Journal of Bioinformatics and Computational Biology
Vol. 16, No. 4 (2018) 1850011 (18 pages)

DOI: 10.1142/S0219720018500117

ProBAPred: Inferring protein–protein binding affinity by incorporating protein sequence and structural features

Bangli Lu*, Chen Li†, Qingfeng Chen*,¶,** and Jiangning Song†,‡,§,‖,**

*School of Computer, Electronic and Information, and
State Key Laboratory for Conservation and
Utilization of Subtropical Agro-Bioresources
Guangxi University, 100 Daxue Road
530004 Nanning, P. R. China

†Infection and Immunity Program
Biomedicine Discovery Institute and
Department of Biochemistry and Molecular Biology
Monash University, VIC 3800, Australia

‡Monash Centre for Data Science
Faculty of Information Technology
Monash University, VIC 3800, Australia

§ARC Centre of Excellence for Advanced Molecular Imaging
Monash University, VIC 3800, Australia

¶qingfeng@gxu.edu.cn

‖Jiangning.Song@monash.edu

Received 11 February 2018
Revised 26 March 2018
Accepted 26 March 2018
Published 29 June 2018

Protein-protein binding interaction is the most prevalent biological activity that mediates a great variety of biological processes. The increasing availability of experimental data of protein–protein interaction allows a systematic construction of protein–protein interaction networks, significantly contributing to a better understanding of protein functions and their roles in cellular pathways and human diseases. Compared to well-established classification for protein–protein interactions (PPIs), limited work has been conducted for estimating protein–protein binding free energy, which can provide informative real-value regression models for characterizing the protein–protein binding affinity. In this study, we propose a novel ensemble computational framework, termed ProBAPred (Protein–protein Binding Affinity Predictor), for quantitative estimation of protein–protein binding affinity. A large number of sequence and structural features, including physical–chemical properties, binding energy and conformation annotations, were collected and calculated from currently available protein binding complex datasets and the literature. Feature selection based on the WEKA package was performed to identify and characterize the most informative and contributing feature subsets. Experiments on the independent test showed that our ensemble method achieved the lowest Mean Absolute Error (MAE; 1.657 kcal/mol) and the second highest correlation coefficient (R-value = 0.467),

** Corresponding authors.

compared with the existing methods. The datasets and source codes of ProBAPred, and the supplementary materials in this study can be downloaded at http://lightning.med.monash.edu/probapred/ for academic use. We anticipate that the developed ProBAPred regression models can facilitate computational characterization and experimental studies of protein–protein binding affinity.

Keywords: Protein–protein binding affinity; regression model; sequence-derived features; structural features; feature selection.

1. Introduction

A variety of biological activities are regulated and mediated by protein–protein interactions (PPIs). Protein–protein interactions underlie a myriad of biological activities, such as activation of DNA synthesis,[1] gene transcription,[2] protein translation,[3] and signal transduction[4] among others. Subtle interruptions within protein interaction networks may result in changes in protein function, thereby leading to human diseases.[5] Protein binding free energy has been used as an important measure for protein interaction.[6] Additionally, in terms of drug design applications, binding free energy can reliably reflect the activity of drugs with other biomolecules,[7] highlighting its potential in guiding drug design and enhancing the efficacy of treatment. In view of the importance of binding free energy, it is highly desirable to develop accurate quantitative models to accurately estimate protein binding free energy.

To date, a variety of computational approaches for PPI prediction have been proposed. Notably, two main research directions for PPI prediction have been established, i.e. PPI classification and PPI binding affinity estimation. The former has been well-established; a variety of computational models addressing this task have been published to date.[8–12] These approaches only generate binary classification results (i.e. prediction of PPI or nonPPI). The binary, less informative outputs of PPI classifiers can only roughly suggest the possibilities of proteins forming interactions. To address this, more and more attention has been recently paid to develop regression models to quantitatively estimate the protein–protein binding affinity. Regression is a type of statistical analysis method that usually takes a group of independent variables as the input to predict an dependent variable, aimed at formulating a mathematical expression between these variables.[13] Compared to PPI classification, protein–protein binding affinity estimation using regression models can provide real values, which are easier to interpret and facilitate experimental validation.

Despite the significance of protein–protein binding affinity estimation, a limited number of computational methods have been published compared to PPI classification. These included PPA_Pred,[14–16] DFIRE,[17] PMF,[18] ICs/NIS,[19] and consensus model.[20] However, the predictive performance of these methods, measured by Mean Absolute Error (MAE) and correlation coefficient, was not satisfactory and require further improvement. Moreover, several published regression models did not provide

web servers or local software tools, making them difficult to be practically applied. In this study, we proposed a novel computational framework for training regression models, termed Protein–protein Binding Affinity Predictor (ProBAPred) for accurate, real-valued estimation of protein–protein binding affinity, by incorporating protein sequence and structural features. We performed systematic feature selection experiments to identify the subsets of most contributing features that led to the best predictive performance based on the WEKA software package,[21] a widely used data mining platform. In particular, at the feature selection step, three effective methods based on attribute evaluator and search methods were employed to identify the optimal feature subsets based on performance comparison among the nine regression models. In addition, we proposed three ensemble frameworks, which were developed based on the combination of three best-performing regression models, i.e. MARS,[22] SMOreg,[23] and Linear regression. The three regression models were further integrated by assigning specific weights according to their MAE and correlation coefficient obtained from the five-fold cross-validation. On the independent test, the unweighted ensemble model (selected as the optimal ProBAPred model) achieved the best performance, as indicated by the lowest MAE (1.657 kcal/mol) and the second highest correlation coefficient value (R-value = 0.467), compared to the existing methods.

