BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 24 2014, pages 3583–3589
doi:10.1093/bioinformatics/btu580

Structural bioinformatics

Advance Access publication August 28, 2014

Protein–protein binding affinity prediction from amino acid sequence

K. Yugandhar and M. Michael Gromiha*

Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai-600036, Tamil Nadu, India

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Protein–protein interactions play crucial roles in many biological processes and are responsible for smooth functioning of the machinery in living organisms. Predicting the binding affinity of protein–protein complexes provides deep insights to understand the recognition mechanism and identify the strong binding partners in protein–protein interaction networks.

Results: In this work, we have collected the experimental binding affinity data for a set of 135 protein–protein complexes and analyzed the relationship between binding affinity and 642 properties obtained from amino acid sequence. We noticed that the overall correlation is poor, and the factors influencing affinity depends on the type of the complex based on their function, molecular weight and binding site residues. Based on the results, we have developed a novel methodology for predicting the binding affinity of protein–protein complexes using sequence-based features by classifying the complexes with respect to their function and predicted percentage of binding site residues. We have developed regression models for the complexes belonging to different classes with three to five properties, which showed a correlation in the range of 0.739–0.992 using jack-knife test. We suggest that our approach adds a new aspect of biological significance in terms of classifying the protein–protein complexes for affinity prediction.

Availability and implementation: Freely available on the Web at http://www.iitm.ac.in/bioinfo/PPA_Pred/

Contact: gromiha@iitm.ac.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 19, 2014; revised on August 14, 2014; accepted on August 21, 2014

1 INTRODUCTION

Protein–protein interactions are crucial prerequisite for many biological interactions involved in cellular signaling, immunity, cellular transport, etc. It is important to identify the interacting pairs of proteins and their strength of interaction for understanding the functions in protein–protein interaction networks. In the past, several studies have been carried out to understand the principle and recognition mechanism of protein–protein complexes (Gromiha *et al.*, 2009), and most of them suggested the importance of amino acid residues located at the interface of the two interacting proteins for the recognition process (Chakrabarti and Janin, 2002; Tuncbag *et al.*, 2011). Recently, Kastritis *et al.* (2014) suggested that along with the interface residues, non-interface residues also play a key role in protein–protein interactions.

Experimentally, protein–protein interactions have been mainly studied with the yeast two-hybrid system, Förster/fluorescence resonance energy transfer, surface plasmon resonance and isothermal titration calorimetry, which provide the affinity of interacting proteins in terms of dissociation constant and binding free energy change, thereby adding a new dimension to the task of protein–protein interaction network analysis. These experimental techniques for measuring the protein–protein binding affinity require expensive experimental setup and heavy man power and are time-consuming. Hence, computational methods have been developed for predicting the binding affinity that has a strong potential to help experimental biologists for selecting protein–protein complexes of interest with respect to the binding affinity.

The problem of binding affinity prediction has been addressed for the past two decades starting with a relatively small dataset of 15 complexes (Horton and Lewis, 1992; Kastritis and Bonvin, 2013). Consequently, various structure-based methods have been proposed using empirical scoring functions (Audie and Scarlata, 2007; Jiang *et al.*, 2002; Ma *et al.*, 2002), knowledge-based methods (Moal *et al.* 2011; Su *et al.*, 2009; Vreven *et al.*, 2012; Zhang *et al.*, 2005) and quantitative structure–activity relationships (Tian *et al.*, 2012; Zhou *et al.*, 2013). The main drawback of these methods is they either considered less number of data or used improper validation procedures. Kastritis *et al.* (2010) set up a benchmark dataset and discussed the performance of several available methods. They reported that the performance of most of the methods was poor on the benchmark dataset. Further, Kastritis *et al.* (2011) refined the dataset, which contains the experimental affinity data for a non-redundant set of 144 protein–protein complexes along with the information about their free proteins. This led to the development of few structure-based methods for protein–protein affinity prediction, and most of them performed well only on the rigid complexes (Kastritis and Bonvin, 2013). Tian *et al.* (2012) developed a method using 378 features, which caused the possibility of over-fitting. Hence, it is necessary to develop accurate and reliable methods to address the task of predicting protein–protein binding affinity. Recently, we have devised a model based on machine learning approaches for distinguishing low- and high-affinity protein–protein complexes using sequence-based parameters (Yugandhar and Gromiha, 2014).

*To whom correspondence should be addressed.

In this work, we have developed a novel approach for predicting the binding affinity of a given protein–protein complex from its amino acid sequence. We have classified the protein–protein complexes based on their function, molecular weight and percentage of binding site residues and analyzed the relationship between sequence and structural properties of interacting free proteins and binding affinity. This imparts biological relevance to the prediction method, as the function of the complex is also taken into consideration. We have classified the complexes into nine groups based on their functions and set up multiple regression equations for predicting the binding affinity. Our method showed a correlation range of 0.836–0.998 and 0.739–0.992 using self-consistency and jack-knife test, respectively.

