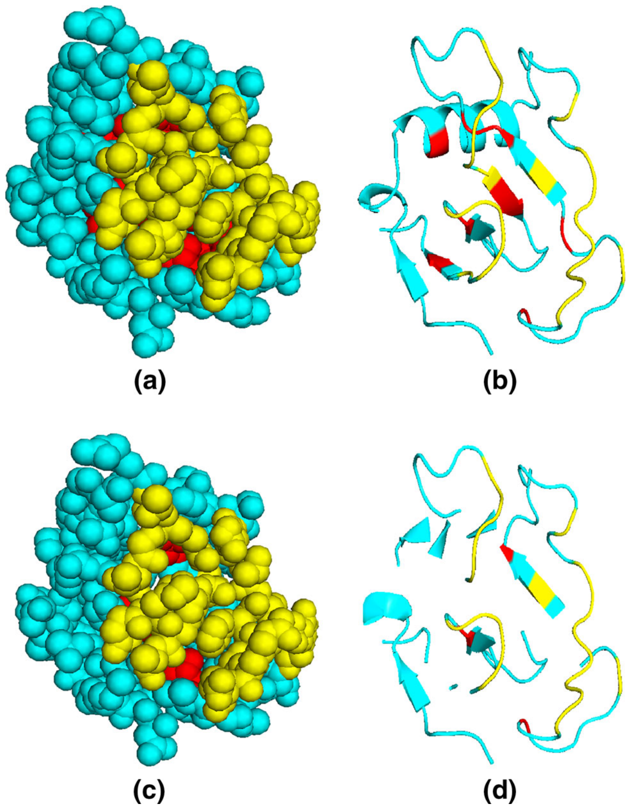

Fig. 2 Visualization of the effect of the proposed DC procedure for protein 1AY7_A. **a** and **b** are images of 3-D structures of 1AY7_A in sphere style and cartoon style, respectively, before the DC procedure; **c** and **d** are images of 3-D structures of 1AY7_A in sphere style and cartoon style, respectively, after the DC procedure. Interactive and non-interactive residues are highlighted in *yellow* and *cyan*, respectively. Marginal residues are highlighted in *red*. The images were generated using PyMOL (DeLano 2002) (Color figure online)

prediction model with a clear classification boundary. As an example, it was observed that 6 out of 9 non-surface residues (highlighted in red) and 5 out of 67 non-interactive residues (highlighted in cyan) were successfully removed after applying the proposed DC procedure on 1AY7_A, as shown in Fig. 2c, d.

It should be noted that there are many other methods for addressing highly imbalanced or skewed dataset, and several of which have been successfully applied to bioinformatics problems (Ertekin et al. 2007a; Ertekin et al. 2007b; Estabrooks et al. 2004; Hong et al. 2007; Hu et al. 2014; Kang and Cho 2006; Laurikkala 2001; Ting 2002; Wang and Japkowicz 2010; Wu and Chang 2005; Zhou and Liu 2010). For example, Liu et al. (2015c) developed a predictor called iDNA-Methyl for identifying DNA methylation sites with a "neighborhood cleaning rule" to address class imbalance; Xiao et al. (2015a) developed iDrug-Target, which can predict the interactions between drug compounds and target proteins using a "synthetic minority over-sampling technique".

Training a PPI Prediction Model on the Cleaned Dataset

Based on the cleaned dataset, we can train a PPI prediction model using any suitable machine learning algorithms. For consistency with the algorithm used in the section titled 'Data-Cleaning Procedure' and considering the fact that random forest algorithms have been demonstrated to be particularly useful for performing PPI predictions (Jia et al. 2015c; Šikić et al. 2009), we also used a random forest algorithm as an engine for constructing a PPI prediction model.

Taking Dset186 as an example, we first obtained a cleaned dataset, denoted Dset186^c, with the proposed DC procedure. As calculated in the section titled 'Data-Cleaning Procedure', the ratio of the number of majority samples to that of minority samples was decreased from 5.56 to 3.13. Nevertheless, a data imbalance still exists. Therefore, we further used a random under-sampling technique to balance the majority and minority samples in the cleaned training dataset, i.e., Dset186^c. In this case, the ratio of the number of majority samples to that of minority samples was set to 1:1. Finally, we could train a random forest algorithm based on the balanced training dataset with all residues represented by the feature vector we developed in the section titled 'Feature Extraction'.

