Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences

Peng Chen,[1] Jinyan Li,[2] Limsoon Wong,[3] Hiroyuki Kuwahara,[1]
Jianhua Z. Huang,[4]* and Xin Gao[1,5]*

[1] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

[2] Advanced Analytics Institute, University of Technology, Sydney, New South Wales, Australia

[3] School of Computing, National University of Singapore, Singapore 117417

[4] Department of Statistics, Texas A&M University, College Station, Texas 77843-3143

[5] Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ABSTRACT

Hot spot residues of proteins are fundamental interface residues that help proteins perform their functions. Detecting hot spots by experimental methods is costly and time-consuming. Sequential and structural information has been widely used in the computational prediction of hot spots. However, structural information is not always available. In this article, we investigated the problem of identifying hot spots using only physicochemical characteristics extracted from amino acid sequences. We first extracted 132 relatively independent physicochemical features from a set of the 544 properties in AAindex1, an amino acid index database. Each feature was utilized to train a classification model with a novel encoding schema for hot spot prediction by the IBk algorithm, an extension of the *K*-nearest neighbor algorithm. The combinations of the individual classifiers were explored and the classifiers that appeared frequently in the top performing combinations were selected. The hot spot predictor was built based on an ensemble of these classifiers and to work in a voting manner. Experimental results demonstrated that our method effectively exploited the feature space and allowed flexible weights of features for different queries. On the commonly used hot spot benchmark sets, our method significantly outperformed other machine learning algorithms and state-of-the-art hot spot predictors. The program is available at http://sfb.kaust.edu.sa/pages/software.aspx.

Proteins 2013; 81:1351–1362

Key words: hot spot residue; protein-protein interaction; physicochemical characteristic; classification; feature selection.

INTRODUCTION

Proteins interact with other molecules to perform their functions. It has been found that the binding energy of proteins is not uniformly distributed over their interaction surfaces.[1] Only a small fraction of interface residues contribute a large portion of the binding energy (e.g., in a typical 1,200 to 2,000 A^2 interface, less than 5% of the interface residues contribute more than 2.0 kcal/mol energy to binding). These few interface residues are called hot spot residues or hot spots.[1,2] In the binding interface, hot spots are packed significantly more tightly than are other residues, and they are also surrounded by residues that are energetically less important.[2] Hot spots are key to understanding binding mechanisms and the stability of protein-protein interactions.[3,4]

Experimentally, a hot spot can be identified if a change in its binding free energy is larger than a certain threshold when the residue is mutated to alanine. There are several databases that include data on changes in the binding free energy from single amino acid mutations in protein-protein interactions. The Alanine Scanning

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: King Abdullah University of Science and Technology (KAUST); Grand numbers: KUS-CI-016-04; GRP-CF-2011-19-P-Gao-Huang.

*Correspondence to: Jianhua Huang; Department of Statistics, Texas A&M University, College Station, TX, 77843-3143, jianhua@stat.tamu.edu
Correspondence to: Xin Gao; Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. E-mail: xin.gao@kaust.edu.sa

Received 7 November 2012; Revised 7 February 2013; Accepted 23 February 2013
Published online 16 March 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24278

Energetics Database (ASEdb) collects data on hot spots identified by alanine scanning mutagenesis experiments.[5] The Binding Interface Database (BID) stores protein interaction information on more than 1,300 mutations of 170 interacting protein pairs.[6] The protein-protein interactions thermodynamic database (PINT) collects various thermodynamic parameters on protein-protein interactions and their mutations.[7] The SKEMPI dataset was recently released. It is a comprehensive database that has collected 3,047 binding free energy changes from 85 protein-protein complexes from the literature.[8] These databases have enabled and greatly facilitated computational prediction of hot spots.

As alanine scanning mutagenesis experiments are costly and time-consuming, there is a need for predicting hot spots *in silico*.[9] A variety of computational prediction methods of hot spots has been proposed, including energy function-based physical models,[10–13] molecular dynamics simulation-based approaches,[14–16] evolutionary conservation-based methods,[4,17,18] graph-based approaches,[19] docking-based methods,[20,21] and machine learning methods that combine features such as solvent accessibility, conservation, sequence profiles, and pairing potential.[22–31]

Over a decade ago, Kortemme and Baker conducted one of the first attempts to computationally predict hot spots.[10] Given the structures of two interacting proteins and their complexes, they applied an all-atom energy function to calculate a change in the binding free energy, $\Delta\Delta G_{bind}$. Their energy function took both backbone and side-chain energy into consideration, including the Lennard-Jones potential, the repulsive term, the hydrogen bond potential, Coulomb electrostatics and a solvent term. They determined that the residues whose changes in binding free energy were greater than 1.0 kcal/mol were hot spots and found a good agreement between the predicted and the experimentally determined binding energy. This pioneering work demonstrated that binding energy can be predicted relatively accurately if the structures of the proteins and their interactions were given. A number of energy-based approaches further improved the prediction accuracy.[10–13,27] Another type of physical approach that has been widely applied to hot spot prediction is molecular dynamics (MD) simulation, where atoms are allowed to interact for a period of time and the mobility of the atoms is calculated from the MD trajectories.[32] A good agreement has been found between hot spots and restricted residue mobility using MD simulation.[14,15] However, MD simulations are computationally expensive and long MD simulations are mathematically ill-conditioned.

The characteristics of hot spots have been well studied. Hot spots are found to be sequentially and structurally more conserved than other residues[17] and packed tightly in buried areas of the interface.[4] Conservation and solvent accessibility have thus been used as two main features to predict hot spots.[4,17,18,25,29] Other features are also found to be correlated with hot spots, including sequence profiles[22] and pairing potential.[25,29] Advances in machine learning provide the possibility to combine these features together. Ofran and Rost trained a neural network model that took the sequence environment, the sequence profile extracted from multiple sequence alignments, to predict secondary structures and solvent accessibility as input features and hot spots as the output.[22] Although their method required only protein sequences to be given, they adopted several programs to predict structural features. Their method worked significantly better when the 3D structure was given. Darnell *et al.* proposed two decision tree-based hot spot predictors, K-FADE and K-CON.[23] Both methods used structural features as inputs. The former used shape-related features, such as shape specificity and residue size, whereas the latter used contact features, such as atomic contacts, hydrogen bonds, and salt bridges. The combined version, knowledge-based FADE and contacts (KFC), demonstrated significant improvements over each individual method. Tuncbag *et al.* encoded four features, accessibility, conservation, pairing potential, and predicted binding free energy from Robetta[10,33] and used different machine learning methods,[25] including decision trees, decision tables, support vector machines (SVMs), naïve Bayes, RBFNetworks, and majority voting. Their conclusions coincided with the ones in[23] that a model with only few key features can predict hot spots to a great extent while a large number of features overfit the problem. Cho *et al.* proposed a hot spot prediction method that used decision trees for feature selection and SVMs for classification. Improving upon Robetta,[10] energetic terms, such as van der Waals potential, solvation potential, hydrogen bonds, and Coulomb electrostatics, were combined by SVMs and Gaussian processes in Ref. 27 to optimize the weights and predict the hot spots. Recently, Wang *et al.* proposed a random forest model that took hybrid features that captured the levels of compactness of the residues.[31]

