Prediction of hot spots in protein interfaces using a random forest model with hybrid features

Lin Wang1, Zhi-Ping Liu2,3, Xiang-Sun Zhang3,4,6 and Luohan Chen2,5,6

1School of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin 300222, China, 2Key Laboratory of Systems Biology, SIBS-Novozyme Translational Research Centre for Pre-Diabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, 3National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China, 4Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and 5Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan

6To whom correspondence should be addressed.
E-mail: zxw@amt.ac.cn (X.-S.Z.); lchen@sibs.ac.cn (L.C.)

Received August 31, 2011; revised November 29, 2011; accepted December 20, 2011

Edited by Roger Williams

Prediction of hot spots in protein interfaces provides crucial information for the research on protein–protein interaction and drug design. Existing machine learning methods generally judge whether a given residue is likely to be a hot spot by extracting features only from the target residue. However, hot spots usually form a small cluster of residues which are tightly packed together at the center of protein interface. With this in mind, we present a novel method to extract hybrid features which incorporate a wide range of information of the target residue and its spatially neighboring residues, i.e. the nearest contact residue in the other face (mirror-contact residue) and the nearest contact residue in the same face (intra-contact residue). We provide a novel random forest (RF) model to effectively integrate these hybrid features for predicting hot spots in protein interfaces. Our method can achieve accuracy (ACC) of 82.4% and Matthew’s correlation coefficient (MCC) of 0.482 in Alanine Scanning Energetics Database, and ACC of 77.6% and MCC of 0.429 in Binding Interface Database. In a comparison study, performance of our RF model exceeds other existing methods, such as Robetta, FOLDEF, KFC, KFC2, MINERVA and HotPoint. Of our hybrid features, three physicochemical features of target residues (mass, polarizability and isoelectric point), the relative side-chain accessible surface area and the average depth index of mirror-contact residues are found to be the main discriminative features in hot spots prediction. We also confirm that hot spots tend to form large contact surface areas between two interacting proteins. Source data and code are available at: http://www.aporc.org/doc/wiki/HotSpot.

Keywords: hot spot/prediction/protein interface/random forest/structural bioinformatics

Introduction

The protein–protein interface is the region of interaction between two non-covalently linked protein molecules. These protein–protein interactions play a key role in signal transduction and metabolic networks (Chen et al., 2009, 2010). Alanine mutation of protein–protein interface residues has shown that the distribution of binding free energy is not average among the interface residues. Instead, there are hot spots that contribute most binding energy (Bogan and Thorn, 1998) in protein interfaces. Structural analysis of the protein–protein surface shows that structurally conserved residues tend to be hot spots, and distinguish between binding sites and exposed surface residues (Hu et al., 2000; Ma et al., 2003). Moreover, several recent studies discovered small molecules bound to hot spots in protein interfaces that can disrupt protein–protein interactions (Wells and McClendon, 2007). It is crucial to identify hot spots in protein interfaces for understanding protein–protein interaction. Uncovering hot spots is also considered as a good starting point for drug design.