2 METHODS

2.1 Dataset

We have set up a dataset of protein–protein complexes based on the following criteria: (i) both the interacting partners are proteins (sequence length of minimum 50 amino acids), (ii) the complex is heterodimeric and (iii) absolute value of binding affinity is known. These criteria yielded a set of 135 complexes (provided in Supplementary Table S1) from the affinity benchmark dataset (Kastritis *et al.*, 2011). For all these complexes, we have derived a set of 615 features based on the amino acid sequence and 640 features based on the structures of two free proteins forming the complex. Considering the gap between the number of currently available protein sequences and their solved structures, sequence-based methods for binding site prediction act as important players in understanding the binding mechanism and analyzing interaction networks. Few methods have been reported for predicting the binding site residues from amino acid sequence using machine learning techniques, such as neural networks and support vector machines (Huang *et al.*, 2013; Ofran and Rost, 2007). For the sequence-based feature calculation, we have obtained the predicted binding site residues (Ofran and Rost, 2007) and property values for the 20 amino acids from AAindex database (Kawashima *et al.*, 2008) and literature (Gromiha, 2005). Similar procedure was followed for calculating the structure-based features, and SPPIDER server (Porollo and Meller, 2007) was used to predict the binding site residues from structures of the free proteins. Other structure-based features include accessible surface area (Hubbard and Thornton, 1993), number of hydrogen bonds (McDonald and Thornton, 1994), non-bonded interaction energy (Gromiha *et al.*, 2009), electrostatic energy and energy due to bond length, bond angle and torsion angle (Guex and Peitsch, 1997). We noticed that several properties are related with each other, which cause a bias in the model. Hence, we have compared the correlation between all possible pairs of properties and reduced the number of properties so that the correlation between any of the properties is not >0.65. This process yielded a set of 113 properties (features).

2.2 Classification of complexes

Using all the complexes together, we obtained the maximum correlation of 0.3 between sequence/structure-based features and their experimental ΔG values. Hence, for understanding the features influencing the binding affinity and improving the prediction efficiency, we have classified the complexes into different groups based on various criteria such as molecular weight, function, predicted percentage of binding site residues and their subgroups (percentage of aromatic and positively charged residues, hydrophobic and hydrophilic residues). The inspection of results obtained for all the classifications by comparing the maximum correlation obtained with single feature and combination of two features showed that the classification based on function is the best one for further analysis (see Section 3).

The function-based classification has been made with seven groups: (i) antigen–antibody (Ag-Ab) complexes, (ii) enzyme–inhibitor (EI) complexes, (iii) other enzymes (OE; complexes of enzymes with proteins other than inhibitors), (iv) G-protein-containing (GC) complexes, (v) receptor-containing (RC) complexes, (vi) non-cognate (NC) complexes and (vii) miscellaneous complexes. Further, the miscellaneous complexes have been subdivided into three different classes based on percentage of predicted binding site residues in the complex (see section 3.3.7), viz., miscellaneous I, II and III (≤4%, 4–8% and >8% of predicted binding site residues, respectively). These nine classes cover the entire dataset of 135 protein–protein complexes.

2.3 Regression models and validation

We have developed independent regression models for all the nine classes by combining more than one feature using multiple regression technique (Grewal, 1987). We have validated the performance of the method using jack-knife test. In this procedure, regression equations have been developed with (n − 1) complexes and used the equation to predict the ΔG of the left out complex. This process has been repeated for 'n' times (where n is the total number of complexes in a given class) and computed the correlation. We have also used support vector regression method available in WEKA machine learning platform (Hall *et al.*, 2009) to compare the obtained results.

3 RESULTS

3.1 Significance of different classifications

We have classified the complexes based on the molecular weights of free proteins, function and predicted number of binding site residues, and computed the correlation between binding affinity and sequence/structure-based properties. The correlation coefficients obtained with protein–protein complexes belonging to different subclasses using sequence-based parameters are presented in Supplementary Table S2. We have repeated the computation with structure-based features, and the results are presented in Supplementary Table S3. We observed that the performance with sequence-based features is marginally better than that with structure-based features. From Supplementary Table S3, it is interesting to note that the receptors and ligands, both having either high or low molecular weights are dominated by the properties influenced by secondary structures (Eathiraj *et al.*, 2005; Li *et al.*, 1998). Other combination of receptors and ligands are dominated by interaction energies, such as hydrophobic, electrostatic and hydrogen bonds (Haspel *et al.* 2008; Lapouge *et al.*, 2000; Lu *et al.*, 1999). Further, we examined the performance of prediction models at different classifications using their absolute single-property correlations and the correlation coefficient obtained with the combination of two properties (Table 1). We found that the classification based on protein functions has the highest absolute single-property correlation of >0.5 in 86% of the subclasses, whereas it is 50–83% in other classifications. In addition, 57% of the subclasses in protein functions showed the correlation of >0.8 using the combination of two properties. Other classifications showed the maximum correlation of only 33%. Hence, we used the classification based on function for further analysis and to refine the model for predicting the binding affinity. Further, the subclass, 'miscellaneous' has the correlation of <0.5 for any of the considered features and we grouped

into three categories based on the percentage of binding site residues at the interface.