During the prediction stage, for each residue in an unseen protein sequence, the trained prediction model first predicts its possibility of being interactive; then, a pre-scribed threshold T is applied to determine whether it is an interactive residue: a residue with a possibility of being larger than T will be predicted as interactive. The value of T is optimized by choosing the one that maximizes the value of Matthews correlation coefficients (MCC) of predictions on Dset186 over leave-one-out cross-validation.

Post-Filtering Procedure

To further improve the PPI prediction performance, a post-filtering (PF) procedure is applied to the initial predictions to reduce potential predicted false positives. This PF procedure is motivated by the following observations made in previous studies (Ofran and Rost 2003; Yan et al. 2004): From the distribution of the number of interactive residues in a window of consecutive residues centered on an interactive residue, approximately 98 % of the observed interactive residues were observed to have at least one additional interactive residue and approximately 76 % had at least four interactive residues in a window of nine consecutive residues (within four residues on either side) (Murakami and Mizuguchi 2010a; Ofran and Rost 2003; Yan et al. 2004). These observations indicate that interactive residues tend to form clusters in sequences (Ofran and Rost 2003; Yan et al. 2004), and neighboring residues of an actual interactive residue have a high potential of being interactive residues (Murakami and Mizuguchi 2010a). Thus, an isolated interactive residue predicted by a model may potentially be a false positive. Therefore, we use a window-based post-filtering (PF) procedure to eliminate those isolated predicted interactive residues to reduce potential false positives.

Figure 2 illustrates the workflow of the proposed PF procedure. More specifically, for each interactive residue predicted by the model, we place a window of size *W* centered on the residue; then, we calculate the number of predicted interactive residues in the window; if this number is less than *m*, the prediction is considered a false positive. In this study, we optimized the values of *W* and *m* by varying *W* from 3 to 11 and *m* from 1 to 5 (Murakami and Mizuguchi 2010a; Ofran and Rost 2003; Yan et al. 2004). In this study, the optimized values of *W* and *m* were set to 11 and 3, respectively (Fig. 3).

We acknowledge that the proposed post-filtering procedure may also reassign a true positive as a false negative. Nevertheless, the following experimental results statistically demonstrate that the overall prediction performance can be further improved by incorporating the proposed post-filtering procedure.

Evaluation Indices

Six routine evaluation indices, i.e., Recall, Precision, Specificity, Accuracy, and *F*-measure, were used to evaluate the prediction performance and are defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

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

where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively.

Although the four abovementioned metrics (Eqs. 3–6) have often been used in the literature to measure the prediction quality of a prediction method, they are no longer the best ones because they lack intuitiveness and are not easy to understand for most biologists, particularly the MCC (the Matthews correlation coefficient). For clarity, we adopt an additional four metrics proposed by Chou (Chou 2001; He et al. 2015; Lin et al. 2014; Liu et al. 2014; Guo et al. 2014; Chen et al. 2013):

$$\text{Recall} = 1 - \frac{N_{-}^{+}}{N^{+}} \tag{9}$$

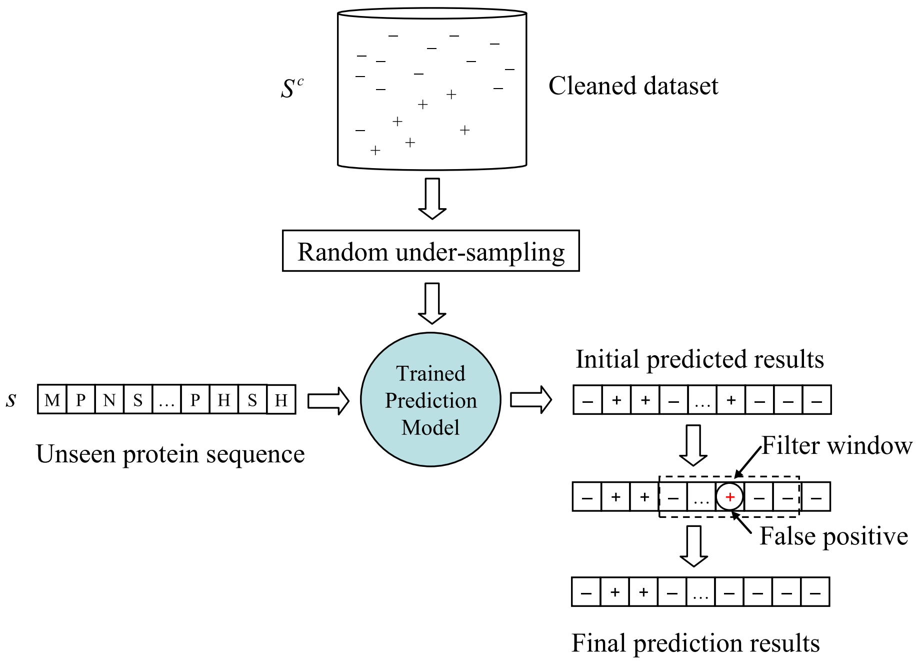

Fig. 3 Workflow of the proposed post-filtering procedure

$$\text{Specificity} = 1 - \frac{N_{+}^{-}}{N^{-}} \tag{10}$$

$$\text{Accuracy} = 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}} \tag{11}$$