2. Material and Methods

2.1. *Overall framework*

The overall framework of ProBAPred is shown in Fig. 1. We can see that there exist four major steps for constructing ProBAPred models and evaluating their predictive performance. These four steps include Data collection and preprocessing, Feature extraction, Feature selection, and regression model establishment and validation. At the first step, two datasets were collected from previous studies[14,15] of protein–protein binding affinity. The first dataset[15] containing 135 PPI complexes with remarked bound complex and its unbound status was used as the benchmark dataset for performing five-fold cross-validation. The other dataset[14] consisting of 39 protein complexes, which did not overlap with the benchmark dataset of 135 PPI complexes, was used as the independent test dataset. At the second step, sequence-derived and structural features were extracted and calculated using several third-party computational tools. Please refer to Secs. 2.3 and 2.4 for more details. Feature selection experiments for sequence-derived features were conducted at the third step using attribute evaluator and search methods provided by the WEKA machine learning platform[21] to characterize the most informative and contributing feature subsets. Due to the limited number of structural features available, we only conducted feature selection on sequence-derived features. More details regarding the calculation and extraction of structural features can be found in Sec. 2.4. During the last step, nine regression models and three ensemble frameworks were built using the WEKA

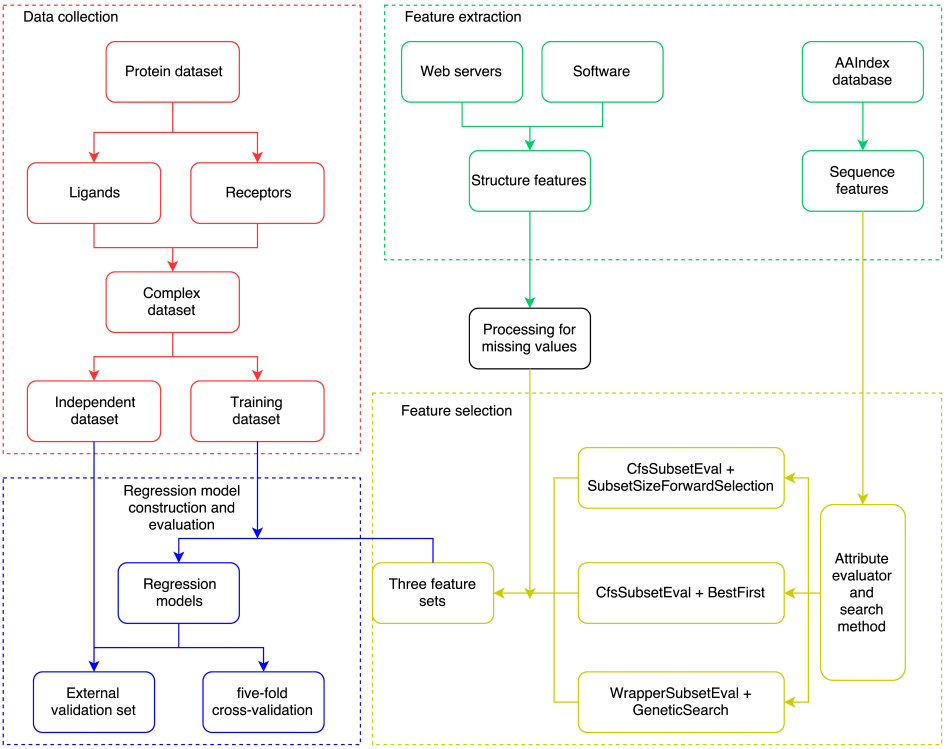


Fig. 1. The overall framework of ProBAPred.

package and MATLAB package, with their predictive performance evaluated and compared with other existing models (i.e. DFIRE and PMF) on both five-fold cross-validation and independent tests.

2.2. *Data preprocessing*

The benchmark dataset which was originally prepared by Kastritis *et al.*[24] and used in the study of Yugandhar *et al.*[15] contained 135 PPI complexes. These complexes (including bound and unbound status) were further categorized nine subgroups according to the function of the proteins: antibody versus antigen (10 complexes), antigen binding antibody (five complexes), enzyme versus inhibitor (31 complexes), enzyme versus substrate (nine complexes), enzyme with a regulatory or accessory chain (11 complexes), noncognate (nine complexes), G protein containing (16 complexes), receptor protein containing (12 complexes), and unclassified (32 complexes). These complexes were included according to the following selection criteria: (1) the sequence length of each binding partner is greater than 50 residues; (2) the binding affinity the complex is known; and (3) the complex is heterodimeric and both interacting partners are proteins. Table S1 in the Supplementary File provides a detailed description of the benchmark dataset.

In order to objectively evaluate the predictive performance, we collected 39 nonoverlapping complexes as an additional validation set, which was prepared by Moal *et al.*[14] to compare the performance of ProBAPred with that of other existing methods. Those binding complexes with partner protein shorter than 50 residues were disregarded. This validation set consists of two antibody–antigen complexes, two enzyme-inhibitor complexes, five enzymes with other complexes, six G-proteins containing complexes, nine receptor proteins containing complexes, and 15 unclassified complexes[14] (Refer to Supplementary Table S2 for more details).

2.3. *Extraction of sequence-derived features*

Sequence feature extraction. Amino acids are the fundamental building blocks of proteins and amino acid sequence determines the protein secondary structure and function.[25] A number of previous studies indicated that when used in combination with other informative features, sequence-derived features contribute to the predictive performance of the machine learning-based models for prediction of protein structural and functional properties.[26–29] Here, we collected 544 amino acid indices from the AAindex database[30] and 48 complementary properties from the literature[31] to calculate sequence features for protein complexes. AAindex is a set of 20 numerical values representing a variety of physicochemical and biological properties of amino acids. The other 48 sequence features extracted from the literature describe physical–chemical, energetic, and conformational properties of amino acids. A complete list of the 544 amino acid indices and 48 sequence features can be found in Table S3. We also computed the frequency of each amino acid of ligands and receptors as an additional type of sequence feature. The derived frequency was further multiplied by the corresponding value of amino acid from the amino acid features to calculate the total value of each amino acid in the ligands and receptors. Then, we divided the length of protein amino acid chain by the total value of each amino acid of ligands and receptors to obtain the average values of amino acid as sequence features. Taken together, a total of 23,650 sequence-derived features were initially extracted and calculated.