Experimental methods such as alanine scanning mutagenesis are time consuming and costly. Therefore, hot spots are available only for a very limited number of complexes. There is a need for computational tools to predict hot spots in protein interfaces (Moreira et al., 2007). So far, several methods have been developed for the prediction, and can be roughly categorized into three groups. The first group includes energy-based methods, such as Robetta (Kortemme and Baker, 2002) and FOLDEF (Guerois et al., 2002), which make a prediction based on an estimate of the energetic contribution to binding for every interface residue. The second group covers knowledge-based methods that try to learn the complex relationship between hot spots and various residue features in training data and predict new hot spots. Ofran and Rost (2007) applied neural networks to predict hot spots with features extracted from sequence environment and evolutionary profile. Darnell et al. (2007) predicted hot spots by decision trees with various features such as atomic contacts, physicochemical properties, shape specificity and computational alanine scanning. Tuncbag et al. (2009) presented an empirical method by combining conservation, solvent accessibility and statistical pairwise residue potentials. Cho et al. (2009) applied support vector machines (SVMs) to predict hot spots with features such as weighted atom packing density, relative accessible surface area, weighted hydrophobicity and molecular interaction types. Xia et al. (2010) and Zhu and Mitchell (2011) also employed SVM classifiers with features such as protrusion index, solvent accessibility and local plasticity. Assi et al. (2010) applied Bayesian networks to predict hot spots with features incorporating structural, evolutionary and energetic information. The third group generally refers to the other methods of predicting hot spots. Shulman-Peleg...
et al. (2007) performed spatial comparisons of physico-chemical interactions shared by functionally similar protein–protein complexes and identified hot spots. Grosdidier and Fernandez-Recio (2008) predicted hot spots by computational docking without protein complex information. del Sol and O’Meara (2005) represented protein complexes as small-world networks, and showed that a relatively small number of highly central residues correspond to hot spots. Li and Liu (2009) used bipartite graph representation of protein complexes. Maximal biclique subgraphs were subsequently identified from the bipartite graphs to locate biclique patterns which might correspond to hot spots. Tuncbag et al. (2010a) analyzed residue contact networks of protein interfaces using minimum cut trees and observed that the highest degree node in each minimum cut tree usually corresponds to a hot spot.

In this work, we propose a novel method for predicting hot spots which extracts a wide variety of features from protein interfaces and employs a random forest (RF) model. Given a protein complex, we first identify interface residues, then classify each interface residue as a hot spot or not. For each target residue, existing machine learning methods extract features only from the target residue. However, hot spots were found to be clustered within locally and tightly packed regions (Keskin et al., 2005). Thus, we extract the features from target residue and its two neighboring residues, i.e. the nearest contact residue in the other face (mirror-contact residue) and the nearest contact residue in the same face (intra-contact residue). We also identify various features underlying the protein interfaces and build an RF model to effectively integrate these hybrid features. The cross-validation prediction results show that our RF predictor achieves overall accuracy (ACC) of 82.4%, Matthew’s correlation coefficient (MCC) of 0.482 and an area under the receiver operating characteristic curve (AUC) of 0.811 in Alanine Scanning Energetics Database (ASEdb, Thorn and Bogan, 2001). In the comparison study, our RF model achieving ACC of 77.6% and MCC of 0.429 outperforms other existing hot spots prediction methods in an independent test set of Binding Interface Database (BID, Fischer et al., 2003). We also evaluate the contribution of every feature to the prediction accuracy. Compared with traditional features, our new features in the model such as the three physico-chemical features of target residues, namely mass, polarizability and isoelectric point, relative side-chain accessible surface area and average depth index of mirror-contact residues are identified as the main discriminative features in the prediction. In particular, the mass of target residues is found to be the most effective feature for predicting hot spots. Moreover, we confirm that hot spots tend to contain large contact surface areas across protein interfaces.

Features of description
To build a predictor that can distinguish hot spots from non-hot spots, we extract various descriptors for every interface residue. Generally, they are categorized into five groups based on their sources and properties.

Category of residues and secondary structure. Electrostatic (including hydrogen bonding) and hydrophobic interactions play key roles in protein–protein interactions. They can be reflected by the dipoles and volumes of the side chains of amino acids, respectively. Similar to that in Shen et al. (2007), we cluster 20 amino acids into six classes based on their dipoles and volumes of the side chains, i.e. Class 1: A, G, V; Class 2: I, L, F, P; Class 3: Y, M, T, S, C; Class 4: H, N, Q, W; Class 5: R, K and Class 6: D, E. Furthermore, residue secondary structure (SS) obtained by dictionary of protein secondary structure algorithm (Kabsch and Sander, 1983) is also used as a descriptor. There are three types of SS, i.e. helix, strand and loop. Thus, we extract two categorical descriptors for every residue.