$$\text{MCC} = \frac{1 - \left(\frac{N_{-}^{+}}{N^{+}} + \frac{N_{+}^{-}}{N^{-}}\right)}{\sqrt{\left(1 + \frac{N_{+}^{-} - N_{-}^{+}}{N^{+}}\right) \cdot \left(1 + \frac{N_{-}^{+} - N_{+}^{-}}{N^{-}}\right)}}, \tag{12}$$

where N^{+} represents the total number of the interacting residues investigated, whereas N_{-}^{+} is the number of interacting residues incorrectly predicted as non-interacting residues; N^{-} represents the total number of non-interacting residues investigated, whereas N_{+}^{-} is the number of non-interacting residues incorrectly predicted as interacting residues. We can find the relations between Eqs. 3–6 and Eqs. 9–12 as follows (Chou 2001; Lin et al. 2014; Guo et al. 2014):

$$\begin{cases} \text{TP} = N^{+} - N_{-}^{+} \\ \text{TN} = N^{-} - N_{+}^{-} \\ \text{FP} = N_{+}^{-} \\ \text{FN} = N_{-}^{+} \end{cases} \tag{13}$$

For a detailed understanding of Eqs. 9–12 and 13), please refer to (Chou 2001; He et al. 2015; Lin et al. 2014; Liu et al. 2014; Guo et al. 2014; Chen et al. 2013). Please note that the set of metrics defined in Eqs. 9–12 is valid only for single-label systems. For multi-label systems, which have become more common in systems biology (Lin et al. 2013) and systems medicine (Xiao et al. 2013), a completely different set of metrics, as defined by (Chou 2013), is needed.

On the one hand, the six evaluation indices (Eqs. 3–8) defined above are threshold dependent, i.e., their values will depend on the threshold chosen for a given prediction model. On the other hand, PPI prediction is a typical imbalanced binary prediction problem; hence, over-pursuing Accuracy is not appropriate (He and Garcia 2009a). Considering that the MCC index provides an overall measurement of the quality of binary predictions (Baldi et al. 2000), we reported these indices for a prediction model with a threshold that maximizes the MCC value of predictions.

Results and Discussion

Effectiveness of Data-Cleaning and Post-filtering Procedures

In this section, we demonstrate the effectiveness of the proposed data-cleaning and post-filtering procedures for improving the prediction performance of PPI predictions.

First, we constructed a benchmark PPIs predictor using a random forest algorithm. To eliminate the severe imbalance in the training dataset, the random under-sampling technique (RUS) was used to balance the majority and minority samples in the training dataset. In this case, the ratio of the number of majority samples to that of minority samples was set to 1:1. For convenience, we termed this benchmark prediction model RF-RUS, which indicates a model trained with the RF algorithm and random under-sampling technique (RUS).

Second, we demonstrate the efficacy of the proposed data-cleaning procedure by incorporating it into the benchmark prediction model. More specifically, we first obtained a cleaned dataset by applying the proposed data-cleaning procedure on the original training dataset; then we performed the benchmark model, i.e., RF-RUS, on the cleaned training dataset. We denoted the RF-RUS procedure featuring a data-cleaning procedure as DC-RF-RUS.

Third, we incorporated the proposed post-filtering procedure into DC-RF-RUS; the resulting procedure was termed DC-RF-RUS-PF.

We performed stringent leave-one-out cross-validation tests on Dset186 for RF-RUS, DC-RF-RUS, and DC-RF-RUS-PF. Performance comparisons between the three methods are listed in Table 2. Please note that in each round of cross-validation for DC-RF-RUS and DC-RF-RUS-PF, only the sequences in the training subsets were cleaned and the sequence in the testing subset was not cleaned.

