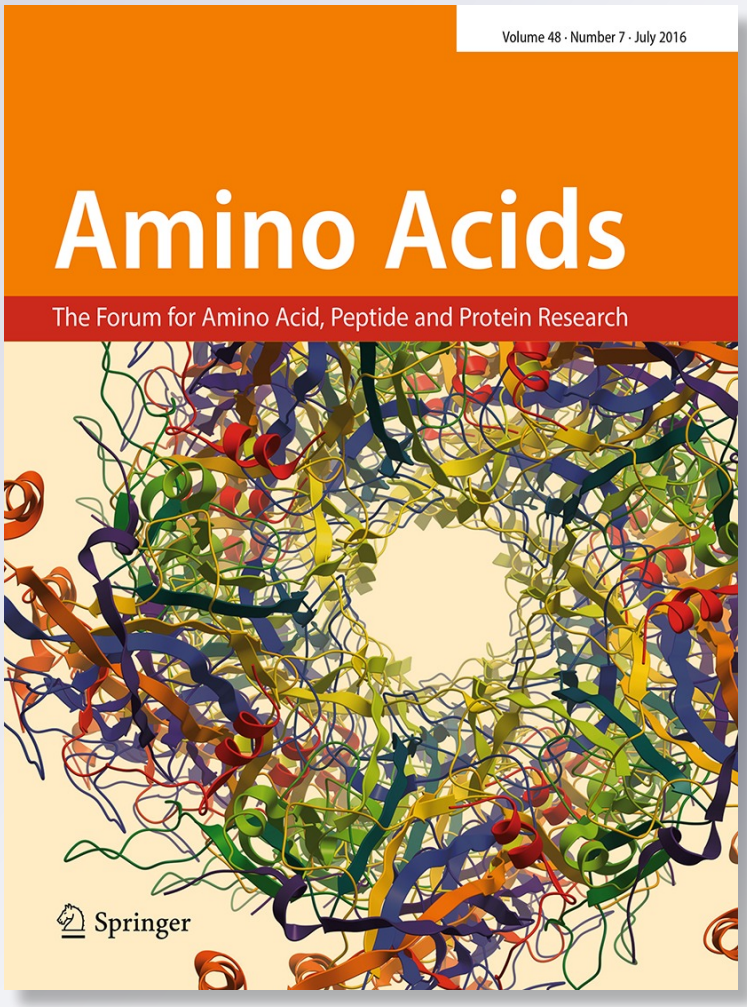
SPAR: a random forest-based predictor for self-interacting proteins with fine-grained domain information

Xuhan Liu, Shiping Yang, Chen Li, Ziding Zhang & Jiangning Song

Amino Acids
The Forum for Amino Acid, Peptide and Protein Research

ISSN 0939-4451
Volume 48
Number 7

Amino Acids (2016) 48:1655-1665
DOI 10.1007/s00726-016-2226-z



Springer

Amino Acids (2016) 48:1655–1665
DOI 10.1007/s00726-016-2226-z

ORIGINAL ARTICLE

SPAR: a random forest-based predictor for self-interacting proteins with fine-grained domain information

Xuhan Liu[1] · Shiping Yang[1] · Chen Li[2] · Ziding Zhang[1] · Jiangning Song[2,3,4]

Received: 21 September 2015 / Accepted: 30 March 2016 / Published online: 13 April 2016

Abstract Protein self-interaction, i.e. the interaction between two or more identical proteins expressed by one gene, plays an important role in the regulation of cellular functions. Considering the limitations of experimental self-interaction identification, it is necessary to design specific bioinformatics tools for self-interacting protein (SIP) prediction from protein sequence information. In this study, we proposed an improved computational approach for SIP prediction, termed SPAR (Self-interacting Protein Analysis serveR). Firstly, we developed an improved encoding scheme named critical residues substitution (CRS), in which the fine-grained domain–domain interaction information was taken into account. Then, by employing the Random Forest algorithm, the performance of CRS was evaluated and compared with several other encoding schemes commonly used for sequence-based protein–protein interaction prediction. Through the tenfold cross-validation tests on a balanced training dataset, CRS performed the best, with the average accuracy up to 72.01 %. We further integrated CRS with other encoding schemes and identified the most important features using the mRMR (the minimum redundancy maximum relevance) feature selection method. Our SPAR model with selected features achieved an average accuracy of 92.09 % on the human-independent test set (the ratio of positives to negatives was about 1:11). Besides, we also evaluated the performance of SPAR on an independent yeast test set (the ratio of positives to negatives was about 1:8) and obtained an average accuracy of 76.96 %. The results demonstrate that SPAR is capable of achieving a reasonable performance in cross-species application. The SPAR server is freely available for academic use at http://systbio.cau.edu.cn/zzdlab/spar/.

Keywords Self-interacting protein · Prediction · Machine learning · Feature selection · Domain–domain interaction

Handling Editor: L. Taher.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-016-2226-z) contains supplementary material, which is available to authorized users.

✉ Ziding Zhang
zidingzhang@cau.edu.cn

✉ Jiangning Song
Jiangning.Song@monash.edu

[1] State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China

[2] Infection and Immunity Program, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia

[3] Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

[4] National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

Introduction

Protein–protein interactions (PPIs) have received much attention due to their important roles in living organisms. Whether and how proteins interact with their protein partners is a fundamental question for their functional studies. Protein self-interaction is a special type of PPI, where two interaction partners are two identical copies expressed by the same gene, and it results in the formation of homo-oligomer. Recent studies have shown that

homo-oligomerization plays a crucial role in a wide range of biological processes, such as gene expression regulation, signal transduction, enzyme activation and immune response (Woodcock et al. 2003; Koike et al. 2009; Baisamy et al. 2005; Hattori et al. 2003; Katsamba et al. 2009).

The biological advantages of homo-oligomer over monomer have been elaborated by Marianayagam et al. (2004). For instance, self-interaction is a key factor in the regulation of protein function through allostery. Through self-interaction, the functional diversity of proteins can be vastly extended without the need of increasing genome size. In addition, self-interaction is conducive to improving the stability and preventing the denaturation of a protein by reducing its surface area (Miller et al. 1987).

Self-interacting proteins (SIPs) have a significant disposition to be located at the hub in protein interaction networks (PINs). In other words, they can interact with a large number of other protein partners, which indicates its functional importance for cellular systems (Ispolatov et al. 2005). Due to high-level expression, SIPs generally have lower aggregation propensities against misfolding (Chen and Dokholyan 2008). In addition, genes that encode SIPs tend to have higher duplicability than others, and they appear to have arisen more often at the whole-genome level rather than at the small scale (Perez-Bercoff et al. 2010). Moreover, exon-shuffling events might promote the acquisition of self-interacting capacity in proteins and the creation of novel PPIs (Cancherini et al. 2010).

