BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 17 2006, pages 2094–2098
doi:10.1093/bioinformatics/btl275

Structural bioinformatics

Intramolecular surface contacts contain information about protein–protein interface regions

Sjoerd J. de Vries and Alexandre M. J. J. Bonvin*

Faculty of Sciences, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584CH, Utrecht, The Netherlands

Received on February 10, 2006; revised on May 14, 2006; accepted on May 29, 2006
Advance Access publication June 9, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Some amino acids clearly show preferences over others in protein–protein interfaces. These preferences, or so-called interface propensities can be used for *a priori* interface prediction. We investigated whether the prediction accuracy could be improved by considering not single but pairs of residues in an interface. Here we present the first systematic analysis of intramolecular surface contacts in interface prediction.

Results: We show that preferences do exist for contacts within and around an interface region within one molecule: specific pairs of amino acids are more often occurring than others. Using intramolecular contact propensities in a blind test, higher average scores were assigned to interface residues than to non-interface residues. This effect persisted as small but significant when the contact propensities were corrected to eliminate the influence of single amino acid interface propensity. This indicates that intramolecular contact propensities may replace interface propensities in protein–protein interface prediction.

Availability: The source code is available on request from the authors.

Contact: a.m.j.j.bonvin@chem.uu.nl

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A large number of studies have established that the residue composition of protein–protein interfaces generally differs from that of the rest of the surface (Ansari and Helms, 2005; Bahadur *et al*., 2003; Bordner and Abagyan, 2005; Glaser *et al*., 2001; Jones and Thornton, 1997; Keskin *et al*., 2004; Lo Conte *et al*., 1999; Ma *et al*., 2003; Zhou and Shan, 2001). When counting the occurrences of each amino acid in interfaces, different numbers are found than expected on the basis of the amino acid composition of the surface. The so-called interface propensity, defined as the ratio of observed over expected counts, has been found to be high for aromatic residues such as phenylalanine, and low for most polar and charged residues, such as lysine. These interface propensities can be utilized for *a priori* interface prediction, by assigning higher probabilities to residues and clusters of residues with high interface propensities. A number of interface prediction methods exist where interface propensities have been taken into account (Bordner and Abagyan, 2005; Bradford and Westhead, 2005; De Vries *et al*., 2006; Gottschalk *et al*., 2004).

In addition, once reasonable models of a complex have been obtained by docking, interface propensities can help in discriminating correct from incorrect solutions. This *a posteriori* filtering can be enhanced by taking the partner protein into account, through the use of intermolecular contacts. As is the case for single amino acids within a specific interface, some contacts are favored across biological interfaces, and intermolecular contact propensities have been derived by a number of groups (Mintseris and Weng, 2003; Moont *et al*., 1999; Ofran and Rost, 2003; Ponstingl *et al*., 2000; Saha *et al*., 2005; Zhou and Zhou, 2002). These have been successfully applied to filter docking solutions (Duan *et al*., 2005; Gottschalk *et al*., 2004; Moont *et al*., 1999; Zhang *et al*., 2005). They have also been used to discriminate crystal contacts from true biological interfaces in X-ray structures (Mintseris and Weng, 2003; Ponstingl *et al*., 2000; Saha *et al*., 2005).

While intermolecular contacts between protein chains have been well-studied, the intramolecular contacts have been neglected. Typically, surface patches are analyzed without regarding the arrangement of individual residues within that patch. This arrangement potentially contains valuable information about protein interfaces.

Therefore, we asked ourselves the following question: are there, in analogy to intermolecular contacts, preferences for specific intramolecular surface contacts, which could be used to distinguish interface regions from non-interface regions? Our goal is not to develop an independent interface predictor, but to investigate whether intramolecular surface contacts can make a valuable contribution, so that they could be incorporated into existing interface predictors.

