Current Bioinformatics, 2013, *8*, 177-182

Current Status of Machine Learning-Based Methods for Identifying Protein-Protein Interaction Sites

Bing Wang[*,1], Wenlong Sun[1], Jun Zhang[2] and Peng Chen[3]

[1]*School of Electrical Engineering and Information, Anhui University of Technology, Maanshan, Anhui 243002, China*

[2]*School of Electronic Engineering & Automation, Anhui University, Hefei, Anhui 230601, China*

[3]*Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui 230031, China*

Abstract: High-throughput experimental technologies continue to alter the study of current system biology. Investigators are understandably eager to harness the power of these new technologies. Protein-protein interactions on these platforms, however, present numerous production and bioinformatics challenges. Some issues like feature extraction, feature representation, prediction algorithm and results analysis have become increasingly problematic in the prediction of protein-protein interaction sites. The development of powerful, efficient prediction methods for inferring protein interface residues based on protein primary sequence or/and 3D structure is critical for the research community to accelerate research and publications. Currently, machine learning-based approaches are drawing the most attention in predicting protein interaction sites. This review aims to describe the state of the whole pipeline when machine learning strategies are applied to infer protein interaction sites.

Keywords: Bioinformatics, machine learning, protein feature, protein interaction site, system biology, whole pipeline.

1. INTRODUCTION

Protein-protein interactions play a pivotal role in live biological cells by controlling the functions that proteins perform, which occur through the formation of complexes, either transient or more long lasting, as a result of a balance between different molecular properties: sequence, shape, charge distribution, entropy and dynamics [1, 2]. Understanding the characteristics of interfacial sites between two interaction proteins is a necessary step to decipher the molecular recognition process and to elucidate protein function and the structure of protein complexes. Only over the past few years, do a vast amount of protein data and the associated data, benefited from rapid development of high-throughput biotechnology, making it possible to investigate the interactions between proteins. However, the residues involved in these interactions are generally not known and the vast majority of the interactions remain to be characterized structurally.

In recent years, some large-scale experimental methods were well-established to analyze protein-protein interactions in a structural view, including mainly the X-ray crystallography, NMR, and site-directed mutagenesis [3]. But, such techniques are tedious, time-consuming and labor-intensive [4], and suffer from high rates of both false positive and false negative predictions [5, 6]. Mrowke *et al.* even estimated that there are 90% protein interactions obtained by Ito and Uetz [7, 8] are not correct [9]. On other hand, current proteomics research generated tremendous protein interaction data which need to be confirmed and annotated by structural information. Therefore, it is becoming more and more important for researchers to seek some good computational approaches, which are much faster and less expensive than most experimental analyses, to predict protein interaction sites.

As a result, many computational techniques have been suggested for predicting potential protein interaction sites. The first try, predicting the individual residues which overlap with interface, was presented by Jones and Thornton in 1997 [10]. After that, many works were reported for inferring protein binding sites in which the investigators attempted to address the problem of predicting protein-protein interactions based on different biological background knowledge, such as detecting the presence of 'proline brackets' [11], solvent-accessible surface area buried upon association [12], free energy changes upon alanine-scanning mutations [13], sequence hydrophobicity distribution [14], evolutionary relationships [15, 16], three-dimensional features [17], sequence properties [18], electrostatic desolvation profiles [19], and so on. These works had achieved many sound scientific results, and benefited the study of protein interaction a lot.