Table 2 shows that DC-RF-RUS outperforms RF-RUS with respect to all six considered evaluation indices except for *Recall*. DC-RF-RUS achieves a 1.7 % improvement on MCC, which is an overall measurement of the quality of binary predictions. The performance comparison between DC-RF-RUS and RF-RUS demonstrates that the proposed data-cleaning procedure improves the data quality, which will facilitate the training of a machine-learning-based PPI prediction model. We also find that the prediction performance is further improved, although not significantly, by incorporating a post-filtering procedure into DC-RF-RUS. We also find that the value of *Specificity* was increased by 2.8 %, which supports the argument we made that the proposed post-filtering procedure can help reduce the number of false positive predictions. We note that the value of *Recall* was reduced by 2.2 %; the underlying reason is that the post-filtering procedure mistakenly reassigned true positive predictions to false negatives. Nevertheless, the overall performances, e.g., *Accuracy*, *F*-measure, and MCC, were still improved by incorporating post-filtering, as shown in Table 2.

Comparisons with Existing PPIs Predictors Over Cross-Validation Test

In this section, we compare the proposed method, i.e., DC-RF-RUS-PF, with two of the most recently released sequence-based PPIs predictors on Dset186 over a stringent

Table 2 Performance comparisons between RF-RUS, DC-RF-RUS, and DC-RF-RUS-PF on Dset186 over leave-one-out cross-validation

Method	MCC	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	*F*-measure (%)
RF-RUS	0.202	28.6	64.3	61.7	61.8	37.3
DC-RF-RUS	0.219	30.1	63.4	63.8	63.3	37.9
DC-RF-RUS-PF	0.229	31.7	61.2	66.6	65.1	38.2

Table 3 Performance comparisons between the proposed method, LORIS, and PSIVER on Dset186 over leave-one-out cross-validation

Method	MCC	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	*F*-measure (%)
Proposed method (DC-RF-RUS-PF)	0.229	31.7	61.2	66.6	65.1	38.2
LORIS (Dhole et al. 2014)	0.221	28.7	69.8	58.6	60.4	38.4
PSIVER (Murakami and Mizuguchi 2010b)	0.151	30.6	41.6	74.3	67.3	35.3

leave-one-out cross-validation test. The first predictor compared is LORIS (Dhole et al. 2014), which identifies PPIs from protein sequences using an L1-regularized logistic regression under the same feature set applied in this study; the second one is PSIVER (Murakami and Mizuguchi 2010b), which is also a sequence-based PPIs predictor but uses a naïve Bayesian classifier with position-specific scoring matrices (PSSM) and predicted accessibility (PA) as feature sources.

Table 3 summarizes the performance comparisons between the proposed method, LORIS, and PSIVER. The table shows that the proposed method significantly outperformed PSIVER with respect to all evaluation indices except for *Specificity*. Improvements of 7.8 and 2.9 % on MCC and *F*-measure, respectively, were achieved by the proposed method relative to the values measured for PSIVER. We also observed that the value of *Specificity* for PSIVER is 7.7 % higher than that of the proposed method. However, the value of Recall of PSIVER is only 41.6 %, which is 19.6 % lower than that of the proposed method, indicating that too many false negatives were incurred during the predictions of PSIVER. Regarding LORIS, which is a recently reported PPI predictor, the proposed method achieves comparable performance. Because the proposed method and LORIS used the same feature set, we can argue that the proposed DC procedure and post-filtering procedure are effective for PPI prediction.

Comparisons with Existing PPIs Predictors Over Independent Validation Test

In this section, we explore the generalization capability of the proposed method, i.e., DC-RF-RUS-PF, by comparing it with existing PPIs predictors on two independent test datasets, i.e., Dtestset72 and PDBtestset164. In addition to LORIS (Dhole et al. 2014) and PSIVER (Murakami and Mizuguchi 2010b), several other existing sequence-based PPIs predictors, including SPRINGS (Singh et al. 2014), ISIS (Ofran and Rost 2007), and SPPIDER (Porollo and Meller 2007), were used.

Table 4 lists the performance comparisons between the proposed method and five other sequence-based PPIs predictors on the independent validation dataset Dtestset72. To fairly compare with previously developed predictors, for Dtestset72, we calculated the individual prediction performances for the 27 rigid body cases, the 6 medium cases, the 3 difficult cases (Murakami and Mizuguchi 2010b), and the overall averaged prediction performance on the entire dataset.

Table 4 shows that the proposed method outperformed five other PPIs predictors for each of the three cases. Regarding the overall prediction performance, average improvements of 2.7 and 1.2 % on MCC and *F*-measure, respectively, were achieved by the proposed DC-RF-RUS-PF compared with LORIS, which is the most recently released sequence-based PPIs predictor.