Atom contacts and atom contact areas. The contact between two atoms (atom_contact) is defined by the CSU program (Sobolev et al., 1999). We exclude any non-attractive contact such as hydrophobic–hydrophilic contacts that destabilize the protein complex, and the remaining types of atom contacts are considered in the following calculation. Atom contacts of a residue (e.g. the ith residue) are computed by summing these atom contacts between the residue and any other residue (e.g. the jth residue) in the protein–protein interface, i.e.

\[
\text{atomic_contacts}(i) = \sum_j \left\{ \sum_{\alpha \in i \cap \beta \in j} \text{atomic_contact}(\alpha, \beta) \right\}
\]  

where atomic_contact (α, β) is equal to 1 if atom α is interacting with atom β, otherwise, it is equal to 0. To represent the interaction area between the target residue and the other face of the interface, atom contact areas of a residue are calculated by summarizing these atom contact areas between the

Materials and methods
Datasets
An interface residue was defined as the one that has atomic contacts with residues that belong to any other chain in the complex. The atomic contacts between residues were described using the CSU program (Sobolev et al., 1999), which is based on inter-atomic distances and the extent of crowding in the environment. A dataset of 318 alanine-mutated interface residues derived from 20 protein complexes was taken from ASEdb (Thorn and Bogan, 2001). The protein complexes are ensured as non-homologous because the sequence identity of at least one protein in each interface is <35% from all other proteins in the dataset. The sequence identity between proteins was calculated using the PISCES sequence culling server (Wang and Dunbrack, 2003). Hot spot was defined when its change of binding energy (ΔΔG) is ≥2.0 kcal/mol upon its mutation to alanine. Thus, these interface residues were divided into 77 hot spots and 241 non-hot spots. The dataset was implemented as our training set. Additionally, the dataset of Cho et al. (2009), which was derived from BID (Fischer et al., 2003), was used as an independent test set. Two residues which are not in protein interfaces have been removed. Moreover, we ensured that these proteins in the test set are not homologous to those in the training set by the same definition in the training set. As a result, the independent test set contained 125 alanine-mutated interface residues in 18 protein complexes with 38 hot spots and 87 non-hot spots. The training and test sets are listed in the Supplementary materials.
Prediction of hot spots in protein interfaces using a RF model with hybrid features

121

residue and any other residue located at the other face, i.e.

\[
\text{atom\_contact\_areas}(i) = \sum_{j \in \text{other face}} \left\{ \sum_{\alpha \in i} \sum_{\beta \in j} \text{atom\_contact\_area}(\alpha, \beta) \right\} 
\]  

(2)

where \( \text{atom\_contact\_area}(\alpha, \beta) \) is the contact area between two interacting atoms \( \alpha \) and \( \beta \).

**Residue contacts and physicochemical features.** Two residues will have residue contact information (residue contact) if there is one pair of contact atoms from them individually. Residue contacts of a residue are calculated by summarizing these residue contacts between the residue and other residues in the interface, i.e.

\[
\text{residue\_contacts}(i) = \sum_{j} \text{residue\_contact}(i, j) 
\]  

(3)

where \( \text{residue\_contact}(i, j) \) is equal to 1 if residue \( i \) interacts with residue \( j \), otherwise, it is equal to 0.

Six more physicochemical features of a residue are also identified as the descriptors, namely, hydrophobicity, hydrophilicity, isolectric point, mass, polarity and polarizability. We obtained the parameters of hydrophobicity from Fauchere and Pliska (1983) and the parameters of other properties from the AAindex database (Kawashima et al., 1999). We define the physicochemical features of a residue by itself and its interacting residues. For instance, the mass of the \( i \)th residue is defined as

\[
\text{mass}(i) = \text{mass\_para}(i) + \sum_{j} \text{mass\_para}(j) 
\]  

(4)

where \( \text{mass\_para}(i) \) and \( \text{mass\_para}(j) \) are the mass parameter of the \( i \)th residue and that of its interacting residue \( j \), respectively.