2.4. *Extraction of structural features*

Six different types of structural features were identified based on the available structures of protein complexes with respect to binding affinity, which included Atomic-accessible Surface Area (ASA), proportion of the interface residues, percentage of predicted binding site residues, normalized interface-packing (NIP) score and normalized surface complementarity (NSC) scores, coarse grain potentials, inter-residue contacts, and noninteracting surface. A total of 53 structural features were then calculated to train the regression models together with the above-selected sequence-derived features (see Table S1 for more details). The calculated structural features are briefly discussed below.

Atomic-accessible Surface Area. ASA was calculated based on a structural window of 1.4Å radius over the surface of the protein using the NACCESS program (http://www.bioinf.manchester.ac.uk/naccess/). ΔASA, which is often used as an effective descriptor to measure the extent of protein binding,[24] represents the absolute value of accessible surface area change between the bound and unbound status.

Proportion of the interface residues. We used DSSP[32] to extract the protein 7-class secondary structure annotations and calculated the proportion of interface residues based on the hydrogen bond estimation, which is considered relevant to the binding free energy.[33] The seven structural descriptors provided by DSSP included 310 helix, α-helix, π-helix, residues in the isolated β-bridge, residue-extending strand participating in β ladder, and hydrogen bonded with turning and bend.

Percentage of predicted binding sites. The percentage of predicted binding sites has been previously reported to be important for understanding the recognition mechanism of protein–protein complexes.[34] The percentage of binding site residues is higher in the protein complexes with higher binding free energy than those with lower binding free energy.[15] To calculate the percentage, we first predicted protein–protein interacting (PPI) residues using the SPPIDER server.[35] Then, we calculated the percentage of predicted binding site residues for each PPI complex as an input feature and examined whether it could impact on the prediction of the complex binding affinity.

NIP and NSC scores. These included normalized interface-packing (NIP) score and normalized surface complementarity (NSC) score. NIP and NSC are two measures for describing the surface complementarity and atom packing based on the calculation of the protein interaction interface area,[36] which are informative for discriminating biological versus nonbiological protein–protein interactions. The interface area was defined as the solvent accessible surface area buried per subunit in the complex, whereas interface packing was defined based on the ratio of enclosed volume and total volume. The surface complementarity score was calculated by the ratio of the total complemented area and the minimum of total triangulated area between the two subunits. These features were suggested as robust measures for assessing geometric properties of protein interaction interfaces and describing selective binding of molecules.[36]

Coarse-grain potentials. Coarse-grain potentials refer to the multiscale coarse-graining (MS-CG) methodology,[37] including four-body potentials, general-four-body potentials, short-range potentials, and 23 two-body contact potentials of protein residues. All these potentials have been widely applied for studying ligand binding and protein design.[38] The values of these 26 coarse grain potentials[38] were calculated using the Potentials R'Us server.[39] They had a high correlation coefficient of 0.34 (refer to Table 1) with binding affinity.

Inter-residue contacts (ICs) and noninteracting surface (NIS). We calculated the number of ICs and the percentage of NIS used in the study of Vangone *et al.*[19] They generated features based on the amino acid type — polar/apolar — for

NIS, and contact types — polar/polar, polar/apolar, charged/charged, charged/apolar — for the ICs.[19] According to the literature, these features are highly correlated with binding affinity.[19] A recent work has proved the effect of the NIS on binding affinity.[40]

2.5. *Feature selection of sequence-derived features*

The initial feature set might include noisy, redundant, and irrelevant features, which can yield to unsatisfactory prediction performance.[41–44] Therefore, feature selection is an essential step to remove such redundant features and generate an optimal feature subset for training regression model. Because there were 53 structural features obtained, we only applied feature selection to select more informative sequence features. In particular, we applied several different feature evaluators in combination with a variety of search methods provided by WEKA to select most informative sequence features. Three combinations of feature evaluator and search methods were used, including CfsSubsetEval + SubsetSizeForwardSelection,[45] CfsSubsetEval + BestFirst,[46] and WrapperSubsetEval[47] + GeneticSearch,[48] which were implemented in WEKA. WEKA used the Pearson's correlation[49] to evaluate the importance of each feature. As a result, 5 features, 39 features, and 2007 features were finally selected by CfsSubsetEval + SubsetSizeForwardSelection, CfsSubsetEval + BestFirst, and WrapperSubsetEval + GeneticSearch, respectively. We then combined three respective sequence feature sets with structural features, and further applied an incremental feature selection (IFS) method[27] to finally determine three optimized feature sets. The prediction performance trained based on the three optimized feature sets is discussed in the Results and Discussion section.

2.6. *Construction and assessment of regression models*

Nine different machine learning algorithms were used to build regression models to predict the binding affinity based on the optimized feature subset after feature selection, which included Support Vector Regression (SVR),[50] RBFNetwork,[51] Linear Regression, SMOreg,[23] Additive Regression,[52] Regression By Discretization,[53] M5P,[54,55] and Multivariate Adaptive Regression Splines (MARS).[22] All these algorithms except MARS, have been implemented in WEKA.[21] MARS was coded using the MATLAB in the ARESLab toolbox. Both five-fold cross-validation (using the benchmark dataset originally prepared by Yugandhar *et al.*[15]) and independent tests (using the additional dataset prepared by Moal *et al.*[14]) were performed to assess the predictive performance of these regression models. Three primary measures MAE, Pearson's correlation coefficient (i.e. PCC or R-value), and p-value were used to assess the predictive performance. MAE and PCC are defined as follows:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |Yi - yi|}{n}, \tag{1}$$