Testing this hypothesis is not trivial. Unlike a residue, which can be classified as either interface or non-interface, there are three classes of intramolecular surface contacts: interface–interface, interface–noninterface and non-interface–non-interface. Both the first class (interface contacts) and the second class (which we will call cross-contacts) can be useful in detecting an interface. The second class may be an average of the first and the third, but not necessarily so: in previous studies it was found that protein interfaces generally consist of a hydrophobic core surrounded by a more hydrophilic rim region (Bahadur *et al*., 2003; Bogan and Thorn, 1998; Chakrabarti and Janin, 2002). Unlike the buried surface area criterion used in these studies, our definition of the true interface is based on residues forming contacts with the partner protein, which causes fewer residues to be defined as interface.

*To whom correspondence should be addressed.

Therefore, we expected the second class to represent core-rim contacts, and hence to be enriched in hydrophilic–hydrophobic contacts relative to the other two classes.

To our knowledge, this study presents the first systematic analysis of these contacts. Only in ProMate (Neuvirth *et al*., 2004) an attempt has been made to use this information by assigning statistically based probabilities to pairs of amino acids. However, the amino acids were pooled into different classes (polar, hydrophobic, aromatic and positively/negatively charged), no distinction between interface contacts and cross-contacts was made.

2 SYSTEM AND METHODS

We calculated all propensities on the test set of Bordner and Abagyan (Bordner and Abagyan, 2005). All interface residues were identified with DIMPLOT (Wallace *et al*., 1995) using default settings (minimum distance from the partner protein 3.9 Å). Proteins for which DIMPLOT could not identify any intermolecular contacts were discarded, resulting in a working set of 1029 protein chains. All residues were designated as surface or non-surface by surface accessibility as calculated by NACCESS (Hubbard and Thornton, 1993). Residues with a mainchain or sidechain relative accessibility of >15 % were designated as surface. All non-surface residues were ignored during the analysis. Two surface residues were defined to be in contact if at least one inter-residue distance between any non-hydrogen atoms was smaller than 5 Å.

3 ALGORITHM

3.1 Calculation of intramolecular contact propensities and single amino acid interface propensities

In general, interface propensity can be defined as the ratio between an interface fraction f_i and a surface fraction f_s:

$$P_i = \frac{f_i}{f_s}. \tag{1}$$

In the case of single amino acid interface propensities, these fractions are amino acid fractions, defined as follows:

$$f(x) = \frac{O(x)}{n}, \tag{2a}$$

where $O(x)$ is the number of observed residues of amino acid x, n is the total number of residues and x is any amino acid.

In the case of intramolecular contact propensities, these fractions are contact fractions, which can be defined in a similar way:

$$f(x,y) = \frac{O(x,y)}{N}, \tag{2b}$$

where $O(x,y)$ is the number of observed intramolecular contacts between a residue of amino acid x and a residue of amino acid y, N is the total number of intramolecular contacts and x and y are any amino acids.

Therefore, Equation (1) can be rewritten as follows:

$$P_i(x) = \frac{O_i(x)n_s}{O_s(x)n_i} \tag{3a}$$

for single interface propensities, and

$$P_i(x,y) = \frac{O_i(x,y)N_s}{O_s(x,y)N_i} \tag{3b}$$

$$P_c(x,y) = \frac{O_c(x,y)N_s}{O_s(x,y)N_c} \tag{3c}$$

for intramolecular contact propensities in the case of interface and cross contacts, respectively.

For each protein in the benchmark set, O and N statistics were computed. All those statistics were pooled for all proteins before applying Equation (3) to calculate the propensities.

3.2 Calculation of corrected intramolecular contact propensities

Corrected intramolecular contact propensities were calculated to eliminate the contribution of single amino acid interface propensities to intramolecular contact preferences.

An additional factor to be eliminated is the fact that amino acids could form a different number of contacts per residue on the interface than on the rest of the surface. Therefore, a contact-forming propensity P^f was computed:

$$P^f(x) = \frac{F_i(x)/f_i(x)}{F_s(x)/f_s(x)}, \tag{4}$$

where $F(x)$ is the contact fraction: the number of contacts where a residue of amino acid x is involved in (counting twice x–x contacts) divided by N.

To eliminate both effects (single amino acid propensities and number of contacts formed), the corrected propensities were calculated in the following way:

$$P_i^{\mathrm{corr}}(x,y) = \frac{P_i(x,y)}{P_i(x)P_i(y)} \cdot \frac{1}{P^f(x)P^f(y)} \cdot \frac{C_s(x,y)}{C_i(x,y)}, \tag{5}$$

where C is an additional correction factor for sampling without replacement: within a protein, the number of interface and surface residues is finite and a residue cannot make a contact with itself.

