Quantitative Biology 2018, 6(2): 163–174
https://doi.org/10.1007/s40484-018-0138-5

RESEARCH ARTICLE

Geometric and amino acid type determinants for protein-protein interaction interfaces

Yongxiao Yang[1], Wei Wang[2], Yuan Lou[1], Jianxin Yin[2] and Xinqi Gong[1,*]

[1] Institute for Mathematical Sciences, Renmin University of China, Beijing 100872, China
[2] School of Statistics, Renmin University of China, Beijing 100872, China
* Correspondence: xinqigong@ruc.edu.cn

Received September 11, 2017; Revised December 13, 2017; Accepted December 13, 2017

***Background*: Protein-protein interactions are essential to many biological processes. The binding site information of protein-protein complexes is extremely useful to obtain their structures from biochemical experiments. Geometric description of protein structures is the precondition of protein binding site prediction and protein-protein interaction analysis. The previous description of protein surface residues is incomplete, and little attention are paid to the implication of residue types for binding site prediction.**

***Methods*: Here, we found three new geometric features to characterize protein surface residues which are very effective for protein-protein interface residue prediction. The new features and several commonly used descriptors were employed to train millions of residue type-nonspecific or specific protein binding site predictors.**

***Results*: The amino acid type-specific predictors are superior to the models without distinction of amino acid types. The performances of the best predictors are much better than those of the sophisticated methods developed before.**

***Conclusions*: The results demonstrate that the geometric properties and amino acid types are very likely to determine if a protein surface residue would become an interface one when the protein binds to its partner.**

Keywords: protein-protein interaction; protein-protein complex interface; geometry feature; residue type; binding site

Author summary: Subtle geometry and chemistry play important roles for protein-protein interactions. The amino acid contact, void and exposure are subtle while significant. Amino acid type additionally determines the behavior when they interacts. We built an integrated approach taking advantage of these features to obtain better protein-protein interface prediction.

INTRODUCTION

Protein-protein recognitions are of great significance in living systems. The interface of protein-protein complex is crucial to understand the principle of protein-protein recognitions. Many studies have been done to obtain more knowledge of the mechanism of protein-protein interactions [1–7]. But the mathematical and physical descriptions of the principle of protein-protein interactions are still ambiguous. It is difficult to find such descriptions directly. So, starting from solving some minor problems may be appropriate.

What makes a surface residue of a protein monomer become an interface one when the monomer binds to the partner? Is it possible to predict at least one true interface residues from the surface residues for all the protein monomers correctly? However, there are many methods for protein-protein interface residue prediction[7–11], and protein-protein interface residue prediction is still an unsolved problem. The two key points of solving this problem are the characterization of protein surface residues and the combination of different characteristics or features to discriminate the true protein-protein interface residues from other surface residues. Because of the structures of surface residues are geometric, it is natural to describe surface residues using geometric features. The widely used geometric features are absolute and relative solvent accessibilities of surface residues [11]. There are

also several commonly used geometric features to describe the protein surface shape such protrusion index which is a measure of concavity or convexity [12]. In terms of geometric description of a single surface residue, the absolute and relative solvent accessibilities are not enough to characterize all the aspects of protein surface residues. It is necessary to use new effective geometric features to describe protein surface residues in detail. Additionally, because of the differences between the geometric structures of different residue types, the discriminants of the 20 commonly occurring amino acids for binding site prediction should be also different. In fact, de Moraes *et al*. found that the performance of the specific amino acid classifiers is better than the results of classifiers without distinction of amino acid types [13]. In their work, a large number of descriptors were selected to characterize different type of amino acids and linear discriminative analysis classifiers were formulated for interface residue prediction.

Here, we found three new geometric descriptors of protein surface residues and trained millions of residue type-specific or non-specific neural network models (nonlinear combinations of different features) for protein binding site prediction. The residue type-specific predictors were integrated to form new models to predict binding site or interface residue of any amino acid type. Here in our paper, residue type-specific and amino acid type-specific were two expressions with the same meaning. The results showed that the new geometric features are effective and the performances of the best predictors are much better than the ones of other methods reported previously.

RESULTS

The comparison of the results of different geometric features

The three new geometric features used to characterize a protein surface residue are exterior contact (EC) area with other residues in protein monomer, exterior void (EV) area which does not contact with other residues and solvent, interior contact (IC) area which is the sum of the contact areas between the atoms of the surface residue (see Methods). The two commonly used geometric features, absolute exterior solvent accessible area (absEA) (see Methods) and relative exterior solvent accessible area (relEA) are used as comparison.

The performances of the five geometric features were evaluated in the whole datasets composed of 134 protein monomers at unbound state. For any protein monomer, the surface residues were ranked by the values of the five geometric features in descending or ascending order respectively. We calculated the **A**rea **U**nder receiver operating characteristic (ROC) **C**urve (AUC) and recorded the rank of the first binding site (RFBS) of different protein monomers for the five geometric features. The higher the AUC and the smaller the RFBS, the better the performance. The perfect result is that all the true binding sites are ranked before other surface residues when AUC is equal to 1. RFBS is a relatively weak measure which is only concerned with the first true binding site.

The boxplot of AUC of different monomers for the five geometric features is shown in Figure 1. The results of absEA, relEA and IC in descending order (absEA_DO, relEA_DO and IC_DO) are superior to the results in ascending order (absEA_AO, relEA_AO and IC_AO) respectively (Figure 1A, B and E). The performances of EV and EC in ascending order (EV_AO and EC_AO) are better than the ones in descending order (EV_DO and EC_DO) respectively (Figure 1C and D). The mean values of AUC of different monomers for absEA_DO, relEA_DO, EC_AO, EV_AO and IC_DO are 0.62, 0.61, 0.55, 0.60 and 0.57 respectively. The boxplot of RFBS of different monomers for the five geometric features is shown in Figure 2. The performances of absEA_DO, relEA_DO, EC_AO, EV_AO and IC_DO are better than the ones of absEA_AO, relEA_AO, EC_DO, EV_DO and IC_AO respectively, which is consistent with the results shown in Figure 1. The mean values of RFBS of different monomers for absEA_DO, relEA_DO, EC_AO, EV_AO and IC_DO are 4.41, 5.16, 5.80, 6.52 and 3.47 respectively.

