BIOINFORMATICS REVIEW Vol. 23 no. 17 2007, pages 2203–2209
doi:10.1093/bioinformatics/btm323

Structural bioinformatics

Interaction-site prediction for protein complexes: a critical assessment

Huan-Xiang Zhou[1,2,3,*] and Sanbo Qin[1,2]

[1]Institute of Molecular Biophysics, [2]School of Computational Science and [3]Department of Physics, Florida State University, Tallahassee, Florida 32306, USA

Received on May 21, 2007; revised and accepted on June 11, 2007

Advance Access publication June 22, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Proteins function through interactions with other proteins and biomolecules. Protein–protein interfaces hold key information toward molecular understanding of protein function. In the past few years, there have been intensive efforts in developing methods for predicting protein interface residues. A review that presents the current status of interface prediction and an overview of its applications and project future developments is in order.

Summary: Interface prediction methods rely on a wide range of sequence, structural and physical attributes that distinguish interface residues from non-interface surface residues. The input data are manipulated into either a numerical value or a probability representing the potential for a residue to be inside a protein interface. Predictions are now satisfactory for complex-forming proteins that are well represented in the Protein Data Bank, but less so for under-represented ones. Future developments will be directed at tackling problems such as building structural models for multi-component structural complexes.

Contact: zhou@sb.fsu.edu

1 INTRODUCTION

Protein interactions are at the center of protein function. Interactions occur through the formation of complexes, either transient or more long lasting. Examples are the replisome, RNA polymerases, the spliceosome, the ribosome, chaperonins and the various complexes formed along signal transduction pathways and during enzyme catalysis and inhibition. They often have multiple components. Structures of protein complexes are the basis for detailed understanding of protein function, and yet their determination is fraught with technical difficulties and lags far behind structure determination for single proteins or protein domains. As anecdotal evidence, CASP (http://predictioncenter.gc.ucdavis.edu/), the biennial exercise for evaluating structure prediction for single proteins or protein domains, has the luxury of ~100 targets. In contrast, the analogous exercise, CAPRI (http://capri.ebi.ac.uk/), for evaluating methods for docking unbound proteins into complexes has had to contend with a total of 25 targets for the last six years. It has been argued that single-domain protein structures are completely represented in the current Protein Data Bank (http://www.rcsb.org/pdb) (Zhang *et al*., 2006). Worldwide structural genomics efforts have so far concentrated on single proteins or protein domains and have only begun to pay attention to protein complexes.

Given this state of affairs, information that provides insight into the structures of protein complexes becomes very valuable. In this category, there is information about the interfaces of protein complexes. The importance of such information motivated an early attempt by Jones and Thornton (1997) to predict surface patches that overlap with interfaces. A method for predicting individual residues in protein–protein interfaces was published in 2001 (Zhou and Shan, 2001). Since then over 20 papers presenting different methods have been published (Bordner and Abagyan, 2005; Bradford and Westhead, 2005; Bradford *et al*., 2006; Burgoyne and Jackson, 2006; Chen and Zhou, 2005; Chung *et al*., 2006; de Vries *et al*., 2006; Fariselli *et al*., 2002; Fernandez-Recio *et al*., 2004; Friedrich *et al*., 2006; Hoskins *et al*., 2006; Koike and Takagi, 2004; Kufareva *et al*., 2007; Landau *et al*., 2005; Li *et al*., 2006; Li *et al*., 2007; Liang *et al*., 2006; Murakami and Jones, 2006; Neuvirth *et al*., 2004; Ofran and Rost, 2003; Porollo and Meller, 2007; Qin and Zhou, 2007a; Res *et al*., 2005; Sen *et al*., 2004; Wang *et al*., 2006a; Wang *et al*., 2006b; Yan *et al*., 2004). This review aims to assess the current status of interface prediction, present an overview of its applications, and project future developments.

2 CHARACTERISTICS OF INTERFACE RESIDUES

Interface prediction relies on characteristics of residues found in interfaces of protein complexes (a number of databases of protein–protein interfaces have been created, e.g. at http://protein3d.ncifcrf.gov/~keskino/, http://dockground.bioinformatics.ku.edu/, and http://www.ces.clemson.edu/compbio/protcom/). Identified by comparing interface and non-interface portions of protein surfaces, the characteristics broaden and reinforce our understanding of proteins in general. Among these the most prominent involve

- *Sequence conservation*. Interface residues are more conserved relative to non-interface surface residues (Zhou and Shan, 2001). Conservation may arise either for functional (Lichtarge *et al*., 1996) or for structural reasons.

*To whom correspondence should be addressed.