Despite significant progress on hot spot prediction, most of the aforementioned methods depend on structural information. For instance, the energy-based[10–12,27] and compactness-based[31] methods all require 3D structures as inputs. However, structural information is much less frequently available compared with sequence information. Although some other methods require only protein sequences as inputs, their performance depends heavily on predicted structural features by existing programs, such as secondary structure prediction and solvent accessibility prediction methods. Therefore, the question of whether hot spots can be accurately predicted from "pure" physicochemical characteristics extracted only from a protein sequence remains unanswered, to the best of our knowledge. In this article, we address this question by exploring the possibility of using only physicochemical features of amino acids in protein sequences to predict hot

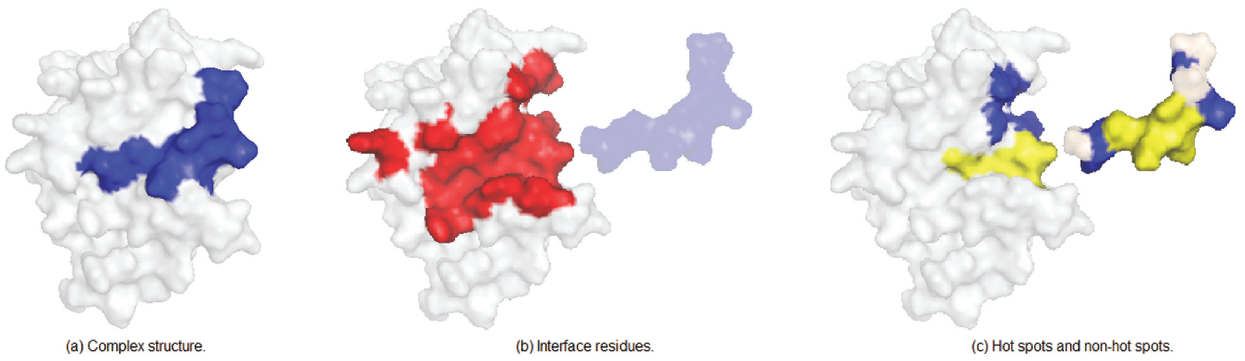


Figure 1

Illustration of interface residues, hot spots, and nonhot spots for PDB ID 1DDM, chains A and B. (a) The structure of the complex (chain A is colored in white and chain B in blue); (b) the interface residues (interfaces in chain A are colored in red and those in chain B are in blue); (c) the hot spots and nonhot spots are colored in yellow and blue, respectively. Only a fraction of interface residues are hot spots and nonhot spots.

spots. Such a method would have a broader range of applications than do structure information-based methods and it would establish simpler principles about binding mechanisms.[22]

Here, we propose a method that predicts hot spots using only physicochemical characteristics extracted from amino acid sequences. We first extract 132 relatively independent physicochemical features from the set of the 544 AAindex1 properties.[34] Each feature is then utilized to train a classification model for hot spot prediction by the IBk algorithm, an extension of the *K*-nearest neighbor algorithm. The combinations of the individual classifiers are explored and the classifiers that appear frequently in the top performing combinations are selected. The hot spot predictor is built based on an ensemble of these classifiers and to work in a voting manner. We then evaluate the performance of the proposed method on two large-scale benchmark datasets to demonstrate that our method effectively exploits the feature space and allows flexible weights of features for different queries. Meanwhile, our method also outperforms the state-of-the-art hot spot prediction methods.

MATERIALS AND METHODS

Definition

Generally speaking, a hot spot is an interface residue in which, if mutated to alanine, the change in the binding free energy is larger than a certain threshold. However, there is no common standard for this threshold in the literature. Among the state-of-the-art works on hot spot prediction, Ref. 23,31,37 used 2.0 kcal/mol as the threshold to differentiate hot spots and nonhot spots, whereas Ref. 22 chose anything above 2.5 kcal/mol to define hot spots and 0 kcal/mol (i.e., no change in binding energy) to define nonhot spots. In Ref. 25, Tuncbag *et al.* defined hot spots to be interface residues with higher than 2.0 kcal/mol and nonhot spots to be the ones with lower than 0.4 kcal/mol. As expected, different definitions for hot spots and nonhot spots change the performance of the prediction methods. Cho *et al.* used two different cutoff values for the definition in Ref. 26, that is, 1.0 and 2.0 kcal/mol. That is, interface residues with binding energies higher than these cutoffs were considered as hot spots and otherwise nonhot spots. They found that prediction on the BID dataset using the 1.0 kcal/mol definition achieved higher F1-scores than did using 2.0 kcal/mol (0.57 vs. 0.52).

In this article, we will mainly use the same definition as the one used in Ref. 25, but we will also show how different definitions influence the outcome in the Results section. Note that in using the definition in Ref. 25, the number of hot spots and nonhot spots may not add up to the total number of interface residues. Figure 1 shows an example of a complex PDB 1DDM,[35] chains A (the phosphotyrosine-binding (PTB) domain of the cell fate determinant Numb) and B (the Numb-associated kinase). Among the 34 interface residues [Fig. 1(b)], nine of them are hot spots and six of them are nonhot spots [Fig. 1(c)].

Dataset

Three state-of-the-art hot spot datasets are used in this work. Following Ref. 25, we used the ASEdb dataset[5] as the training set. Protein sequences in ASEdb were filtered so that the sequence identity between any pair of sequences was below 35%. The binding free energy of each interface residue was extracted. The interface residues with binding energies higher than 2.0 kcal/mol were considered as hot spots and those below 0.4 kcal/mol were considered as nonhot spots. This resulted in 149 residues in the training set, including 58 hot spots and 91 nonhot spots.

Similar to Ref. 25, we used the BID dataset[6] as one of our test sets. This dataset was filtered by the same sequence identity threshold and hot spots and nonhot spots were extracted using the same binding energies as in the training set. As a result, the BID test set consisted of 112 residues with 54 hot spots and 58 nonhot spots. The data in the training and test sets came from different complexes and were mutually exclusive.

In addition, we adopted a recently released, comprehensive dataset, the SKEMPI dataset,[8] as the second test set to test our method on large-scale data. The SKEMPI set contains 3,047 mutations collected from the literature. We chose the alanine mutations and defined hot spots and nonhot spots in the same way, which resulted in a test set with 196 hot spots and 380 nonhot spots. Furthermore, we used this dataset to test the performance of our method with different definitions for hot and nonhot spots. When we set the definition of hot and nonhot spots to residues with changes in binding free energy higher and lower than 2.0 kcal/mol, respectively, there were 196 hot spots and 777 nonhot spots in the set. With the cutoff value of 1.0 kcal/mol, we had 378 hot spots and 595 nonhot spots.