Previous studies have shown that self-interaction is mediated by specific protein domains. To study the mechanisms of self-interaction, Akiva et al. (2008) used self-interacting domains as a model and found that enabling/disabling loops uncovered at surface could be considered as determinants of the interactions. In addition, Hashimoto and Panchenko (2010) observed that insertions and deletions play a critical role in maintaining different oligomeric states of SIPs. Up to date, several molecular mechanisms of self-interaction have been proposed (Hashimoto et al. 2011), including domain swapping, ligand-induced, residue substitution or post-translational modifications (PTMs) at the interfaces, insertions and deletions.

Although a variety of experimental and computational methods have been designed for PPI identification (Zhou et al. 2012; Zahiri et al. 2013; Zaki et al. 2009), these methods have certain limitations when being applied to protein self-interaction identification. Firstly, due to biological artifacts and design limitations, two common types of high-throughput protein interaction assays, i.e. Y2H and TAP/MS, have limited capacity for detecting SIPs (Gibson and Goldberg 2009). Secondly, a majority of computational methods for PPI prediction often consider the correlational information between protein pairs, such as co-expression, co-localization and coevolution. However, such information is unavailable when dealing with two identical protein partners. Moreover, the datasets used to construct most of the methods did not include the PPIs between identical partners, which makes them not suitable for SIP prediction. In a previous study, Liu et al. (2013) integrated multiple representative known properties to construct a prediction model and developed an online predictor named SLIPPER for human SIP prediction. Note that SLIPPER is not a pure sequence-based predictor, as the PIN feature (i.e. network degree) made the largest contribution to its performance. Therefore, the major drawback of this method is that it cannot deal with the proteins that are not covered by the current human interactome.

Given the aforementioned limitations of existing methods, it is desirable to develop a more effective computational approach for protein self-interaction prediction based on protein sequences. Using the Random Forest (RF) algorithm (Breiman 2001), in this study, we designed a RF-based approach, termed SPAR (Self-interacting Protein Analysis serveR), for predicting SIPs based on an improved sequence-encoding scheme, which exploits the fine-grained domain–domain interaction (DDI) information from the 3did database (Mosca et al. 2014).

Materials and methods

Data collection and dataset construction

A total of 20,199 curated human protein sequences were downloaded from the UniProt database (version 2015.04.01) (UniProt 2015). The PPI data were collected from a variety of resources, including DIP (version 20150101) (Salwinski et al. 2004), BioGRID (version 3.3.123) (Chatr-Aryamontri et al. 2015), IntAct (version 2015.04.09) (Orchard et al. 2014), InnateDB (version 2015.04.20) (Breuer et al. 2013) and MatrixDB (version 2014.12.16) (Launay et al. 2015). Here, we only extracted those PPIs for which the two interaction partners were identical and whose interaction type was annotated as 'direct interaction' in relevant databases. As a result, we obtained 2994 human protein self-interaction instances.

In order to train a prediction model and evaluate its performance properly, a golden standard (GS) dataset was constructed. In the first step, the short (<50 residues) and long (>5000 residues) proteins were removed from the whole human proteome. Then, for the GS positive dataset (GSP), we refined the protein self-interaction data by ensuring that each sample in GSP must be a high-quality SIP, which satisfies one of the following conditions: (1) the self-interaction is detected by at least two kinds of large-scale experiments or one small-scale experiment; (2) the

protein is annotated as homo-oligomer (including homodimer and homotrimer) in UniProt; (3) the self-interaction is reported by at least two publications. Here, we mainly referred to the BioGRID database (Chatr-Aryamontri et al. 2015) for the definition of experimental methods. To construct the GS negative dataset (GSN), all types of the SIPs were removed from the whole human proteome (including proteins annotated as 'direct interaction' and more extensive 'physical association'). In addition, the predicted SIPs annotated in UniProt were also removed. The resulting GS dataset included 1441 SIPs as positives and 15,938 non-SIPs as negatives.

Furthermore, to investigate the cross-species performance of SPAR, we also used the same strategy as described above to construct the yeast GS dataset, which contained 710 positive samples and 5511 negative samples.

Sequence-encoding schemes

In this study, we proposed an improved feature-encoding scheme, named critical residues substitution (CRS). Firstly, we used HMMER3 (Finn et al. 2011) to identify the significant Pfam domains (Finn et al. 2014) contained in each protein sequence. All the parameters were set as default. After that, we could extract the self-interacting domain information, by querying the DDI information in the 3did database (version 2015.03.11). Furthermore, we could identify the possible sites at the binding interface through the alignment between the query sequence and the domain consensus sequence in HMMER3 results. As shown in Fig. 1, our encoding scheme involved in two types of residue substitutions: the mostly conserved residues and the residues at DDI interface. They were regarded as critical information for protein self-interaction and were extracted to construct our feature vector. Because there are a total of 41 possible types (which correspond to 20 types of conserved amino acids represented by uppercase, 20 types of non-conserved amino acids represented by lowercase and the gap '-') in consensus sequence and 21 types in query sequence (which correspond to 20 types of amino acids and the gap '-') in each sequence alignment, there exist $41 \times 21 - 1 = 860$ possible substitutions (it is impossible for the gap to occur at one site in both the consensus sequence and the query sequence). For each substitution X $\rightarrow$ Y from the consensus sequence to the sample sequence, the uppercase letters for X in the consensus sequence denote conserved sites, while the others are not conserved. All the Ys from the sample sequence are represented by uppercase. Accordingly, each sequence is encoded as a 860-dimensional vector as follows:

$$\left(\frac{N_{\mathrm{AA}}}{N_{\mathrm{total}}}, \frac{N_{\mathrm{AC}}}{N_{\mathrm{total}}}, \ldots, \frac{N_{\mathrm{A-}}}{N_{\mathrm{total}}}, \frac{N_{\mathrm{aA}}}{N_{\mathrm{total}}}, \ldots, \frac{N_{\mathrm{-Y}}}{N_{\mathrm{total}}}\right)_{860}$$

where N_{total} is the total count of all substitutions, and N_{XY} is the count of each type of substitution in a single protein sample. Each dimension in this vector stands for the frequency of the corresponding substitution.