3.2 Prediction of binding affinity

The single-property correlation for the protein–protein complexes belonging to nine subclasses is presented in Table 2. Inspection of the results presented in this table revealed that the absolute correlation is >0.6 in most of the subclasses and the correlation is >0.8 for three of them. Further, we have performed a step-wise least square fit using multiple regression technique for identifying the combinations of features to predict the binding affinity at high accuracy. For avoiding over-fitting, we limited the combinations up to 5 from the selected list of 113 properties. The number of properties used in each subclass and the correlation coefficient with binding affinity is shown in Table 2. We noticed that the correlation lies in the range of 0.836–0.998 in the subclasses. The *P*-value for all the models is <0.028 (data not shown), suggesting that the results are statistically significant. Further, the models were evaluated with jack-knife test, and the results are included in Table 2. The combination of features showed the correlation in the range of 0.739–0.992. The results are comparable with the self-consistency results, and this suggests that there is no over-fitting in our models. The relationship between experimental and predicted ΔG for all the 135 complexes is shown in Figure 1. We showed that the binding affinities of 79% of the complexes have been accurately predicted within a deviation of 1 kcal mol^{-1} and 90% within 2 kcal mol^{-1} from the experimental values.

Table 1. Comparison of correlation results among different classes of protein–protein complexes

Classification	% of classes with maximum single property $\|r\| > 0.5$	% of classes with maximum 2-combination $r > 0.8$
Functions	**85.7**	**57.1**
Molecular weight	50.0	00.0
% of predicted binding site residues	50.0	33.3
% of aromatic and positively charged residues in predicted binding sites	50.0	00.0
% of hydrophobic residues in predicted binding sites	66.7	33.3
% of hydrophilic residues in predicted binding sites	83.3	16.7

Note: The highest percentage is shown in bold.

3.3 Prediction results in different subclasses based on functions

We have analyzed the performance of the method for predicting the binding affinity in different subclasses of 'function', and the details are discussed below. The regression equations obtained for all the subclasses are given in Supplementary Table S4a and b.

3.3.1 Ag-Ab complexes Ag-Ab complexes are formed between antibodies and proteins that act as specific antigens to them. The binding free energy of these complexes shows a narrow range of

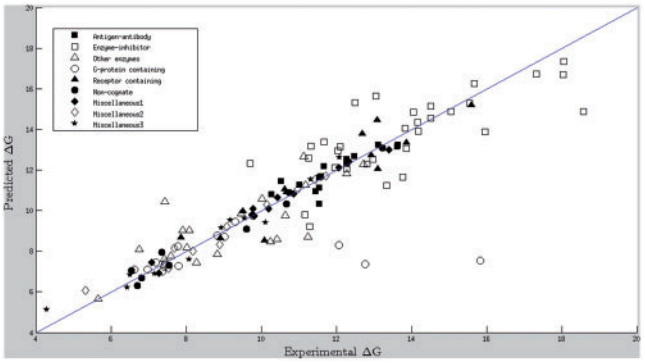

Fig. 1. Scatterplot for experimental and predicted ΔG with jack-knife test

Table 2. Correlation between amino acid properties and binding affinity

Sr. No	Class	Number of complexes	Correlation coefficient (r)			Mean absolute error
			single property	Self-consistency[a]	Jack-knife test	
1	Ag-Ab	15	0.593	0.926 (3)	0.854	0.400
2	EI	31	–0.600	0.836 (5)	0.739	1.169
3	OE	20	0.669	0.883 (5)	0.765	0.965
4	GC	16	–0.749	0.971 (3)	0.953	0.236
5	RC	12	0.711	0.976 (3)	0.931	0.696
6	NC	09	–0.939	0.996 (3)	0.986	0.333
7	M1	11	–0.752	0.998 (3)	0.992	0.186
8	M2	10	–0.828	0.994 (3)	0.983	0.278
9	M3	11	0.859	0.994 (4)	0.980	0.380

[a]Number of features is given in parentheses.

10–14 kcal mol^{-1}. The combination of three properties is able to predict the binding affinity of these complexes to the correlation of 0.93 and 0.85, respectively, using self-consistency and jack-knife tests.