- *Proportions of the 20 types of amino acids*. In protein–protein interfaces, hydrophobic (and aromatic) residues and arginine are enriched whereas other charged residues are depleted (Conte *et al*., 1999; Zhou and Shan, 2001). The enrichment of arginine has been attributed to cation-π interactions (Crowley and Golovin, 2005). There also appears to be a stronger tendency for hydrophobic residues in interfaces to cluster relative to those in non-interface surfaces (Neuvirth *et al*., 2004).
- *Secondary structure*. Interfaces seem to favor β-strands while disfavor α-helices; loops in interfaces also tend to be longer (Neuvirth *et al*., 2004).
- *Solvent accessibility*. Interface residues have higher solvent accessibilities than non-interface surface residues (Chen and Zhou, 2005; Jones and Thornton, 1997). The latter residues do not have intermolecular interaction partners upon complex formation and will thus tend to maximize intramolecular interactions, reducing their solvent accessibilities. Solvent accessibilities can be predicted from protein sequence; these methods are typically trained on datasets in which interface residues are grossly under-represented, and thus will tend to under-predict the solvent accessibilities of interface residues. It was found that the difference between predicted and actual solvent accessibilities is more discriminating than the actual solvent accessibility itself (Porollo and Meller, 2007). For each type of amino acid, further classification of solvent accessibility according to secondary structure may also improve the discriminatory power of solvent accessibility (Hoskins *et al*., 2006).
- *Side-chain conformational entropy*. Interface residues appear to be less likely to sample alternative side-chain rotamers (Cole and Warwicker, 2002; Liang *et al*., 2006), perhaps to minimize entropic cost upon complex formation.

Of all these attributes, the ones related to protein structures typically are not particularly sensitive to local conformational changes that accompany complex formation. It should be noted that many parts of a protein surface can be epitopes for antibody binding through antibody maturation. Thus, protein antigens are not suitable targets for interface prediction.

3 METHODS OF INTERFACE PREDICTION

The characteristics distinguishing interface residues generally are rather weak. Multiple types of input data from multiple residues are needed to classify a single surface residue into the interface or non-interface group. The multiple residues typically are spatial neighbors of the residue under consideration (Zhou and Shan, 2001), since an interface is formed by spatially neighboring residues [note, however, methods that use only protein sequences have also appeared (Ofran and Rost, 2003; Res *et al*., 2005)]. Broadly speaking, interface prediction methods can be divided into numerical value based and probabilistic. Both types of methods depend on training over a dataset. In general, only surface residues are considered for interface prediction.

In the first type of methods, the input data, $\mathbf{x}_i$, of the residue i under consideration, and the corresponding quantities $\mathbf{x}_j$ of spatially neighboring residues $j \in \mathbf{n}_i$ (the neighbor list of residue i) are used to define a function

$$S_i = f(\mathbf{x}_i, \mathbf{x}_{j \in \mathbf{n}_i}, c) \tag{1}$$

where **c** is a set of coefficients to be determined by training. The 'state' of residue i, which can either I for interface or N for non-interface residue, is determined by the numerical value of S_i. For example, I is predicted if S_i is above a threshold value; N is predicted otherwise. The goal of training is to minimize the difference between predicted and actual states in the training dataset.

Numerical value-based methods vary widely in details. Below the strengths and limitations of these methods are highlighted.

- *Linear regression* (Kufareva *et al*., 2007; Li *et al*., 2006). In this method, Equation (1) is a linear function of input data such as solvent accessibilities, with **c** as coefficients. The strength of this method is its simplicity, which can be useful, e.g. for optimizing the selection of input data included for interface prediction. The performance of linear regression generally lags behind other methods.
- *Scoring function* (Burgoyne and Jackson, 2006; de Vries *et al*., 2006; Hoskins *et al*., 2006; Landau *et al*., 2005; Liang *et al*., 2006; Murakami and Jones, 2006). Scoring functions are modeled after empirical energy functions, with terms for contributions for different input data. The functional forms of the individual terms are usually far more complicated than linear, allowing for better discrimination, but their introduction requires physical insight.
- *Support vector machine* (Bordner and Abagyan, 2005; Bradford and Westhead, 2005; Chung *et al*., 2006; Koike and Takagi, 2004; Res *et al*., 2005; Wang *et al*., 2006a; Wang *et al*., 2006b; Yan *et al*., 2004). In this method, the input data are non-linearly mapped to a feature space, in which a hyper-plane is obtained that optimally separates the data points corresponding to the I state from those corresponding to the N state. The price for improved performance over linear regression is the loss of transparency in the prediction method.
- *Neural network* (Chen and Zhou, 2005; Fariselli *et al*., 2002; Ofran and Rost, 2003; Porollo and Meller, 2007; Zhou and Shan, 2001). Here the input data are linearly combined into a single input to a node, which performs a non-linear transform. Typically there is a layer of intermediate ('hidden') nodes, whose output data are again fed to a final output node. The coefficients, or weight, of the linear combinations in forming input to nodes are optimized on a training dataset to minimize the difference between predicted output value (ranging from, e.g. 0 to 1) and the value coding the actual state (1 for I and 0 for N). Again, there is a trade off between performance and transparency.

The goal of probabilistic methods is to find the conditional probability $p(s|x_1, \ldots, x_k)$, where $s = \mathrm{I}$ or N, x_1 to x_k are input data for a residue under consideration. Interface is predicted, e.g. when $p(\mathrm{I}|x_1, \ldots, x_k)$ is greater than a threshold value. There are different levels of sophistication in obtaining the conditional probability from the training dataset.