In order to benchmark the prediction performance of this proposed encoding scheme, six other encoding schemes that are commonly used for sequence-based PPI prediction were also involved in our study, i.e. auto-covariance (AC) (Guo et al. 2008), Moran autocorrelation (MAC), Geary

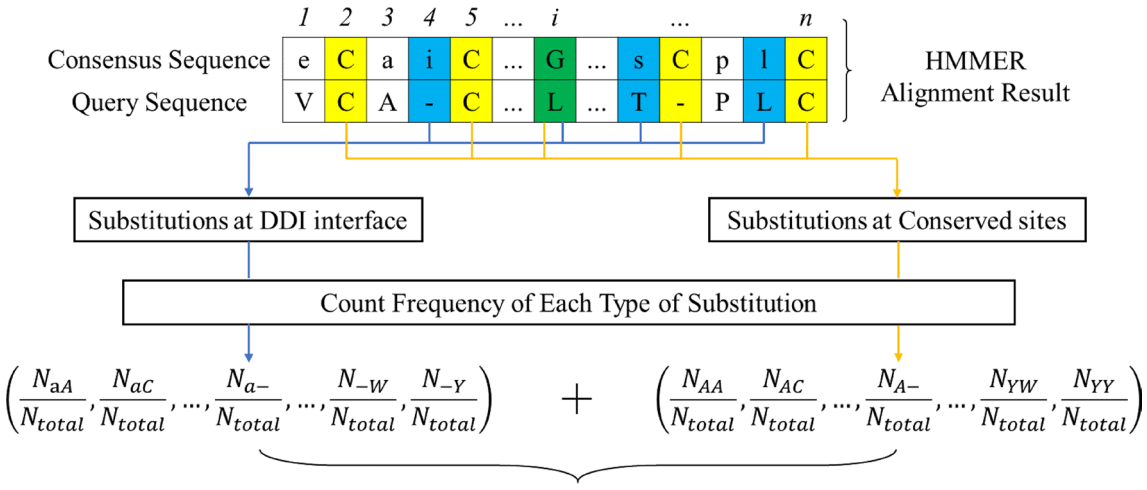


Fig. 1 Computational framework of the proposed CRS-encoding scheme. For the consensus sequence, the *uppercase* means the corresponding site is conserved and the *lower case* denotes that the corresponding site is non-conserved. The critical residues at the DDI interface are colored by *yellow*, while the critical residues at conserved sites are colored by *blue*. In addition, a residue would be colored by *green*, if it is located at both conserved site and the DDI interface

autocorrelation (GAC) (Xia et al. 2010), Moreau–Broto autocorrelation (MBMAC) (Feng and Zhang 2000), conjoint triad (CT) (Shen et al. 2007) and local descriptor (LD) (Yang et al. 2010). These encoding schemes could transform a protein sequence to a feature vector. These six feature schemes can be calculated automatically online using the PROFEAT website (Rao et al. 2011). For AC, MAC, GAC and MBMAC, six physicochemical properties of amino acids used in a previous PPI prediction study (You et al. 2013) were taken into account, including hydrophobicity (H), volumes of side chains (VSC), polarity (P1), polarizability (P2), solvent-accessible surface area (SASA) and net charge index of side chains (NCISC). The original values of these properties are given in Supplementary Table S1. The corresponding feature dimensions generated by these schemes were 180, 180, 180, 180, 343 and 630, respectively.

Feature selection

After integrating the extracted features from sequence, mRMR method was applied to rank and select more important features (Peng et al. 2005). The mRMR method ranks features in a two-step procedure based on the mutual information (MI). In the first step, features are ranked based on the relevance D between each feature f and label variables l in the feature set T, which can be calculated by:

$$D = \mathrm{MI}(f, l)$$

In the second step, each feature in T will be moved into S one by one based on the mRMR value, which is defined as:

$$\max_{f_j \in T} \left(D_j - \frac{1}{|s|} \sum_{f_i \in S} \mathrm{MI}(f_j, f_i) \right)$$

It will be calculated for $|T|$ rounds, and for each round, only one feature can be selected to move from T to S. Finally, the ranking of all features is obtained where the earlier features put into S would be placed on the top. The mRMR program can be downloaded from http://penglab.janelia.org/proj/mRMR/.

Model construction and evaluation

In our study, RF was adopted to build the prediction model, given the fact that it is popular in the field of bioinformatics and has been shown to provide competitive performance compared with other machine learning techniques in many applications. RF is an ensemble learning algorithm that consists of a certain number of decision trees. In the training process, RF constructs a probability distribution model by assembling the decision trees. For a given input feature vector, all decision trees in the trained model can give vote to predict which class it belongs to. We also compared our model with other machine learning algorithms, such as support vector machine (SVM), Naïve Bayes (NB) and k-nearest neighbor (KNN). Here, we employed the *scikit-learn* (Pedregosa et al. 2011), a Python language package, as an implementation of these algorithms. For the RF algorithm, the number of trees was set as 1000, and other parameters were set as default. The maximum of feature number of the RF classifier was set as the square root of the total number of features in each tree, and the criterion of the quality of a split was based on the Gini impurity by default. Regarding the SVM algorithm, the radial basis function (RBF) kernel was used. The range of parameters C and γ were set as $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{5}]$. After optimization, they were set as 1.0 and 0.25, respectively. For the KNN algorithm, the Euclidean distance was used to measure the distance between any two samples and k was set as 5 (by default).

In order to train and evaluate our model, the GS dataset was further partitioned into the training set and independent test set. We randomly selected 1/6 of the samples from both positive and negative human GS datasets as the independent test set. Given that the number of negative instances is much larger than the positive ones in GS dataset, we randomly chose negative samples from the remaining human GS negative dataset to construct the training set with the ratio of 1:1. In order to ensure the reliability of performance evaluation, the negative training set was repeatedly constructed ten times, and the results were reported in the form of "mean ± SD (standard deviation)".

Furthermore, we also performed tenfold cross-validation tests to benchmark and compared the performance of various models based on different types of feature-encoding schemes on the training dataset. The independent test was used for avoiding the over-fitting of feature selection by mRMR. In addition to the human GS dataset, the yeast GS dataset was also used to assess the predictive capability of SPAR for the cross-species prediction of SIPs. Five evaluation measures, accuracy (Ac), sensitivity (Sn), specificity (Sp), F_{measure} and Matthew correlation coefficient (MCC), were used to assess the performance. They are defined as:

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$F_{\mathrm{measure}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

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

where TP, TN, FP, FN represent true positive, true negative, false positive and false negative, respectively. When the above measures were used for performance comparison, all of them were calculated at the threshold where the maximum of MCC was obtained. Receiver operating characteristic (ROC) curves were also plotted to comprehensively evaluate the models' performance. The area under the ROC curves (AUC) was also calculated to quantify the performance.