Further, the property based on percentage of predicted charged residues at the interface has been identified as one of the important factors. This is supported by the previous studies that aromatic and charged residues at the protein–protein interface play key role in complex formation (Monaco-Malbet *et al.*, 2000; Pruett and Air, 1998). Moreover, the mutation of charged residues at the binding sites showed a large reduction in the affinity and is important for the recognition process (Pruett and Air, 1998). Interestingly, our method could accurately predict the binding affinity of 94% of the complexes with a deviation of 1 kcal mol^{-1} using jack-knife test.

3.3.2 EI complexes The interacting partners in this class of protein–protein complexes are enzymes and inhibitors. The binding free energy of the 31 complexes in this class has the variation of 10 kcal mol^{-1}. CDK2-CKSHS1 protein complex has the lowest free energy of 9.7 kcal mol^{-1} and the complex between Colicin E9 nuclease and Im9 immunity protein has the highest free energy of 18.6 kcal mol^{-1}. However, the combination of five features could predict the binding free energy with the correlation of 0.836. The structural analysis of EI complexes revealed that the binding site residues and β-sheet tendency are important for specificity (Albeck and Schreiber, 1999; Ghosh *et al.*, 2007). Interestingly, these two features have been selected as important players in determining the protein–protein binding affinity. Our method could accurately predict the binding free energy for 23 of the 31 complexes within the deviation of 2 kcal mol^{-1} using jack-knife test.

3.3.3 Other enzymes This class includes complexes formed by enzymes with proteins other than inhibitors. The binding free energy of these complexes follows a range of 6–13 kcal mol^{-1}. The combination of five properties improved the correlation up to 0.88 and 0.77, respectively, using self-consistency and jack-knife tests. These properties include weights for α-helix (Qian and Sejnowski, 1988) and percentage of predicted binding site residues. Interestingly, these two properties are reported to be important for understanding the recognition mechanism of protein–protein complexes (Chakrabarti and Janin, 2002). The binding free energy for 18 and 13 of these 20 complexes has been accurately predicted within the deviation of 2 and 1 kcal mol^{-1}, respectively, using jack-knife test.

3.3.4 GC complexes G-proteins are an important family of proteins that act as effectors in many signal transduction events. These proteins work together with G-protein-coupled receptors (GPCRs) to transmit signals from many hormones, neurotransmitters and other signaling factors (Campbell and Reece, 2002). The complexes, which have G-proteins as one of the interacting partners are included in this class, and the average ΔG is 8.98 kcal mol^{-1}. GC complexes are mainly dominated by hydrophobic interactions and secondary structural elements such as α-helices (Eathiraj *et al.*, 2005; Lapouge *et al.*, 2000). Our regression model used three properties to predict the binding affinity. It is noteworthy that the number of predicted hydrophobic residues at the binding sites and helical propensity (Geisow and Roberts, 1980) was selected in our model. This result demonstrates the agreement between experimental observations and computational analysis. A set of 13 among 16 complexes in this class was predicted with high accuracy, with three outliers, and the deviation between experimental and predicted binding affinity in 13 complexes is 0.24 kcal mol^{-1}. The binding free energy of 81% of the complexes was predicted within the deviation of 0.5 kcal mol^{-1} (Fig. 1).

3.3.5 RC complexes This class consists of protein–protein complexes having one of the interacting partners that function as receptor in biological processes. It contains 12 complexes with an average ΔG value of 11.58 kcal mol^{-1}. The combination of three properties including a property based on thermodynamic nature of the proteins enhanced the correlation up to 0.98. In a previous study, Maenaka *et al.* (2001) suggested that the binding affinities of RC complexes are attributed with thermodynamic quantities, enthalpy and entropy. This observation supports our results and reiterates the importance of thermodynamic properties in governing the binding affinity. Interestingly, the binding free energies for 92% of the complexes have been accurately predicted within the deviation of 1.5 kcal mol^{-1}.

3.3.6 NC complexes The concept of cognate and NC protein–protein complexes arises when two complexes have different binding affinity values in spite of having a similar geometry. In this case, the complex having more affinity is considered as cognate because in most cases this is the one having biological relevance than the other (NC) (Kastritis *et al.*, 2011). Hence, we assumed that the features influencing the binding affinity of NC complexes may be different from their cognate counter parts and included them in this separate class that consists of nine complexes and the ΔG values shows a range of 7–13 kcal mol^{-1}. The combination of three properties is able to predict the binding affinity of these complexes to the correlation of 0.996 and 0.986, respectively, using self-consistency and jack-knife tests. Further, features such as weights for α-helix and the percentage of predicted binding site residues have been identified as important factors influencing the binding affinity. This is supported by previous studies on NC complexes that binding site residues at the interface and helical tendencies could influence the binding affinity (Erman *et al.*, 1997; Li *et al.*, 1998). Our method could accurately predict the binding affinity of all the complexes within the deviation of 0.6 kcal mol^{-1} (Fig. 1).

