Amino Acids (2010) 38:263–270
DOI 10.1007/s00726-009-0245-8

ORIGINAL ARTICLE

Identifying protein–protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area

Rong Liu · Wenchao Jiang · Yanhong Zhou

Received: 5 October 2008 / Accepted: 21 January 2009 / Published online: 12 February 2009

Abstract Transient protein–protein interactions play a vital role in many biological processes, such as cell regulation and signal transduction. A nonredundant dataset of 130 protein chains extracted from transient complexes was used to analyze the features of transient interfaces. It was found that besides the two well-known features, sequence profile and accessible surface area (ASA), the temperature factor (*B*-factor) can also reflect the differences between interface and the rest of protein surface. These features were utilized to construct support vector machine (SVM) classifiers to identify interaction sites. The results of threefold cross-validation on the nonredundant dataset show that when *B*-factor was used as an additional feature, the prediction performance can be improved significantly. The sensitivity, specificity and correlation coefficient were raised from 54 to 62%, 41 to 45% and 0.20 to 0.29, respectively. To further illustrate the effectiveness of our method, the classifiers were tested with an independent set of 53 nonhomologous protein chains derived from benchmark 2.0. The sensitivity, specificity and correlation coefficient of the classifier based on the three features were 63%, 45% and 0.33, respectively. It is indicated that our classifiers are robust and can be applied to complement experimental techniques in studying transient protein–protein interactions.

Keywords Protein–protein interactions · Transient interface · Sequence profile · Temperature factor · Accessible surface area · Support vector machine

Introduction

Protein–protein interactions are critical to many biological processes. The so-called interaction sites or functional sites play a crucial role in protein–protein interactions. Identifying these pivotal sites is useful to get a better understanding of molecular recognition process at the residual and atomic level, to uncover the mechanism of metabolic and signal transduction networks, and to gain important clues for rational drug design (Chelliah et al. 2004).

According to their lifetime, protein–protein interactions can be divided into permanent interactions and transient interactions (Jones and Thornton 1996). Due to the structural stability of permanent complexes, permanent interactions are much easier to study by experimental methods, such as X-ray crystallography and NMR spectroscopy. On the other hand, since transient interactions often neither form stable crystals nor give good NMR structures, transient complexes are notoriously hard to study experimentally (Szilágyi et al. 2005). Nevertheless, transient interactions are the focus of significant interest owing to their biological importance, particularly with respect to cell regulation and signal transduction (Hoskins et al. 2006). Thus, computational methods are needed to assist in finding the features of transient protein–protein interfaces and identifying residues in these interfaces.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-009-0245-8) contains supplementary material, which is available to authorized users.

R. Liu · W. Jiang · Y. Zhou (✉)
Hubei Bioinformatics and Molecular Imaging Key Laboratory,
College of Life Science and Technology,
Huazhong University of Science and Technology,
430074 Wuhan, China
e-mail: yhzhou@hust.edu.cn

R. Liu
e-mail: Liurong116@tom.com

By statistically analyzing different types of protein–protein interfaces, some common features have been attained and used to identify interaction sites. It was found that there are distinct differences in amino acid composition between interface and noninterface, as well as between different types of interfaces. Compared with noninterfaces, permanent interfaces always contain more hydrophobic residues (Glaser et al. 2001). Although some transient interfaces are also hydrophobic, they are rich in aromatic residues and depleted in charged residues (Lo Conte et al. 1999). Evolutionary conservation of residues is another important feature for the identification of interaction sites. Generally, interface residues are more conservative than noninterface residues during evolution. Transient interfaces tend to evolve at a relatively higher rate than permanent interfaces (Mintseris and Weng 2005). Previous studies have demonstrated that interface residues are more solvent accessible than noninterface residues. Solvent accessibility is one of the most effective features used to predict homodimer interfaces (Jones and Thornton 1997a, b). It has been suggested that interface residues have lower temperature factors (*B*-factors) than the exterior of protein, which contributes to less flexibility of the interfacial regions (Jones and Thornton 1995). In addition, secondary structure (Neuvirth et al. 2004; Ansari and Helms 2005) and side-chain conformational entropy (Cole and Warwicker 2002; Liang et al. 2006) can also be used to distinguish interface residues from noninterface residues. Thus, these features are valuable for identifying interaction sites.