To compare the performances of the five geometric features, the percentages of protein monomers in the dataset were recorded when the RFBS of any one of these monomers is smaller than N. These protein monomers are called positive monomers. When N is from 1 to 30, the results are shown in Figure 3. IC is superior to the other four geometric features. When the number of retained surface residues for any protein monomer is 3, the percentages of positive monomers is 59.0, 50.0, 50.7, 44.0 and 67.9 for absEA, relEA, EC, EV and IC respectively. All the 134 protein monomers are positive ones when the number of retained surface residues are 22, 24 and 18 for absEA, relEA and IC respectively.

The best performance of models without distinction of residue types

Generally, the models are trained, validated and tested in the training, validation and test sets respectively. To find the best model with strong generalization, a new criteria is designed for protein-protein interface residue prediction. When the number of retained surface residues is from 1 to 10 and the standard deviation of percentages of positive monomers in the three sets (training, validation and test

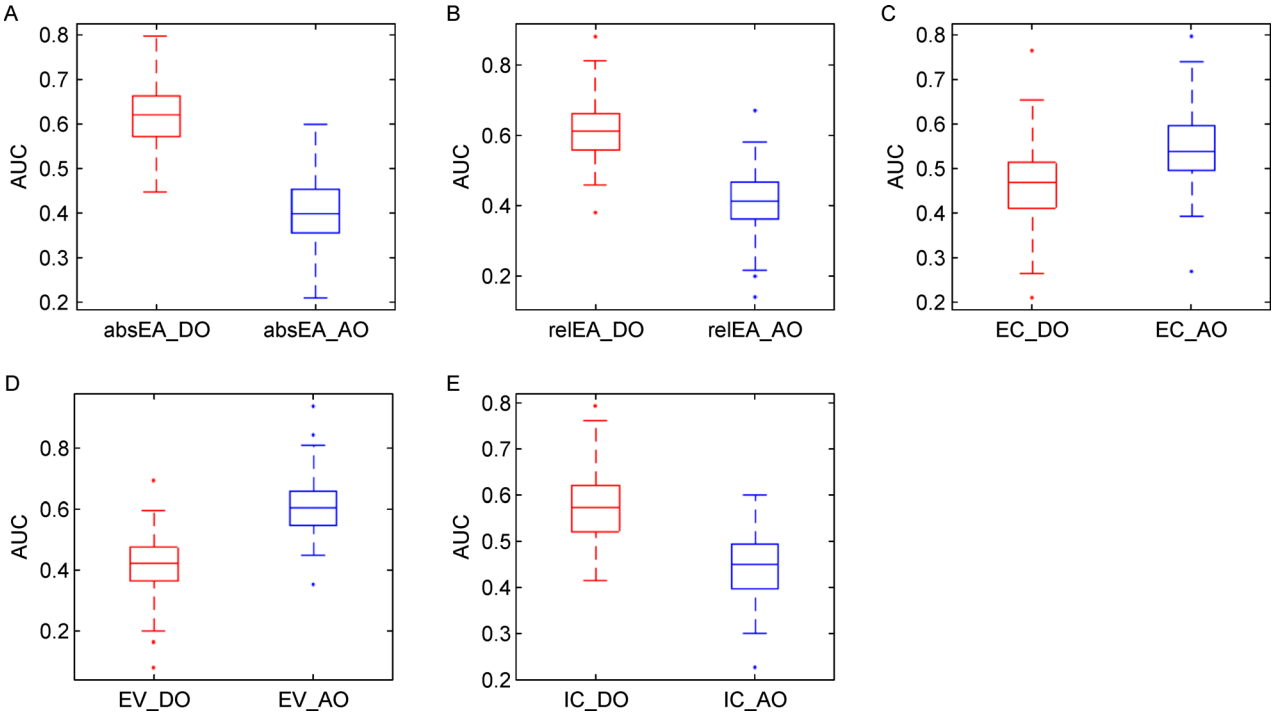

Figure 1. Comparison of the performances of the geometric features in descending and ascending orders with AUC. The results of absEA, relEA, EC, EV and IC are shown in (A), (B), (C), (D) and (E) respectively. DO represents descending order and AO represents ascending order. They are shown in red and blue respectively. The dots are the outliers.