Again, statistics from the whole benchmark set were pooled before the calculation of actual fractions and propensities. Calculations for the corrected cross-contact propensities were performed in the same way. For the derivation of Equation (5) and the computation of C, see Supplementary Materials.

3.3 Scoring using propensities

The propensities were tested on the benchmark of Mintseris *et al*. (2005) using unbound structures. Antibody–antigen structures and structures with missing residues in the interface region were discarded, resulting in a test set of 116 protein chains. To ensure unbiased testing, proteins with a sequence identity of >30 % to any protein in the test set were removed from the working set and the propensities were re-calculated: this resulted in virtually no change in the four sets of propensities ($r^2 = 0.955$–0.997).

In the test set, interface and surface residues were defined using the same criteria as for the working set above. For each surface residue, a circular patch was defined consisting of the closest surface neighbours. It has previously been shown that the number of interface residues depends on the number of surface residues by a power law (Chen and Zhou, 2005). We made therefore a log–log plot of the number of interface and surface residues and found this power to be 0.25, which is somewhat lower than the value of 0.35 found by Chen and Zhou. However, we found the correlation to be

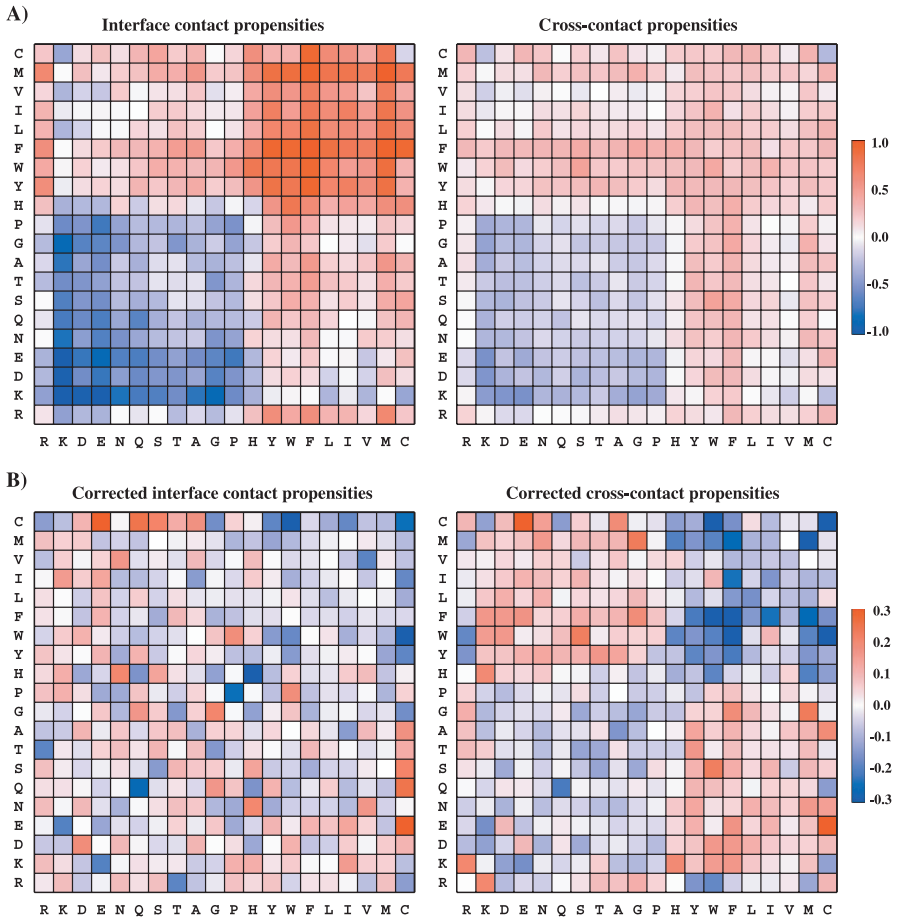

Fig. 1. (**A**) Pairwise intramolecular contact propensities (natural logarithm of each propensity) and (**B**) pairwise intramolecular contact propensities corrected for single amino acid interface propensity (natural logarithm of each propensity). Left, interface contact propensities. Right, cross-contact propensities.