The features mentioned above have been combined to predict interaction sites in different types of complexes, which is based on a wide range of machine learning methods (Zhou and Shan 2001; Koike and Takagi 2004; Landau et al. 2005; Li et al. 2006; Bradford et al. 2006; Friedrich et al. 2006; Li et al. 2007). However, only a few studies have chosen the interaction sites in transient complexes as prediction objects. Ofran and Rost (2003) developed a neural network that identifies transient protein–protein interfaces from local sequence information. Neuvirth et al. (2004) utilized a naive Bayesian method with 13 features to identify the interfaces of unbound structures of transient heterodimers at a patch level. Liang et al. (2006) presented a linear combination of energy score, interface propensity and residue conservation score to predict interface residues of the unbound structures used by Neuvirth et al. (2004). Dong et al. (2007) input binary profile interface propensity, sequence profile and accessible surface area (ASA) to support vector machine (SVM) for recognizing interaction sites in transient complexes. Although the existing prediction methods have achieved success at different levels, the prediction of interface residues in transient complexes is still at its primary stage.

The purpose of our research is to focus on the identification of protein–protein interaction sites in transient complexes. By analyzing the features of transient interfaces, we found that besides the two well-known features, sequence profile and ASA, *B*-factor can also reflect the differences between interface and the rest of protein surface. Then, *B*-factor, sequence profile, ASA or the combinations of them were used to construct SVM classifiers to recognize interface residues. The results show that *B*-factor plays a key role in identifying the interaction sites in transient complexes, and that utilizing the complementarity of the three features is favorable for improving the prediction performance.

Materials and methods

Dataset

The experimental data in this study were derived from the dataset used by Ansari and Helms (2005). This dataset contains 170 transient protein–protein interaction pairs, not including antigen–antibody interactions. The corresponding transient complexes were extracted from the protein data bank (PDB) (Berman et al. 2000). To further advance the quality of experimental data, the dataset was filtered strictly. The complexes having multiple models solved by NMR spectroscopy were discarded. The pairs containing chains less than 50 residues were eliminated to filter out small molecules. For the chains that interact with multiple partners, the one including the most interface residues was selected as a representative. After filtering the dataset, there were 117 transient protein–protein interaction pairs, namely 234 protein chains. Finally, 234 chains were clustered to remove redundant chains using the BLASTCLUST program (Altschul et al. 1990) with identity threshold of 30% and length coverage threshold of 90%. As a result, a nonredundant dataset composed of 130 protein chains was used in this research.

Definition of surface residues and interface residues

In this study, the method of Fariselli et al. (2002) was adopted to define surface residues and interface residues. A residue was considered as a surface residue if its ASA is at least 16% of its nominal maximum area (Rost and Sander 1994). The DSSP program (Kabsch and Sander 1983) was used to calculate the ASA of each residue in unbound chain. The atom coordinates of a single chain were derived from the corresponding PDB file. A surface residue was defined as an interface residue if the distance between its Cα atom and any residue's Cα atom from its partner chain is less than 1.2 nm. According to this

definition, our dataset contained 16,056 surface residues, about 29% of which were interface residues.

Feature extraction

B-factor

B-factor is a measure of atomic thermal motion and disorder. The *B*-factor of Cα atom was used to represent the flexibility of each residue and normalized by the following equation (Yuan et al. 2003):

$$\mathrm{NB_r} = \frac{B_\mathrm{r} - (B)}{\sigma(B)} \tag{1}$$

where B_r is the *B*-factor of residue r, (B) and $\sigma(B)$ are the mean value and the standard deviation of the *B*-factors for the chosen chain, respectively.

Sequence profile

Sequence profile was generated by three iterations of PSI-BLAST searches (Altschul et al. 1997) against NCBI nonredundant database with the BLOSUM62 substitution matrix and E-value threshold of 0.001. The profile value was scaled between 0 and 1 by the following equation (Kim and Park 2003):

$$f(x) = \begin{cases} 0.0 & \text{if } x \leq -5 \\ 0.5 + 0.1x & \text{if } -5 < x < 5 \\ 1.0 & \text{if } x \geq 5 \end{cases} \tag{2}$$

where x is the original profile value.