3.3.7 Miscellaneous complexes The complexes that do not fall in any of the above six classes are termed as miscellaneous complexes. Generally, the binding site residues are believed to play important roles in governing the binding affinity. However, the non-interface residues are also reported to be crucial for specific and efficient complex formation (Kiel et al., 2004). Hence, we considered features accounting for interface information alone (number of binding site residues) as well as interface residues normalized with total residues (% of binding site residues). For three of the six classes, percentage of binding site residues is selected as one of the best features in our analysis. Hence, we have further divided the complexes in this class into three subclasses based on the percentage of predicted binding site residues.

These subclasses have the binding site residues in the ranges of 0–4% (miscellaneous1, M1), 4–8% (miscellaneous2, M2) and >8% (miscellaneous3, M3), respectively (Ofran and Rost, 2007). M1 contains 11 complexes, and the binding free energy of these complexes shows a range of 7–14 kcal mol^{-1}. Our method identified the percentage of aromatic and positively charged residues, and the number of predicted hydrophobic residues at the interface as important features to predict the binding affinity. This result agrees with the previous observations that aromatic, charged and hydrophobic interactions are important for the affinity of protein–protein complexes (Andersen et al., 1999). Further, West et al. (2001) reported that four of the five substitutions that resulted in significant reduction in binding affinity were of aromatic/charged residues, suggesting their importance in protein–protein interactions. The combination of three properties raised the correlation up to 0.99, and the deviation is 0.186 kcal mol^{-1}. M2 consists of 10 complexes and the average ΔG is 9.1 kcal mol^{-1}. The combination of three properties including the weights to β-sheet enhanced the correlation up to 0.98. In an earlier work, Bonsor et al. (2007) suggested the importance of β-sheet at the interface of a protein–protein complex. Interestingly, this feature was shown to be crucial for discriminating protein–protein complexes with respect to their binding affinities (Yugandhar and Gromiha, 2014). The binding affinities of all the complexes have been predicted within the deivation of 0.8 kcal mol^{-1}. M3 includes 11 complexes, which have more number of binding site residues, and the ΔG of these complexes shows a range of 4–12 kcal mol^{-1}. Our method could identify a property related to electrostatic interactions (Charton and Charton, 1983) as an influencing factor, and this observation in accordance with the report by Haspel et al. (2008) that electrostatic interactions could be driving force in protein–protein complex formation and binding properties are influenced by electrostatic changes. Binding free energy for all the miscellaneous complexes could be predicted with a deviation of 0.85 kcal mol^{-1} using jack-knife test by our method. Further, we have analyzed the features influencing the binding affinity of ordered and disordered protein–protein complexes. Of the 32 complexes in this class, 23 have at least one of the interacting proteins in disordered state. We performed least square fit for the disordered and ordered set of complexes separately to identify the features influencing binding affinity. Interestingly, percentage of predicted binding site residues and net charge of the protein are selected as the best features for disordered and ordered sets, respectively. These features have also been identified for discriminating high- and low-affinity protein–protein complexes (Yugandhar and Gromiha, 2014), and this result shows the importance of electrostatic interactions in ordered complexes (Sheinerman and Honig, 2002).

3.4 Affinity prediction for homodimeric complexes

As the principles governing the binding mechanism and specificity for homodimeric complexes differ from that of heterodimeric complexes, we developed a separate model for homodimeric complexes using the available affinity data for 23 complexes (Luo *et al.*, 2014). A combination of six features yielded a correlation of 0.914 and 0.827 using self-consistency and jack-knife test, respectively. The selected features account for the importance of hydrophobic and electrostatic interactions at the interface (Bahadur *et al.*, 2003).

3.5 Analysis on importance of binding site residues of protein–protein complexes

As reported in Section 3, the percentage of predicted binding site residues was selected as one of the features in three classes, viz., EI, OE and NC through feature selection process. Further, the classification based on binding site residues improved the correlation in the subclass of miscellaneous protein–protein complexes. These observations emphasize the importance of binding site residues along with the role of charged residues for understanding the binding specificity of protein–protein complexes. We illustrate their importance with specific examples from each class mentioned above [CDK2-CKSHS1 (1BUH) and Trypsin-BPTI (2PTC) complexes (ΔG: 9.7 and 18.04 kcal mol^{-1}, respectively) from EI, Glycerol kinase–Glucose-specific protein IIIGlc (1GLA) and Von Willebrand factor-Botrocetin (1IJK) (ΔG: 6.8 and 10.4 kcal mol^{-1}, respectively) from OE class and Cytochrome C peroxidase-Cytochrome C (2PCB) and RNase SA-Barstar (1AY7) complexes (ΔG: 6.8 and 13.2 kcal mol^{-1}, respectively) from NC]. The binding site residues have been identified from the complex structures available in Protein Data Bank (Rose *et al.* 2013) using PDBsum (Laskowski, 2009) and related with ΔG values. We have also predicted the binding site residues using amino acid sequence (Ofran and Rost, 2007), which shows a similar trend to those obtained with structural data.