Among the computational technologies which were applied into the protein interaction sites prediction, machine learning-based methods are becoming the most efficient ways [15, 20-36]. Several machine learning approaches, such as Bayesian network [22, 30], neural networks [24, 27, 31-34, 36, 37], support vector machines [20, 21, 23, 25, 26, 28, 35], and conditional random field [29], have been proposed to address the problem of protein interaction site prediction. Typically, these methods classify each target residue into either the interface or non-interface residues group based on a sliding window strategy by which some sequence or spatial neighborhood residues can be involved in classifiers as input. Clearly, the performance of prediction is heavily

*Address correspondence to this author at the School of Electrical Engineering and Information, Anhui University of Technology, Maanshan, Anhui 243002, China; Tel: 1-(615)322-6552; E-mail: wangbing@ustc.edu

2212-392X/13 $58.00+.00

dependent on the feature selection or extraction. If the features are distinguishing between protein non-interface and interface, the residues present in protein interfaces should be easier to identify. Because none of the individual property is sufficient to describe protein interface, a combination of some of them may be a good way to improve the prediction results. The second factor is the design of the classifier, for none of machine learning method can work very well in all situations, and the performance of each of them will be varied on different data sets and different algorithm architectures.

Recently, three comprehensive reviews have been published for providing insight into the prediction of protein-protein interaction sites [38-40]. While these works reviewed the fundamentals of protein binding and docking, and gave a whole picture of the prediction of interface residue, our review mainly focuses on machine learning-based methods and their application in the prediction of protein interaction sites. In this review, some important components in the entire pipeline will be described when machine learning approaches are adopted to address the problem of inferring interaction sites, from protein feature extraction, feature representation, algorithm design to performance evaluation.

2. DEFINITIONS

2.1. Protein Complexes

Protein interactions are the basis of the formation of protein complexes, which are groups of two or more associated polypeptide chains. According to the interaction strength, protein complexes can be divided into obligatory and transient complexes. Obligatory interactions occur between the proteins which have high shape complementarity, and are characterized by the presence of hydrophobic residues [41]. Protein docking, therefore, is a good alternative of the prediction of obligatory interaction which model two or more known structures mainly based on protein surface complementarity and electrostatics [42]. Transient complexes are weakly associated between protein interfaces which have lower geometrical complementarily. Generally, the interface areas between proteins in transient complexes are relative small, and tend to be formed by polar residues.

Protein complexes can also be divided into homo-complexes and hetero-complexes based on sequence identity. Interaction sites in homo-complexes are characterized by hydrophobic, and can be distinguished by a relatively large interface of non-polar residues [43]. Many studies, therefore, focus on inferring the interaction sites in hetero-complexes, because some of them are transient complexes which are more difficult to be identified by experimental methods [15, 16, 18, 24, 29, 34, 35, 37].

The previous studies showed that there are different amino acid compositions among homo-obligomer, hetero-obligomer, homo-transient and hetero-transient complexes, which indicated that different types of protein interface can be differentiated based on the suitable features [41]. When machine learning methods are adopted for the prediction of protein interaction sites, the interface type is important and should be taken into account because the same features have different capabilities in distinguishing the interface from non-interface residue in different interaction types. Homo- and hetero- complexes can be easily differentiated based on sequence identity. However, sometimes some interactions can not be clearly assigned into a definite type of transient and permanent complex, which causes the prediction task to be more difficult.

2.2. Surface and Interface Residues

Proteins interact with each other through interfaces which are composed of some surface residues. Therefore, the prediction qualities of all machine learning-based predictors are to a large extent dependent on how the surface residues are defined. Usually, a residue is considered to be a surface residue if its relative accessible surface area (RSA) is higher than a ratio cut-off of its nominal maximum area, whose value was defined by Rost and Sander [44]. The accessible surface area (ASA) can be calculated for each residue in each protein chain using the DSSP program [45]. Different studies set up different cutoff values, which may be 5%, 10%, 16% or 25% [15, 18, 24, 34, 46].