Accessible surface area (ASA)

ASA was calculated in the process of defining surface residues with the DSSP program and scaled between 0 and 1 by the following equation (Wang et al. 2008):

$$\mathrm{NASA_r} = \frac{\mathrm{ASA_r}}{\max(\mathrm{ASA_r})} \tag{3}$$

where $\mathrm{ASA_r}$ is the ASA of residue r, $\max(\mathrm{ASA_r})$ is the nominal maximum area of residue r.

Classifier construction

In our experiment, SVM classifiers (Vapnik 1995) were used to identify whether a surface residue was located at the interface or not. SVM classifiers were constructed using *B*-factor, sequence profile, ASA or the combinations of them. Each classifier input a window containing a target residue and its ten spatially nearest surface residues. As a result, each residue was represented by an 11-component vector if *B*-factor or ASA was used and by a 220-component vector if sequence profile was used. In this study, SVM classifiers were implemented using the LIBSVM package (Chang and Lin 2001) with the radial basis function as kernel. For each classifier, we used a grid search to determine the optimal values of C and γ so as to maximize the correlation coefficient (CC) of cross-validation.

Cross-validation

Threefold cross-validation was used to train and test the classifiers. The whole dataset was randomly divided into three subsets with an approximately equal number of chains. In each validation, one subset was used for testing while the rest were used for training. In our dataset, only 29% of surface residues were interface residues. If all noninterface residues were used for training, the classifiers would prefer to classify a target residue as a noninterface residue. Therefore, for each run, the classifiers were trained using all interface residues and an equal number of non-interface residues extracted randomly from the training set and this procedure was repeated five times. A residue was classified as an interface residue if it was predicted to be positive at least three times, otherwise a noninterface residue.

Evaluation measures of classifier performance

In this study, four widely used measures, sensitivity, specificity, accuracy and CC, were adopted to evaluate the performances of different classifiers. These evaluation measures are defined as follows:

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{6}$$

$$\text{Correlation coefficient (CC)} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{7}$$

where TP, FP, TN and FN represent the numbers of true positives, false positives, true negatives and false negatives, respectively.

Results and discussion

Features of transient protein–protein interface

The residue distributions in the interface and noninterface are shown in Fig. 1a. It is clear that 11 residue types were

enriched in the interface, six types (Phe, Ile, Met, Leu, Val, Trp) of which were hydrophobic residues. In addition, Tyr and Arg that are potential hot spots were also overrepresented in the interface. The similar phenomena have been observed by Ansari and Helms (2005). The overrepresented 11 residue types in our study included the seven types in their research. Moreover, there were more hydrophobic residues in our results, which was probably owing to the different definitions of interface residues and the different sizes of datasets.

Residues exhibiting relatively low *B*-factors are generally those participating in forming secondary structures, neighboring disulfide bridges, or are involved in ligands binding (Tseng and Liang 2007). As shown in Fig. 1b, by calculating and comparing the mean values of the *B*-factors of residues in the interface and noninterface, we found that the mean values of the interface residues were all significantly lower than those of the noninterface residues.

It has been long demonstrated that interface residues are more conservative than noninterface residues during evolution (Mintseris and Weng 2005). We followed the method of Zhou and Shan (2001) to average the diagonal elements of sequence profile for all residue types in the interface and compared them against the corresponding averages in the noninterface. Figure 1c shows that except for Thr and Trp, the averages over the interface residues were all higher than those over the noninterface residues. The results confirm that residues in the transient interface are more conservative.

Previous findings have suggested that interface residues are more solvent accessible than noninterface residues (Jones and Thornton 1997a, b). From Fig. 1d, the average ASAs of the interface residues and noninterface residues reveal that except for Asp, Gly and Ala, the solvent accessibilities of the other 17 residue types in the transient interface were stronger.