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(Yi - \overline{Y})(yi - \overline{y})}{\sqrt{\sum_{i=1}^{n}(Yi - \overline{Y})^2}\sqrt{\sum_{i=1}^{n}(yi - \overline{y})^2}}. \quad (2)$$

3. Results and Discussion

3.1. *Selected optimal features and their contribution*

We applied three combinations of feature evaluator and search methods in WEKA, including CfsSubsetEval + SubsetSizeForwardSelection, CfsSubsetEval + BestFirst, and WrapperSubsetEval + GeneticSearch, as sequence feature selection methods in this study. As a result, we obtained three sequence feature sets after feature selection, which contained 5, 39, and 2007 sequence features, respectively. Refer to Supplementary Table S3 for more detail. As aforementioned, we calculated six types of structural features, including ASA, proportion of the interface residues, percentage of PPI residues, NIP and NSC scores, coarse grain potentials, and ICs/NIs, all of which have been previously shown to be relevant to protein–protein binding affinity. The maximum correlation coefficient values between each type of features and the experimental binding affinity in the benchmark dataset are shown in Table 1. We can see that except for the percentage of PPI residues, and NIP and NSC scores, all the other structural features had significant p-values (0.05 was used as the statistical significance threshold).

To characterize the importance of each feature in the structural feature set, we evaluated the correlation coefficient using nine regression models, by adding a feature each time from the initial structural feature set until all structural features (53 features in total) had been added once. Three correlation coefficient curves (Figs. 2(a)–2(c)), were plotted, corresponding to three feature sets containing 5, 39, and 2007 sequence feature sets, respectively. We found that the values of correlation coefficient values increased first and then declined after reaching the peak in the cases of certain regression models, e.g. RegressionByDiscretization, M5P, and MARS. After the number of added structural features reached 47, 53, and 47, respectively, the correlation coefficients achieved their maximum values in the case of MARS across all the three sequence feature sets. As a comparison, SMOreg and Linear

Table 1. The maximum correlation coefficients and the corresponding p-value between each type of features and experimental binding affinity.*

Structural feature type	Correlation coefficient	p-value
ASA	0.2217	0.005
Proportion of the interface residues	0.2075	0.016
Percentage of PPI residues	0.1166	0.18
NIP and NSC	0.0118	0.448
Coarse grain potentials	0.3358	<0.001
ICs and NIS	0.3632	<0.001

Note: *Correlation coefficients were reported as absolute values.

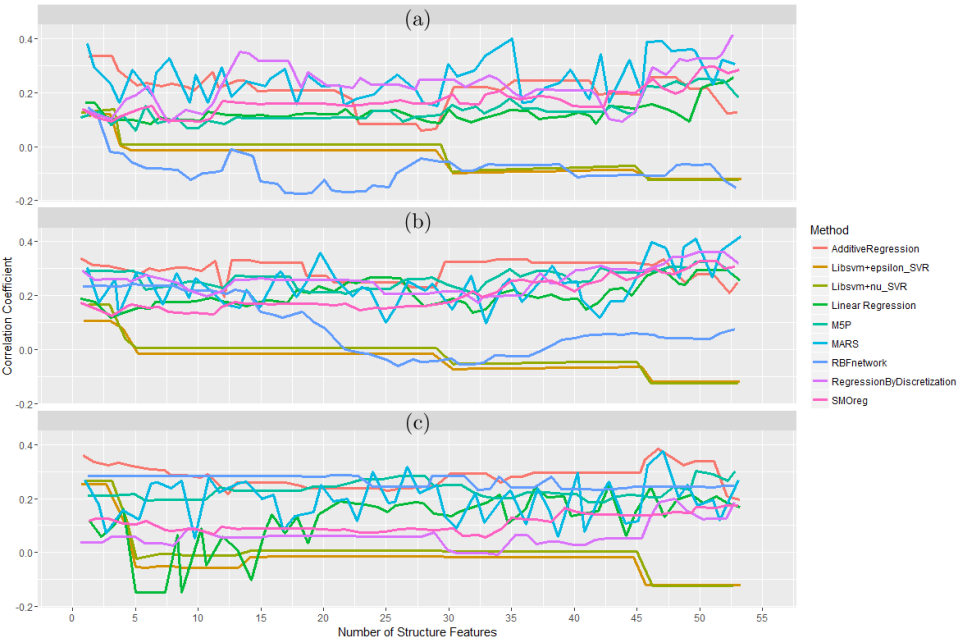


Fig. 2. Relationship between the predictive performance (evaluated by correlation coefficient) of different algorithms and the number of added structural features, based on: (a) feature set containing 2007 sequence features and structural features selected by IFS; (b) feature set containing 39 sequence features and structural features selected by IFS, and (c) feature set containing five sequence features and structural features selected by IFS. The x-axis indicates the number of incremental structural features added, and the y-axis means correlation coefficient between the experimental and predicted binding affinities based on nine respective regression models on five-fold cross-validation.

Regression showed a general trend for improvement, despite the presence of some fluctuation in the curve. Therefore, it is not surprising that the number of added features could impact on the performance of the regression models.

Table 2 provides the maximum correlation coefficient values and the corresponding numbers of structural features added. It can be seen that the correlation coefficient of MARS was 1.017–3.538 times greater than that of the other methods tested on the same feature set. Another observation was that several regression models could achieve their highest correlation coefficient values when all structural features (53 in total) were used to train the models. Therefore, in the following experimental studies, all structural features were used together with different sequence feature sets to build the regression models.

3.2. *Performance of regression models trained using sequence and structural features on the five-fold cross-validation*

In this section, we combined the three types of selected sequence-derived features (i.e. containing 5, 39, and 2007 sequence features) together with structural features to

Table 2. The maximum correlation coefficient (absolute values) achieved by different algorithms based on different numbers of added structural features.