rather weak ($r = 0.43$). A relation between interface and surface sizes clearly exists since the correlation coëfficient is still highly significant ($P \ll 0.01$). However, the exact nature of the relation cannot be established, since fitting a linear relationship yields nearly the same correlation coefficient ($r = 0.40$). We chose to use a power law for interface estimation in this study, using the following formula:

$$N_i = 3.68 N_s^{0.25}, \tag{6}$$

where N_i is the number of interface residues and N_s is the number of surface residues.

Scoring using contact propensities was done as follows. For each contact within a patch and between a patch and a non-patch residue, the interface and cross-contact propensities, respectively, were determined. All propensities were summed into a score for that patch. Scoring using corrected contact propensities was done in the same way. Scoring using single amino acid interface propensities was done by summing the interface propensities of all residues in the patch.

All scores were normalized by dividing by the standard deviation of the scores of all residues in the test set, resulting in a Z-score for each residue.

4 RESULTS AND DISCUSSION

4.1 Calculation of propensities

Intramolecular interface contact and cross-contact propensities were calculated for the protein–protein test set of Bordner and Abagyan (2005). Since a propensity is always the ratio of an observed and an expected value, the choice of an appropriate random model for expected counts is crucial. We chose a model that for each protein effectively reshuffles the partners of all surface contacts, since this takes into account the residue composition of the surface and the fact that some amino acids may form more contacts than others.

Figure 1A show the propensities for interface contacts and cross-contacts. In general, these propensities correlate well with each other ($r = 0.79$). It is observed that hydrophobic contacts are preferred for interface contacts and to a lesser degree for cross-contacts. However a number of hydrophilic–hydrophobic contacts, such as Lys–Phe and Arg–Cys, are enriched in cross-contacts but not in interface contacts.

4.2 Scoring using propensities

The obtained propensities were tested for *a priori* predictive power on the benchmark set of protein–protein complexes of

Mintseris *et al*. (2005). True interface and non-interface residues were defined from the crystal structures of the complexes (see System and Methods). After that, blind predictions using contact propensities were made for each surface residue of the unbound structures.

Normally, a residue can be scored by testing the likelihood of a hypothesis against the available data. (What is the probability of this residue being in the interface given these propensities?) However, since a contact involves two residues and a residue forms multiple contacts, using contact propensities for *a priori* prediction requires that a hypothesis is made for multiple residues simultaneously. Therefore, we evaluated for each residue the hypothesis that it is at the heart of an interface patch.

Within that patch, all interface contacts were evaluated against the interface contact propensities and all cross-contacts of patch residues with surrounding non-interface residues were evaluated against the cross-contact propensities. The score corresponding to the sum of all evaluations was assigned to the residue forming the heart of the patch.

Figure 2A shows the distribution of the scores obtained for interface and non-interface residues. Although there is a large overlap between the two (58.2 %), it is clear that interface residues score better than non-interface residues on average. This indicates that intramolecular contacts propensities can be used to identify interface regions on protein surfaces, although intramolecular contacts alone are not sufficient for accurate prediction.

However, it must be verified that the observed preference is not merely an effect of single amino acid interface propensities. Since interfaces are enriched in hydrophobic residues, hydrophobic contacts are expected to be enriched as well even when contacts are distributed at random. Therefore, the same proteins were tested for prediction using single interface propensities by the same design: each residue was hypothesized to form the heart of a circular patch and the sum of interface propensities of the residues in the patch was calculated and assigned to that residue. These results are shown in Figure 2B. Comparing Figure 2A and B shows that intramolecular contact propensities are superior in performance to single amino acid interface propensities: the overlap between the curves is 60.2 % for single interface propensities compared with only 58.2 % for pairwise propensities.

To assess the significance between pairwise and single residue interface propensities, the average rank of interface residues among the test set was computed (Table 1). The average rank was found to be significantly better for pairwise propensities compared with single propensities (Student's paired t test, $P = 0.0075$).