Extracting physicochemical features

AAindex1 is a database that stores 544 physicochemical properties of 20 types of amino acids.[34] However, some properties are highly correlated and may degrade the performance of classifiers. Therefore, we extracted a subset of relatively independent properties in which the correlation coefficient between any two properties was below 0.7. In this work, we applied a sequential filtering approach to achieve this goal. We randomly chose an initial property from the 544 AAindex1 properties, computed its correlation with all other properties and removed the properties that had greater than 0.7 correlation with this one. Then, in the remaining property set, we randomly chose another property from the remaining ones and repeated the filtering procedure using the same criterion. This process was repeated until no pair in the remaining subset had greater than 0.7 correlation, which resulted in a subset of 132 properties (Supporting Information A.4). Our approach stochastically determined the order of property selection; however, by running our approach several times, we found that the hot spot prediction performance was quite robust against this randomness. This was largely expected because, conceptually, the approach is equivalent to clustering the properties according to their pairwise correlations and selecting the representative ones.

For a given residue, a sliding window of 11 residues in the sequence centered at this residue is used to encode the features for the residue. In previous works, we proposed a novel encoding schema integrating the hydrophobic scale and the sequence profile to describe a residue.[36,37] The sequence profile for one residue was calculated using the BLAST tool and then multiplied by an AAindex1 scale, which is a vector containing values of a physicochemical property for 20 amino acid types. For instance, the sequence profile SP_i, for residue i, and the AAindex1 scale $AA1_j$, are both vectors with 1×20 dimensions. Therefore, $MSK_i^j = SP_i^* AA1_j$ for residue i, where $*$ represents the element-wise product of two vectors. We demonstrated that the standard deviation, S_i^j, of MSK_i^j was an informative, evolutionary feature.[36] Therefore, for each AAindex1 scale j, residue i is represented by a feature vector of S_{i+k}^j for the sliding window centered at i, where $k \in [-5, +5]$. In total, we identified a set of 132 feature vectors for each residue of the protein. More details about the feature encoding schema can be found in Supporting Information A.1.1.

Constructing classifiers by mining physicochemical features

Instead of identifying a single classifier from the training data, we aimed to identify an ensemble of classifiers. To search the feature space of 132 properties, we proposed the following three-step approach.

First, for each of the 132 AAindex1 descriptors, we trained a classifier. Each classifier was trained by the IBk algorithm, a variant of the K-nearest neighbor classifier.[38] IBk has lower memory complexity than the traditional nearest neighbor algorithm and can partially adapt the high noise level in the training set. Parameter K can be selected by cross-validation.

After the 132 classifiers were trained, we searched for a combination of fewer than six of these classifiers that gave good prediction results on the training data. As suggested by previous studies,[23,25] a small number of key features should be enough for hot spot prediction whereas a large number of features usually overfit the problem. We thus explored the feature combinations with few classifiers only. All combinations of no more than five classifiers were enumerated. For each combination, a residue was predicted to be a hot spot if at least half of the classifiers in the combination predicted it to be a hot spot. All the combinations were thus ranked according to their prediction accuracy.

Finally, the top ranked r combinations were considered and the most frequent p classifiers in these combinations were selected to form the final ensemble of classifiers. Since these classifiers were considered to be discriminant, a residue was predicted to be a hot spot if at least q of these classifiers predicted it to be a hot spot. The settings of the parameters r, p, and q are discussed in the Results section.

Evaluation criteria

In this work, we adopted six evaluation measures to assess the performance of the hot spot predictors. These criteria are sensitivity (Sen), specificity (Spec), accuracy (Acc), precision (Prec), F1-score (F1), and Matthews correlation coefficient (MCC). The definitions of these measures can be found in Supporting Information A.1.2.

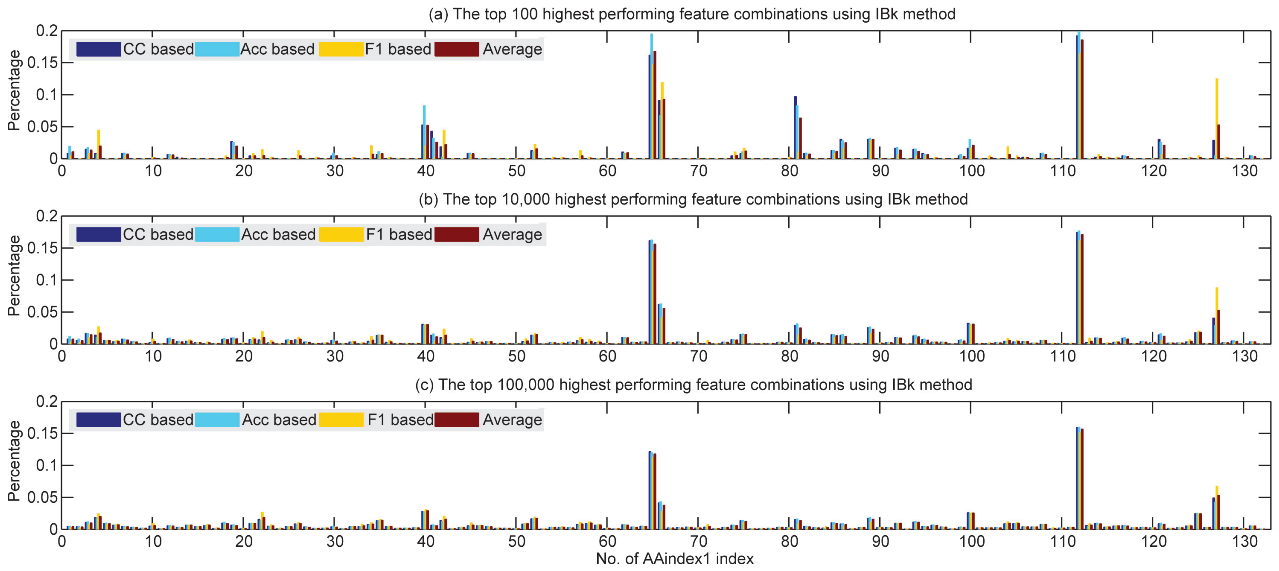


Figure 2

Frequency of feature occurrence in the top performing combinations of classifiers. The *x*-axis represents the 132 classifiers that correspond to the 132 relatively independent AAindex1 descriptors. The combinations of fewer than six classifiers are ranked according to correlation coefficient (CC), accuracy (Acc), F1-score (F1) and the average of the three, and the top (a) 100, (b) 10,000, and (c) 100,000 performing combinations are used to count the frequency.