There is no consistent definition for interface residues in the protein interaction community. In some studies, the interface residues are defined if calculated ASA in the complex (CASA) of this residue is less than that in the monomer (MASA) by a threshold, i.e., 1 Å2 [18]. In this case, the corresponding unbound monomers should be provided in the database because the conformation will be changed when this protein chain forms a complex with the other protein chain(s). However, the available set of bound and unbound structure of the same protein chain is very small, which limit the application of this definition of the interface residue. Therefore, many works adopted a relatively simple strategy by which a residues can be seen as the interface residue if the spatial distance between its α-carbon (CA) atom and random CA atoms in the other protein chains in the complex is less than 1.2 nm [15, 27]. Also, there is a similar definition of interface residue if any of the heavy atoms within a cut-off distance, such as 6 Å, or any atoms in its interaction partner chain [47].

For the definitions of surface and interface residues, the selection of different cut-offs is very important because they decide the datasets on which machine learning-based predictor will be trained. The main advantage of machine learning-based methods in the prediction of interaction sites is to learn some interaction-related information from known data, and then apply the learned information as experience to analyze the unknown data. Apparently, more strictly the cut-offs are selected, more confident the dataset will be, but it will also increase the false negative rate of the surface and interface residues.

3. PROTEIN FEATURES

3.1. Feature Extraction

The successes of machine learning application in prediction of protein interaction sites depend on the protein features which can differentiate interface from other surface residues. Protein interaction surfaces are composed of interface residues and nearby residues. Each interface has its

own spatial structure, amino acid composition, local chemical and physical environment. Therefore, as a result of a combinational influence of different factors, protein interfaces can be represented by different protein features. The features with distinct expression levels between interface and non-interface will be good choices for prediction of protein interaction site. Currently, the following features have been used for protein interaction sites researches.

Amino Acid Composition: As basic information, amino acid composition had been used in many researches to predict the protein interaction sites. Jones and Thornton [10, 43] firstly used this features and found that ratios of amino acids between protein interfaces and non-interfaces are different. Hydrophobic residues and arginine are present in a relatively high frequency in protein interfaces, and can be used to predict interface residues in some studies, while other studies show that the amino acid composition is not enough to identify interfaces [30, 36, 48, 49].

Sequence Entropy: Protein sequence entropy is a conservation score which can estimate sequence variability [50, 51]. This score ranks the frequencies of the occurrence of 20 type's amino acid in protein families, usually normalized over the range of 0-100 by which the lowest values are corresponding to the most conserved positions. Some studies group 20 amino acids into different subgroups based on their physical and chemical properties by considering different mutation tendencies from one amino acid to another [15]. The HSSP database is a good resource to extract sequence entropy [52].

Solvent Accessible Area: Interface residues are likely to be accessible to solvent in the unbound state. Non-interface residues tend to maximize intra-molecular interaction, and therefore reduce their solvent accessibilities. The difference between predicted and actual solvent accessibilities are already used as a discriminative feature in some predictors [24, 34, 53].

Secondary Structure: The most common proteins secondary structures are alpha helices and beta sheets, where the latter one is the favorite of interfaces. Ofran and Rost found that the prediction result of secondary structure could improve the performance of protein interaction site predictors [32].

Evolutionary Conservation: Many methods for protein interaction sites prediction use evolutionary conservation as a primary indicator of the location of interface residues because it reflects evolutionary selection at the interaction sites to maintain the functions of protein families [30, 36, 54-57]. Residue conservation at the interfaces is observed to be higher than those of general surface residues, and therefore has discriminatory power for protein interaction sites [58, 59].

There are also some other protein features that have been used in current studies, such as the position-specific scoring matrices (PSSMs) from multiple sequence alignment (MSA) [50], 3D-motifs [60, 61], interface propensities [26, 62], and so on. Because none of the single feature carries sufficient information of protein interactions, and the relation among those features is not clear for investigators yet, a combination of some of them is found to be an effective way to improve the prediction performance in machine learning methods-based predictors [15, 63].