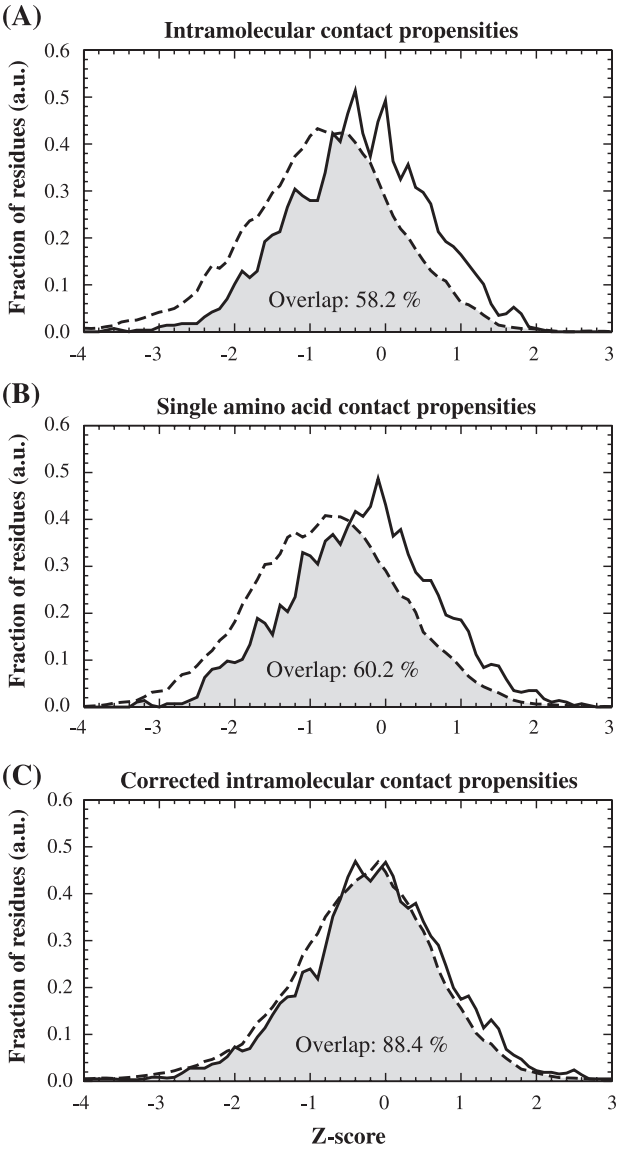

Fig. 2. Normalized distributions of scores for interface residues (continuous line) and non-interface residues (dashed line) for different scoring schemes: (**A**) intramolecular pairwise propensity scoring, (**B**) single amino acid interface propensity scoring and (**C**) intramolecular pairwise propensity corrected for single amino acid interface propensity. The overlap between the curves is indicated by the gray area together with the corresponding percentage of the total area.

4.3 Calculation of corrected propensities

The significance of the difference between Figure 2A and B is hard to assess. Since the surface patches used for scoring are highly overlapping, the residue scores within each protein are strongly correlated with those of their structural neighbors, and the degrees of freedom cannot be determined.

Therefore, we did a more rigorous comparison by eliminating the influence of interface propensity on the intramolecular contact propensity. This was accomplished by not taking surface contacts as a random model for reshuffling, but instead only the interface contacts and cross-contacts, respectively. The resulting corrected propensity totally eliminates all effects of residue composition within and around the interface. If intramolecular contact propensities and single interface propensities are equivalent, this correction would result in corrected propensities very close to zero, and scoring using those propensities would result in a random predictor, with full overlap between surface and interface residue scores.