We acknowledge that ISIS exhibits the best performance in terms of *Accuracy*. However, under the class imbalance scenario, over-pursuing overall accuracy is not appropriate and can be deceiving in evaluating the performance of a predictor/classifier (He and Garcia 2009b; Yu et al. 2013b, c). As shown in Table 4, the *Recall* value of ISIS is only 35.0 %, which is the lowest of the six consider predictors. In other words, ISIS predicts too many false negatives, leading to a very small value of MCC, which is the overall measurement of the quality of binary predictions.

Table 5 lists the performance comparisons between the proposed method and other four sequence-based PPIs predictors on the independent validation dataset PDBtestset164. Table 5 shows that the DC-RF-RUS-PF method again achieved the best performance on PDBtestset164. Moreover, DC-RF-RUS-PF is significantly better than SPRINGS (Singh et al. 2014), PSIVER (Murakami and Mizuguchi 2010b), and SPPIDER (Porollo and Meller

Table 4 Performance comparisons between the proposed method and other sequence-based PPIs predictors on the independent validation dataset Dtestset72

Method	MCC	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	*F*-measure (%)
Rigid body cases (27)						
DC-RF-RUS-PF	0.193	24.7	62.2	63.8	63.3	32.4
LORIS (Dhole et al. 2014)	0.175	23.2	63.8	60.3	60.9	32.0
SPRINGS (Singh et al. 2014)	0.167	23.5	59.2	62.5	62.1	31.3
PSIVER (Murakami and Mizuguchi 2010b)	0.127	23.9	46.5	68.8	65.5	27.3
ISIS (Ofran and Rost 2007)	0.110	22.0	37.9	75.7	70.9	25.9
SPPIDER (Porollo and Meller 2007)	0.087	20.4	44.7	65.2	62.9	24.4
Medium cases (6)						
DC-RF-RUS-PF	0.256	29.1	64.5	68.1	67.5	36.7
LORIS (Dhole et al. 2014)	0.187	25.0	60.9	63.4	63.3	32.9
SPRINGS (Singh et al. 2014)	0.197	26.2	59.1	65.6	64.9	33.7
PSIVER (Murakami and Mizuguchi 2010b)	0.171	28.9	43.5	75.3	70.2	27.1
ISIS (Ofran and Rost 2007)	0.050	18.4	23.0	82.6	75.2	19.0
SPPIDER (Porollo and Meller 2007)	0.055	19.4	36.1	68.4	62.7	18.4
Difficult cases (3)						
DC-RF-RUS-PF	0.231	29.8	64.5	66.2	65.3	39.0
LORIS (Dhole et al. 2014)	0.174	26.5	61.1	62.7	61.8	35.5
SPRINGS (Singh et al. 2014)	0.143	24.9	57.7	62.3	60.3	32.8
PSIVER (Murakami and Mizuguchi 2010b)	0.139	26.9	53.2	61.9	62.8	33.2
ISIS (Ofran and Rost 2007)	0.001	17.8	33.5	67.7	62.4	23.0
SPPIDER (Porollo and Meller 2007)	0.070	22.1	70.4	41.3	49.3	32.7
Overall average performance (72)						
DC-RF-RUS-PF	0.204	25.6	62.7	64.6	64.0	33.6
LORIS (Dhole et al. 2014)	0.177	23.8	63.1	61.0	61.4	32.4
SPRINGS (Singh et al. 2014)	0.170	24.1	59.0	63.0	62.4	31.8
PSIVER (Murakami and Mizuguchi 2010b)	0.135	25.0	46.5	69.3	66.1	27.8
ISIS (Ofran and Rost 2007)	0.091	21.0	35.0	76.2	70.9	24.5
SPPIDER (Porollo and Meller 2007)	0.081	20.4	45.4	63.7	61.7	24.1

Table 5 Performance comparisons between the proposed method and other predictors on the independent validation dataset PDBtestset164

Method	MCC	Precision (%)	Recall (%)	Specificity (%)	Accuracy (%)	*F*-measure (%)
Proposed method (DC-RF-RUS-PF)	0.148	32.4	52.6	65.3	61.1	36.0
LORIS (Dhole et al. 2014)	0.111	26.3	53.8	60.9	58.8	32.3
SPRINGS (Singh et al. 2014)	0.108	26.8	40.7	64.8	60.6	31.1
PSIVER (Murakami and Mizuguchi 2010b)	0.078	25.3	46.4	63.4	59.6	29.5
SPPIDER (Porollo and Meller 2007)	0.015	23.1	16.2	85.1	71.6	12.9

2007). Compared with LORIS (Dhole et al. 2014), which outperformed other existing methods, DC-RF-RUS-PF also makes an improvement of 3.7 % on MCC and F-measure.