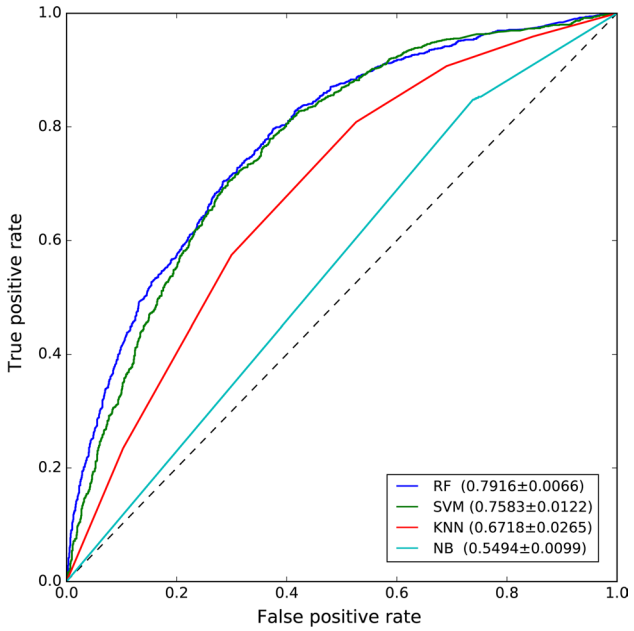


Fig. 2 ROC curves of the different types of machine learning algorithms (i.e. RF, SVM, KNN and NB) based on the results of tenfold cross-validation on the human GS dataset

Results and discussion

Results of tenfold cross-validation

Here, four different kinds of machine learning methods, RF, SVM, KNN and NB, were used to train prediction models and were compared with each other with the CRS-encoding scheme. By tenfold cross-validation on the human GS dataset, we observed that the performance of RF method was better than others, for the AUC reached up to 0.7916 (Fig. 2). Therefore, the RF algorithm was justified to be an excellent machine learning method to construct our prediction model in this work.

Then, we constructed different RF models by using the CRS and other six encoding schemes and evaluated their prediction performance on the human GS dataset (Fig. 3a). Table 1 also illustrates the prediction results via tenfold cross-validation. Compared with other feature-encoding schemes, CRS outperformed other schemes in terms

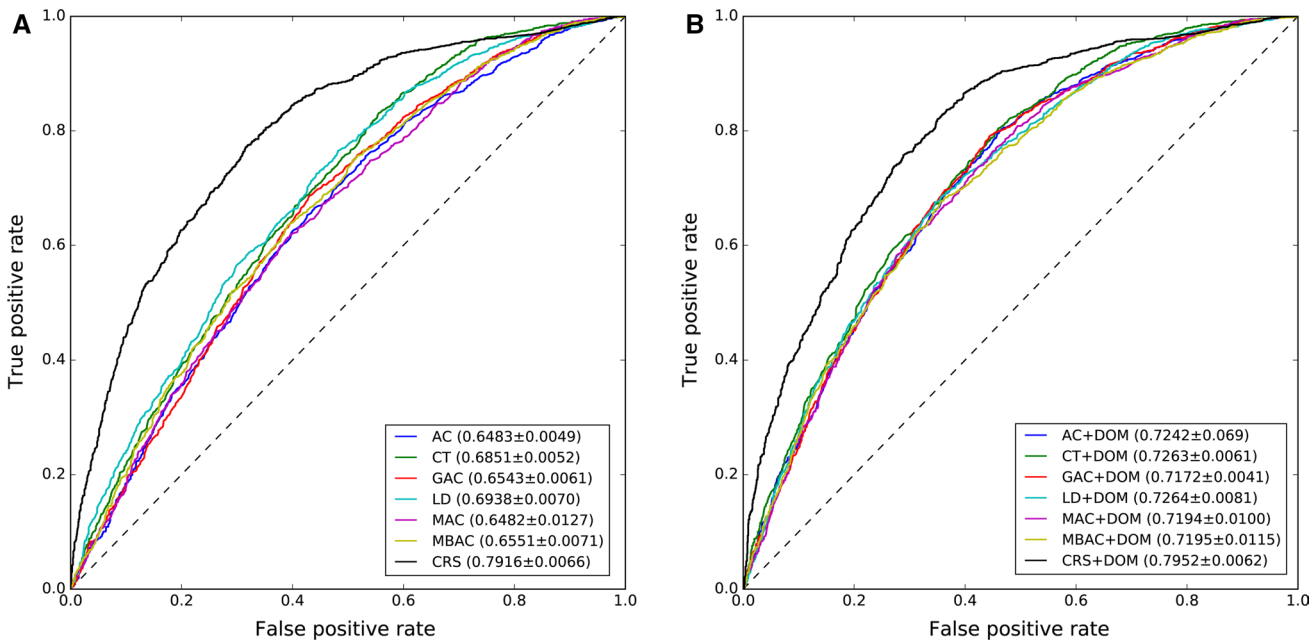


Fig. 3 ROC curves of the different types of feature-encoding schemes based on the results of tenfold cross-validation on the human GS dataset. The performances were obtained (**a**) without domain features and (**b**) with domain features, respectively. *Parameters* in *brackets* are the AUC values of each prediction model

Table 1 Performance comparison of the CRS method and the other six encoding schemes based on tenfold cross-validation on the human GS dataset

Schemes	Ac (%)	Sp (%)	Sn (%)	MCC	$F_{measure}$ (%)
AC	61.47 ± 0.81	51.37 ± 9.04	71.57 ± 7.51	0.2373 ± 0.0082	64.87 ± 1.94
CT	62.91 ± 1.75	37.93 ± 10.04	87.89 ± 6.78	0.3045 ± 0.0150	70.28 ± 0.87
GAC	61.81 ± 0.68	55.41 ± 10.63	68.19 ± 10.74	0.2432 ± 0.0124	63.73 ± 4.10
LD	64.26 ± 1.03	46.28 ± 6.62	82.25 ± 4.86	0.3078 ± 0.0118	69.68 ± 0.89
MAC	61.29 ± 1.53	48.34 ± 11.32	74.24 ± 8.60	0.2385 ± 0.0177	65.59 ± 1.88
MBAC	60.45 ± 1.26	34.77 ± 7.62	86.13 ± 5.35	0.2466 ± 0.0135	68.50 ± 0.84
CRS	72.01 ± 0.87	61.73 ± 6.64	82.29 ± 5.23	0.4528 ± 0.0107	74.58 ± 0.92