In all the three examples, we found that the strong binding protein–protein complexes have higher percentage of residues in the binding sites compared with those of weak binding complexes. This implies the significance of binding site residues in governing binding specificity and affinity. Our results support the previous reports in the literature suggesting the importance of binding site residues in protein–protein complexes (Chakrabarti and Janin, 2002). Further, it has been reported that interface hydrogen bonds might contribute to the stability of protein–protein complexes and hence are important in protein–protein interactions (Sheinerman *et al.*, 2000). Among the six complexes, hydrogen bonds formed in four complexes at the interface have at least one charged residue as the donor or acceptor, and the remaining two complexes 2PTC and 1AY7 have 78.57 and 81.82% of their interface hydrogen bonds formed by the involvement of charged residues. Interestingly, there is no salt bridge formed at the interface in all these six complexes. This observation reiterates the importance of charged residues in binding sites and their role as anchor residues in the binding pockets of protein–protein complexes mainly by forming inter molecular hydrogen bonds apart from salt bridges (Arkin and Wells, 2004; Rajamani *et al.*, 2004; Sheinerman *et al.*, 2000; Stites, 1997). The comparison of interface residues for two complexes having different binding free energies are shown in Supplementary Figure S1 and it reveals that the complex having higher percentage of binding site residues (Supplementary Figure S1a) has the higher binding free energy than the other one with lower percentage of binding site residues (Supplementary Figure S1b). The charged residues at the binding sites that are known to be playing key role especially by forming

hydrogen bonds at protein–protein interfaces are represented using different colors.

3.6 Comparison of multiple regression and SVM-based regression methods

We have compared the results obtained in the present study using multiple regression technique with support vector regression, and the details are given in Supplementary Table S5. We observed that the performance of the present method is better than or similar to that obtained with support vector regression in most of the classes. Further, the average correlation is 0.91 and 0.88, respectively, for multiple regression and support vector regression. The average MAE for these models is 0.52 and 0.60. This shows the better performance of the present method than support vector regression.

3.7 Comparison with previous methods

The present work is the first sequence-based method for predicting the binding affinity of protein–protein complexes. Hence, it is not appropriate to compare with other existing structure-based methods. However, the comparison of features and performance provides additional information for using different methods. The available structure-based methods used 3–378 features for predicting the binding affinity (Audie and Scarlata, 2007; Horton and Lewis, 1992; Jiang *et al.*, 2002; Ma *et al.*, 2002; Moal *et al.* 2011; Su *et al.*, 2009; Tian *et al.*, 2012; Vreven *et al.*, 2012; Zhang *et al.*, 2005; Zhou *et al.*, 2013). The main shortcomings of these methods are (i) the requirement of structural information, (ii) usage of relatively small number of data, (iii) good performance only on rigid complexes and (iv) possibility of over-fitting (Kastritis and Bonvin, 2013). On the other hand, in our previous work we have developed a machine learning-based model for discriminating protein–protein complexes in terms of their binding affinity, and it was a two-state classification problem (Yugandhar and Gromiha, 2014). The present method has several advantages such as (i) no structural information is required, (ii) we classified the complexes based on function, that imparts biological relevance to this method, (iii) properly validated using jack-knife test. Nevertheless, there are certain limitations in our method such as (i) being a sequence-based method, it cannot be used to check the binding affinity of different binding poses of a single protein pair, (ii) the method is based on a relatively small dataset and (iii) classification has been done on a broader level of functions. In future, when new experimental data for protein–protein binding affinity get accumulated, the method can be refined and/or make additional sub-classifications to increase its reliability.

Further, we have compared the performance of our method (PPA-Pred) with other structure-based methods, viz., DFIRE (Liu *et al.*, 2004), PMF (Su *et al.*, 2009) and consensus approach (Moal *et al.*, 2011) using a dataset of 137 complexes along with its subsets with flexible and rigid complexes [Complexes with interface RMSD (c_α atom) <1 Å are considered as rigid and the remaining as flexible]. Our method showed better performance than all of them, and it could predict efficiently for both rigid as well as flexible complexes (Fig. 2). The improvement of correlation using our method is 0.25, 0.44 and 0.32 in rigid, flexible and all complexes, respectively.