3.2. Feature Representation

Currently, the methods of the predicting protein interaction sites can be divided into to subgroups: patch-based or residue-based predictors [39]. Patch-based methods define and analyze a series of residue patches on the surface of protein structures to predict interaction patches using a combined score based on some characters [10, 22, 43]. Machine learning-based methods mainly adopt residue-based strategy by which a consistent feature vector can be easily represented. A sliding window technique is mostly used in order to involve the association among neighboring residues. The length of the sliding window is an important parameter which can affect the prediction results. Most studies use a random setup of window length. Recently, Sikic *et al.* [2] proposed an entropy-based method to determinate the window length as follows:

$$-\sum_{i=1}^{L} p_i \times \log_2 p_i - \log_2 L \quad (1)$$

where L is the length of a window, and p_i is the frequency appearance of ith interacting residues in a window of L residues. The maximum difference can be obtained when the length of the window is set up to 9 in their work. Another study revealed that the contributions of the residues inside the window are different, and a corresponding coefficient for each residue has been assigned based on a assumption of normal distribution：

$$r_i = e^{-0.5(i-(L+1)/2)^2/\sigma^2} \quad (2)$$

where standard deviation σ^2 is a parameter of normal distribution which can also be calculated by i and L [18]. These two methods will be very helpful to decide a reasonable length of the sliding window and the contribution of each residue inside window.

4. PREDICTION METHODS

Machine learning methods have been applied in predicting the protein-protein interaction sites for decades. Based on the protein feature mentioned above, machine learning approaches, such as Bayesian network, artificial neural networks, and support vector machines, treat the protein interaction sites prediction as a classification task, i.e., interface and non-interface residues. These methods have been reviewed by Zhou and Qin [38]. We will mainly focus on some novelty methods in this part.

Chen and Jeong [64] divided some selected features into three groups based on their sources, i.e., physicochemical features and evolutionary conservation score, amino acid distance, and PSSMs. Then, an integrative random forest model had been constructed on each group of feature, and the final classification results will be obtained by a majority vote strategy. This method can avoid the imbalanced data problem, which is common in prediction of interaction sites because non-interface residues are much more than interface residues.

Another work [50] also handled the imbalanced problem by bootstrap resampling technique, and used SVM-based fusion classifiers to increase the accuracy of prediction perfor-mance of protein interaction sites. They presented one compo-nent ensemble classifier based on the SVM method for each of the eight different feature spaces firstly, and then combined them with a weighted voting to make the final decision.

Chen and Li [18] applied a SVM ensemble strategy to identify protein interface residues based on an integrative profile by combining hydrophobic and evolutionary information. In their work, a novelty method of residue sequence profile construction had been developed, in which the contribution of every residue in sequence window is different. They also used self-organizing map (SOM) technique to investigate the interacting relationship of the residues, and achieved a relatively high prediction performance, i.e., a precision of 81.96% and a specificity of 96.35%.

There are also some other prediction approaches to improve the performance, such as using genetic algorithm [65], ensemble predictor of radial basis function neural network [37], and so on. It can be found that a combination of some protein features or/and predictor ensemble is an effective way to improve the accuracy of prediction of protein interaction sites. The reason is that the machine learning-based predictors always have a functional tendency in some extent to the features by which the predictors were trained. Predictor's ensemble can take into account the complementarities of different protein features.

5. PERFORMANCE EVALUATION

As discussed in many literatures, prediction accuracy, which is the ratio of the number of correctly predicted protein interface residues to the total number of predicted interface residues, is not enough to evaluate prediction performance due to the imbalance of interface and non-interface data sets. For example, there are over 70,000 protein structures collected in PDB [66], however only 23,759 entries can be found interaction information in the current DIP database [67]. That means most protein interactions cannot be available now. Also, there is a big imbalance between interface and non-interface residues because only a small part of the surface is seen as interface in almost every dataset used in the protein interface prediction. Usually, there are other three measures have been used:

$$\begin{aligned} \mathrm{Re}\,call &= \frac{TP}{TP+FN} \\ \mathrm{Pr}\,ecision &= \frac{TP}{TP+FP} \\ MCC &= \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \end{aligned} \tag{3}$$

where TP (True Positive) is the number of true positives, i.e., residues predicted to be interface residues that actually are interface residues; FP (False Positive) is the number of false positives, i.e., residues predicted to be interface residues that are in fact not interface residues; TN (True Negative) is the number of true non-interface residues; and FN (False Negative) is the number of false non-interface residues. The MCC is a measure of how well the predicted class labels correlate with the actual class labels. The range of MCC value is from -1 to 1, where a correlation coefficient of 1 corresponds to perfect predictions, while a correlation coefficient of 0 corresponds to random guessing.