**Relative accessible surface area and relative side-chain accessible surface area.** It is known that the relative accessible surface area of residues is closely related to the protein–protein interactions (Cho et al., 2009; Tuncbag et al., 2009; Xia et al., 2010). Here, we compute the relative accessible surface area (rel_ASA) of the \( i \)th residue as

\[
\text{rel\_ASA}(i) = \frac{\text{ASA}_{\text{mono}}(i) - \text{ASA}_{\text{comp}}(i)}{\text{ASA}_{\text{mono}}(i)} 
\]  

(5)

where \( \text{ASA}_{\text{mono}}(i) \) and \( \text{ASA}_{\text{comp}}(i) \) are the buried surface areas of the \( i \)th residue in monomer and complex, respectively.

The side chain distinguishes one amino acid from another and dictates its specifically structural and functional properties. We also calculate the relative side-chain accessible surface area (rel_s-ch_ASA) as (5) simultaneously.

**Depth index.** The depth of an atom refers to the distance from its closest solvent accessible atom. Hot spots are usually deeply buried inside the protein interface to form local interactions (Moreira et al., 2007). For the \( i \)th residue in complex state, several depth indices were calculated by the PSAIA program (Mihel et al., 2008), i.e. average depth index (ave_dpx), mean depth of all atoms in a residue, standard deviation of depth index (sd_dpx), side-chain average depth index (s-ch_ave_dpx), mean depth of all side-chain atoms) and standard deviation of side-chain depth index (sd_s-ch_dpx). Relative depth index (rel_dpx) and relative side-chain depth index (rel_s-ch_dpx) of the \( i \)th residue upon complex formation are calculated as

\[
\text{rel\_dpx}(i) = \frac{\text{ave\_dpx}_{\text{comp}}(i) - \text{ave\_dpx}_{\text{mono}}(i)}{\text{ave\_dpx}_{\text{comp}}(i)} 
\]  

(6)

\[
\text{rel\_s\_ch\_dpx}(i) = \frac{\text{s\_ch\_ave\_dpx}_{\text{comp}}(i) - \text{s\_ch\_ave\_dpx}_{\text{mono}}(i)}{\text{s\_ch\_ave\_dpx}_{\text{comp}}(i)} 
\]  

(7)

Totally, there are 19 descriptors for each residue. For every target residue that we are predicting, the features are encoded from its descriptors and those of its two neighbor residues, i.e. mirror-contact residue and intra-contact residue. The mirror-contact residue of a target residue is defined as its nearest contact residue in the other face of this interface, and the intra-contact residue is the nearest contact residue in the same face. The distance between two residues is defined as the shortest Euclidean distance between their atoms. Consequently, there are 57 features for every target residue.

**Classification algorithm**

RF is an ensemble classification algorithm that employs a collection of decision trees to reduce the output variance of individual trees and thus improves the stability and accuracy of classification (Breiman, 1992; Liu et al., 2010). Typically, an RF model is created by two main ideas of bagging and random selection of features. After drawing several bootstrap samples from these original ones, each unpruned tree is trained on each bootstrap sample individually. At each node of a tree, RF randomly selected a constant number of features and the one with the maximum decrease in Gini index is chosen for the split when growing the tree. After training, the prediction is determined by a majority voting scheme of individual trees (Breiman, 1992). In this work, the randomForest R package (Liaw and Wiener, 2002) was implemented. We built an RF model by setting the number of randomly selected features and the number of trees to be 11 and 68, respectively. As part of the model, RF provides the measurement of feature importance, which is computed as the average decrease of classification accuracy when the values of a particular feature are randomly permuted on these out-of-bag (OOB) samples. We used the measure to evaluate the importance of various features.