	Correlation coefficient		
Regression model	53 + 5*	53 + 39*	53 + 2007*
LibSVM + epsilon-SVR (regression)	0.1206(3)	0.1187(46–53)	0.2527(3)
LibsSVM + nu-SVR (regression)	0.1374(3)	0.1659(1–2)	0.2668(1–3)
RBFnetwork	0.1709(19)	0.2433(5)	0.2856(21)
Linear Regression	0.2552(**53**)	0.2954(52)	0.2491(43)
SMOreg	0.297(51)	0.3288(52)	0.1756(**53**)
AdditiveRegression	0.3354(3)	0.3378(1)	0.3839(47)
RegressionByDiscretization	0.4131(**53**)	0.3612(52)	0.1997(48)
M5P	0.2506(50–51)	0.3292(51)	0.3026(50)
MARS	0.389(47)	0.42(**53**)	0.376(47)

Note: *Number of structural features + number of selected sequence features.

build and evaluate the nine regression models using the benchmark dataset on five-fold cross-validation. Note that MARS could only be trained using 118 complexes in the benchmark dataset, whereas the other 17 complexes could not be processed by MATLAB due to the presence of missing values. Table 3 shows the MAEs and R-values of all nine regression models. As shown, MARS achieved the overall highest R-value (representing the correlation coefficient) and relatively lower MAE (MAE = 2.36 kcal/mol; $R = 0.42$) using the feature set that contained 53 structural and 39 selected sequence-derived features.

We further evaluated and compared the performance of other existing models for predicting binding affinity based on five-fold cross-validation using the benchmark dataset. The result shows that the MAE of three models (i.e. MARS, Linear Regression and SMOreg) outperformed DFIRE and PMF. Note that some existing methods could not be applied in the experiment. For example, the consensus model was inaccessible at the time of evaluation. The ICs/NIS model was trained using a

Table 3. Correlation coefficient values and MAEs of nine regression algorithms with different feature sets based on five-fold cross-validation. The values marked in bold represent the overall lowest MAEs and highest R-values for the regression models trained using corresponding feature sets.

	53 + 5*		53 + 39*		53 + 2007*	
Regression model	MAE	R-value	MAE	R-value	MAE	R-value
Libsvm + epsilon_SVR (regression)	2.3304	−0.1187	2.3304	−0.1187	2.3304	−0.1187
Libsvm + nu_SVR (regression)	2.3229	−0.1233	2.3229	−0.1233	2.3229	−0.1233
RBFnetwork	2.337	−0.1528	2.3032	0.0756	2.1701	0.247
Linear Regression	2.5994	0.2552	2.8	0.2538	2.4819	0.1687
SMOreg	2.3998	0.2841	2.474	0.3077	3.1787	0.1756
AdditiveRegression	2.6012	0.1252	2.4744	0.249	2.6745	0.1971
RegressionByDiscretization	**2.2113**	**0.4131**	2.4164	0.319	2.5908	0.1874
M5P	2.4216	0.1808	2.3344	0.2764	**2.3424**	**0.302**
MARS	2.565	0.3066	**2.36**	**0.42**	2.68	0.27
Average	2.42	0.13	2.424	0.1844	2.53	0.145

dataset that contained 77 overlapping protein complexes with our benchmark dataset. PPA-Pred[14–16] was also not suitable for being included comparison as the benchmark dataset in this study was identical with the training dataset of PPA-Pred. The result shows that DFIRE achieved an MAE of 4.82 and an R-value of 0.341; whereas PMF achieved an MAE of 3.22 and an R-value of 0.34. It is noteworthy that MARS outperformed both DFIRE and PMF in terms of both MAE and R-value.

3.3. *Performance comparison with existing methods based on independent test dataset*

An external validation set was originally prepared by Moal *et al.* as the independent test dataset, which included 39 protein complexes. This independent test dataset did not overlap with the benchmark dataset used in this study, and thus is suitable for performing independent test of different methods. Four existing methods, including PPA-Pred, DFIRE, PMF, and ICs/NIS were compared with the nine regression models developed in Secs. 3.1 and 3.2 on this dataset.

Table 4 presents the predictive performance (evaluated in terms of MAE and R-value) of all the methods based on the independent test dataset using the $53 + 39$ feature set. Although certain methods such as DFIRE and PMF achieved higher correlation coefficient values, the MAE values of these methods were high. As a comparison, the two regression models developed using LibSVM generally achieved lower MAEs but attained lowest correlation coefficient values of 0. As can be seen in Table 4, Linear Regression and SMOreg achieved lower MAEs and higher correlation coefficient values (Linear Regression: $R = 0.4626, \text{MAE} = 1.7763\,\text{kcal/mol}$;

Table 4. Performance comparison between different regression models in terms of MAE and correlation coefficients (R-value) evaluated on the independent test dataset.

Regression model	MAE	R-value
PPA-Pred[14–16]	2.7	0.07
DFIRE[17]	6.904	0.424
PMF[18]	3.095	0.407
ICs/NIS	**1.869**	**0.499**
LibSVM + epsilon_SVR (regression)	1.790	0
LibSVM + nu_SVR (regression)	1.711	0
RBFnetwork	1.775	0.167
AdditiveRegression	2.037	0.203
RegressionByDiscretization	2.329	0.176
M5P	1.980	0.366
Linear Regression	1.776	0.463
SMOreg	1.947	0.458
MARS	1.958	0.205
Weighted_MAE	**1.66**	**0.463**
Weighted_R	**1.662**	**0.449**
Unweighted	**1.657**	**0.467**

SMOreg: $R = 0.458, \text{MAE} = 1.9474\,\text{kcal/mol}$) compared to other models (except for ICs/NIS).