Clearly, the test results demonstrate that the generalization capability of the proposed method outperforms that of the previously reported methods. The good performance on an independent test further demonstrates the effectiveness of the proposed method for protein–protein interaction prediction.

Conclusions

In this study, we developed a DC-RF-RUS-PF algorithm for protein–protein interaction prediction. Experimental results obtained for benchmark datasets demonstrate the superiority of the proposed method over the existing PPI predictors. The good performance of the proposed method is derived from the use of the combined discriminative feature of protein residues and the powerful RF classification algorithm, particularly a data-cleaning procedure that can remove marginal targets and a post-processing procedure that can potentially reduce predicted false positives. Because PPI prediction is a typical imbalanced learning problem, our main focus is on cleaning 'harmful' non-interactive residues. In this respect, the aim is to eliminate classification bias in training models caused by class overlap attributed to class imbalance. The results of cross-validation and independent validation tests indicate the efficacy of the method. Our method is not limited to PPI prediction and can be applied to other bioinformatics problems in which serious class imbalance exists. As demonstrated in a series of recent publications (Chen et al. 2014, 2015; Ding et al. 2014; Jia et al. 2015b; Lin et al. 2014; Liu et al. 2015b, c; Guo et al. 2014; Chen et al. 2013), in developing new prediction methods, user-friendly and publicly accessible web servers will significantly enhance the effects of this imbalance (Chou 2015). Therefore, we will make efforts in our future work to provide a web server for the prediction method presented in this paper. Nevertheless, the benchmark datasets and the source code for the proposed method and have been made available at http://csbio.njust.edu.cn/bioinf/PPIS for free academic use.

Our future work will focus on further enhancing the accuracy with which protein–protein interaction sites are predicted by incorporating new discriminative features and powerful classification algorithms. Currently, the proposed method requires approximately 100 s for predicting a sequence with 300 residues. We will further optimize the method to improve the computational efficiency.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 61373062, 61233011, and 61222306), the Jiangsu Postdoctoral Science Foundation (No. 1201027C), the Natural Science Foundation of Jiangsu (No. BK20141403), the China Postdoctoral Science Foundation (No. 2014T70526 and 2013 M530260), the Fundamental Research Funds for the Central Universities (No. 30920130111010), and "The Six Top Talents" of Jiangsu Province (No. 2013-XXRJ-022).

References

Agrawal NJ, Helk B, Trout BL (2014) A computational tool to predict the evolutionarily conserved protein–protein interaction hot-spot residues from the structure of the unbound protein. FEBS Lett 588:326–333

Ahmed Z, Tetlow IJ, Ahmed R, Morell MK, Emes MJ (2015) Protein–protein interactions among enzymes of starch biosynthesis in high-amylose barley genotypes reveal differential roles of heteromeric enzyme complexes in the synthesis of A and B granules. Plant Sci 233:95–106

Ako-Adjei D, Fu W, Wallin C, Katz KS, Song G, Darji D, Brister JR, Ptak RG, Pruitt KD (2015) HIV-1, human interaction database: current status and new features. Nucleic Acids Res 43:D566–D570

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424

Betel D, Breitkreuz KE, Isserlin R, Dewar-Darch D, Tyers M, Hogue CW (2007) Structure-templated predictions of novel protein interactions from sequence information. PLoS Comput Biol 3:e182

Bock JR, Gough DA (2001) Predicting protein–protein interactions from primary structure. Bioinformatics 17:455–460

Bradford JR, Westhead DR (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 21:1487–1494

Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein–protein interfaces using a Bayesian network prediction method. J Mol Biol 362:365–386

Breiman L (2001) Random forests. Mach Learn 45:5–32

Burgoyne NJ, Jackson RM (2006) Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces. Bioinformatics 22:1335–1342

Chen X-W, Jeong J-C (2009) Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 25:585–591

Chen C-T, Peng H-P, Jian J-W, Tsai K-C, Chang J-Y, Yang E-W, Chen J-B, Ho S-Y, Hsu W-L, Yang A-S (2012) Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. PLoS One 7:e37706

Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68

Chen W, Feng P-M, Deng E-Z, Lin H, Chou K-C (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 462:76–83