**Measurements of prediction performance**

The prediction accuracy (ACC), sensitivity (SE), precision (PR), specificity (SP) and MCC were used for assessing prediction performance. These measurements are defined as

\[
\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} 
\]  

(8)

\[
\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} 
\]  

(9)

\[
\text{PR} = \frac{\text{TP}}{\text{TP} + \text{FP}} 
\]  

(10)

\[
\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} 
\]  

(11)

\[
\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} 
\]  

(12)
where TP, FP, TN and FN are the numbers of true-positive, false-positive, true-negative and false-negative residues in the prediction, respectively. Furthermore, we performed the receiver operating characteristic (ROC) analysis for the prediction and comparison. The predictive performance is usually evaluated in terms of the area under the ROC curve (AUC). However, AUC might not be optimal when the training set has a high ratio of negative samples because a low specificity in actual number could translate into a large number of false positive. In those cases, one should focus on the region of the ROC curve with high specificity, which is often of prime interest. Hence, we complemented the AUC measure with partial AUC (pAUC) which is calculated only in the specificity ranges from 1 to 0.8 (Robin et al., 2011).

### Results and discussion

**Cross validation in the training set**

We tested the RF classifier by the 10-fold cross validation in the dataset. The dataset was divided into 10 folds randomly with almost the same size. The 9-fold was used as a training set and the remaining 1-fold was used as a test set. The process was repeated 10 times for each fold as a test set. It predicted these hot spots with ACC of 82.4% and MCC of 0.482. The overall prediction performance is shown in Table I. Our RF model extracts features from the target residue and its two neighbor residues, i.e. mirror-contact residue and intra-contact residue. In contrast, the prediction result by the features only from the target residue is also shown in Table I, i.e. RF_target method, which achieved an ACC of 79.6% and MCC of 0.385. We found that the features extracted from the neighbor residues indeed improve the prediction accuracy. SVM is often selected as an alternative classifier. For comparison, we also implemented an SVM algorithm in the training set. It achieved an ACC of 80.8% and MCC of 0.457 in the 10-fold cross validation.

Robetta is an alanine scanning method for hot spots prediction. We used the estimated ΔΔG value as the classification score and 2.0 kcal/mol as the threshold. In the training set, Robetta predicted hot spots with ACC of 78.6% and MCC of 0.397 individually. The detailed measures of these different predictors are also listed in Table I. The results indicate that our RF model has better prediction performance.

We also performed the ROC analysis on these different models. The ROC curves are shown in Fig. 1. The points corresponding to the defaulted score thresholds are also plotted. The AUC values are 0.811, 0.793, 0.804 and 0.789 for RF, RF_target, SVM and Robetta, respectively. Furthermore, our RF model outperforms other models in the left high-sensitivity part of the ROC curves. The pAUC values are 0.711, 0.655, 0.690 and 0.677 for RF, RF_target, SVM and Robetta, respectively. These results provide more evidence for the effectiveness of the proposed RF model of predicting hot spots in protein interfaces with hybrid features.

**Prediction in the independent test set**

To remove the possible biases in the comparison with other methods, we implemented the prediction in an independent test set derived from BID by our trained RF predictor. We compared our RF method with several related works that include alanine scanning methods such as Robetta and FOLDEF (Guerois et al., 2002), decision tree methods such as KFC (Darnell et al., 2007) and three recently published methods MINERVA (Cho et al., 2009), HotPoint (Tuncbag et al., 2009, 2010b) and KFC2 (Zhu and Mitchell, 2011). The samples in the test set were predicted by the available methods individually. The prediction performance of these methods is shown in Table II. The trained RF achieved the ACC of 77.6% and MCC of 0.429. In Table II, our method shows the best prediction performance compared with other methods.
methods in the independent set. Furthermore, we validated the prediction performance of our method by statistics-based assessment. The details are shown in the Supplementary materials.