6. CONCLUSIONS

The rise of powerful high-throughput experimental technologies has fundamentally changed the study of current system biology, which already benefited the study of protein interactions a lot. However, it also presents some serious challenges for prediction of protein interaction sites, such as protein feature extraction and representation, prediction methods, results analysis and performance evaluation. Machine learning, as a branch of artificial intelligence, had been adopted for addressing these problems and demonstrated good performance in prediction. It should be noted that there is still room to improve the prediction performance of all the studies, such as more interaction-related protein feature mining, integration of other biological resource like genome and transcriptome information, and ensemble different classifiers. Some new computational techniques are also helpful if they can be employed into the protein interaction field [68], It is clear that machine learning algorithms hold incredible promise for protein interaction research; their capabilities in the hands of investigators will undoubtedly accelerate our understanding of the mechanism of cell to perform their functions.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (Nos. 61272269).

REFERENCES

[1] Alberts B, Molecular biology of the cell. 4th ed. 2002, New York: Garland Science.

[2] Sikic M, Tomic S, Vlahovicek K. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol 2009; 5: e1000278.

[3] Leon O, Roth M. Zinc fingers: DNA binding and protein-protein interactions. Biol Res 2000; 33: 21-30.

[4] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999; 402: 86-90.

[5] Legrain P, Selig L. Genome-wide protein interaction maps using two-hybrid systems. FEBS Lett 2000; 480: 32-36.

[6] Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 2002; 18: 529-536.

[7] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 2001; 98: 4569-4574.

[8] Uetz P, Giot L, Cagney G, *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000; 403: 623-627.

[9] Mrowka R, Patzak A, Herzel H. Is there a bias in proteome research? Genome Res 2001; 11: 1971-1973.