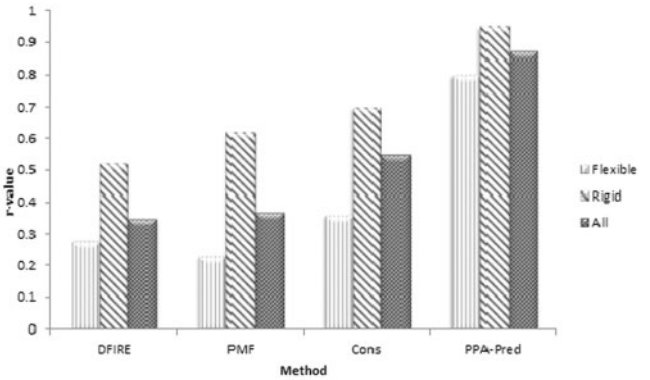

Fig. 2. Comparison of the performance of our method (PPA-Pred) with previous methods

3.8 Prediction on the Web

We have developed a Web server named 'Protein-Protein Affinity Predictor (PPA-Pred)' for predicting the binding affinity of protein–protein complexes from amino acid sequence. It takes the functional information and amino acid sequence in FASTA format as input. The output includes the predicted value of binding affinity, ΔG and K_d. The K_d value is derived using the following equation.

$$\ln K_d = -\frac{\Delta G}{RT} \quad (1)$$

In the above equation, ΔG is the dissociation free energy, K_d is the dissociation constant, R is the gas constant (1.987×10^{-3} kcal mol^{-1} K^{-1}), and T is the temperature (assumed to be room temperature i.e. 25°C). The Web server is freely accessible at http://www.iitm.ac.in/bioinfo/PPA_Pred/.

4 CONCLUSION

We have developed the first sequence-based method for predicting the binding affinity of protein–protein complexes using a robust methodology based on the functional classification. We obtained a mean correlation and MAE of 0.91 and 0.52, respectively, using jack-knife test. Further, we have systematically analyzed the importance of selected features in each class and related with experimental observations. It is evident that the percentage of binding site residues plays an important role in governing protein–protein binding affinity. We suggest that our method (PPA-Pred) could be used as an efficient tool for protein–protein interaction network analysis for specific diseases.

ACKNOWLEDGEMENTS

The authors thank Prof. Taguchi for fruitful discussions and the Bioinformatics facility, Indian Institute of Technology Madras and High performance computing environment (HPCE) for computational facilities. K.Y. thanks Paruchuri Anoosha for the help in developing the Web server and the University Grants Commission (UGC), Government of India for providing research fellowship. M.M.G. thanks Chuo University for a

Visiting Fellowship. The authors acknowledge the reviewers for their constructive comments.

Funding: The work was partly supported by the Department of Science and Technology, Government of India to M.M.G. (SR/SO/BB-0036/2011).

Conflict of interest: none declared.

REFERENCES

Albeck,S. and Schreiber,G. (1999) Biophysical characterization of the interaction of the beta-lactamase TEM-1 with its protein inhibitor BLIP. *Biochemistry*, **38**, 11–21.

Andersen,P.S. *et al*. (1999) Role of the T cell receptor α chain in stabilizing TCR-superantigen-MHC class II complexes. *Immunity*, **10**, 473–483.

Arkin,M.R. and Wells,J.A. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.*, **3**, 301–317.

Audie,J. and Scarlata,S. (2007) A novel empirical free energy function that explains and predicts protein-protein binding affinities. *Biophys. Chem.*, **129**, 198–211.

Bahadur,R.P. *et al*. (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Bonsor,D.A. *et al*. (2007) Molecular mimicry enables competitive recruitment by a natively disordered protein. *J. Am. Chem. Soc.*, **129**, 4800–4807.

Campbell,N.A. and Reece,J.B. (2002) *Biology*. 6th edn. Benjamin Cummings, San Francisco.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.

Charton,M. and Charton,B.I. (1983) The dependence of the Chou-Fasman parameters on amino acid side chain structure. *J. Theor. Biol.*, **102**, 121–134.

Eathiraj,S. *et al*. (2005) Structural basis of family-wide Rab GTPase recognition by rabenosyn-5. *Nature*, **436**, 415–419.

Erman,J.E. *et al*. (1997) Cytochrome c/cytochrome c peroxidase complex: Effect of binding-site mutations on the thermodynamics of complex formation. *Biochemistry*, **36**, 4054–4060.

Geisow,M.J. and Roberts,R.D. (1980) Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromolecules*, **2**, 387–389.

Ghosh,M. *et al*. (2007) The nuclease a-inhibitor complex is characterized by a novel metal ion bridge. *J. Biol. Chem.*, **282**, 5682–5690.

Grewal,P.S. (1987) *Numerical Methods of Statistical Analysis*. Sterling Publishers, New Delhi.

Gromiha,M.M. (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J. Chem. Inf. Model.*, **45**, 494–501.

Gromiha,M.M. *et al*. (2009) Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Mol. Biosyst.*, **5**, 1779–1786.

Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

Hall,M. *et al*. (2009) The WEKA data mining software: an update. *SIGKDD Explor.*, **11**, 10–18.