Because MARS performed well on five-fold cross-validation (Table 3), we were interested in exploring the possibility of combining three models (i.e. Linear Regression, SMOreg, and MARS) by assigning different weight to each model based on MAE (Weighted_MAE) and R-value (Weighted_R) to further improve the predictive performance. For the Weighted_MAE model, the weights were assigned based on their MAE values obtained on five-fold cross-validation using the benchmarking dataset (the weights of MARS, Linear Regression, and SMOreg were 0.358, 0.301, and 0.341, respectively). While for the Weighted_R model, the weights were assigned based on their R-values obtained on five-fold cross-validation using the benchmarking dataset (the weights of MARS, Linear Regression, and SMOreg were 0.428, 0.259, and 0.313, respectively). Moreover, we also built another ensemble model, called Unweighted, which assigned equal weights to all the three individual models (i.e. the weight was 1/3 for each). The predicted MAEs and R-values of these three ensemble models are shown in Table 4. It is notable that the **Unweighted** model trained using the $53 + 39$ feature set achieved the lowest MAE values amongst all other methods on the independent test, with the R-values ranked as the second best in general. Altogether, considering both the MAE and R-value, we concluded that the **Unweighted** was the optimal model for predicting the binding affinity and used as the final model for ProBAPred.

3.4. *Case study*

To further test the predictive performance of ProBAPred, we presented a case study of two protein complexes from the independent dataset and applied ProBAPred to predict their binding affinity. The PDB ID of the first protein complex is 3GC3.[56] Its ligand is important for receptor desensitization and is required for triggering "alternative" signals and regulating agonist-mediated G-protein coupled receptor (GPCR) signaling by mediating both receptor desensitization and resensitization processes.[57] The receptor of 3GC3 is a major protein of the polyhedral coat of coated pits and vesicles, and has been proposed to contribute to stabilization of kinetochore fibers of the mitotic spindle by acting as inter-microtubule bridge.[58] The experimental and predicted binding affinities of 3GC3 were $-7.75\,\text{kcal/mol}$ and $-7.803\,\text{kcal/mol}$, respectively. The MAE for this case study complex was only $0.053\,\text{kcal/mol}$, indicating a good predictive result by ProBAPred. The PDB ID of the second protein complex is 3AJB.[59] Its ligand is necessary for peroxisomal membrane proteins in the peroxisomes which are involved in peroxisome biosynthesis and integrity.[60] The receptor of 3AJB is thus essential for peroxisomal membrane proteins. It can bind and stabilize newly synthesized PMPs in the cytoplasm by interacting with their hydrophobic membrane-spanning domains.[61] The experimental and predicted binding affinities of this complex (3AJB) were $-10.08\,\text{kcal/mol}$ and $-10.04\,\text{kcal/mol}$, respectively. ProBAPred also achieved a lower MAE of

0.04 kcal/mol for this complex. In summary, the comparison between the experimental binding affinity and the predicted value indicate that ProBAPred is capable of estimating the protein–protein binding affinity.

3.5. *Software availability*

We have made publicly available an online webpage of ProBAPred, which is available at http://lightning.med.monash.edu/probapred/. The source codes, user instructions, supplementary documents, and examples can be downloaded from this website. All these materials are freely available for academic purposes. The source codes were written in MATLAB and Java programming languages, and accordingly dependency check is recommended prior to the installation. We also provided a detailed user instruction document within the downloaded source codes with regards to essential procedures of how to run ProBAPred locally. In addition, we also provided our contact detail so that users can report potential bugs and help improve ProBAPred.

4. Conclusion

Protein–protein binding and interaction is a fundamental biological process. Accurate estimation of protein–protein binding affinity can significantly contribute to characterization of novel interacting partner, facilitation of mechanistic studies and generation of new biological hypothesis. In this paper, we proposed ProBAPred, a useful ensemble regression-based computational framework by integrating a variety of protein sequence-derived and structural features to accurately estimate the binding affinity. A number of feature selection methods were used to select optimized feature sets. We generated nine individual regression models and three ensemble models based on weighting strategies and benchmarked the performance of these models with other existing methods. Five-fold cross-validation, independent test, and case studies showed that ProBAPred could accurately quantify the relationship between protein sequence/structural features and the binding affinity. We hope that ProBAPred can be employed as a useful tool for accurate prediction of protein–protein binding affinity and guide biologists with their hypothesis generation and experimental validation.

Conflicts of Interest

There are no conflicts of interest to declare.

Acknowledgments

This work was financially supported by grants from the National Natural Science Foundation of China (NSFC) (61363025 and 61751314), a key project of Natural Science Foundation of Guangxi (2017GXNSFDA198033) and a key research and

development plan of Guangxi (AB17195055), the Australian Research Council (ARC) (LP110200333 and DP120104460), the National Health and Medical Research Council of Australia (NHMRC) (4909809), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965), and a Major Inter-Disciplinary Research (IDR) project awarded by Monash University. Bangli Lu and Chen Li contributed equally to this work.

Author Contributions Statement

Q. C. and J. S. conceived and designed the project; C. L. and B. L. performed data collection, feature selection, model construction, evaluation and drafted the manuscript; J. S. and Q. C. provided critical feedback for the study and participated in the discussion for data analysis; All authors revised and approved the manuscript.