Performance of SVM classifiers

The results of threefold cross-validation on 130 chains are given in Table 1. As can be seen from Table 1, all the classifiers can predict significantly better than the random predictions (shown in parentheses). When single feature was used, SVM_B can identify residues in the transient interfaces most effectively, SVM_P was second to SVM_B, and SVM_A was relatively inferior. For the classifiers using the combination of two features, SVM_{B+P} obtained the best CC of 0.262. Although SVM_{P+A} and SVM_{B+A} did not perform as good as SVM_{B+P}, they were still superior to the classifiers with single feature. However, SVM_{B+P+A} achieved a much better performance than the above classifiers based on two features. Especially, compared with SVM_{P+A}, the CC was raised from 0.198 to 0.290. These results indicate that *B*-factor plays a vital role in identifying

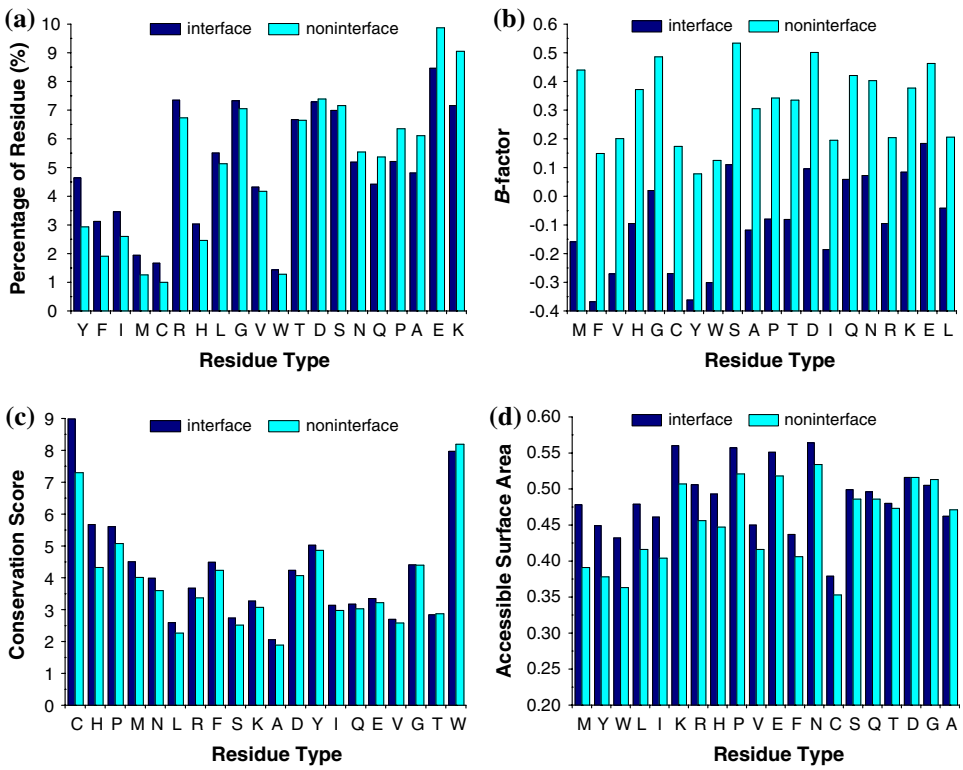

Fig. 1 Comparison between interface residues and noninterface residues. **a** residue distributions, **b** *B*-factors, **c** conservation scores, **d** accessible surface areas

Table 1 The results of threefold cross-validation on 130 chains

Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	CC
SVM_A	43.4 (55.3)[a]	33.7 (28.7)	58.5 (47.0)	0.077 (−0.011)
SVM_P	52.8 (51.4)	39.5 (29.6)	62.5 (50.2)	0.181 (0.010)
SVM_B	59.7 (47.7)	40.7 (28.2)	63.0 (49.8)	0.220 (−0.015)
SVM_{P+A}	53.9 (51.4)	40.6 (29.5)	63.3 (49.9)	0.198 (0.007)
SVM_{B+A}	59.3 (47.4)	41.9 (29.3)	64.0 (51.7)	0.234 (0.008)
SVM_{B+P}	60.3 (50.1)	43.6 (28.9)	65.7 (50.4)	0.262 (0.005)
SVM_{B+P+A}	61.8 (49.2)	45.4 (28.8)	67.1 (49.7)	0.290 (−0.008)

The subscripts are defined as follows: *B* *B*-factor, *P* sequence profile, *A* ASA

[a] Random predictions were obtained by randomly shuffling the labels of samples in training sets and retraining the classifiers to predict test sets