RESULTS

Prediction of hot spot residues

Feature selection

We explored all combinations of fewer than six classifiers. Each combination was used as a majority voting-based hot spot predictor. The prediction performance was evaluated by the cross-validation evaluation of the accuracy, the correlation coefficient, or the F1 measure on the training data.[39] The top combinations were thus ranked according to different criteria. We then counted the occurrence of each feature in the different ranks. Figure 2 shows the frequency of the 132 classifiers when the top 100, 10,000, and 100,000 combinations are considered when using MCC, Acc, or F1 as the performance measure. The figure suggests that different criteria are quite consistent in terms of the frequency of the features. The 112nd, 65th, and 127th features appear at similar frequency levels when different numbers of top combinations are considered, while the 42nd, 66th, and 81st features appear frequently when the top 100 combinations are considered but less frequently when more combinations are taken into account. Table I shows the average prediction performance of the top combinations of classifiers. In this work, we extracted the most frequent features based on the F1 measure.

Table I

Average Prediction Performance of the Top Combinations of Classifiers

	Top no.	Sen	Spe	Acc	MCC	Prec	F1	Frequent features	Cutoff
Acc based	25	0.67	0.79	0.74	0.47	0.74	**0.70**	65 112 81 66 40 3 19 100 86 89	3.25%
	100	0.65	0.80	0.73	0.46	0.74	0.69	112 65 40 81 66 41 89 100 86 19	2.54%
	10,000	0.68	0.71	0.69	0.39	0.67	0.67	112 65 66 81 100 40 127 89 125 3	1.63%
	100,000	0.72	0.61	0.66	0.33	0.62	0.66	112 65 66 127 40 100 125 89 4 52	1.61%
CC based	25	0.67	0.79	0.74	0.47	0.74	**0.70**	65 112 66 81 86 41 89 3 40 94	3.2%
	100	0.67	0.78	0.73	0.46	0.74	0.70	112 65 81 66 40 41 86 89 121 127	2.82%
	10,000	0.70	0.68	0.69	0.39	0.66	0.68	112 65 66 127 100 40 81 89 125 3	1.61%
	100,000	0.73	0.60	0.66	0.33	0.61	0.66	112 65 127 66 40 100 125 4 89 52	1.62%
F1 based	25	0.85	0.60	0.71	0.45	0.65	**0.73**	66 112 127 65 42 4 75 89 3 21	1.6%
	100	0.84	0.59	0.71	0.44	0.64	0.73	112 65 127 66 4 42 89 52 34 40	2.01%
	10,000	0.80	0.55	0.67	0.37	0.61	0.69	112 65 127 66 40 100 4 42 125 22	1.9%
	100,000	0.78	0.52	0.64	0.31	0.59	0.67	112 65 127 40 66 22 100 4 125 42	1.99%

For each criterion (Acc, CC, or F1), the average performance of the top 25, 100, 10,000, and 100,000 combinations is reported. The performance is evaluated in terms of sensitivity, specificity, accuracy, MCC, precision, and F1-score in columns 3–8, respectively. The 10 most frequent classifiers in the top combinations, each of which corresponds to an AAindex1 descriptor, are reported in column 9. The minimal frequency among the 10 classifiers is reported in the last column.
The bold value of F1 for each criterion denotes the best performance among the four types of combinations.

Table II

The Prediction Performance of the Top 25 Combinations of Classifiers

Rank	F1	Classifier combination				
1	0.746	4	65	66	89	127
2	0.740	62	65	66	112	127
3	0.738	65	66	74	112	127
4	0.736	45	65	66	81	112
5	0.733	42	66	106	112	127
6	0.733	52	65	86	112	127
7	0.732	65	66	75	89	112
8	0.732	4	42	65	66	112
9	0.732	21	65	66	112	127
10	0.732	26	65	66	112	127
11	0.732	52	65	66	112	127
12	0.730	21	65	66	75	112
13	0.730	42	66	85	112	127
14	0.730	3	65	100	112	127
15	0.730	4	22	65	66	115
16	0.730	4	65	66	112	127
17	0.729	22	65	66	85	127
18	0.729	42	66	75	112	127
19	0.729	42	66	112	114	127
20	0.729	4	40	65	66	127
21	0.727	3	65	75	112	127
22	0.727	28	65	66	112	127
23	0.727	42	62	66	112	127
24	0.727	42	66	89	112	127
25	0.727	57	58	89	100	112
Average	0.732	112	66	127	65	42

In each combination, the indices of the AAindex1 descriptors for the classifiers are listed. The last row lists the average F1-score and the most frequent AAindex1 descriptors.

For the F1-score-based measure, we list the top 25 feature combinations in Table II. Each of the 25 combinations consists of five classifiers. The highest F1-score is 0.75 for the top ranked combination. More detailed analysis can be found in Supporting Information Table S1. As shown in Table II, the most frequent features are the 112nd, 66th, 127th, 65th, and 42nd AAindex1 descriptors.

Hot spot prediction by the 10 most frequent classifiers

The 10 most frequent classifiers were selected from the top combinations that had F1-scores of at least 0.6, which resulted in 249,518 combinations. We therefore set r to be 249,518 in our work. The threshold of the F1-score is comparable to those of the existing methods. These 10 classifiers are shown in the top part of Table III. It can be seen that most of the features associated with these classifiers are among the top 10 frequent features when a small number of combinations are considered, as shown in Table I. The 10 classifiers represent features related to the secondary structures of proteins, the hydrophobicity of amino acids, the distribution of amino acids in proteins and the property of surface area accessibility of amino acids to the solvent. More details about the 10 most frequent features can be found in Supporting Information A.3.

An ensemble of these 10 classifiers was built as a new hot spot predictor. Given a query residue, each of the 10 classifiers is applied to predict whether it is a hot spot or a nonhot spot. The residue is predicted to be a hot spot if at least q of the 10 classifiers predict it to be a hot spot residue. Table III shows the prediction performance of the 10 most frequent classifiers and of the ensemble of classifiers. For the individual classifiers, the most accurate one, the 112nd AAindex1 descriptor, can achieve an accuracy value of 0.66 and an F1-score of 0.64. When q is set to three, that is, a residue is predicted to be a hot spot if at least three out of the 10 classifiers predict it to be so, the ensemble can achieve an accuracy value of 0.73 and F1-score of 0.76. It is shown in Table III that the performance of the ensemble predictor is not monotonic with respect to q. This is due to the fact that the 10 classifiers were selected from the 132 AAindex1 descriptors that are relatively independent of each other, suggesting that although they still contain shared information, they contribute to the overall prediction accuracy complementarily.