References

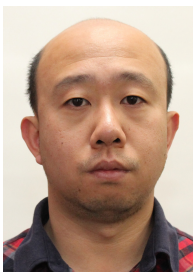
1. Shibutani S, Takeshita M, Grollman AP, Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG, *Nature* **349**:431–434, 1991, doi:10.1038/349431a0.
2. Forsythe JA *et al.*, Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1, *Mol Cell Biol* **16**:4604–4613, 1996.
3. Ma Y, Hendershot LM, Delineation of a negative feedback regulatory loop that controls protein translation during endoplasmic reticulum stress, *J Biol Chem* **278**:34864–34873, 2003, doi:10.1074/jbc.M301107200.
4. Apel K, Hirt H, Reactive oxygen species: Metabolism, oxidative stress, and signal transduction, *Annu Rev Plant Biol* **55**:373–399, 2004, doi:10.1146/annurev.arplant.55.031903.141701.
5. Vidal M, Cusick ME, Barabasi AL, Interactome networks and human disease, *Cell* **144**:986–998, 2011, doi:10.1016/j.cell.2011.02.016.
6. Massova I, Kollman P, Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies, *J Am Chem Soc* **121**:11, 1999.
7. Ajay A, Murcko MA, Computational methods to predict binding free energy in ligand-receptor complexes, *J Med Chem* **38**:4953–4967, 1995.
8. Sun T, Zhou B, Lai L, Pei J, Sequence-based prediction of protein–protein interaction using a deep-learning algorithm, *BMC Bioinf* **18**:277, 2017, doi:10.1186/s12859-017-1700-2.
9. Perovic V *et al.*, TRI_tool: A web-tool for prediction of protein-protein interactions in human transcriptional regulation, *Bioinformatics* **33**:289–291, 2017, doi:10.1093/bioinformatics/btw590.
10. Mirabello C, Wallner B, InterPred: A pipeline to identify and model protein-protein interactions, *Proteins* **85**:1159–1170, 2017, doi:10.1002/prot.25280.
11. Garcia-Garcia J *et al.*, iFrag: A protein–protein interface prediction server based on sequence fragments, *J Mol Biol* **429**:382–389, 2017, doi:10.1016/j.jmb.2016.11.034.
12. Wang YB *et al.*, Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network, *Mol Biosyst* **13**:1336–1344, 2017, doi:10.1039/c7mb00188f.

13. Armstrong JS, Illusions in regression analysis, *Int J Forecasting* **28**:689–694, 2012, doi:10.1016/j.ijforecast.2012.02.001.
14. Moal IH, Fernandez-Recio J, Comment on 'protein–protein binding affinity prediction from amino acid sequence', *Bioinformatics* **31**:614–615, 2015, doi:10.1093/bioinformatics/btu682.
15. Yugandhar K, Gromiha MM, Protein–protein binding affinity prediction from amino acid sequence, *Bioinformatics* **30**:3583–3589, 2014, doi:10.1093/bioinformatics/btu580.
16. Yugandhar K, Gromiha MM, Response to the comment on 'protein–protein binding affinity prediction from amino acid sequence', *Bioinformatics* **31**:978, 2015, doi:10.1093/bioinformatics/btu821.
17. Liu S, Zhang C, Zhou H, Zhou Y, A physical reference state unifies the structure-derived potential of mean force for protein folding and binding, *Proteins* **56**:93–101, 2004, doi:10.1002/prot.20019.
18. Su Y, Zhou A, Xia X, Li W, Sun Z, Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction, *Protein Sci* **18**:2550–2558, 2009, doi:10.1002/pro.257.
19. Vangone A, Bonvin AM, Contacts-based prediction of binding affinity in protein-protein complexes, *Elife* **4**:e07454, 2015, doi:10.7554/eLife.07454.
20. Moal IH, Agius R, Bates PA, Protein–protein binding affinity prediction on a diverse set of structures, *Bioinformatics* **27**:3002–3009, 2011, doi:10.1093/bioinformatics/btr513.
21. Hall M *et al.*, The WEKA data mining software: An update, *ACM SIGKDD Explorations Newsletter* **11**:9, 2009.
22. Friedman JH, Multivariate adaptive regression splines, *Ann Stat* **19**:1–67, 1991, doi:10.1214/aos/1176347963.
23. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK, Improvements to the SMO algorithm for SVM regression, *IEEE Trans Neural Netw* **11**:1188–1193, 2000, doi:10.1109/72.870050.
24. Kastritis PL *et al.*, A structure-based benchmark for protein–protein binding affinity, *Protein Sci* **20**:482–491, 2011, doi:10.1002/pro.580.
25. Sadowski MI, Jones DT, The sequence-structure relationship and protein function prediction, *Curr Opin Struct Biol* **19**:357–362, 2009, doi:10.1016/j.sbi.2009.03.008.
26. Ng PC, Henikoff S, SIFT: Predicting amino acid changes that affect protein function, *Nucl Acids Res* **31**:3812–3814, 2003.
27. Li F *et al.*, GlycoMine: A machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome, *Bioinformatics* **31**:1411–1419, 2015, doi:10.1093/bioinformatics/btu852.
28. Wang Y *et al.*, Knowledge-transfer learning for prediction of matrix metalloprotease substrate-cleavage sites, *Sci Rep* **7**:5755, 2017, doi:10.1038/s41598-017-06219-7.
29. Song J *et al.*, PROSPERous: High-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, *Bioinformatics* **34**:684–687, 2017, doi:10.1093/bioinformatics/btx670.
30. Kawashima S *et al.*, AAindex: Amino acid index database, progress report 2008, *Nucleic Acids Res* **36**:D202–205, 2008, doi:10.1093/nar/gkm998.
31. Gromiha MM, A statistical model for predicting protein folding rates from amino acid sequence with structural class information, *J Chem Inf Model* **45**:494–501, 2005, doi:10.1021/ci049757q.
32. Kabsch W, Sander C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**:2577–2637, 1983, doi:10.1002/bip.360221211.