Haspel,N. *et al*. (2008) Electrostatic contributions drive the interaction between Staphylococcus aureus protein Efb-C and its complement target C3d. *Protein Sci.*, **17**, 1894–1906.

Horton,N. and Lewis,M. (1992) Calculation of the free energy of association for protein complexes. *Protein Sci.*, **1**, 169–181.

Huang,J. *et al*. (2013) MetaPIS: a sequence-based meta-server for protein interaction site prediction. *Protein Pept. Lett.*, **20**, 218–230.

Hubbard,S.J. and Thornton,J.M. (1993) *NACCESS 2.1.1*. Department of Biochemistry and Molecular Biology, University College, London.

Jiang,L. *et al*. (2002) Potential of mean force for protein-protein interaction studies. *Proteins*, **46**, 190–196.

Kastritis,P.L. and Bonvin,A.M. (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.*, **9**, 2216–2225.

Kastritis,P.L. *et al*. (2011) A structure-based benchmark for protein-protein binding affinity. *Protein Sci.*, **20**, 482–491.

Kastritis,P.L. and Bonvin,A.M. (2013) On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface*, **10**, 20120835.

Kastritis,P.L. *et al*. (2014) Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface. *J. Mol. Biol.*, **426**, 2632–2652.

Kawashima,S. *et al*. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.

Kiel,C. *et al*. (2004) Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proc. Natl Acad. Sci. USA*, **101**, 9223–9228.

Lapouge,K. *et al*. (2000) Structure of the TPR Domain of p67phox in Complex with Rac· GTP. *Mol. Cell*, **6**, 899–907.

Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.

Li,W. *et al*. (1998) Dual recognition and the role of specificity-determining residues in colicin E9 DNase-immunity protein interactions. *Biochemistry*, **37**, 11771–11779.

Liu,S. *et al*. (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.

Lu,W. *et al*. (1999) Probing intermolecular backbone H-bonding in serine protease-protein inhibitor complexes. *Chem. Biol.*, **6**, 419–427.

Luo,J. *et al*. (2014) A functional feature analysis on diverse protein–protein interactions: application for the prediction of binding affinity. *J. Comput. Aided Mol. Des.*, **28**, 619–629.

Ma,X.H. *et al*. (2002) A fast empirical approach to binding free energy calculations based on protein interface information. *Protein Eng.*, **15**, 677–681.

Maenaka,K. *et al*. (2001) The human low affinity Fcγ receptors IIa, IIb, and III bind IgG with fast kinetics and distinct thermodynamic properties. *J. Biol. Chem.*, **276**, 44898–44904.

McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.

Moal,I.H. *et al*. (2011) Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, **27**, 3002–3009.

Monaco-Malbet,S. *et al*. (2000) Mutual conformational adaptations in antigen and antibody upon complex formation between an Fab and HIV-1 capsid protein p24. *Structure*, **8**, 1069–1077.

Ofran,Y. and Rost,B. (2007) Interaction sites identified from sequence. *Bioinformatics*, **23**, e13–e16.

Porollo,A. and Meller,J. (2007) Prediction-based fingerprints of protein-protein interactions. *Proteins*, **66**, 630–645.

Pruett,P.S. and Air,G.M. (1998) Critical interactions in binding antibody NC41 to influenza N9 neuraminidase: amino acid contacts on the antibody heavy chain. *Biochemistry*, **37**, 10660–10670.

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Rajamani,D. *et al*. (2004) Anchor residues in protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **101**, 11287–11292.

Rose,P.W. *et al*. (2013) The rcsb protein data bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.

Sheinerman,F.B. *et al*. (2000) Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.*, **10**, 153–159.

Sheinerman,F.B. and Honig,B. (2002) On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.*, **318**, 161–177.

Stites,W.E. (1997) Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chem. Rev.*, **97**, 1233–1250.

Su,Y. *et al*. (2009) Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. *Protein Sci.*, **18**, 2550–2558.

Tian,F. *et al*. (2012) Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. *Amino Acids*, **43**, 531–543.

Tuncbag,N. *et al*. (2011) Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. *Phys. Biol*, **8**, 035006.

Vreven,T. *et al*. (2012) Prediction of protein-protein binding free energies. *Protein Sci.*, **21**, 396–404.

West,A.P. *et al*. (2001) Mutational analysis of the transferrin receptor reveals overlapping HFE and transferrin binding sites. *J. Mol. Biol.*, **313**, 385–397.

Yugandhar,K. and Gromiha,M.M. (2014) Feature selection and classification of protein–protein complexes based on their binding affinities using machine learning approaches. *Proteins*, **82**, 2088–2096.

Zhang,C. *et al*. (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.

Zhou,P. *et al*. (2013) Biomacromolecular quantitative structure-activity relationship (BioQSAR): a proof-of-concept study on the modeling, prediction and interpretation of protein–protein binding affinity. *J. Comput. Aided Mol. Des.*, **27**, 67–78.