**Feature analysis**

The RF algorithm has the capability to estimate the importance of a specific feature. It was evaluated by calculating the average decrease of classification accuracy on the OOB samples when the values of a particular feature are randomly permuted. Table III shows the mean decrease of accuracy of these particular features when it is >15%. The permutation-based mean decrease in accuracy was used to assess the contribution of each feature to the classification. We found that the features extracted from target residues and their mirror-contact residues play an important role in the prediction of hot spots. This implies the strong relationship between the cooperativity of these residues and interaction events in protein interface (Schreiber and Fersht, 1995). Moreover, hot spots contribute most of the conserved physicochemical interactions across the interfaces (Shulman-Peleg et al., 2007). The importance of features provides insights for their discrimination of specificity in hot spots and non-hot spots. To show the feature differences in hot spots and non-hot spots, we chose four features which are relatively high-discrimination power features from four feature groups, respectively, i.e. the mass and the atom contact areas of target residues, the relative side-chain accessible surface area and the average depth index of mirror-contact residues.

**Mass of target residues.** The three physicochemical features of target residues, i.e. mass_target, polarizability_target and isopoint_target, were found to be the top features with high discrimination power. In Table III, the mass of target residues (mass_target) in (4) appears to be the most important feature. The loss of prediction accuracy on the OOB samples is 40.3% when values of mass_target are randomly permuted. Figure 2A shows the box plots of mass_target between different residues in the training set. The mean value is 1383 Da in hot spots, 1013 Da in non-hot spots and 1102 Da in all residues. The difference of mass_target between hot spots and non-hot spots is statistically significant (Mann–Whitney U test, \( P \text{ value} = 4.74 \times 10^{-14} \)). Therefore, mass_target is identified as a feature with discrimination power in distinguishing hot spots from other residues. Our result suggests that the regions around the tightly packed hot spots tend to have larger mass than other regions (Bogan and Thorn, 1998). Thus, protein binding may occur at these locations with more flexibility owing to large side chains (Ahmad et al., 2004).

**Relative side-chain accessible surface area and average depth index of mirror-contact residues.** In Table III, we found that the decrease of prediction accuracy on the OOB samples are 25.3 and 24.8% when the encoded feature of the relative side-chain accessible surface area (rel_s-ch_ASA_mirror) and that of the average depth index of mirror-contact residues (ave_dpx_mirror) are permuted, respectively. This provides evidence for the contact effects of the interface residues locating at the other face. The hot spot of one face packs against the hot spot of the other face and establishes a region determinant for binding in protein complex. Protein–protein interactions are generally dominated by hydrogen bonds, salt bridges and hydrophobic contacts across the interface. Complementary protein surface patterns underlying local interactions need to be desolvated, densely packed and hence deeply buried to make contribution to the binding free energy (Moreira et al., 2007). Compared to other residues, it has been found that hot spot generally has a larger relative side-chain accessible surface area and a larger average depth index (Xia et al., 2010). Similarly, we identified that the other face of hot spot, i.e. mirror-contact residue, has a larger relative side-chain surface area and a larger average depth index. Figure 2B and C show the statistics of rel_s-ch_ASA_mirror and ave_dpx_mirror in different residues, respectively. For rel_s-ch_ASA_mirror, the means are 0.85 of hot spots, 0.71 of non-hot spots and 0.74 of all residues. The means are found as 1.24 Å of hot spots, 0.79 Å of non-hot spots and 0.90 Å of all residues for ave_dpx_mirror. Both of them have statistically significant difference in hot spots and non-hot spots (\( P \text{ value} = 1.88 \times 10^{-7} \) and \( P \text{ value} = 6.91 \times 10^{-7} \), respectively).

**Atom contact areas of target residues.** As shown in Table III, the loss of prediction accuracy is 17.9% on the OOB samples when the features of atom contact areas of target residues (atom_contact_areas_target) are randomly permuted. It implies the contribution of atom contact area for detecting hot spots in protein interfaces, which are often tight regions optimally characterized by complementary shapes. The energy of protein–protein binding is closely

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean decrease of accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass of target residues</td>
<td>40.3</td>
</tr>
<tr>
<td>Polarizability of target residues</td>
<td>31.1</td>
</tr>
<tr>
<td>Isoelectric point of target residues</td>
<td>29.3</td>
</tr>
<tr>
<td>Relative side-chain accessible surface area of target residues</td>
<td>26.0</td>
</tr>
<tr>
<td>Relative accessible surface area of target residues</td>
<td>26.0</td>
</tr>
<tr>
<td>Relative side-chain accessible surface area of mirror-contact residues</td>
<td>25.3</td>
</tr>
</tbody>
</table>