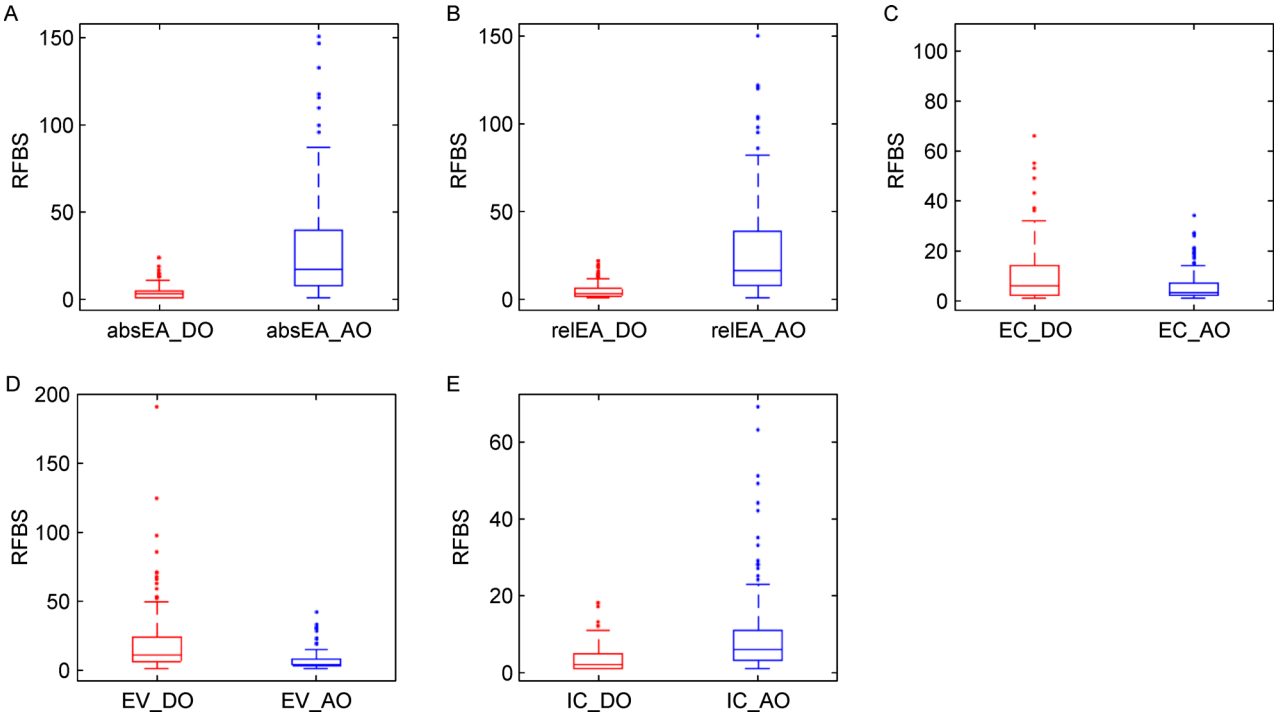

Figure 2. Comparison of the performances of the geometric features in descending and ascending orders with rank of the first binding site (RFBS). The results of absEA, relEA, EC, EV and IC are shown in (A), (B), (C), (D) and (E) respectively. DO represents descending order and AO represents ascending order. They are shown in red and blue respectively. The dots are the outliers.

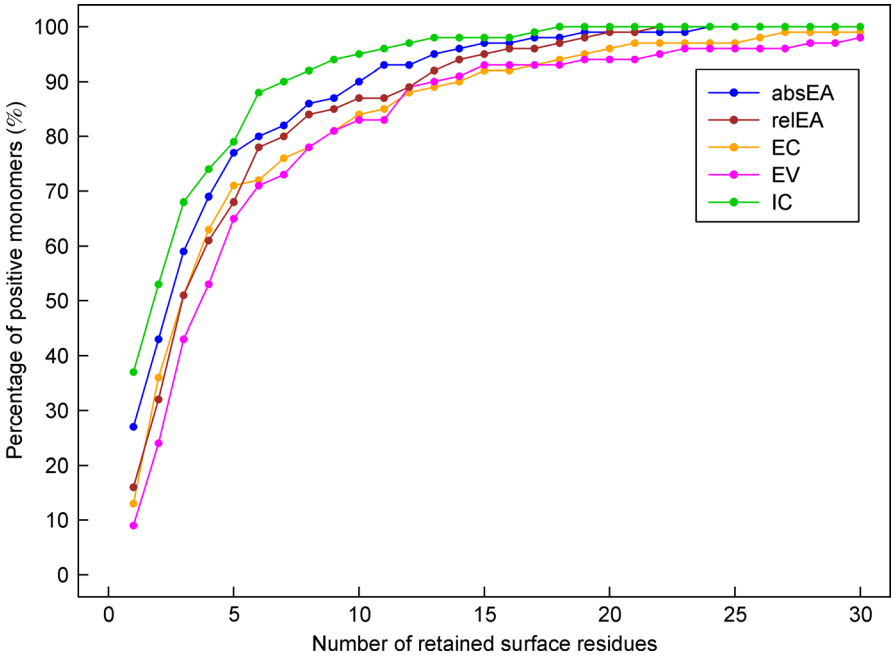

Figure 3. Comparison of the performances of the geometric features. The results are the ones of absEA in descending order, relEA in descending order, EC in ascending order, EV in ascending order and IC in descending order. The curves of absEA, relEA, EC, EV and IC are shown in blue, red, orange, purple and green respectively. The horizontal coordinate is the number of retained surface residues for any protein monomer. The vertical coordinate is the percentage of positive monomers for which there exist at least one true binding sites among the retained surface residues.

sets) is equal or smaller than 10, the highest mean percentage of positive monomers in the three sets and the corresponding feature combination are recorded.

As shown in Table 1, when the number of retained surface residues are 1, 3, 5 and 10, the mean percentages of positive monomers are 52.6, 79.4, 92.0 and 100 respectively. There are 14 kinds of feature combinations which can achieve the best performances. All the 14 kinds of feature combinations incorporate IC, and IC reflects the self-condition of a surface residue, which indicates that IC is the most important feature among the nine features. The second frequent feature is EC. EC is the only connection between a surface residue and the protein monomer, and its indispensability is foreseeable.

The results of residue type-specific predictors

For a specific amino acid type, when all the protein monomers which possess interface residues of the amino acid type are positive ones, the minimum number of retained surface residues is recorded. For the best models with the minimum of retained surface residues, the percentage of interface residues among the retained residues are calculated, the feature combination with the highest percentage of interface residues is also recorded.

As shown in Table 2, among the feature combinations corresponding different amino acid types, EC and IC are the most frequent features, which is in accord with the results of models without distinction of amino acid types. The minimum number of retained surface residues for different amino acid types is from 1 to 9, which is better than the results of models without distinction of amino acid types mentioned above. The correlation coefficient between minimum number of retained surface residues and the percentage of interface residues is -0.90, the minimum number of retained surface residues becomes smaller with the increasing of the percentage of interface residues, which indicates that enhancing the percentage of interface residues is an effective way to reduce the difficulty of interface residue prediction such as decomposing the dataset into different amino acid type-specific dataset in this work. EC and IC are the most frequent features among the 20 best feature combinations, which suggests that their importance is not influenced by amino acid type.