Chen W, Feng P, Ding H, Lin H, Chou K-C (2015) iRNA-methyl: identifying N 6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 490:26–33

Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256:705–708

Chou K (2001) Using subsite coupling to predict signal peptides. Protein Eng 14:75–79

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247

Chou KC (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol Biosyst 9:1092–1100

Chou K-C (2015) Impacts of bioinformatics to medicinal chemistry. Med Chem 11:218–234

Cukuroglu E, Gursoy A, Nussinov R, Keskin O (2014) Non-redundant unique interface structures as templates for modeling protein interactions. PLoS One 9:e86738

DeLano WL (2002) The PyMOL molecular graphics system, http://www.pymol.org

Dhole K, Singh G, Pai PP, Mondal S (2014) Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. J Theor Biol 348:47–54

Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H, Chen W, Chou K-C (2014) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res Int. doi:10.1155/2014/286419

Drewes G, Bouwmeester T (2003) Global approaches to protein–protein interactions. Curr Opin Cell Biol 15:199–205

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18:529–536

Ertekin S, Huang J, Bottou L, Giles L (2007a). Learning on the border: active learning in imbalanced data classification. In: ACM Conference on Information and Knowledge Management, pp 127–136

Ertekin S, Huang J, Giles CL (2007b) Active learning for class imbalance problem. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Amsterdam, pp 823–824

Estabrooks A, Jo TH, Japkowicz N (2004) A multiple resampling method for learning from imbalanced data sets. Comput Intell 20:18–36

Fariselli P, Pazos F, Valencia A, Casadio R (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269:1356–1361

Friedrich T, Pils B, Dandekar T, Schultz J, Müller T (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics 22:2851–2857

Fry DC (2015) Targeting protein-protein interactions for drug discovery. Protein Protein Interact Methods Appl 1278:93–106

Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) A fast method to predict protein interaction sites from sequences. J Mol Biol 302:917–926

Gromiha MM, Yokota K, Fukui K (2009) Energy based approach for understanding the recognition mechanism in protein–protein complexes. Mol Biosyst 5:1779–1786

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529

Hall DA, Ptacek J, Snyder M (2007) Protein microarray technology. Mech Ageing Dev 128:161–167

He H-B, Garcia EA (2009a) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21:1263–1284

He H, Garcia EA (2009b) Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 21:1263–1284

He X, Han K, Hu J, Yan H, Yang J-Y, Shen H-B, Yu D-J (2015) TargetFreeze: identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. J Membr Biol 19(1):1–10

Hong X, Chen S, Harris CJ (2007) A kernel-based two-class classifier for imbalanced data sets. IEEE Trans Neural Networks 18:28–41

Hu L, Huang T, Shi X, Lu W-C, Cai Y-D, Chou K-C (2011) Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS One 6:e14556

Hu J, He X, Yu D-J, Yang X-B, Yang J-Y, Shen H-B (2014) A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. PLoS One 9(9):107676

Hubbard SJ, Thornton JM (1993) Naccess. Computer Program, vol 2, Department of Biochemistry and Molecular Biology, University College, London

Hwang H, Pierce B, Mintseris J, Janin J, Weng Z (2008) Protein–protein docking benchmark version 3.0. Proteins Struct Function Bioinform 73:705–709

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci 97:1143–1147

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci 98:4569–4574

Jia J, Liu Z, Xiao X, Liu B, Chou K-C (2015a) Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J Biomol Struct Dyn. doi:10.1080/07391102.2015.1095116

Jia J, Liu Z, Xiao X, Liu B, Chou K-C (2015b) iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 377:47–56

Jia J, Xiao X, Liu B (2015c) Prediction of protein-protein interactions with physicochemical descriptors and wavelet transform via random forests. J Lab Autom. doi:10.1177/2211068215581487

Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63:31–65

Jones S, Thornton JM (1997a) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272:121–132

Jones S, Thornton JM (1997b) Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272:133–143

Joo K, Lee SJ, Lee J (2012) Sann: solvent accessibility prediction of proteins by nearest neighbor method. Proteins Struct Function Bioinform 80:1791–1797

Kang PS, Cho SZ (2006) EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems. Neural Inf Process Proc 4232(1):837–846

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. Artif Intell Med Proc 2101:63–66

Lin WZ, Fang JA, Xiao X, Chou KC (2013) iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst 4:634–644

Lin H, Deng E-Z, Ding H, Chen W, Chou K-C (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42:12961–12972

Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou K-C (2014) iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9(9):e106691

Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C (2015a) Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10:e0121501

Liu B, Fang L, Wang S, Wang X, Li H, Chou K-C (2015b) Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J Theor Biol 385:153–159

Liu Z, Xiao X, Qiu W-R, Chou K-C (2015c) iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem 474:69–77

Marceau AH, Bernstein DA, Walsh BW, Shapiro W, Simmons LA, Keck JL (2013) Protein interactions in genome maintenance as novel antibacterial targets. PLoS One 8(3):e58765

Mihel J, Šikić M, Tomić S, Jeren B, Vlahoviček K (2008) PSAIA–protein structure and interaction analyzer. BMC Struct Biol 8:21

Murakami Y, Mizuguchi K (2010a) Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. Bioinformatics 26:1841–1848

Murakami Y, Mizuguchi K (2010b) Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics 26:1841–1848

Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544:236–239

Ofran Y, Rost B (2007) ISIS: interaction sites identified from sequence. Bioinformatics 23:e13–e16

Porollo A, Meller J (2007) Prediction-based fingerprints of protein–protein interactions. Proteins Struct Function Bioinform 66:630–645

Russell RB, Aloy P (2008) Targeting and tinkering with interaction networks. Nat Chem Biol 4:666–673

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Sharon M, Sinz A (2015). Studying protein–protein interactions by combining native mass spectrometry and chemical cross-linking. Analyzing biomolecular interactions by mass spectrometry, pp 55–79

Šikić M, Tomić S, Vlahoviček K (2009) Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 5:e1000278

Singh G, Dhole K, Pai PP, Mondal S (2014) SPRINGS: prediction of protein-protein interaction sites using artificial neural networks. PeerJ 1:7

Skrabanek L, Saini HK, Bader GD, Enright AJ (2008) Computational prediction of protein–protein interactions. Mol Biotechnol 38:1–17

Sudha G, Nussinov R, Srinivasan N (2014) An overview of recent advances in structural bioinformatics of protein–protein interactions and a guide to their principles. Prog Biophys Mol Biol 116:141–150

Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng 14:659–665

Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417:399–403

Wang BX, Japkowicz N (2010) Boosting support vector machines for imbalanced data sets. Knowl Inf Syst 25:1–20

Wang B, Chen P, Huang D-S, J-j Li, Lok T-M, Lyu MR (2006) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 580:380–384

Wu G, Chang EY (2005) KBA: kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng 17:786–795

Xiao X, Wang P, Lin WZ, Jia JH, Chou KC (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436:168–177

Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C (2015a) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J Biomol Struct Dyn 33(10): 2221–2233

Xiao X, Zou H-L, Lin W-Z (2015b) iMem-Seq: a multi-label learning classifier for predicting membrane proteins types. J Membr Biol 248:745–752

Xu Y, Wen X, Wen L-S, Wu L-Y, Deng N-Y, Chou K-C (2014) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9:e105018

Yan C, Dobbs D, Honavar V (2003) Identification of surface residues involved in protein-protein interaction—a support vector machine approach, intelligent systems design and applications. Springer, Berlin, pp 53–62

Yan C, Dobbs D, Honavar V (2004) A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 20:i371–i378

Yu D-J, Shen H-B, Yang J-Y (2011) SOMRuler: a novel interpretable transmembrane helices predictor. IEEE Trans NanoBiosci 10:121–129

Yu D-J, Hu J, Wu X-W, Shen H-B, Chen J, Tang Z-M, Yang J, Yang J-Y (2013a) Learning protein multi-view features in complex space. Amino Acids 44:1365–1379

Yu DJ, Hu J, Huang Y, Shen HB, Qi Y, Tang ZM, Yang JY (2013b) TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. J Comput Chem 34:974–985

Yu DJ, Hu J, Tang ZM, Shen HB, Yang J, Yang JY (2013c) Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. Neurocomputing 104:180–190

Yugandhar K, Gromiha MM (2014a) Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches. Proteins Struct Funct Bioinform 82:2088–2096

Yugandhar K, Gromiha MM (2014b) Protein-protein binding affinity prediction from amino acid sequence. Bioinformatics 30(24):3583–3589

Zhou ZH, Liu XY (2010) On multi-class cost-sensitive learning. Comput Intell 26:232–257

Zhou H-X, Shan Y-B (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins Struct Funct Bioinform 44:336–343

Zou H-L, Xiao X (2015) A new multi-label classifier in identifying the functional types of human membrane proteins. J Membr Biol 248:179–186