*The average decrease of classification accuracy when the values of a particular feature are randomly permuted on the out-of-bag samples.
related to the buried hydrophobic surface area, and large atom contact area is required to make a water-tight seal around a critical set of energetically favorable interactions (Bogan and Thorn, 1998). Figure 2D shows the box plots of atom_contact_areas_target in different residues. The means of hot spots, non-hot spots and all residues are 77.31, 43.72 and 51.85 Å², respectively. The feature is significantly different in hot spots and non-hot spots (P value = 9.46 × 10⁻¹¹). The result indicates that hot spots tend to contain large contact surface areas between two interacting proteins.

Case studies

Thrombin binding with thrombomodulin. Serine proteinase alpha-thrombin (PDBID:1dx5, chain N and chain B) causes blood clotting and the anticoagulation is activated when it binds to thrombomodulin (PDBID: 1dx5, chain J) (Fuentes-Prior et al., 2000). Three hot spots (ARG67:N, TYR76:N and GLU80:N) and 13 non-hot spots have experimentally been determined in thrombin. In these 16 alanine-mutated residues, our method identified two residues (ARG67:N, TYR76:N) as hot spots and the rest as non-hot spots. Two of the three hot spots and all the non-hot spots were correctly predicted (shown in Supplementary Fig. S1). In contrast, MINERVA predicted two residues (GLN38:N, ILE82:N) as hot spots and the others as non-hot spots. None of the 3 hot spots and 11 non-hot spots were correctly predicted. KFC identified two residues (PHE34:N, TYR76:N) as hot spots and the rest as non-hot spots. KFC can correctly predict 1 of the 3 hot spots and 12 of the 13 non-hot spots.

Our method and KFC correctly predicted TYR76:N as a hot spot residue which is colored by green in Fig. 3. It closely contacts the other face of the interface with a high value of atom contact areas of the target residue (atom_contact_areas_target = 97 Å²). Hot spot residue ARG67:N, colored by yellow in Fig. 3, was only identified by our method. Both the residue and its mirror-contact ILE414:J (colored by pink in Fig. 3) are deeply buried with high values of average depth indices (ave_dpx_target = 3.36 Å and ave_dpx_mirror = 3.23 Å). Although residue ARG67:N loosely contacts the other face with a low value of atom contact areas of target residue (atom_contact_areas_target = 16.8 Å²), the side chain of its mirror-contact residue ILE414:J stretches into its face of the interface and forms many deeply buried atomic contacts containing a high value of atom contact areas of mirror-contact residue (atom_contact_areas_mirror = 102.7 Å²). Our method successfully identified ARG67:N by using the RF model of integrative features from its mirror-contact residue ILE414:J.

Calmodulin binding with a peptide of smooth muscle myosin light chain kinase. Calmodulin (CaM) is a versatile Ca²⁺-binding protein that regulates the activity of numerous effector proteins in response to Ca²⁺ signals (Meador et al., 1992). Calcium-bound calmodulin (Ca²⁺-CaM) (PDBID:...
1cdl, chain A) can relieve the autoinhibition of smooth-muscle myosin light chain kinase by binding to a peptide of the kinase (PDBID: 1cdl, chain E) (Meador et al., 1992).

The defined hot spot residues are PHE92:A, TRP800:E, GLY804:E, ILE810:E, ARG812:E and LEU813:E in their interface. Furthermore, PHE12:A, PHE19:A, LYS799:E, LYS802:E, ARG808:E and GLY811:E were found to be non-hot spots. Our method correctly predicted five of the six hot spot residues, i.e. PHE92:A, TRP800:E, GLY804:E, ILE810:E and LEU813:E (shown in Supplementary Fig. S2). As comparisons, MINERVA correctly predicted four hot spots and KFC identifies three of them. In addition, our method identified four of the six non-hot spots, i.e. PHE12:A, LYS802:E, ARG808:E and GLY811:E, while MINERVA and KFC predicted four and five non-hot spots, respectively. The example provides more evidence of the advantages by integrating various features in our RF method.