Additionally, the AUC in the validation and test sets were calculated for the best 20 amino acid type-specific predictors to investigate the differences of the performances of these models. As shown in Table 3, TRP, LEU, PHE, and CYS are among the 5 amino acid type with the highest mean AUC, which is consistent with the previous study [13]. The mean AUC of the best model for TRP is

Table 1 The best performance of models without distinction of residue types

NRSR	MPPM (%)	FC
1	52.6	absEA, relEA, EC, IC, pK_a2
2	69.6	relEA, EC, EV, IC, H1
3	79.4	absEA, relEA, EC, EV, IC, H1
4	87.2	EC, IC, H2, pK_a1
5	92.0	absEA, IC, H1, pK_a1, pK_a2
6	95.2	absEA, EV, IC, pK_a2
7	97.2	EC, IC, pK_a1
8	97.8	EV, IC, H1, H2, pK_a1, pK_a2
9	99.0	relEA, EC, EV, IC, H1, pK_a1, pK_a2
10	100	EV, IC, H1, H2, pK_a1
10	100	relEA, EC, IC, H1
10	100	absEA, EV, IC, pK_a2
10	100	absEA, relEA, EC, IC, H2, pK_a2
10	100	absEA, relEA, EC, EV, IC, H1, pK_a1

NRSR: Number of retained surface residues; MPPM: mean percentage of positive monomers; FC: feature combination.

the highest among the 20 amino acid types. One of the reasons may be that the percentage of interface residues among surface residues for TRP is the highest. The third highest mean AUC is the one of the best model for LEU,

Table 2 The best performance of residue type-specific predictors

AAT	NM	FC	MNRSR
TRP	1,302,000	absEA, relEA, EC, IC	1
MET	1,302,000	relEA, IC	2
CYS	2,646,000	absEA, relEA, EV, IC, pK_a1	2
HIS	2,646,000	absEA, relEA, EC, IC	3
PHE	1,302,000	relEA, EC, EV, IC	5
TYR	2,646,000	relEA, EC, IC, pK_a1	5
ASN	1,302,000	absEA, EV, IC	5
GLN	1,302,000	relEA, EC	5
GLY	1,302,000	relEA, EC, IC	6
LEU	1,302,000	absEA, EC, EV, IC	6
PRO	1,302,000	absEA, EC, EV, IC	6
ARG	2,646,000	absEA, relEA, EC, IC, pK_a1	6
ALA	1,302,000	absEA, relEA, EC, EV, IC	7
VAL	1,302,000	relEA, EC, IC	7
GLU	2,646,000	absEA, EC, EV	7
THR	1,302,000	absEA, EC, IC	7
ASP	2,646,000	relEA, EC, EV, IC	8
SER	1,302,000	absEA, EC, EV, IC	8
ILE	1,302,000	relEA, EC, EV, IC	2
LYS	2,646,000	absEA, EC, pK_a1,	9

AAT: amino acid type; NM: Number of models; FC: feature combination; MNRSR: minimum number of retained surface residues.

Table 3 Comparison of residue type-specific predictors with AUC

AAT	MV	SD
TRP	0.740691	0.026848
MET	0.709371	0.000889
LEU	0.691242	0.020786
PHE	0.681853	0.004221
CYS	0.673883	0.001095
VAL	0.655965	0.020019
ARG	0.643709	0.010168
HIS	0.623985	0.006025
ILE	0.619012	0.024085
GLY	0.617001	0.021070
TYR	0.605156	0.030080
PRO	0.603635	0.049412
GLN	0.598281	0.003730
LYS	0.591322	0.091578
SER	0.585913	0.002099
ASN	0.576256	0.020634
GLU	0.569275	0.087888
THR	0.558572	0.060780
MET	0.557027	0.048983
ASP	0.538908	0.046054

AAT: amino ccid type; MV: mean value of AUC in the validation and test sets; SD: standard deviation of AUC in the validation and test sets.

but the percentage of interface residues for LEU is very low (Supplementary Table S1). It may reflect that the structures of interface residues and non-interface residues for LEU are very different. ASN, GLU, ASP are among the 5 amino acid type with the lowest mean AUC, which is also consistent with the previous work [13].

The performance of models generated by integrating residue type-specific predictors

To break through the limitation mentioned above, the best models of different amino acid types were integrated to generate new models for predicting interface residues without distinction of amino acid types. There were 294,000 models trained in this step (see Methods).

As shown in Table 4, when the number of retained surface residues are 3 and 5, the mean percentages of positive monomers are 78.4 and 91.6 respectively. They are a little worse than the results of models trained directly in the set without distinction of amino acid types. However, it just need to retain 9 surface residues when all the protein monomers are positive ones. On the whole, the way used to integrate the models here does not work as effectively as expected. In the future, this is a direction that can be improved.

Table 4 The best performances of the integrated amino acid-specific predictors

NRSR	MPPM (%)
1	33.9
2	47.8
3	78.4
4	86.7
5	91.6
6	94.1
7	97.5
8	99.2
9	100.0

NRSR: number of retained surface residues; MPPM: mean percentage of positive monomers.

Comparison with other methods

In order to compare the performances of our method (IAASP, integrated amino acid-specific predictor) and IC with the ones of other sophisticated methods such as meta-PPISP[14], VORFFIR[15] and PredUs [16], the percentages of positive monomers for these methods were calculated in the dataset composed of the validation and test sets. The reason why we only chose IC rather than other features was because IC was the most frequent and important feature in both models with and without distinction of residue types.

As shown in Figure 4, IC is the best predictor when the number of retained residues is from 3 to 7. IAASP is the best one with the number of retained residues higher than 7. VORFFIR and PredUs are better when only 1 or 2 residues are retained.

The main reason of the differences may be that the other methods are developed to predict all the interface residues correctly. The aim of method proposed here is predicting at least one interface residues for every protein monomer correctly. The experimental biologists may be preferred to use our method.

An example (PDB code: 4H03) in the test set predicted by our method is shown in Figure 5. The percentages of interface residues of the receptor and ligand are 5.3 and 7.2 respectively. It is relatively difficult to predict the interface residues for these two monomers correctly. The rank of the first true interface residues for the two monomers are 4, which illustrates the effectiveness of our method.