Performance comparison for ensembles with different numbers of classifiers

To evaluate the parameter p, that is, the number of top classifiers that are needed in the ensemble, we built two other ensembles of classifiers for comparisons with the one discussed above. The two ensembles contain six and four of the most frequent classifiers, respectively. The performance of the two ensembles is tabulated in Supporting Information Tables S2 and S3. By comparing the two tables with Table III, we see that the best q changes for different p. For the ensemble with 10

Table III

Prediction Performance of the Top 10 Features and those of the Ensemble

	No.	Sen	Spe	Acc	MCC	Prec	F1
Individual	112	0.65	0.67	0.66	0.32	0.63	**0.64**
	65	0.54	0.72	0.63	0.26	0.62	0.58
	127	0.60	0.57	0.58	0.16	0.54	0.57
	66	0.56	0.67	0.62	0.23	0.59	0.57
	81	0.31	0.87	0.61	0.21	0.67	0.42
	40	0.48	0.67	0.58	0.15	0.56	0.52
	4	0.08	0.98	0.56	0.15	0.80	0.14
	52	0.17	0.95	0.59	0.20	0.75	0.28
	89	0.25	0.85	0.57	0.13	0.59	0.35
	86	0.04	1.00	0.55	0.14	1.00	0.07
Ensemble	1	0.98	0.03	0.47	0.04	0.47	0.63
	2	0.94	0.30	0.60	0.31	0.54	0.69
	3	0.92	0.57	0.73	0.52	0.65	**0.76**
	4	0.60	0.82	0.71	0.43	0.74	0.66
	5	0.29	0.97	0.65	0.35	0.88	0.43
	6	0.08	1.00	0.57	0.21	1.00	0.14

The top part shows the performance of each individual classifier, which corresponds to an AAindex1 descriptor. The bottom part shows the performance of the ensemble of these 10 classifiers when a different q is set. For instance, the first row in the bottom part means that the ensemble predicts a residue to be a hot spot if at least one classifier in the ensemble predicts it to be a hot spot.
The bold numbers respectively denote the highest F1 among the individual classifiers and among the Ensemble classifiers.

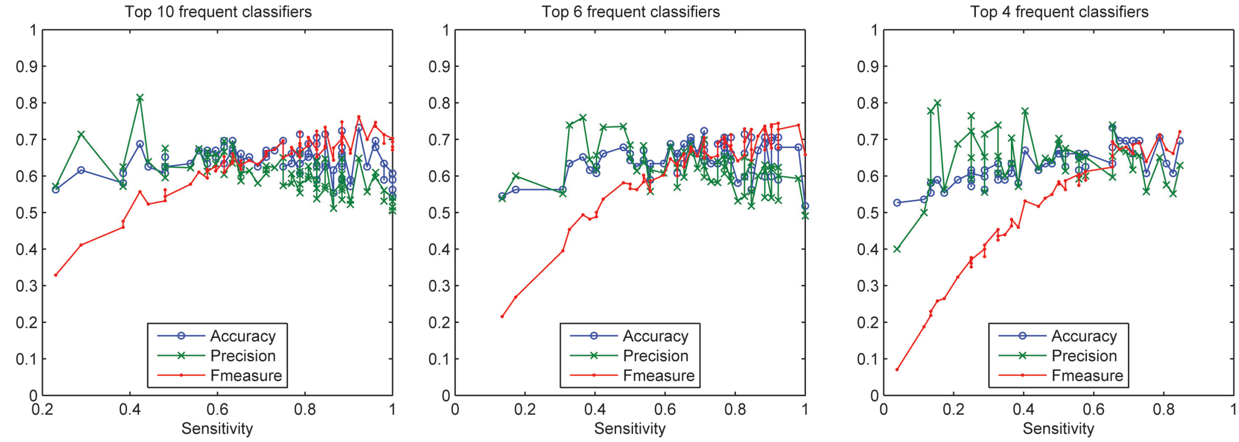


Figure 3

Performance of the prediction models with the ten, six, and four most frequent classifiers. The specificity (blue), precision (green), and F1-score (red) are shown for different sensitivity values.

classifiers, three classifiers are needed to vote for a hot spot, whereas for the ensembles with six or four classifiers, two are needed. It can be seen that the best tradeoff between recall and precision is achieved when p is set to 10 and q is set to three. This implies that the 10 most frequent classifiers are all informative. An error by one classifier could be offset by the other classifiers. Moreover, different classifiers may play important roles in determining whether a particular residue is a hot spot or not. Our approach of extracting an ensemble of classifiers as the predictor can provide flexibility to deal with this. In the traditional classification techniques, one single classifier is trained to assign weights to different features that are then used for any query. A feature with a high weight will therefore always play a dominant role for any query residue and a feature with a low weight will play only a minor role.

Figure 3 shows the changes of specificity, precision and F1-score for different sensitivity values of the best prediction model with the top ten, six and four classifiers, respectively. Since there are only 112 test instances, these curves are not smooth. It is clear that the prediction model with 10 features is much more stable with respect to precision and F1-score than are the other two models. For the same sensitivity level, the model with 10 features always has a higher F1-score than the other models have.

Performance on SKEMPI dataset with different definitions of hot and nonhot spots

To further evaluate the prediction performance of our method, we applied it to the SKEMPI dataset.[8] In this analysis, we considered four different test settings. In the first experiment (Experiment 1), we utilized ASEdb as the training dataset and SKEMPI as the validation dataset while using the same threshold values for the definition for hot spots and nonhot spots as in our previous analysis on BID dataset (i.e., 2.0 kcal/mol for hot spots and 0.4 kcal/mol for nonhot spots). The other three experiments (Experiment 2 to Experiment 4) were all based on 10-fold cross-validation on the SKEMPI dataset, each with a different definition to differentiate hot spots and nonhot spots. Experiment 2 followed the same definition of hot spots and nonhot spots as Experiment 1. In both Experiment 3 and Experiment 4, we changed the definition and considered each interface residue as a hot spot if its binding free energy change was higher than a given threshold and as a nonhot spot otherwise. The threshold value of the binding free energy change was 2.0 kcal/mol in Experiment 3, whereas it was 1.0 kcal/mol in Experiment 4.

In each of the experiments, we used the 10 most frequent features that we had extracted and built a hot spot predictor based on an assemble of the 10 most frequent classifiers. The performance results of our predictors from the four experiments are shown in Table IV. From Experiment 1 and Experiment 2, we obtained F1-scores of 0.56 and 0.63, respectively. Since the size of ASEdb is much smaller than that of SKEMPI, these results suggest that an increase in training instances can improve the prediction performance of our hot spot predictor. In Experiment 3, we obtained an F1-score of 0.40. Compared with this experiment, Experiment 4 resulted in a higher prediction performance with an F1-score of 0.61. In these two settings, we found a large change in the class distribution; whereas hot spots only accounted for 20% of the data in Experiment 3, they accounted for 39% of the data in Experiment 4. These suggest that a significant decrease in the prediction performance in Experiment 3 was contributed by a highly imbalanced dataset which was largely skewed towards nonhot spots. Among these experiments, Experiment 2 resulted in the highest F1-score, while Experiment 4 was a close second.