The corrected propensities are shown in Figure 1B, and they significantly differ from zero (χ^2, $P < 0.025$ for interface contacts, $P \ll 0.001$ for cross-contacts). There are only a few trends in corrected interface propensities: Cys and Met preferentially contact hydrophilic residues, and His and Pro preferentially contact other residues than His and Pro. In contrast, in the cross-contacts,

Table 1. Difference in average interface residue rank between pairwise or single residue propensity scoring

Number of residues		
Interface	1431	
Total surface	19 688	
	Pairwise propensity	Single propensity
Interface residue rank		
Average	7236	7392
Standard deviation	±5341	±5409
Standard error	±141.2	±143.0
Difference		
t statistic	2.675	
Degrees of freedom	1430	
P-value	0.0075	

Ranks were computed over the whole test set for all interface residues: rank 1 corresponds to the best-scoring residue in the test set.

a clear depletion of hydrophilic–hydrophilic and especially hydrophobic–hydrophobic contacts is observed, again in favor of hydrophilic–hydrophobic contacts. These results confirm the hydrophobic core—hydrophilic rim hypothesis.

4.4 Scoring using corrected propensities

The corrected propensities were used for prediction in the same manner as before (Fig. 2C). Although the predictive power was greatly reduced, there is still a difference between interface and non-interface residues (overlap 88.4 %). Moreover, this difference turned out to be highly significant: for each protein, the average score of the interface residues was computed, and it was shown to be significantly higher than the average score of the non-interface residues (Student's paired *t*-test, $n = 116$, $P = 0.0185$, one-tailed). Therefore, preferences for intramolecular contacts persist even when the influence of interface propensity is removed, and these preferences do have predictive power on their own.

4.5 Conclusion

Interface prediction is a multi-faceted discipline, combining information from diverse sources such as conservation, interface propensity and spatial arrangement of atoms. In this study we provide the first systematic analysis of a hitherto neglected source of information, intramolecular contact propensities. Using a classification into interface–interface, non-interface–non-interface and interface–non-interface (cross) contacts, and an analysis at the amino acid level, we show that intramolecular surface contacts contain information about putative protein–protein interfaces. The difference with single amino acid interface propensities is small but significant.

Considering their increased predictive power, intramolecular surface contacts will be able to replace the single amino acid interface propensities currently used in interface predictors. We are planning to implement intramolecular surface contact propensities into our interface prediction program WHISCY (De Vries *et al*., 2006).

Conflict of Interest: none declared.

REFERENCES

Ansari,S. and Helms,V. (2005) Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, **61**, 344–355.

Bahadur,R.P. *et al.* (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Bordner,A.J. and Abagyan,R. (2005) Statistical analysis and prediction of protein–protein interfaces. *Proteins*, **60**, 353–366.

Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.

Chen,H. and Zhou,H.X. (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins*, **61**, 21–35.

De Vries,S.J. *et al.* (2006) WHISCY: WHat Information does Surface Conservation Yield? Application to data-driven docking *Proteins*, **63**, 479–489.

Duan,Y. *et al.* (2005) Physicochemical and residue conservation calculations to improve the ranking of protein–protein docking solutions. *Protein Sci.*, **14**, 316–328.

Glaser,F. *et al.* (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.

Gottschalk,K.E. *et al.* (2004) A novel method for scoring of docked protein complexes using predicted protein–protein binding sites. *Protein Eng. Des. Sel.*, **17**, 183–189.

Hubbard,S.J. and Thornton,J.M. (1993) *NACCESS*. Department of Biochemistry and Molecular Biology, University College London.

Jones,S. and Thornton,J.M. (1997) Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

Keskin,O. *et al.* (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Lo Conte,L. *et al.* (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

Ma,B. *et al.* (2003) Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.

Mintseris,J. and Weng,Z. (2003) Atomic contact vectors in protein–protein recognition. *Proteins*, **53**, 629–639.

Mintseris,J. *et al.* (2005) Protein–protein docking benchmark 2.0: an update. *Proteins*, **60**, 214–216.

Moont,G. *et al.* (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364–373.

Neuvirth,H. *et al.* (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces,. *J. Mol. Biol.*, **325**, 377–387.

Ponstingl,H. *et al.* (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.

Saha,R.P. *et al.* (2005) Interresidue contacts in proteins and protein–protein interfaces and their use in characterizing the homodimeric interface. *J. Proteome. Res.*, **4**, 1600–1609.

Wallace,A.C. *et al.* (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.

Zhang,C. *et al.* (2005) A knowledge-based energy function for protein–ligand, protein–protein and protein–DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.

Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.

Zhou,H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.