DISCUSSION

The meaning of new geometric features

Protein binding site information is closely related to protein function. Protein function is determined by protein geometric structure. So, geometry could drive

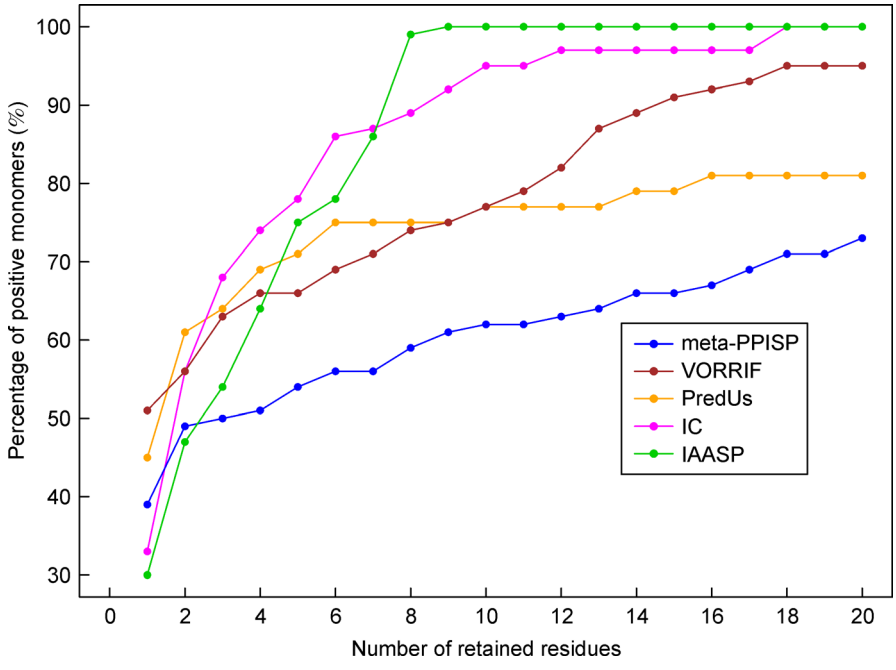

Figure 4. Comparison of the performances of the best integrated amino acid-specific predictor (IAASP) with other methods. The curves of meta-PPISP, VORRIF, PredUs, IC and IAASP are shown in blue, red, orange, purple and green respectively. The horizontal coordinate is the number of retained residues for any protein monomer. The vertical coordinate is the percentage of positive monomers for which there exist at least one true binding sites among the retained residues.

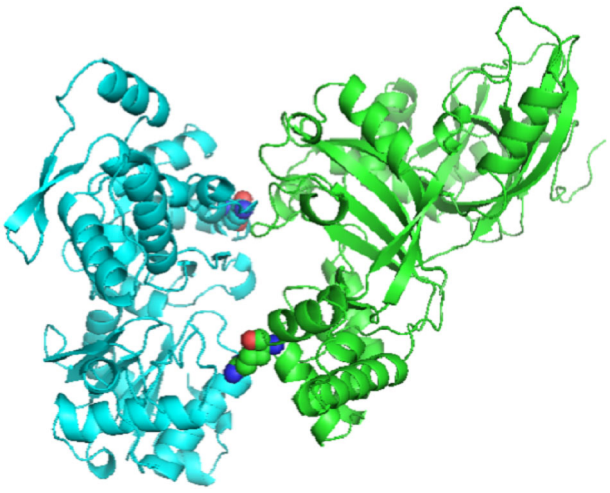

Figure 5. An example predicted by our method (PDB code: 4H03). The colors of the receptor and ligand are green and cyan respectively. The first true interface residues among the retained residues are shown in sphere model.

the discovery of protein binding site, and the geometric description of protein surface residues is of great importance for protein binding site prediction.

The new geometric features reflect different aspects of protein surface residues. EC is the "bridge" between a surface residue and protein monomer. The bridge will become stable with the increasing of EC. On the other hand, EC can be regarded as a measure of the strength of the structural restriction imposed by protein monomer on the surface residue. EV is a part of the residue surface and uncommitted by other molecules. IC is the internal condition of a surface residue. A higher value of IC corresponds to a more compact structure. The new geometric features and the frequent used ones such as absEA and relEA influence each other and characterize a surface residue together.

The implication of residue types

Generally, there are two factors influencing the performance of protein binding site predictor. One is the percentage of interface residues. If the percentage of interface residues is 100, any protein binding site predictor including random sampling can achieve the perfect results. Another one is the divergence degree between the interface and non-interface residues. The higher divergence degree will make the prediction easier.

Both of the two aspects mentioned above for the 20 residue types are different. The root cause is that the intrinsic geometric structures of different residue types are different. Every residue type may have its own discriminant for binding site prediction. Protein binding site prediction is relatively easy for some residues types such as TRP, MET and CYS because of the high percentages of interface residues. Although the percentage of interface residues for LEU is not high, the mean value of AUC for LEU is higher than those of many other residue types, which suggests that the geometric structures of interface and non-interface residues may be obviously different. Some residue types may be more suitable for functional signal recognition.

Why train millions of models?

Millions of models were trained in this work. There are two concrete reasons for training so many models. One is the large number of feature combinations. In order to find the most effective features and feature combination for protein binding site prediction, all the possible feature combinations were considered to find their own best models. Of course, there are many effective methods for feature selection [17]. The enumeration method adopted here is a little stupid, but effective. It is allowed by the computational resources available. Another one is the large number of structures of neural networks. The number of hidden layers and the number of nodes in every layer are adjustable, and the optimal structure of neural network is unknown. The only way is training models in some simple structures with one or two hidden layers and a small number of nodes. Even so, there are still hundreds of neural network structures. Additionally, because of the limitation of optimization algorithm, 100 models were trained for the fixed feature combination and neural network structure. It indicates that human and computer are not so clever as nature, so we have to try millions of times to find the rule of nature.

Indeed, there are many available deep architectures for pattern recognition. We have conducted some work about the application of deep learning on protein-protein interface residue pair recognition. Different from that work, this work is conducted to investigate the importance of geometric properties and amino acid types on protein-protein interface residue prediction. The method adopted here and the results can illustrate the determinative role of these two kind of factors. These results suggest that geometric properties and amino acid types may determine protein-protein interface residues.

CONCLUSIONS

In this work, different models with or without distinction of amino acid types for interface residue prediction were trained and compared, the amino acid type-specific predictors are superior than the models without distinction of amino acid types. Enhancing the percentage of interface residues is an effective way to make interface

prediction less difficult. No matter among the best amino acid type-specific or non-specific models, EC and IC are the most important features. Integrating the amino acid type-specific predictors generates new models to predict interface residues of any amino acid type. When the number of retained surface residues is 9, the percentage of positive proteins is 100. The result of the best model is much better than the ones of the other methods developed before. These results suggest that geometric properties and amino acid types may determine protein-protein interface residues.