Tenfold cross-validation tests on the human GS dataset were used to conduct the performance comparison

Table 2 Performance comparison of the CRS method and the other six encoding schemes after incorporating the domain features (DOM)

Schemes	Ac (%)	Sp (%)	Sn (%)	MCC	$F_{measure}$ (%)
AC + DOM	67.18 ± 0.63	53.17 ± 2.50	81.19 ± 2.08	0.3583 ± 0.0126	71.21 ± 0.63
CT + DOM	66.35 ± 0.93	47.36 ± 7.13	85.34 ± 5.61	0.3569 ± 0.0105	71.68 ± 1.05
GAC + DOM	66.93 ± 0.42	55.60 ± 4.26	78.25 ± 3.85	0.3487 ± 0.0078	70.26 ± 0.98
LD + DOM	66.63 ± 0.93	55.85 ± 8.86	77.42 ± 7.43	0.3453 ± 0.0104	69.78 ± 1.69
MAC + DOM	67.13 ± 0.73	57.71 ± 4.31	76.54 ± 3.34	0.3497 ± 0.0121	69.94 ± 0.79
MBAC + DOM	66.55 ± 0.92	62.11 ± 7.27	70.99 ± 7.94	0.3360 ± 0.0214	67.79 ± 2.89
CRS + DOM	72.50 ± 0.53	62.03 ± 5.28	82.97 ± 5.33	0.4628 ± 0.0136	75.05 ± 1.38

Tenfold cross-validation tests on the human GS dataset were used to conduct the performance comparison

of the average values of Ac = 72.01 %, Sp = 61.73 %, Sn = 82.29 % and MCC = 0.4528 with smaller SD values. These results also demonstrated the robustness of CRS. These six encoding schemes have been proved to be effective for PPI prediction, especially the interaction between different partners. For instance, Guo et al. (2008) applied the AC-encoding scheme to construct an SVM model and achieved an average Ac of 88.09 % in their PPI prediction study. However, we noted that the performance of the AC-encoding scheme tested on our dataset for prediction of protein self-interaction was not remarkable (with an average Ac of only 61.47 %, Table 1).

It is well known that protein self-interactions are determined by self-interacting domains. Indeed, the domain information is taken into account in the CRS-encoding scheme to some extent, but it is completely missed in the other six encoding schemes under investigation. Therefore, to examine whether inclusion of the domain information could help to improve the performance, we further extracted a set of conventional domain information features represented as DOM. The DOM features include the numbers of homo-domain and hetero-domain interactions, each of which includes three types: intra-chain, inter-chain and both. Then we incorporated these six DOM features to each encoding scheme and conducted another tenfold cross-validation to examine the potential effect of domain interaction information on the prediction performance. As expected, the DOM features can greatly improve the prediction performance across almost all the encoding schemes except CRS (Table 2; Fig. 3b). Another important observation is that the performance of CRS was still robust and performed the best compared with other encoding schemes after incorporating the DOM features. In addition, possibly because CRS had contained the fine-grained domain information, the performance of CRS was not significantly improved further after integrating DOM features. These results suggest that the substitutions occurring at both conserved sites, and the DDI interface provides fine-grained domain information in comparison to the DOM features. Therefore, the CRS-encoding scheme is very powerful in the prediction of protein self-interactions.

Feature analysis

Recently, the DDI information was extensively used for PPI prediction. For instance, Shatnawi and Zaki developed a new method for structural domain identification. They predicted whether two proteins could interact or not, based on the identified interacting domain pairs in these two proteins (Shatnawi and Zaki 2015). In contrast, we used the fine-grained domain information in our SIP prediction, in which the substitution of the key residues that play important roles in protein self-interaction was captured by the CRS features. Here, we conducted a brief analysis of the self-interacting domain distribution and CRS features in our dataset.

In the human GS dataset, the percentage of the positive samples containing more than one self-interacting

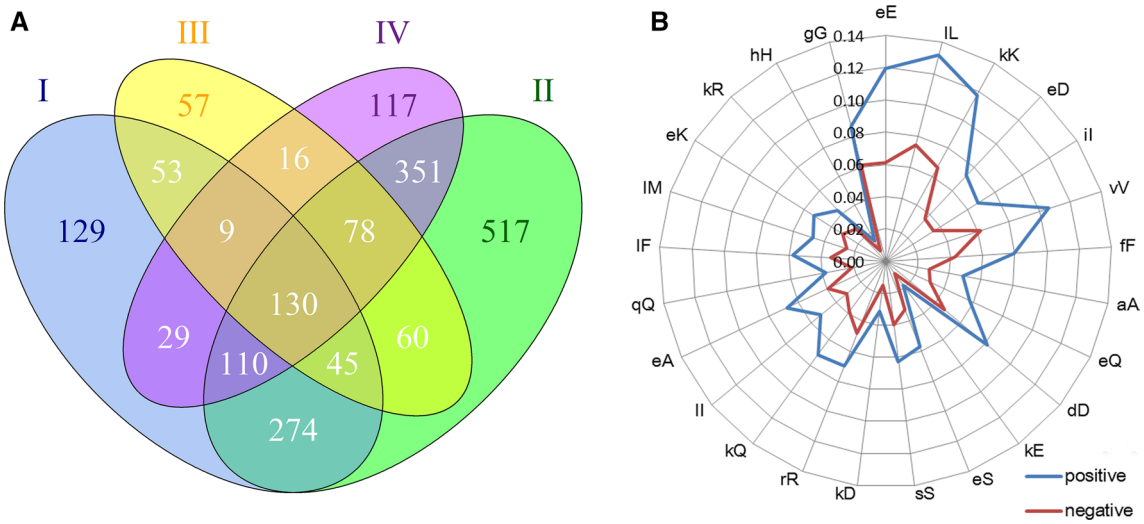


Fig. 4 Venn diagram of the distribution of self-interacting domain in the human and yeast GS datasets (**a**) and radar diagram of the top 25 significant different features compared between positive and negative samples in the human GS dataset (**b**). In *panel* **a**, *I* and *II* denote the positive and negative samples in the human GS dataset, respectively; *III* and *IV* denote the positive and negative samples in the yeast GS dataset, respectively. I, II, III and IV are colored by *blue*, *green*, *violet* and *yellow*, respectively. In *panel* **b**, since all of these features come from the CRS method, the feature names were represented as two characters XY, meaning the substitution from HMM consensus sequence to sample sequence. If the first character is lowercase, it means that it was not a mostly conserved site