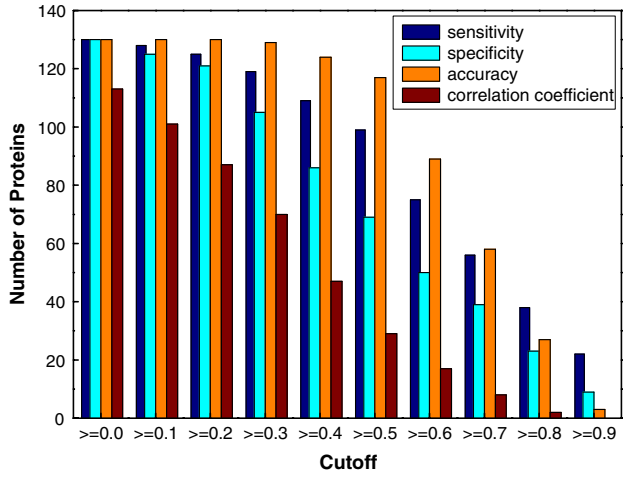

Fig. 2 The distributions of evaluation measure values of SVM_{B+P+A} for 130 chains

interaction sites in transient complexes, and that utilizing the complementarity of the three features is favorable for improving the prediction performance.

For each evaluation measure, setting different cutoffs from 0 to 1 with a 0.1 increment each time, the corresponding numbers of chains were obtained. The distributions of evaluation measure values of SVM_{B+P+A} are exhibited in Fig. 2. It can be observed that the sensitivity values of 99 (76%) chains exceeded 50%, and 69 (53%) chains had the specificity values exceeding the same cutoff. The distributions of accuracy values show that the accuracy values were greater than 20% for all chains, 117 (90%) chains of which achieved the values over 50%. In addition, it can be found that the CC values were not less than 0 for 113 (87%) chains, which suggests that our method is effective.

Evaluation of the predictions using three-dimensional structure

To further illustrate the effectiveness of our method, the prediction results of the protein complex 1ABR (PDB ID) chosen from our dataset were visualized using the PyMOL package (DeLano 2002). The complex 1ABR that is a type II ribosome-inactivating protein is composed of an A-chain (1ABR:A) linked by a disulfide bond to a B-chain (1ABR:B) (Tahirov et al. 1995). As can be seen from Fig. 3, the classifiers with single feature can identify part of interface residues in 1ABR:B, but incorrectly predicted many false positives and false negatives. However, when the three features were combined, the classifier not only identified more interface residues, but also reduced the number of false predictions. It is indicated again that combining the three features can improve the prediction performance.

Independent testing

Benchmark 2.0 is a nonredundant dataset for testing protein–protein docking algorithms (Mintseris et al. 2005). This dataset (excluding antigen-antibody) contains 62 transient protein complexes. The chains sharing more than 30% sequence identity with anyone of the 130 chains in our dataset were eliminated. After this process, we got a nonhomologous set consisting of 37, 11 and 5 chains with minor (rigid body), medium (medium difficult) and large (difficult) conformational changes. Then, we used our dataset as a training set to train the classifiers with combined features, and predicted the interaction sites contained by all chains and nonhomologous chains in benchmark 2.0, respectively. In order to balance positive and negative samples, all interface residues and a same number of randomly sampled noninterface residues from our dataset were extracted for training and this procedure was repeated five times.

The results of different classifiers tested on the whole set (shown in parentheses) and the nonhomologous set are displayed in Table 2. It can be seen that the prediction abilities of the four classifiers were consistent with the results attained by threefold cross-validation on 130 chains. SVM_{B+P+A} got the best performance not only for all chains, but also for nonhomologous chains. In the classifiers based on two features, SVM_{B+P} was the best, SVM_{B+A} was second to SVM_{B+P}, and SVM_{P+A} was relatively inferior. Especially, as the magnitude of conformational changes increases, only combining sequence profile and ASA can not favorably identify the interface residues. However, when *B*-factor was input as an additional feature, the prediction performance was obviously improved. It is interesting that the classifiers utilizing *B*-factor as a feature got better performance on the difficult set than on the other