**Spatial neighbor residues tend to cooperate with the target residue to form a hot spot cluster**

In the RF model, we integrated various features from the target residue and its two spatial neighbor residues, namely mirror-contact residue and intra-contact residue. The importance of features as well as their differences in hot spots and non-hot spots has been evaluated. The combined information from neighbor residues improved the prediction accuracy. Our following analysis suggests that the mirror-contact residue and intra-contact residue tend to be hot spot residues in protein interface simultaneously. They contribute the binding affinity and form a hot spot cluster as determined by the change in the free energy of binding ($\Delta AG$) upon mutation of the residues to alanine.

We calculated $\Delta AG$ of the mirror-contact residue for every residue in the training set. The mirror-contact residues whose $\Delta AG$s are known will be kept. We totally collected 179 alanine-mutated mirror-contact residues, which were divided into two groups with 37 mirror-contact residues of hot spots and 142 mirror-contact residues of non-hot spots. Figure 4 shows the distribution of $\Delta AG$ in the 179 mirror-contact residues. The mean $\Delta AG$s were 3.18 of hot spots and 1.40 of non-hot spots (Mann–Whitney U test, $P$ value $= 2.93 \times 10^{-6}$). The result displays the differences of $\Delta AG$s in hot spots and non-hot spots for these mirror-contact residues. It also indicates that most of the mirror-contact residues with large $\Delta AG$s of the hot spot residues are hot spots simultaneously. Similarly, we also found the $\Delta AG$ difference is significant in hot spots and non-hot spots for the intra-contact residues ($P$ value $= 9.95 \times 10^{-8}$). Hot spots tend to form a small cluster of residues which are tightly packed at the center of the interface between two interacting proteins (Keskin et al., 2005). The results provide new clues for the improvement of prediction performance in our RF model by combining features from spatial neighbor residues of the target residue.

**Conclusions**

In this work, we identified various features from target residues, mirror-contact residues and intra-contact residues, and developed an RF model to effectively integrate these features for predicting interaction hot spots in protein interface. The performance of our model was first evaluated by the 10-fold cross validation and further validated in an independent test set. Clearly, our method shows better overall prediction performance than other methods in the comparison study. Moreover, the hybrid features were quantitatively assessed by the RF-based importance measure and the results show that the features are effective in describing the differences and contributing to the protein-binding events. We considered the features of spatial neighbors of target residues and identified some region specificity underlying the local structures of protein binding. This provides more insights for detecting hot spots in protein interface and identifying the interaction mechanism of protein complexes. Due to the time consumption and labor intensity in experimental determination of binding free energy for alanine-mutated residues, there are still a limited number of known alanine-mutated residues.
Our prediction model offers a powerful tool for detecting candidate residues in the studies of alanine scanning mutagenesis for functional protein interaction sites. Our method is applicable to the case that the 3D structure of a protein complex is available. To extend our method to single-molecular structures or even sequences will enlarge the application domain and it is an important research topic.

**Supplementary data**

Supplementary data are available at PEDS online.

**Acknowledgements**

We appreciate the anonymous reviewers and the editor for their valuable suggestions which considerably improved our primary manuscript.

**Funding**

This work was supported by the Startup Research Fund for Introduced Talents in Tianjin University of Science and Technology (Grant No. 201000404); Shanghai Natural Science Foundation (Grant No. 11ZR1443100); sponsored by National Natural Science Foundation of China (NSFC) (Grant Nos. 31100949, 60873205, 61134013, 61072149 and 91029301) and Shanghai Pujiang Program; and also supported by the Chief Scientist Program of Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS) (Grant No. 2009CSP002), by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2009CSP002, by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2011KIP203, and by the support of SA-SIBS Scholarship Program.

**References**