[10] Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997; 272: 133-143.
[11] Kini RM, Evans HJ. Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. FEBS Lett 1996; 385: 81-86.
[12] Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol 1997; 4: 973-974.
[13] 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.
[14] Pazos F, Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins-Struct Funct Bioinform 2002; 47: 219-227.
[15] Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett 2006; 580: 380-384.
[16] Li JJ, Huang DS, Wang B, Chen P. Identifying protein-protein interfacial residues in heterocomplexes using residue conservation scores. Int J Biol Macromol 2006; 38: 241-247.
[17] Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004; 336: 943-955.
[18] Chen P, Li J. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. BMC Bioinformatics 2010; 11: 402.
[19] Fiorucci S, Zacharias M. Prediction of protein-protein interaction sites using electrostatic desolvation profiles. Biophys J 2010; 98: 1921-1930.
[20] Res I, Mihalek I, Lichtarge O. An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics 2005; 21: 2496-2501.
[21] Bordner AJ, Abagyan R. Statistical analysis and prediction of protein-protein interfaces. Proteins 2005; 60: 353-366.
[22] Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR. Insights into protein-protein interfaces using a Bayesian network prediction method. J Mol Biol 2006; 362: 365-386.
[23] Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 2005; 21: 1487-1494.
[24] Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins 2005; 61: 21-35.
[25] Chung JL, Wang W, Bourne PE. Exploiting sequence and structure homologs to identify protein-protein binding sites. Proteins 2006; 62: 630-640.
[26] Dong Q, Wang X, Lin L, Guan Y. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. BMC Bioinformatics 2007; 8: 147.
[27] Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein--protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002; 269: 1356-1361.
[28] Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel 2004; 17: 165-173.
[29] Li MH, Lin L, Wang XL, Liu T. Protein-protein interaction site prediction based on conditional random fields. Bioinformatics 2007; 23: 597-604.
[30] Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004; 338: 181-199.
[31] Ofran Y, Rost B. Predicted protein-protein interaction sites from local sequence information. Febs Lett 2003; 544: 236-239.
[32] Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics 2007; 23: e13-16.
[33] Pettit FK, Bare E, Tsai A, Bowie JU. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J Mol Biol 2007; 369: 863-879.
[34] Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007; 66: 630-645.
[35] Wang B, Wong HS, Huang DS. Inferring protein-protein interacting sites using residue conservation and evolutionary information. Protein Pept Lett 2006; 13: 999-1005.
[36] Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001; 44: 336-343.
[37] Wang B, Chen P, Wang P, Zhao G, Zhang X. Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes. Protein Pept Lett 2010; 17: 1111-1116.
[38] Zhou HX, Qin S. Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics 2007; 23: 2203-2209.
[39] Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. Progress and challenges in predicting protein-protein interaction sites. Brief Bioinform 2009; 10: 233-246.
[40] de Vries SJ, Bonvin AM. How proteins get in touch: interface prediction in the study of biomolecular complexes. Curr Protein Pept Sci 2008; 9: 394-406.
[41] Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003; 325: 377-387.
[42] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 1992; 89: 2195-2199.
[43] Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997; 272: 121-132.
[44] Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994; 20: 216-226.
[45] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22: 2577-2637.
[46] Qin S, Zhou HX. meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 2007; 23: 3386-3387.
[47] Ahmad S, Keskin O, Sarai A, Nussinov R. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. Nucleic Acids Res 2008; 36: 5922-5932.
[48] Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999; 285: 2177-2198.
[49] Crowley PB, Golovin A. Cation-pi interactions in protein-protein interfaces. Proteins 2005; 59: 231-239.
[50] Deng L, Guan J, Dong Q, Zhou S. Prediction of protein-protein interaction sites using an ensemble method. BMC Bioinformatics 2009; 10: 426.
[51] Sen TZ, Kloczkowski A, Jernigan RL, *et al.* Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. BMC Bioinformatics 2004; 5: 205.
[52] Schneider R, Sander C. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res 1996; 24: 201-205.
[53] Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. Protein Sci 2006; 15: 1017-1029.
[54] Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996; 257: 342-358.
[55] Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002; 12: 21-27.
[56] Madabushi S, Yao H, Marsh M, *et al.* Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002; 316: 139-154.
[57] Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 2004; 336: 1265-1282.
[58] Guharoy M, Chakrabarti P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics 2010; 11: 286.
[59] Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 2004; 13: 190-202.
[60] Henschel A, Winter C, Kim WK, Schroeder M. Using structural motif descriptors for sequence-based binding site prediction. BMC Bioinformatics 2007; 8 Suppl 4: S5.
[61] Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 2005; 347: 565-581.
[62] Liu B, Wang X, Lin L, Dong Q. Exploiting three kinds of interface propensities to identify protein binding sites. Comput Biol Chem 2009; 33: 303-311.
[63] Yan C, Dobbs D, Honavar V. A two-stage classifier for identification of protein-protein interface residues. Bioinformatics 2004; 20(Suppl 1): I371-I378.

[64] Chen XW, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 2009; 25: 585-591.

[65] Du X, Cheng J, Song J. Improved prediction of protein binding sites from sequences using genetic algorithm. Protein J 2009; 28: 273-280.

[66] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003; 10: 980.

[67] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 2004; 32: D449-451.

[68] You ZH, Lei YK, Gui J, Huang DS, Zhou X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics 26: 2744-2751.

Received: June 9, 2011 Revised: October 31, 2011 Accepted: December 5, 2011