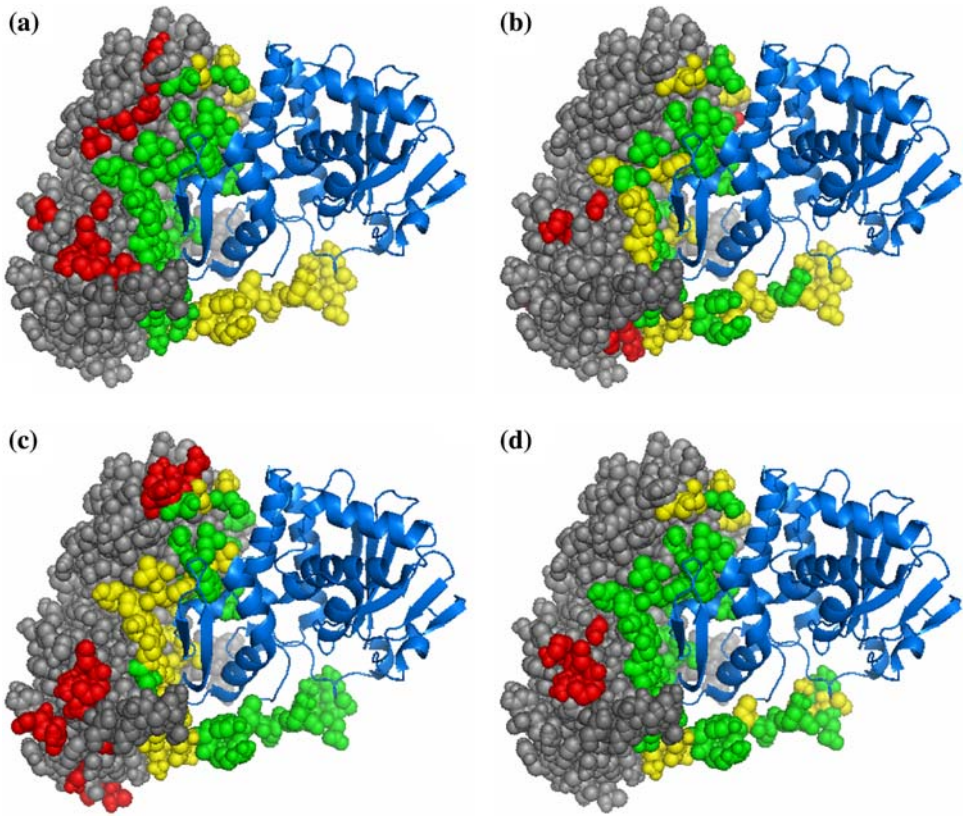

Fig. 3 Visualization of prediction results for complex 1ABR (PDB ID). **a** SVM_B, **b** SVM_P, **c** SVM_A, **d** SVM_{B+P+A}. The colors of different residues are defined as follows: *green* denotes true positives (TP), *red* denotes false positives (FP), *yellow* denotes false negatives (FN)

Table 2 The results of independent testing on benchmark 2.0

Subset	No. of chains	Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	CC
Rigid body	37 (86)[a]	SVM_{P+A}	55.4 (59.6)	38.6 (38.9)	63.4 (66.0)	0.200 (0.248)
		SVM_{B+A}	65.8 (61.8)	40.2 (40.7)	63.6 (67.4)	0.257 (0.278)
		SVM_{B+P}	65.9 (66.6)	44.3 (44.9)	67.8 (70.8)	0.312 (0.348)
		SVM_{B+P+A}	67.6 (67.8)	44.9 (46.7)	68.3 (72.2)	0.328 (0.374)
Medium difficult	11 (24)	SVM_{P+A}	34.5 (44.7)	35.7 (35.4)	69.3 (67.6)	0.149 (0.180)
		SVM_{B+A}	58.8 (58.0)	42.8 (38.5)	71.2 (68.1)	0.308 (0.259)
		SVM_{B+P}	50.0 (56.0)	45.1 (41.7)	73.3 (71.1)	0.297 (0.290)
		SVM_{B+P+A}	49.0 (56.3)	47.9 (42.9)	74.9 (71.9)	0.318 (0.303)
Difficult	5 (14)	SVM_{P+A}	27.6 (43.5)	31.1 (29.9)	73.3 (62.6)	0.129 (0.107)
		SVM_{B+A}	85.2 (73.3)	39.4 (40.5)	70.8 (68.2)	0.423 (0.343)
		SVM_{B+P}	73.9 (69.3)	40.2 (40.5)	72.8 (68.6)	0.385 (0.326)
		SVM_{B+P+A}	70.0 (68.4)	39.3 (40.7)	72.4 (68.9)	0.359 (0.325)
All	53 (124)	SVM_{P+A}	46.9 (54.1)	37.4 (37.0)	66.5 (65.8)	0.187 (0.214)
		SVM_{B+A}	66.6 (62.6)	40.6 (40.2)	66.7 (67.7)	0.294 (0.284)
		SVM_{B+P}	63.1 (64.7)	43.8 (43.6)	70.0 (70.5)	0.320 (0.332)
		SVM_{B+P+A}	63.4 (65.4)	44.6 (45.0)	70.6 (71.6)	0.331 (0.351)