METHODS

Constructing datasets

The procedure for constructing datasets is shown in Figure 6. The datasets are constructed based on protein-protein docking benchmark version 5.0 [18]. Protein-protein docking benchmark version 5.0 is the recent version of a public docking benchmark which is non-redundant and reliable [18]. Although the benchmark is constructed for developing and evaluating docking methods, it can be also used to develop interface prediction methods because the interfaces of the complexes in the benchmark are representative for different protein functional categories. There are 143 dimers among the 230 complexes in protein-protein docking benchmark version 5.0. In unbound state, there are 17 dimers in which there exists at least one position that are indeterminable, 57 dimers in which at least one interface residue is mutated or missing, and 2 dimers in which some backbone atoms are missing. So, only 67 dimers (134 monomers) satisfy our request. These dimers are divided into three subsets (training, validation and test sets) according to the version of benchmark. The benchmark version 5.0 was updated from benchmark version 3.0 and 4.0 [19,20]. Although the way of dividing the dataset is not unique, the proteins updated in different versions are not similar, the dividing way used here is one of reasonable ways for which the expandability and applicability of interface residue prediction method can be tested in the updated datasets.

Then, the information of interface and surface residues are obtained for all the monomers, and the values of different features are calculated for these residues. Finally,

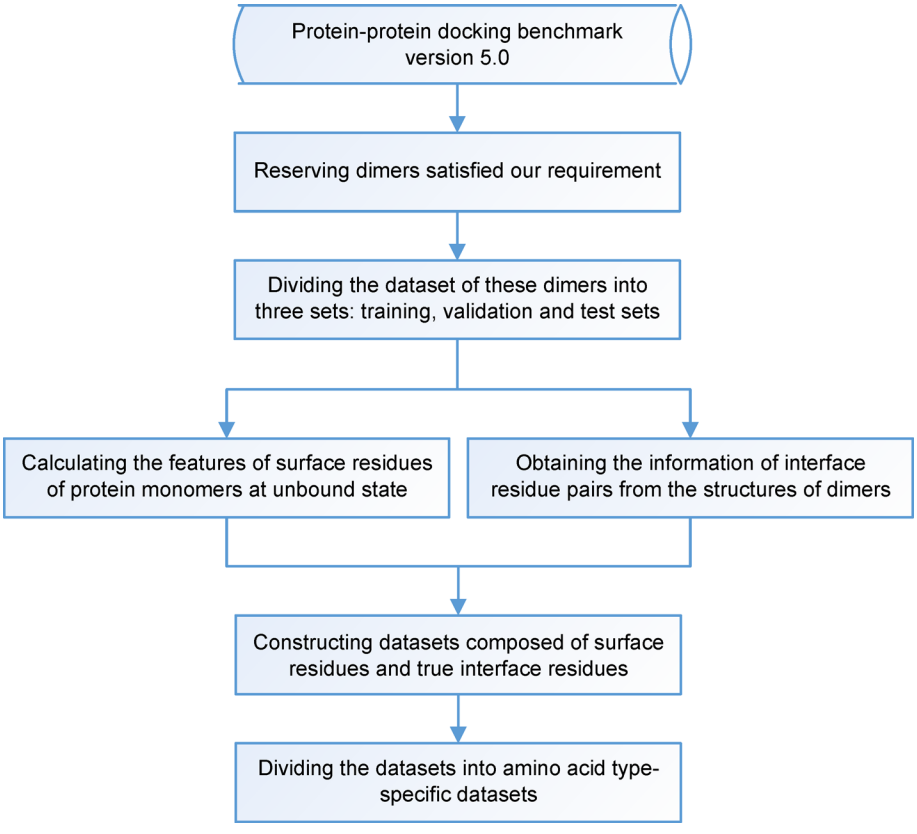

Figure 6. Procedure for constructing datasets. Firstly, protein dimers with high quality in protein-protein docking benchmark version 5.0 [18] are selected to obtain the interface and surface residues. Then, the dataset composed of interface and surface residues are divided into amino acid type-specific ones.

the different information are integrated and the datasets are divided into amino acid type-specific subsets.

Obtaining the information of surface and interface residues

When the absolute exterior solvent accessible area of a residue in a protein monomer is not zero, the residue is defined as a surface residue. If a surface residue of a protein monomer contact with the partner, the residue is defined as interface residue.

There are 68, 40 and 26 monomers in training, validation and test sets respectively (Supplementary Table S2). The number of surface residues of the 134 monomers in benchmark version 5.0 is 26,934, the number of interface residues is 3,575. The percentages of interface residues among the surface residues are 13.8, 13.9 and 11.1 in training, validation and test sets respectively. The percentage of interface residues among the surface residues in the whole set is 13.3. The probability that there exist at least one true interface residues among the randomly sampled 10 surface residues is 0.76, and the probability is almost zero that there exist true interface residues for all the 134 protein monomers when only 10 surface residues are randomly sampled for each of them.

The data information for different amino acid types are shown in Supplementary Table S1. Number of existing-surface-residue-monomers (ESRM) is the number of protein monomers which possess surface residues of specific amino acid type. Number of existing-interface-residue-monomers (EIRM) is the number of protein monomers which possess interface residues of specific amino acid type. Not all the 134 monomers have surface residues and interface residues for every amino acid type. The number of interface residues and surface residues in the sets of EIRM for different amino acid types are also shown in Supplementary Table S1. The percentages of interface residues among surface residues for the 20 amino acid types are from 12.9 to 33.3. For every amino acid type, the ESRM are also divided into three subsets to train, validate and test amino acid type-specific interface residue predictors according to the benchmark version as the datasets without distinction of amino acid types. The mean percentages of interface residues among surface residues in the three subsets are from 19.5 to 51.8.