Table IV
Prediction Performance of the Top 10 Features on SKEMPI Data Set

			Test					
Experiments	Train/test	Threshold	Hot spots	Nonhot spots	Ratio	Rec	Prec	F1
Experiment 1	ASEdb/SKEMPI	≥2/<0.4	196	380	0.34	0.86	0.41	0.56
Experiment 2	SKEMPI/SKEMPI					0.70	0.58	0.63
Experiment 3	(cross-validation)	≥2/<2	196	777	0.20	0.31	0.55	0.40
Experiment 4		≥2/<1	378	595	0.39	0.73	0.52	0.61

The third column denotes the definitions for hot and nonhot spots. The 4 and 5 columns are for the number of hot spots to that of nonhot spots. The "Ratio" column denotes the ratio of the number of hot spots to the whole data set.

By comparing the results from these two, we found that a larger gap in the change in binding free energies between the hot spots and the nonhot spots can help increase the prediction performance. The average changes in binding energy for the hot spots class and the nonhot spots class were 3.67 and 0.15 in Experiment 2, respectively, while the same were 2.6 and 0.34 in Experiment 4, respectively. Our results suggest that a larger gap in $\Delta\Delta G$ between these two classes in Experiment 2 reduced the occurrence of false positives and false negatives, which, in turn, increased precision at the expense of recall and overall increased the F1-score.

Comparison with other methods

Here, we first compare our method that uses IBk as the classification algorithm for each feature with the same framework but using other classification algorithms. We then compare our prediction method with the state-of-the-art hot spot predictors.

To validate the use of the IBk algorithm as a classifier for each feature, we explore the same framework by replacing IBk by other machine learning algorithms. SVMs with linear kernel functions (SVM), naïve Bayes (NaïveBayes) and the one rule algorithm (OneR) with default parameters were implemented and trained on the same training set. Table V shows the performance of these different algorithms on the BID test set. Our method, referred to as IBk10, consistently outperforms the other machine learning algorithms in terms of precision. It also outperforms all the other methods in recall, except for the one rule algorithm, which has high recall at the cost of a large number of false positives. The overall performance of our method is significantly better than the other methods, as indicated by the improved F1-score.

Table V
Comparison with Other Machine Learning Algorithms. IBk10 is Our Prediction Model; it is an Ensemble of the 10 Most Frequent Classifiers

Method	Rec	Prec	F1
IBk10	0.92	0.65	0.76
SVM	0.88	0.57	0.69
OneR	0.96	0.51	0.71
NaïveBayes	0.88	0.58	0.70

Our method was then compared with other hot spot predictors on the same datasets. Table VI shows the comparison of our models, that is, IBk10, IBk6, and IBk4, with other methods, including Tuncbag's method,[25] Robetta,[10] ISIS,[22] KFCA (KFC+Robetta),[23] and KFC[23] on the BID test set. As expected, IBk10 outperforms IBk6 and IBk4, while all of them perform significantly better in terms of both recall and F1 than the other hot spot prediction methods. In particular, the IBk10 model can identify at least 22% more hot spot residues than any other hot spot predictor, while maintaining at least an 11% higher F1 measure than the others. As observed from Figure 3, the precision and F1-score are noisy with respect to sensitivity. We thus evaluated IBk10 for ranges of practically useful sensitivity values, that is, above 0.9 and above 0.6. As shown in Table VI, IBk10 still outperforms all the other hot spot predictors in terms of the F1 measure.

Case studies

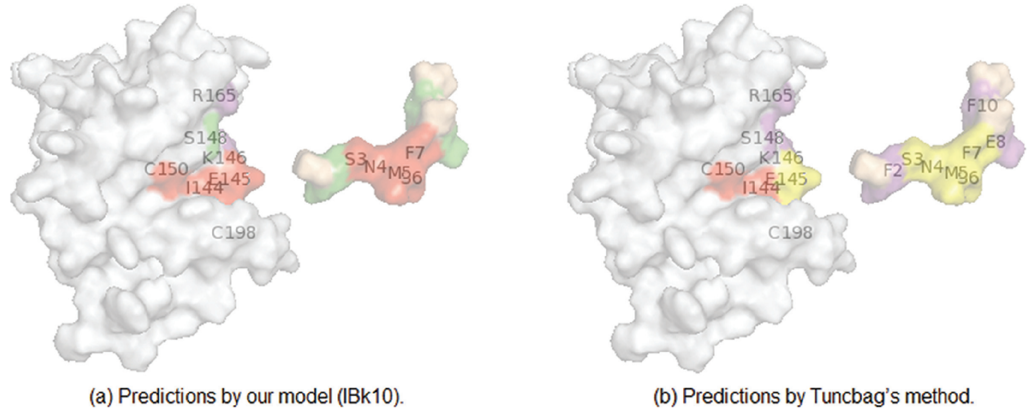
PTB domain-numb-associated kinase

We randomly chose a complex, PDB ID 1DDM, chains A and B, as a case study to demonstrate the performance of our prediction of hot spot residues compared with Tuncbag's method.[25] It is a complex of the PTB domain of the cell fate determinant Numb (chain A) with a Numb-associated kinase (Nak) peptide (chain B), which

Table VI
Comparison with Other Hot Spot Predictors

Method	Rec	Prec	F1
IBk10	0.92	0.65	0.76
IBk6	0.88	0.61	0.72
IBk4	0.79	0.65	0.71
IBk10^{R1}	[0.9, 1.0]	0.55	0.70
IBk10^{R2}	[0.6, 1.0]	0.59	0.68
Tuncbag	0.59	0.73	0.65
Robetta	0.57	0.63	0.60
ISIS	0.70	0.48	0.57
KFCA	0.48	0.53	0.51
KFC	0.36	0.51	0.42

IBk10, IBk6, and IBk4 signify our prediction models, which are ensembles of the ten, six, and four most frequent classifiers, respectively. *IBk10*^{R1} lists the average precision and F1-score for IBk10 for the sensitivity range above 0.9. *IBk10*^{R2} lists the average precision and F1-score for IBk10 for the sensitivity range above 0.6.



(a) Predictions by our model (IBk10).

(b) Predictions by Tuncbag's method.