The subscripts are defined as follows: *B* *B*-factor, *P* sequence profile, *A* ASA

[a] The numbers of all chains in different categories are in parentheses

two sets. The results further illuminate that *B*-factor is crucial to identify the interaction sites of the chains with large conformational changes. In addition, except for SVM_{B+A}, the performances of the other three classifiers tested on the nonhomologous set were not so good as on the whole set. Even so, SVM_{B+P+A} achieved a CC of 0.331 on the nonhomologous set, which was better than the value of threefold cross-validation on 130 chains. The prediction results confirm that our classifiers are robust, and that using more training samples can acquire better performance.

Comparison with cons-PPISP

A direct comparison with other methods is difficult due to the differences in the definitions of surface residues and interface residues and the preparations of datasets. We made an attempt to compare our method with cons-PPISP, because they were both tested on the protein–protein docking benchmark set. Cons-PPISP that used sequence profile and solvent accessibility as input to neural networks was developed by Chen and Zhou (2005). Their method was tested on 68 unique chains of 40 complexes in benchmark 1.0. The sensitivity and specificity were 50% and 50% for the enzyme-inhibitor category, 28 and 31% for other category, and 38 and 42% for the whole 68 chains, respectively. Our method was tested on 95 unique chains of 62 complexes in benchmark 2.0, which achieved a sensitivity and specificity of 69 and 51% for the enzyme-inhibitor category, 64 and 42% for other category, and 67 and 44% for the whole 95 chains, respectively. Our method achieving better performance probably depends on three factors. First, *B*-factor was used as an additional feature in our method, which led to the obvious improvement of prediction performance. When sequence profile and ASA were combined, the sensitivity and specificity of our method for 95 chains were only 55 and 36%. Second, we used a balanced training set to train the classifiers, which may result in a relatively high sensitivity. Third, owing to the training set of Chen and Zhou including some other types of complexes, the features extracted from these complexes may not be suitable for predicting interface residues in transient complexes.

Conclusion

Transient protein–protein interactions play a vital role in many biological processes. Due to the limitation of experimental methods, the knowledge of these interactions is inadequate. In this research, transient interfaces were chosen as study objects and the features of these interfaces were analyzed. It was found that besides sequence profile and ASA, *B*-factor can also distinctly reflect the differences between interface and noninterface. We converted these features into input vector and used SVM classifiers to predict residues in the interface. It is indicated that the incorporation of *B*-factor is important to identify interaction sites in transient complexes, and that the information contained within these features are complementary. Therefore, our method can complement experimental techniques in studying transient protein–protein interactions. Incorporation of our method with more physicochemical properties and structural attributes will prompt the study of protein–protein interactions.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 90608020, 30370354, and 90203011), NCET-060651, the National Platform Project of China (Grant No. 2005DKA64001), and the Ministry of Education of China (Grant Nos. 20050487037 and 505010).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389

Ansari S, Helms V (2005) Statistical analysis of predominantly transient protein–protein interfaces. Proteins 61:344–355. doi:10.1002/prot.20593

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235

Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein–protein interfaces using a Bayesian network prediction method. J Mol Biol 362:365–386. doi:10.1016/j.jmb.2006.07.028

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at: (http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Chelliah V, Chen L, Blundell TL, Lovell SC (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. J Mol Biol 342:1487–1504. doi:10.1016/j.jmb.2004.08.022