Calculating the values of different features

Nine features are used to characterize the surface residues. The schematic diagram of geometric features of a surface residue are shown in Figure 7. The absEA (Figure 7B) and relEA are calculated by NACCESS [21]. EC is the sum of the areas of a surface residue contacting with other residues; IC is the sum of the contact areas between the atoms of a surface residue (Figure 7C). They are computed by Qcontacts [22]. The exterior void area (EV, Figure 7D) are the whole exterior solvent accessible area minus the sum of absEA and EC. The whole exterior solvent accessible area is calculated based on the

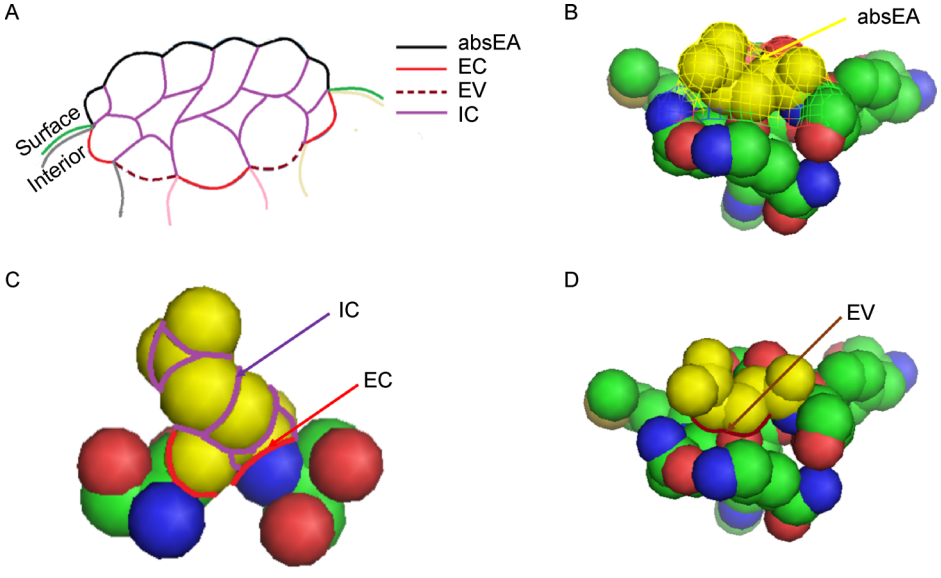

Figure 7. Schematic diagram of the geometric features. (A) Schematic façade in 2D of the geometric features. The colors of absolute exterior solvent accessible area (absEA), exterior contact area with other residues (EC), exterior void area (EV) and interior contact area of a surface residue (IC) are black, red, brown and purple respectively. (B) The absolute exterior solvent accessible area of a surface residue. The absEA is displayed in yellow mesh. (C) EC and IC of a surface residue. EC and IC are colored by red and purple respectively. (D) EV of a surface residue. EV is colored by brown.

coordinates of the surface residue which are extracted from the protein structure.

In order to describe a protein surface residue better, two versions of hydropath index (H1 and H2) [23,24] and the computing and standard pK_a (pK_a1 and pK_a2) [25] were also used as the descriptors. The four features are also related to geometric structures and residue types of surface residues. The nine features are shown as follows:

(1) absEA: **abs**olute **E**xterior solvent accessible **A**rea
(2) relEA: **rel**ative **E**xterior solvent accessible **A**rea
(3) EC: **E**xterior **C**ontact area with other residues
(4) EV: **E**xterior **V**oid area
(5) IC: **I**nterior **C**ontact area
(6) H1: **H**ydropathy index, version **1**
(7) H2: **H**ydropathy index, version **2**
(8) pK_a1: computation
(9) pK_a2: standard

Training neural network models

Models without distinction of residue types

The models were generated based on the training set without distinction of amino acid types. They were evaluated in the validation and test sets. The best models were selected according to the performances in the three sets. The workflow is shown Supplementary Figure S1.

In order to explore the best performance of the nine features which can form 511 kinds of feature combinations, we trained a lot of pattern recognition neural networks and applied the back-propagation mechanism to learn the weight of the network with scaled conjugate gradient algorithm[26–28], and the hyperbolic tangent sigmoid transfer function was adopted in the hidden layers, the softmax transfer function in the output layer. The formation of the hyperbolic tangent sigmoid transfer function is as follows:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \quad (1)$$

The target output of the network was set to 0 or 1, 0 represents the non-interface residue pairs, 1 represents the interface residue pairs. The network could return a real number between 0 and 1.

For any one of the 511 kinds of feature combinations, the number of the hidden layers could be 1 or 2, the number of neurons in every hidden layer can be from 1 to 10, so there are 110 different neural network architectures. In order to search the optimal model in the situation that the initial weights of the neurons are randomly assigned, 100 models were trained when the feature combination and the network architecture were fixed, which was restricted by our computing resources. Finally, 11,000 models were generated for every kind of feature combination. So, there are 5,621,000 models for predicting interface residues.

Although millions of neural network models were trained in the training set, they were uncorrelated with the validation and test sets, the models with overfitting can be excluded according to the performance in the validation and test sets.

Models with distinction of residue types

Because of the differences of geometric properties of the 20 residue types, their protein binding site predictors should differ from one another. The models with distinction of amino acid types were trained based on the amino acid type-specific training set. They were evaluated in the corresponding amino acid type-specific validation and test sets. The best models were selected according to the performances in the three sets. The workflow is shown Supplementary Figure S2.

Because the two versions of hydropathy index and pK_a2 are constant for specific amino acid type, they do not make differences on the performance of the amino acid type-specific models. Additionally, the values of pK_a1 are zeros for some amino acid types. So, there are five or six features employed to describe a surface residue. They can form 31 or 62 kinds of feature combinations.

In order to find out the best performance of the features for specific amino acid type, a large number of pattern recognition neural networks were trained, the procedure was similar as the one used above. The difference is that the number of neurons in every hidden layer can be from 1 to 20, there are 420 different neural network architectures because of less number of features. We let the number of nodes of models with distinction of residue types be 20 in exchange under the conditions of computing resources. 42,000 models are generated for each kind of feature combination. Finally, there are 1,302,000 or 2,646,000 models for a specific amino acid type to predict interface residues. In total, there are 7,896,000 models for the 20 amino acid types. These models are called amino-acid-specific predictors (AASP).

Integrating residue type-specific predictors

For every amino acid type, there exist some models whose performance are the best. The performance is evaluated in the set of the monomers which possess interface residues of the amino acid type. The workflow is shown Supplementary Figure S3.

Because it is not necessary to know all the interface information in biochemical experiments, and theoretically predicting all the interface residues correctly is very difficult, the performance of a predictor on a protein monomer is estimated by the existence of true interface

residues among the retained surface residues. If there are at least one interface residues for the monomer among the top N retained surface residues, the protein is called positive monomer. The criterion is designed based on the real requirement of biological experiments. Additionally, signals on protein surface for protein-protein binding are only a small part of the residues on protein-protein interface. In other words, only a small fraction of protein-protein interface residues play a key role in the formation of protein-protein complex. It may be easier to recognize these key residues than other interface residues theoretically.