domain is 90.07 %, which is significantly higher than that of the negative samples (p value $<2.2 \times 10^{16}$, Wilcoxon rank test). Similar phenomenon was observed in the yeast dataset. From this perspective, the domain features can be exploited as useful discriminative features for distinguishing from proteins that cannot self-interact. All the DDI annotations in the 3did database were collected from 3D protein structures; the coverage of DDI information may not be complete since the increase of 3D structure data cannot keep pace with the sequence data. Therefore, a minority of orphan proteins without domain information cannot be handled appropriately by using domain features. The distribution of self-interacting domains in the human and yeast GS datasets is provided in Fig. 4a. It can be seen that both positive and negative samples contained a variety of self-interacting domains, which might be distinct between different species. It might further imply that only domain features cannot fully capture the distinctive characteristics for all kinds of SIPs, and thus, it is difficult to achieve satisfactory performance for cross-species prediction, given that different species might have evolved to use different types of self-interacting domains.

Using Wilcoxon rank tests, we investigated all statistically significant features between the positive and negative samples in human GS dataset. Figure 4b displays the mean values of top 25 features, all of which were encoded by the CRS method. This also rationalized why CRS achieved a better performance for protein self-interaction prediction. Besides, all of these 25 features represent features that describe the substitution at the DDI interface, rather than the substitution at conserved positions inferred from the multiple sequence alignments of domains. A number of these features represent no substitution, such as 'eE', 'lL' and 'kK', while the others are the substitutions between residues with similar physicochemical properties, such as 'eD', 'iL' and 'eQ'.

Incremental feature selection

Despite achieving a better predictive performance compared with other encoding schemes, CRS does have the following limitation. In a few cases, a possible SIP does not contain any self-interacting domain. Since the CRS method depends on the self-interacting domain information, it has no capability to accurately predict the SIPs without self-interacting domain information. In order to address this issue and improve the performance of the predictive model, we integrated the CRS with other six encodings together and ranked them by the mRMR algorithm on the whole human GS training set. Then features were stepwise selected from top to bottom in the ranked feature set. For each incremental feature selection, the prediction model was constructed on the human GS training set and subsequently evaluated on the human-GS-independent test set and yeast GS dataset (Fig. 5).

As shown in Fig. 5, the maximum MCC in human and yeast dataset were 0.3836 and 0.2585 when the top 292 and 275 features in mRMR rank were selected, respectively. Here, the top 292 features (maximum MCC value in human GS dataset) were chosen to train our final prediction model

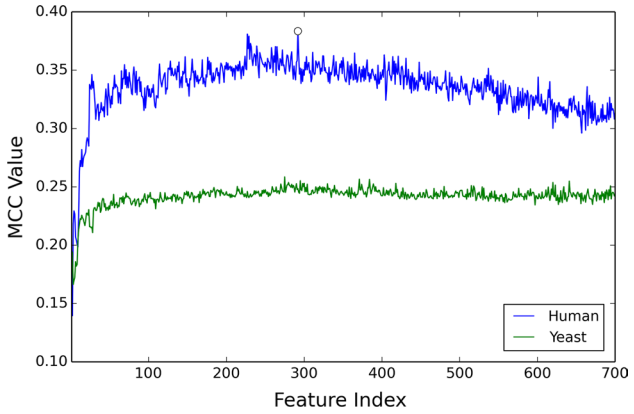


Fig. 5 MCC values obtained by a stepwise incremental feature selection. The maximum MCC values on the human and yeast datasets were 0.3836 and 0.2585, when the top 292 and 275 features were selected by mRMR, respectively

in the human GS training dataset. Among these selected features, the numbers of features from AC, MAC, GAC, MBAC, CT, LD and CRS encoding schemes were 7, 2, 2, 62, 13, 67 and 139, respectively.

Comparison with other methods

To further benchmark our method, we also compared the performance of our final model (named SPAR) with one existing SIP predictor SLIPPER and three PPI predictors DXECPPI (Du et al. 2014), PPIevo (Zahiri et al. 2013) and LocFuse (Zahiri et al. 2014) based on the human- and yeast-GS-independent sets. When compared with SLIPPER, we directly enquired the results on its web server by the gene name of each sample. Although the performance of SLIPPER was superior to SPAR (Table 3; Fig. 6), SLIPPER also contained some limitations. Firstly, it cannot predict a protein based on its sequence, but only allows for querying the predicted results, which had been stored in their database, by gene names. Secondly, it integrated a great deal of known knowledge, such as GO terms, PINs, drug targets and enzymes. Particularly, the degree of a protein in the PINs made a great contribution for SIP prediction. However, for an unknown or artificial protein in real applications, all of the information was very difficult to be accessible directly. Therefore, our SPAR was necessary for the improved SIP prediction as long as its protein sequence was known.

To compare with DXECPPI, the protein sequences in the human- and yeast-independent sets were submitted to its web server to perform the prediction. As a result, we found that the comprehensive performance of SPAR was better than DXECPPI (Table 3; Fig. 6), although the MCC value of our approach was lower than that of DXECPPI on the yeast-independent dataset (Table 3). Since the two interacting protein partners were identical, the traditional PPI predictor which utilized correlation information between two proteins, such as coevolution, co-expression and co-localization, could not work effectively for the SIP prediction. Therefore, our method can be an essential complement for the PPI prediction. In addition, the results of the prediction on the yeast GS dataset proved that our SPAR has a prominent scalability when applied to cross-species SIP prediction.

Zahiri et al. have recently developed two sequence-based PPI predictors [i.e. PPIevo (Zahiri et al. 2013) and LocFuse (Zahiri et al. 2014)]. They proved that the position-specific scoring matrix (PSSM) generated by PSI-BLAST (Altschul et al. 1997) could be converted into effective features for PPI prediction. In PPIevo, a 420-D vector for each protein sequence was constructed by extracting feature from PSSM. After that, they updated the PPIevo encoding scheme by reducing the dimensionality, and further proposed a "protein sequence and consensus sequence hybridization (SCH)" encoding scheme in LocFuse, which resulted in a 648-D feature vector in total. To compare our method with the feature vectors constructed by PPIevo and LocFuse, we first extracted the

Table 3 Performance comparison of the models trained by CRS features with mRMR-selected features (SPAR), two PSSM-based features (PPIevo and LocFuse) and two other predictors (SLIPPER, DXECPPI) evaluated on both human and yeast independent sets