33. Noskov SY, Lim C, Free energy decomposition of protein–protein interactions, *Biophys J* **81**:737–750, 2001, doi:10.1016/S0006-3495(01)75738-4.
34. Chakrabarti P, Janin J, Dissecting protein–protein recognition sites, *Proteins* **47**:334–343, 2002.
35. Porollo A, Meller J, Prediction-based fingerprints of protein–protein interactions, *Proteins* **66**:630–645, 2007, doi:10.1002/prot.21248.
36. Mitra P, Pal D, New measures for estimating surface complementarity and packing at protein–protein interfaces, *FEBS Lett* **584**:1163–1168, 2010, doi:10.1016/j.febslet.2010.02.021.
37. Izvekov S, Voth GA, A multiscale coarse-graining method for biomolecular systems, *J Phys Chem B* **109**:2469–2473, 2005, doi:10.1021/jp044629q.
38. Pokarowski P *et al.*, Inferring ideal amino acid interaction forms from statistical protein contact potentials, *Proteins* **59**:49–57, 2005, doi:10.1002/prot.20380.
39. Feng Y, Kloczkowski A, Jernigan RL, Potentials 'R' Us web-server for protein energy estimations with coarse-grained knowledge-based potentials, *BMC Bioinf* **11**:92, 2010, doi:10.1186/1471-2105-11-92.
40. Kastritis PL, Rodrigues JP, Folkers GE, Boelens R, Bonvin AM, Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface, *J Mol Biol* **426**:2632–2652, 2014, doi:10.1016/j.jmb.2014.04.017.
41. Li Y *et al.*, Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features, *Sci Rep* **4**:5765, 2014, doi:10.1038/srep05765.
42. Wang M *et al.*, Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets, *Bioinformatics* **30**:71–80, 2014, doi:10.1093/bioinformatics/btt603.
43. Song J *et al.*, PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection, *Sci Rep* **7**:6862, 2017, doi:10.1038/s41598-017-07199-4.
44. Saeys Y, Inza I, Larranaga P, A review of feature selection techniques in bioinformatics, *Bioinformatics* **23**:2507–2517, 2007, doi:10.1093/bioinformatics/btm344.
45. Derksen S, Keselman H, Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables, *Br J Math Stat Psychol* **45**:18, 1992.
46. Selvakuberan K, Indradevi M, Rajaram R, Combined feature selection and classification — A novel approach for the categorization of web pages, *J Inf Comput Sci* **3**:7, 2008.
47. Devi MI, Rajaram R, Selvakuberan K, Generating best features for web page classification, *Webology* **5**:377–384, 2008.
48. Hajela P, Lin C, Genetic search strategies in multicriterion optimal design, *Struct Optim* **4**:9, 1992.
49. Hunt RJ, Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability, *J Dent Res* **65**:128–130, 1986, doi:10.1177/00220345860650020701.
50. Chang CC, Lin CJ, LIBSVM: A library for support vector machines, *ACM Trans Intel Syst Tec* **2**, 2011, doi:Artn 2710.1145/1961189.1961199.
51. Hardy RL, Multiquadric equations of topography and other irregular surfaces, *J Geophys Res* **76**:1905, 1971, doi:10.1029/JB076i008p01905.
52. Stone CJ, Additive regression and other nonparametric models, *Ann Stat* **13**:689–705, 1985, doi:10.1214/aos/1176349548.
53. Breiman L, Random forests, *Mach Learn* **45**:5–32, 2001, doi:10.1023/A:1010933404324.
54. Quinlan RJ, *5th Australian Joint Conf Artificial Intelligence* (World Scientific, Singapore), pp. 343–348, 1992.
55. Wang Y, Witten I, *9th European Conf Machine Learning* (Springer), 1997.

56. Kang DS *et al.*, Structure of an arrestin2-clathrin complex reveals a novel clathrin binding domain that modulates receptor trafficking, *J Biol Chem* **284**:13, 2009.
57. Hoffmann C, Ziegler N, Reiner S, Krasel C, Lohse MJ, Agonist-selective, receptor-specific interaction of human P2Y receptors with beta-arrestin-1 and -2, *J Biol Chem* **283**:30933–30941, 2008.
58. Booth DG, Hood FE, Prior IA, Royle SJ, A TACC3/ch-TOG/clathrin complex stabilises kinetochore fibres by inter-microtubule bridging, *EMBO J* **30**:14, 2014.
59. Sato Y *et al.*, Structural basis for docking of peroxisomal membrane protein carrier Pex19p onto its receptor Pex3p, *EMBO J* **29**:4083–4093, 2010, doi:10.1038/emboj.2010.293.
60. Ghaedi K, Tamura S, Okumoto K, Matsuzono Y, Fujiki Y, The peroxin pex3p initiates membrane assembly in peroxisome biogenesis, *Mol Biol Cell* **11**:18, 2000.
61. Jones JM, Morrell JC, Gould SJ, PEX19 is a predominantly cytosolic chaperone and import receptor for class 1 peroxisomal membrane proteins, *J Cell Biol* **164**:11, 2004.

Bangli Lu received his M.E. degree from Guang Xi University, China. His research interests are bioinformatics, machine learning, data mining, and big data processing.



Chen Li received his Ph.D. degree in bioinformatics from Monash University, Australia. He is currently a postdoctoral research fellow at the Monash Biomedicine Discovery Institute, Monash University. His research interests include systems immunology, proteomics, immune-peptidomics, systems biology, and data mining.

Qingfeng Chen received the B.Sc. and M.Sc. degrees in mathematics from Guangxi Normal University, China, in 1995 and 1998, respectively, and the Ph.D. degree in Computer Science from the University of Technology Sydney, in September 2004. He is now a professor with Guangxi University, China, and a hundred talent program of Guangxi. His research interests include bioinformatics, data mining, and artificial intelligence. He has published 40 refereed papers and two monographs by Springer, including the IEEE Transactions on Knowledge and Data Engineering and the Data Mining and Knowledge Discovery. He has been serving as an associate editor for Engineering Letters, and was invited to be a guest editor of two special issues for Current Protein & Peptide Science, and co-chairs for several international conferences.

Jiangning Song received his B.E. and DEng degrees from Jiangnan University, China. He is affiliated with the Monash Centre for Data Science, Faculty of Information Technology and Monash Biomedicine Discovery Institute, Monash University, Melbourne, Australia. His research interests include bioinformatics, computational biology, machine learning, data mining, and pattern recognition.