Figure 4

Performance of our method (a) and Tuncbag's method (b) for PDB ID 1DDm, chains A and B. Chain B is drawn away from chain A for the purpose of illustration. For chain A, red residues (I144, E145, C150, and C198 for our method; I144 and C150 for Tuncbag) are the correctly predicted hot spots; the green residue (S148) is a nonhot spot that is predicted to be a hot spot; magenta residues (K146 and R165 for ours; K146, S148, and R165 for Tuncbag) are nonhot spots that are predicted correctly; and yellow residues (E145 and C198 for Tuncbag's method) are hot spots that are predicted incorrectly. For chain B, red residues (S3, N4, M5, S6, and F7) are the five hot spots that are predicted correctly; and green residues (F2, E8, and F10) are nonhot spots that are predicted to be hot spots by our method. In addition, for chain B, magenta residues (F2, E8, and F10) are nonhot spots that are predicted correctly; and yellow residues (S3, N4, M5, S6, and F7) are hot spots that are predicted incorrectly by Tuncbag's method.

forms a beta-turn at the "NMSF" (a four residue-motif extracted from the Nak) site followed by another turn near the C-terminus.[35] The complex contains nine hot spot residues (S3, N4, M5, S6, and F7 from chain B, and I144, E145, C150, and C198 from chain A) and six nonhot spot residues (F2, E8, and F10 from chain B, and K146, S148, and R165 from chain A). The Nak peptide, that is, chain B, consists of 11 residues, all of which are considered as interface residues. Figure 4(a) shows our prediction results. For chain A, all four hot spots were predicted correctly. We correctly identified all five hot spots for chain B, but wrongly predicted the three nonhot spots of chain B. One possible explanation is that residues F2, E8, and F10 are located at the ends of chain B. Our model, as a sliding window-based predictor, did not have enough information to make correct predictions about them. In contrast, Tuncbag's method only made correct predictions on two hot spots and six nonhot spots, as shown in Figure 4(b).

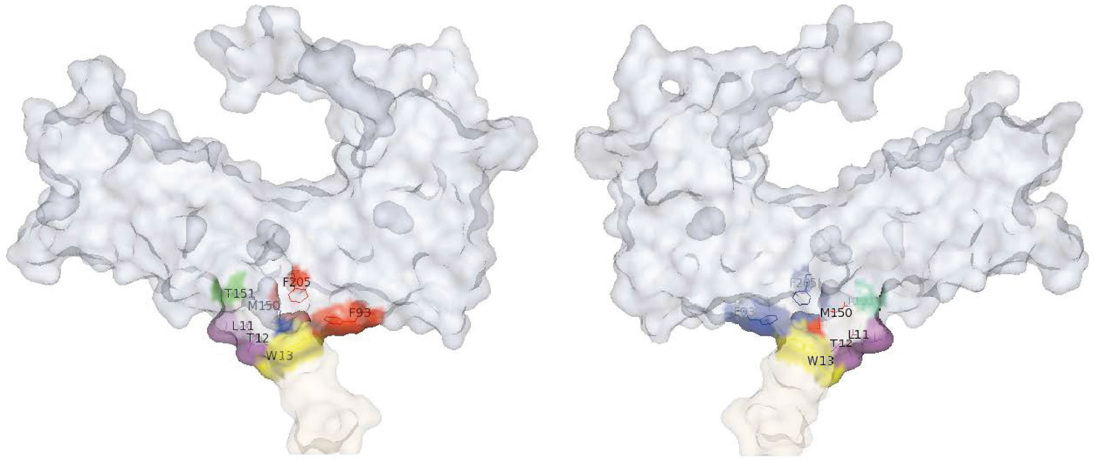
Erythropoietic receptor-erythropoietin mimetic peptide 1

We chose another complex, PDB ID 1EBP, chains A and C, as a case study to compare our method with KFC.[23] 1EBP is a complex of agonist erythropoietin (EPO) mimetic peptides with the extracellular domain of EPO receptors.[40] Experiments identified four hot spots (F93, M150, and F205 of chain A and W13 of chain C) and three nonhot spots (T151 of chain A, and L11 and T12 of chain C).[1] Our method correctly identified two hot spots and two nonhot spots. Two hot spots and one nonhot spot were wrongly predicted by our method, as shown in Figure 5(a). The prediction of the KFC method is shown in Figure 5(b). It correctly identified only one hot spot (M150).

DISCUSSION

Among the 10 most frequent AAindex1 descriptors, four of them, the 40th, 52nd, 89th, and 127th, are secondary structure-related features. Ofran and Rost found that both hot spots and nonhot spots prefer flexible structural regions of proteins, that is, loops,[22] which makes sense because the flexibility of the loop regions makes it possible to adapt conformational changes during interactions. However, hot spots and nonhot spots have different preferences for loops. For instance, in our training set (i.e., the ASEdb set), the abundance of loops for hot spots was 57%, whereas that for nonhot spots was 50%. These different preferences imply that secondary structure information may be useful in differentiating hot spots from nonhot spots. Previous research[22] reported a loop abundance of 57% for hot spots and that of 43% for nonhot spots on the same ASEdb set. The main reason for this different abundance of loops is the different definitions of hot spots and nonhot spots in these two works. In Ref. 22, the definition of hot spots is when the change in the binding energy is above 2.5 kcal/mol and that for nonhot spots is when there is no change in the binding energy. In this work, the definition for hot spots is when the change in the binding energy is above 2.0 kcal/mol and that for nonhot spots is when the change is below 0.4 kcal/mol.

Three of the 10 most frequent descriptors, numbers 65, 66, and 81, are features related to the hydrophobicity of amino acids, which is consistent with previous studies.



(a) Predictions by our method (IBk10).

(b) Predictions by KFC2's method.

Figure 5

Performance of our method (a) and KFC method (b) for PDB ID 1EBP, chains A and C. For chain A, red residues are the hot spots that are predicted correctly; green residues are nonhot spots that are predicted to be hot spots; and blue residues are nonhot spots that are predicted correctly. For chain C, red residues are hot spots that are predicted correctly; yellow residues are nonhot spots that are predicted to be hot spots; and magenta residues are nonhot spots that are predicted correctly.

It has been found that most of the residues at interfaces are hydrophobic, and Trp, Tyr, and Arg are particularly likely to be hot spots.[2,41] Such residues are largely surrounded by hydrophobic rings, probably to occlude bulk solvents.[2,42] This so-called O-ring theory argues that water exclusion may lead to a better interacting environment of the energetically hot residues rather than directly providing thermodynamic stability.[2,17,42,43] A number of previous works used the hydrophobicity of amino acids to study protein-binding sites and hot spots. A receptor-binding domain was predicted by analyzing the hydrophobicity distribution on protein sequences.[44] The web server PPI-PRED applied SVM methods to predict protein-protein interactions based on different features, including surface shape, solvent-accessible surface area (ASA), conservation, electrostatic potential, hydrophobicity, and interface residue propensity.[45] The SHARP2 server combined solvation potential and hydrophobicity calculations with other geometric descriptors and propensity scores for predicting protein-binding sites.[46]

Two other descriptors among the most frequent 10, the 4th and 112nd AAindex1 descriptors, encode the distribution of amino acids in proteins. Previous studies showed that residue composition is one of the fundamental physicochemical properties for protein folding[47] and protein interactions.[48,49] Amino acid composition has been broadly used in comparative analysis on different segments within protein complex structures, including both intraprotein comparison[49–51] and interprotein comparison.[22,52] Ofran and Rost studied the preferences of residue contacts for six types of protein interfaces, that is, intradomain, domain-domain, homo-obligomers, homocomplexes, hetero-obligomers, and heterocomplexes.[52] They found that the preferences differed remarkably between the six types of interfaces. Here, we demonstrated that such information is also useful in hot spot prediction.