Scheme	Human					Yeast				
	Ac (%)	Sp (%)	Sn (%)	MCC	$F_{measure}$ (%)	Ac (%)	Sp (%)	Sn (%)	MCC	$F_{measure}$ (%)
SLIPPER	91.10	95.06	47.26	0.4197	46.82	71.90	72.18	69.72	0.2842	36.16
DXECPPI	30.90	25.83	87.08	0.0825	17.28	87.46	94.93	29.44	0.2825	34.89
PPIevo	78.04	25.82	87.83	0.2082	27.73	66.28	87.46	60.14	0.1801	28.92
LocFuse	80.66	80.50	50.83	0.2026	27.65	66.66	68.10	55.49	0.1577	27.53
CRS	91.54	96.72	34.17	0.3633	36.83	72.69	74.37	59.58	0.2368	33.05
SPAR	92.09	97.40	33.33	0.3836	41.13	76.96	80.02	53.24	0.2484	34.54

The ratio of positive to negative samples in the human- and yeast-independent sets was approximately 1:11 and 1:8, respectively

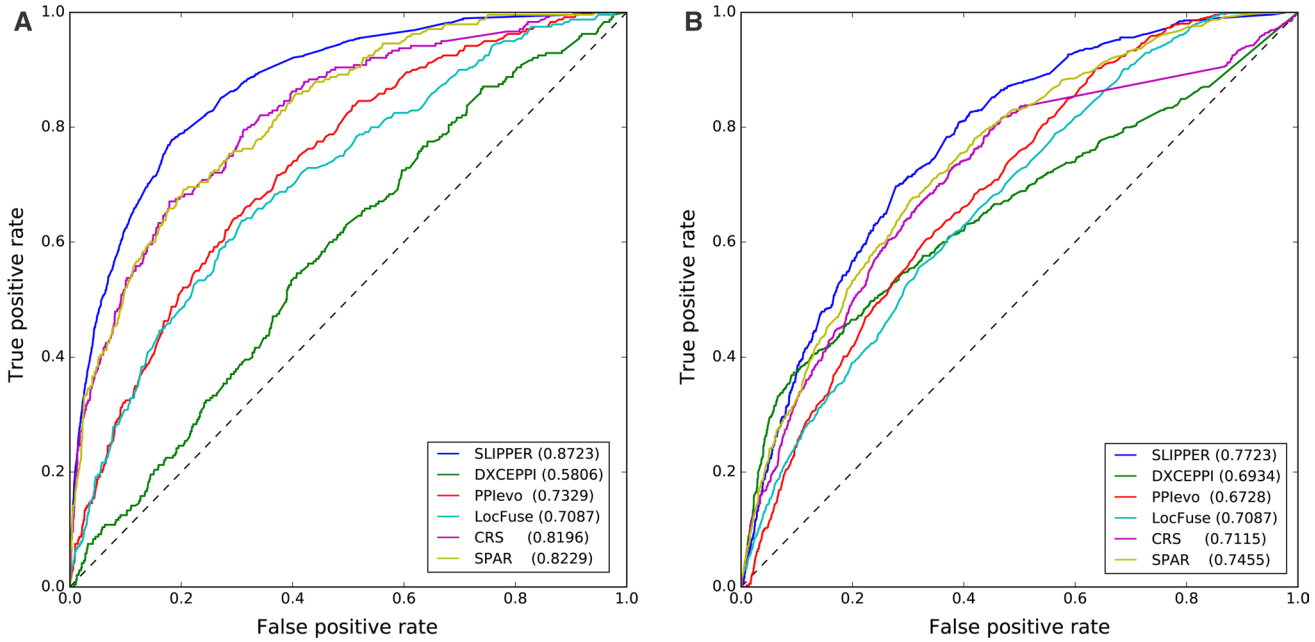


Fig. 6 ROC curves of CRS, SPAR, the PPIevo-encoding, the LocFuse-encoding and two other predictors (SLIPPER and DXEDPPI) based on the results of independent test. The performance was based on (**a**) the human-GS-independent test set and (**b**) the yeast GS dataset, respectively. *Parameters* in *brackets* are the AUC values of each prediction model

feature vectors from the PSSM for each sample by running these two feature construction programs downloaded from http://lbb.ut.ac.ir/Download/LBBsoft/. Then, we retrained our RF models by using the human training dataset based on these two encoding schemes. In general, the models based on these two encoding schemes also achieved a good performance on both human- and yeast-independent datasets, although they were not as effective as the features generated by the CRS method (Table 3; Fig. 6). In terms of the future method development, these two encoding schemes may serve as valuable features to further improve our SIP predictor.

Web server implementation

To facilitate high-throughput prediction of SIPs, a web server of SPAR has been made freely available at http://systbio.cau.edu.cn/zzdlab/spar/ to the research community. At the input webpage, users can submit their protein by pasting a FASTA-formatted sequence into the text box. After pressing the submit button, the server will automatically predict whether the query protein can self-interact based on the calculated probability score, which will be shown in the output page instantaneously. Typically, it takes a few seconds for the server to process a task. Moreover, the server provides the details about self-interacting domain information which is generated by integrating HMMER3 alignment results and DDI annotations in 3did database. Besides, users will receive a job ID, which can be saved for the future query. All the prediction tasks will be stored for 1 month.

Conclusion

In this study, we have developed a new computational method named SPAR for predicting SIPs based on sequence information. By taking advantage of the DDI information in the 3did database, we proposed an improved feature-encoding scheme named CRS and constructed models based on the RF algorithm. We evaluated the performance of this method on the human SIP dataset and also compared with other popular feature-encoding methods commonly used for PPI prediction, such as AC, CT, LD, MAC, GAC and MBMAC. The empirical results on the benchmark dataset showed that the domain information made a significant contribution to the performance improvement of CRS. Moreover, we found that the substitutions that occur at conserved sites and the DDI interface might play a critical role in determining whether these self-interacting domains can result in protein self-interactions. Finally, an optimized model was constructed by integrating all the important features ranked by the mRMR algorithm. The performance of the final model achieved an accuracy of 92.09 and 76.96 % on the human- and yeast-independent datasets, respectively. We anticipate that SPAR can serve as

an important tool to facilitate the high-throughput prediction analysis of protein self-interactions.