When all the monomers in the set are positive ones, the minimum number of retained residues and the corresponding models are recorded. In those models, the best model is the one whose percentage of interface residues among the retained residues is the highest. The predicted values by the best model are regarded as the combined feature of surface residues of the amino acid type.

The procedure of constructing new datasets, generating and analyzing new models is shown in Supplementary Figure S3. The combined feature, H1 and H2 which can form 7 kinds of feature combinations are used to describe all the surface residues without distinction of amino acid types. As we know, the models with distinction residue type used 5 geometric features or 6 features combining 5 geometric features and pK_a1. So the three features contains both geometric and hydropathy information were enough for modeling. According to the same procedure of training models with distinction of amino acid types, there are 42,000 models generated for each kind of feature combination. Finally, there are 294,000 models to predict interface residues for any protein. These models are called Integrated Amino Acid-Specific Predictors (IAASP).

Accuracy measures

Interface residue prediction can be regarded as a binary classification problem, each residue in protein monomer could be either interface (positive, P) or non-interface (negative, N) one. Many accuracy measures are constructed base on the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

AUC is a metric to evaluate the discriminative ability of the method. Its values range from 0 to 1, where 1 corresponds to a perfect prediction, 0 to a perfectly inverse prediction. ROC curve represents the relationship between False Positive Rate (FPR) and True Positive Rate (TPR). TPR is the fraction of correctly predicted interface residues:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \quad (2)$$

FPR is the fraction of incorrectly predicted interface residues:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{FN}}. \quad (3)$$

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.1007/s40484-018-0138-5.

ACKNOWLEDGEMENTS

Experiments run on Renda Xing Cloud that currently has 64 physical nodes. This research was supported by the National Natural Science Fundation of China (Nos. 31670725 and 91730301), and the State Key Laboratory of Membrane Biology to Xinqi Gong.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yongxiao Yang, Wei Wang, Yuan Lou, Jianxin Yin and Xinqi Gong declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Gao, M. and Skolnick, J. (2010) Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proc. Natl. Acad. Sci. USA, 107, 22517–22522
2. Chothia, C. and Janin, J. (1975) Principles of protein-protein recognition. Nature, 256, 705–708
3. Jones, S. and Thornton, J. M. (1996) Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA, 93, 13–20
4. Keskin, O., Gursoy, A., Ma, B. and Nussinov, R. (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem. Rev., 108, 1225–1244
5. Koshland, D. E. (1995) The key-lock theroy and the induced fit theroy. Angew. Chem. Int. Ed., 33, 2375–2378
6. Teichmann, S. A. (2002) Principles of protein-protein interactions. Bioinformatics, 18, S249
7. Zhang, Q. C., Petrey, D., Norel, R. and Honig, B. H. (2010) Protein interface conservation across structure space. Proc. Natl. Acad. Sci. USA, 107, 10896–10901
8. Aumentado-Armstrong, T. T., Istrate, B. and Murgita, R. A. (2015) Algorithmic approaches to protein-protein interaction site prediction. Algorithms Mol. Biol., 10, 7
9. Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J. C. and Deane, C. M. (2016) Progress and challenges in predicting protein interfaces. Brief. Bioinformatics, 17, 117–131
10. Maheshwari, S. and Brylinski, M. (2015) Predicting protein interface residues using easily accessible on-line resources. Brief. Bioinform., 16, 1025–1034
11. Xue, L. C., Dobbs, D., Bonvin, A. M. and Honavar, V. (2015) Computational prediction of protein interfaces: a review of data driven methods. FEBS Lett., 589, 3516–3526

12. Pintar, A., Carugo, O. and Pongor, S. (2002) CX, an algorithm that identifies protruding atoms in proteins. Bioinformatics, 18, 980–984
13. de Moraes, F. R., Neshich, I. A., Mazoni, I., Yano, I. H., Pereira, J. G., Salim, J. A., Jardine, J. G. and Neshich, G. (2014) Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. PLoS One, 9, e87107
14. Qin, S. and Zhou, H. X. (2007) meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics, 23, 3386–3387
15. Segura, J., Jones, P. F. and Fernandez-Fuentes, N. (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics, 12, 352
16. Zhang, Q. C., Deng, L., Fisher, M., Guan, J., Honig, B. and Petrey, D. (2011) PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res., 39, W283–W287
17. Wang, L., Wang, Y. and Chang, Q. (2016) Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods, 111, 21–31
18. Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jimenez-Garcia, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. and Weng, Z. (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J. Mol. Biol. 427, 3031–3041
19. Hwang, H., Vreven, T., Janin, J. and Weng, Z. (2010) Protein-protein docking benchmark version 4.0. Proteins, 78, 3111–3114
20. Hwang, H., Pierce, B., Mintseris, J., Janin, J. and Weng, Z. (2008) Protein-protein docking benchmark version 3.0. Proteins, 73, 705–709
21. Hubbard, S.J. and Thornton, M. (1993) Naccess Version 2.1.1. Department of Biochemistry and Molecular Biology, University College, London
22. Fischer, T. B., Holmes, J. B., Miller, I. R., Parsons, J. R., Tung, L., Hu, J. C. and Tsai, J. (2006) Assessing methods for identifying pair-wise atomic contacts across binding interfaces. J. Struct. Biol., 153, 103–112
23. Eisenberg, D. (1984) Three-dimensional structure of membrane and surface proteins. Annu. Rev. Biochem., 53, 595–623
24. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157, 105–132
25. Olsson, M. H., Søndergaard, C. R., Rostkowski, M. and Jensen, J. H. (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pK_a predictions. J. Chem. Theory Comput., 7, 525–537
26. Møller, M. F. (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw., 6, 525–533
27. Kishore, R. and Kaur, M. T. (2012) Backpropagation algorithm: an artificial neural network approach for pattern recognition. Inter. J. Sci. & Engin. Res., 3, 1–4
28. Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning representations by back-propagating errors. Nature, 323, 533–536