Chen H, Zhou HX (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins 61:21–35. doi:10.1002/prot.20514

Cole C, Warwicker J (2002) Side-chain conformational entropy at protein–protein interfaces. Protein Sci 11:2860–2870. doi:10.1110/ps.0222702

DeLano WL (2002) The PyMOL molecular graphics system. Software available at: (http://www.pymol.org)

Dong Q, Wang X, Lin L, Guan Y (2007) Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. BMC Bioinformatics 8:147. doi:10.1186/1471-2105-8-147

Fariselli P, Pazos F, Valencia A, Casadio R (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 269:1356–1361. doi:10.1046/j.1432-1033.2002.02767.x

Friedrich T, Pils B, Dandekar T, Schultz J, Müller T (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics 22:2851–2857. doi:10.1093/bioinformatics/btl486

Glaser F, Steinberg DM, Vakser IA, Ben-Tal N (2001) Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 43:89–102. doi:10.1002/1097-0134(20010501)43:2<89::AID-PROT1021>3.0.CO;2-H

Hoskins J, Lovell S, Blundell TL (2006) An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. Protein Sci 15:1017–1029. doi:10.1110/ps.051589106

Jones S, Thornton JM (1995) Protein–protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63:31–65. doi:10.1016/0079-6107(94)00008-W

Jones S, Thornton JM (1996) Principles of protein–protein interactions. Proc Natl Acad Sci USA 93:13–20. doi:10.1073/pnas.93.1.13

Jones S, Thornton JM (1997a) Analysis of protein–protein interaction sites using surface patches. J Mol Biol 272:121–132. doi:10.1006/jmbi.1997.1234

Jones S, Thornton JM (1997b) Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 272:133–143. doi:10.1006/jmbi.1997.1233

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637. doi:10.1002/bip.360221211

Kim H, Park H (2003) Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 16:553–560. doi:10.1093/protein/gzg072

Koike A, Takagi T (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 17:165–173. doi:10.1093/protein/gzh020

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302. doi:10.1093/nar/gki370

Li JJ, Huang DS, Wang B, Chen P (2006) Identifying protein–protein interfacial residues in heterocomplexes using residue conservation scores. Int J Biol Macromol 38:241–247. doi:10.1016/j.ijbiomac.2006.02.024

Li MH, Lin L, Wang XL, Liu T (2007) Protein–protein interaction site prediction based on conditional random fields. Bioinformatics 23:597–604. doi:10.1093/bioinformatics/btl660

Liang S, Zhang C, Liu S, Zhou Y (2006) Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 34:3698–3707. doi:10.1093/nar/gkl454

Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein–protein recognition sites. J Mol Biol 285:2177–2198. doi:10.1006/jmbi.1998.2439

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. Proc Natl Acad Sci USA 102:10930–10935. doi:10.1073/pnas.0502667102

Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z (2005) Protein–protein docking benchmark 2.0: an update. Proteins 60:214–216. doi:10.1002/prot.20560

Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 338:181–199. doi:10.1016/j.jmb.2004.02.040

Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544:236–239. doi:10.1016/S0014-5793(03)00456-3

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226. doi:10.1002/prot.340200303

Szilágyi A, Grimm V, Arakaki AK, Skolnick J (2005) Prediction of physical protein–protein interactions. Phys Biol 2:S1–S16. doi:10.1088/1478-3975/2/2/S01

Tahirov TH, Lu TH, Liaw YC, Chen YL, Lin JY (1995) Crystal structure of abrin-a at 2.14 Å. J Mol Biol 250:354–367. doi:10.1006/jmbi.1995.0382

Tseng YY, Liang J (2007) Predicting enzyme functional surfaces and locating key residues automatically from structures. Ann Biomed Eng 35:1037–1042. doi:10.1007/s10439-006-9241-2

Vapnik VN (1995) The nature of statistical learning theory. Springer, New York

Wang Y, Xue Z, Shen G, Xu J (2008) PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids 35:295–302. doi:10.1007/s00726-007-0634-9

Yuan Z, Zhao J, Wang ZX (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 16:109–114. doi:10.1093/proeng/gzg014

Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44:336–343. doi:10.1002/prot.1099