The 86th descriptor, the nonbonded energy composition, is closely related to the surface area accessibility property of amino acids to the solvent. Previous studies reported that there was a certain relationship between binding energy changes and decreases in the ASA of individual residues as a consequence of complexation.[20] Such features have also been used to predict hot spot residues. Guney *et al.* combined solvent accessibility with conservation in an empirical function to identify hot spots computationally.[24] Tuncbag *et al.* proposed a method to predict hot spots based on solvent accessibility and statistical pairwise residue potentials of the interface residues.[25] interaction sites identified from sequence (ISIS) is an approach to identify interface residues from sequences without knowledge of interaction partner[52] and is further applied to the identification of hot spot residues,[22] with features of evolutionary profile, predicted secondary structure, and accessibility to the solvent.

The features selected by our method are thus the ones that have been demonstrated to be useful by different studies, which demonstrates the effectiveness of our feature selection approach. Furthermore, our classification approach, which builds an ensemble of classifiers and asks them to vote, has demonstrated better performance than the state-of-the-art hot spot predictors. Our method allows the features to play different roles for different query residues, whereas the traditional methods assign a fixed weight to each feature.

CONCLUSIONS

In this article, we proposed a method to explore physicochemical characteristics to differentiate hot spots from nonhot spots. We identified 10 informative and complementary classifiers, each of which corresponds to a classification model learned by the IBk algorithm based on the sliding-window encoding of a physicochemical feature. An ensemble of the 10 classifiers provides accurate prediction of hot spots. Our method coincides with previous studies in the sense that the 10 features are found to be important for hot spot or interface residue prediction. Meanwhile, it significantly improves the state-of-the-art hot spot prediction methods by allowing the features to be flexible based on the different query residues. Such a method can be potentially used to mine features for other classification problems in bioinformatics. Our program is available at http://sfb.kaust.edu.sa/pages/software.aspx.

REFERENCES

1. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
2. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
3. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci USA 2002;99:14116–14121.
4. Keskin O, Ma B, Nussinov R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 2005;345:1281–1294.
5. Thorn KS, Bogan AA. Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17:284–285.
6. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 2003;19:1453–1454.
7. Kumar MDS, Gromiha MM. PINT: protein-protein interactions thermodynamic database. Nucleic Acids Res 2006;34:D195–D198.
8. Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 2012;28:2600–2607.
9. DeLano WL. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 2002;12:14–20.
10. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci 2002;99:14116–14121.
11. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
12. Gao Y, Wang R, Lai L. Structure-based method for analyzing protein-protein interfaces. J Mol Model 2004;10:44–54.
13. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The foldx web server: an online force field. Nucleic Acids Res 2005; 33:W382–W388.
14. Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem 2002;23:15–27.
15. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proc Natl Acad Sci USA 2004;101:11287–11292.
16. Gonzlez-Ruiz D, Gohlke H. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. Curr Med Chem 2006;13:2607–2625.
17. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003; 100:5772–5777.
18. del Sol A, O'Meara P. Small-world network approach to identify key residues in protein-protein interaction. Proteins 2005;58:672–682.
19. Brinda KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 2002;15:265–277.
20. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 2005;102:15447–15452.
21. Grosdidier S, Fernandez-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinform 2008;9:447.
22. Ofran Y, Rost B. Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol 2007;3:e119.
23. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins 2007;68:813–823.
24. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res 2008;36:D662–D666.
25. Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics 2009;25:1513–1520.
26. Cho Ki, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res 2009;37:2672–2687.
27. Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinform 2009;10:365.
28. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinform 2010;11:174.
29. Tuncbag N, Keskin O, Gursoy A. HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res 2010;38:W402–W406.
30. Lise S, Buchan D, Pontil M, Jones DT. Predictions of hot spot residues at protein-protein interfaces using support vector machines. PLoS One 2011;6:e16774.
31. Wang L, Liu ZP, Zhang XS, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Eng Des Sel 2012;25:119–126.
32. Alder BJ, Wainwright TE. Studies in molecular dynamics. I. general method. J Chem Phys 1959;31:459–466.
33. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. Sci STKE 2004:219:pl2.
34. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–D205.
35. Zwahlen C, Li SC, Kay LE, Pawson T, Forman-Kay JD. Multiple modes of peptide recognition by the PTB domain of the cell fate determinant numb. EMBO J 2000;19:1505–1515.

36. Chen P, Li J. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. BMC Bioinform 2010;11:402.
37. Chen P, Wong L, Li J. Detection of outlier residues for improving interface prediction in protein heterocomplexes. IEEE/ACM Trans Comput Biol Bioinform 2012;9:1155–1165.
38. Aha DW, Kibler DF, Albert MK. Instance-based learning algorithms. Mach Learn 1991;6:37–66.
39. Chen YW, Lin CJ. Combining SVMS with various feature selection strategies. In: Guyon I, Nikravesh M, Gunn S, Zadeh L, editors. Feature extraction, Vol. 207 of Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer; 2006, pp 315–324.
40. Livnah O, Stura EA, Johnson DL, Middleton SA, Mulcahy LS, Wrighton NC, Dower WJ, Jolliffe LK, Wilson IA. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 A. Science 1996;273:464–471.
41. Tsai CJ, Nussinov R. Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. Protein Sci 1997;6:1426–1437.
42. Liu Q, Li J. Protein binding hot spots and the residue-residue pairing preference: a water exclusion perspective. BMC Bioinform 2010;11:244.
43. Moreira IS, Fernandes PA, Ramos MJ. Hot spots–a review of the protein-protein interface determinant amino-acid residues. Proteins 2007;68:803–812.
44. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000;302:917–926.
45. Bradford JR, Westhead DR. Improved prediction of proteincprotein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–1494.
46. Murakami Y, Jones S. SHARP2: protein-protein interaction predictions using patch analysis. Bioinformatics 2006;22:1794–1795.
47. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
48. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89–102.
49. Saha RP, Bahadur RP, Chakrabarti P. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. J Proteome Res 2005;4:1600–1609.
50. Gao X, Bu D, Xu J, Li M. Improving consensus contact prediction via server correlation reduction. BMC Struct Biol 2009;9:28.
51. Gao X. Towards automating protein structure determination from NMR data. PhD dissertation, University of Waterloo,2009.
52. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.