Acknowledgments We thank Dr. Yuan Zhou at China Agricultural University for helpful discussions on this work. This work was supported by grants from the National Natural Science Foundation of China (31271414, 31471249, 61202167, 61303169).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain–domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci USA 105(36):13292–13297. doi:10.1073/pnas.0801207105

Altschul SF, Madden TL, Schaffer 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(17):3389–3402

Baisamy L, Jurisch N, Diviani D (2005) Leucine zipper-mediated homo-oligomerization regulates the Rho-GEF activity of AKAP-Lbc. J Biol Chem 280(15):15405–15412. doi:10.1074/jbc.M414440200

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. doi:10.1023/A:1010933404324

Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock RE, Brinkman FS, Lynn DJ (2013) InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res 41(Database issue):D1228–D1233. doi:10.1093/nar/gks1147

Cancherini DV, Franca GS, de Souza SJ (2010) The role of exon shuffling in shaping protein–protein interaction networks. BMC Genom 11(Suppl 5):S11. doi:10.1186/1471-2164-11-S5-S11

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res 43(Database issue):D470–D478. doi:10.1093/nar/gku1204

Chen Y, Dokholyan NV (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol Biol Evol 25(8):1530–1533. doi:10.1093/molbev/msn122

Du X, Cheng J, Zheng T, Duan Z, Qian F (2014) A novel feature extraction scheme with ensemble coding for protein–protein interaction prediction. Int J Mol Sci 15(7):12731–12749. doi:10.3390/ijms150712731

Feng ZP, Zhang CT (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19(4):269–275

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39(Web Server issue):W29–W37. doi:10.1093/nar/gkr367

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230. doi:10.1093/nar/gkt1223

Gibson TA, Goldberg DS (2009) Questioning the ubiquity of neofunctionalization. PLoS Comput Biol 5(1):e1000252. doi:10.1371/journal.pcbi.1000252

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36(9):3025–3030. doi:10.1093/nar/gkn159

Hashimoto K, Panchenko AR (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. Proc Natl Acad Sci USA 107(47):20352–20357. doi:10.1073/pnas.1012999107

Hashimoto K, Nishi H, Bryant S, Panchenko AR (2011) Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. Phys Biol 8(3):035007. doi:10.1088/1478-3975/8/3/035007

Hattori T, Ohoka N, Inoue Y, Hayashi H, Onozaki K (2003) C/EBP family transcription factors are degraded by the proteasome but stabilized by forming dimer. Oncogene 22(9):1273–1280. doi:10.1038/sj.onc.1206204

Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. Nucleic Acids Res 33(11):3629–3635. doi:10.1093/nar/gki678

Katsamba P, Carroll K, Ahlsen G, Bahna F, Vendome J, Posy S, Rajebhosale M, Price S, Jessell TM, Ben-Shaul A, Shapiro L, Honig BH (2009) Linking molecular affinity and cellular specificity in cadherin-mediated adhesion. Proc Natl Acad Sci USA 106(28):11594–11599. doi:10.1073/pnas.0905349106

Koike R, Kidera A, Ota M (2009) Alteration of oligomeric state and domain architecture is essential for functional transformation between transferase and hydrolase with the same scaffold. Protein Sci 18(10):2060–2066. doi:10.1002/pro.218

Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S (2015) MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res 43(Database issue):D321–D327. doi:10.1093/nar/gku1091

Liu Z, Guo F, Zhang J, Wang J, Lu L, Li D, He F (2013) Proteome-wide prediction of self-interacting proteins based on multiple properties. Mol Cell Proteomics 12(6):1689–1700. doi:10.1074/mcp.M112.021790

Marianayagam NJ, Sunde M, Matthews JM (2004) The power of two: protein dimerization in biology. Trends Biochem Sci 29(11):618–625. doi:10.1016/j.tibs.2004.09.006

Miller S, Lesk AM, Janin J, Chothia C (1987) The accessible surface area and stability of oligomeric proteins. Nature 328(6133):834–836. doi:10.1038/328834a0

Mosca R, Ceol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 42(Database issue):D374–D379. doi:10.1093/nar/gkt887

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42(Database issue):D358–D363. doi:10.1093/nar/gkt1115

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238. doi:10.1109/TPAMI.2005.159

Perez-Bercoff A, Makino T, McLysaght A (2010) Duplicability of self-interacting human genes. BMC Evol Biol 10:160. doi:10.1186/1471-2148-10-160

Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 39(Web Server issue):W385–w390. doi:10.1093/nar/gkr284

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32(Database issue):D449–D451. doi:10.1093/nar/gkh086

Shatnawi M, Zaki NM (2015) Novel domain identification approach for protein–protein interaction prediction. In: Computational intelligence in bioinformatics and computational biology (CIBCB), 2015 IEEE (conference on, 12–15 Aug 2015), pp 1–8. doi:10.1109/CIBCB.2015.7300340

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H (2007) Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci USA 104(11):4337–4341. doi:10.1073/pnas.0607879104

UniProt C (2015) UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204–D212. doi:10.1093/nar/gku989

Woodcock JM, Murphy J, Stomski FC, Berndt MC, Lopez AF (2003) The dimeric versus monomeric status of 14-3-3zeta is controlled by phosphorylation of Ser58 at the dimer interface. J Biol Chem 278(38):36323–36327. doi:10.1074/jbc.M304689200

Xia JF, Han K, Huang DS (2010) Sequence-based prediction of protein–protein interactions by means of rotation forest and autocorrelation descriptor. Protein Pept Lett 17(1):137–145

Yang L, Xia JF, Gui J (2010) Prediction of protein–protein interactions from protein sequence using local descriptors. Protein Pept Lett 17(9):1085–1090

You ZH, Lei YK, Zhu L, Xia J, Wang B (2013) Prediction of protein–protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinform 14(Suppl 8):S10. doi:10.1186/1471-2105-14-S8-S10

Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A (2013) PPIevo: protein–protein interaction prediction from PSSM based evolutionary information. Genomics 102(4):237–242. doi:10.1016/j.ygeno.2013.05.006

Zahiri J, Mohammad-Noori M, Ebrahimpour R, Saadat S, Bozorgmehr JH, Goldberg T, Masoudi-Nejad A (2014) LocFuse: human protein–protein interaction prediction via classifier fusion using protein localization information. Genomics 104(6 Pt B):496–503. doi:10.1016/j.ygeno.2014.10.006

Zaki N, Lazarova-Molnar S, El-Hajj W, Campbell P (2009) Protein–protein interaction based on pairwise similarity. BMC Bioinform 10:150. doi:10.1186/1471-2105-10-150

Zhou Y, Zhou YS, He F, Song J, Zhang Z (2012) Can simple codon pair usage predict protein–protein interaction? Mol BioSyst 8(5):1396–1404. doi:10.1039/c2mb05427b