- *Naïve Bayesian* (Neuvirth *et al*., 2004). In this method, the different input data, x_1 to x_k, are assumed to be independent, leading to

$$p(s|x_1, \ldots, x_k) = p(s) \prod_{l=1}^{k} \frac{p(x_l|s)}{p(x_l)} \tag{2}$$

where $p(s)$ is the fraction of state s in the training dataset, $p(x_l)$ is the probability density of input data x_l in the whole dataset, and $p(x_l|s)$ is the corresponding quantity in the subset with a given state s.

- *Bayesian network* (Bradford *et al.*, 2006). When two input data, e.g. x_1 and x_2, are known to be dependent, then their contributing factor to $p(s|x_1,\ldots,x_k)$ is no longer $p(x_1|s)p(x_2|s)$, but the joint probability $p(x_1, x_2|s)$.
- *Hidden Markov model* (Friedrich *et al.*, 2006). This method involves a chain of states, such as 'matching to an I position in a multiple sequence alignment', 'matching to an N position', insertion and deletion. Each state can emit an amino acid from the alphabet of 20 or be silent (like in a deletion state). The chain of states is hidden but the chain of amino acids, i.e. the protein sequence, is observed. The hidden Markov model gives the probability, $p(s_i = \mathrm{I}|\mathbf{a})$, that residue i within the protein sequence $\mathbf{a}$ is in the interface state.
- *Conditional random field* (Li *et al.*, 2007). In this method, each position along the protein sequence is assigned a state label, either I or N. Given the protein sequence $\mathbf{a}$, the probability that the state-label sequence is $\mathbf{s}$ takes the form

$$p(\mathbf{s}|\mathbf{a}) \propto \exp[\sum_l \lambda_l \sum_i f_l(s_{i-1},s_i,\mathbf{a}) + \sum_l \mu_l \sum_i g_l(s_i,\mathbf{a})] \tag{3}$$

The feature functions f_l and g_l can be viewed as contributing scores for residue i (and i −1) within sequence $\mathbf{a}$ to be labeled as s_i (and $s_{i\ -\ 1}$). After training, in which the weights λ_l and μ_l are fixed, one predicts the state-label sequence as the one which maximizes $p(\mathbf{s}|\mathbf{a})$.

A clustering process is often applied to screen residues showing strong indications for interface. This serves to weed out isolated residues and to select the most probable cluster(s) of residues as the final prediction.

The different methods are suited for different types of input data; indeed, in their implementations different types of data have been included. This presents opportunities for improving prediction performance through combining results from individual predictors. A few studies (de Vries *et al.*, 2006; Qin and Zhou, 2007a; Sen *et al.*, 2004) have shown that such metamethods hold great promises. A final method worth mentioning is one in which protein–protein poses generated by docking is used to predict interface (Fernandez-Recio *et al.*, 2004). This is the reverse of applying interface prediction to help solving the docking problem, which will be discussed later.

4 EVALUTION OF WEB SERVERS

A number of method developers have set up web servers, allowing users to obtain predicted interface residues by submitting protein structures. The servers also allow for an objective comparison among different methods.

Interface prediction has to fulfill two competing demands. The predictions should cover as many of the real interface residues as possible, but at the same time should predict as few false positives as possible. These two demands are measured by coverage and accuracy, respectively. If the number of real interface residues is RI and among all predictions of interface residues, the numbers of true and false positives are TP and NP, respectively, then coverage is

$$\mathrm{Cov} = \mathrm{TP}/\mathrm{RI} \tag{4}$$

and accuracy is

$$\mathrm{Acc} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{5}$$

In the literature, coverage has been referred to as recall or sensitivity, and accuracy has been referred to as precision or specificity. In some studies, performance assessment has involved non-interface (i.e. negative) predictions. Since the number of real interface residues is usually far outweighted by the number of non-interface residues, such an assessment tends to favor methods which make only a small number of positive predictions and is thus not recommended.

The following web servers are assessed:

- cons-PPISP (http://pipe.scs.fsu.edu/ppisp.html), a method using PSI-Blast sequence profile and solvent accessibility as input to a neural network (Chen and Zhou, 2005).
- Promate (http://bioportal.weizmann.ac.il/promate), a naïve Bayesian method based on properties, such as secondary structure, atom distribution, amino-acid pairing and sequence conservation (Neuvirth *et al.*, 2004).
- PINUP (http://sparks.informatics.iupui.edu/PINUP/), a method based on an empirical scoring function consisting of a side-chain energy term, a term proportional to solvent accessible area, and a term accounting for sequence conservation (Liang *et al.*, 2006).
- PPI-Pred (http://bioinformatics.leeds.ac.uk/ppi-pred), a support vector machine method taking six properties (including surface shape and electrostatic potential) as input (Bradford and Westhead, 2005).
- SPPIDER (http://sppider.cchmc.org/), a neural-network method that includes predicted solvent accessibility as input (Porollo and Meller, 2007).
- Meta-PPISP (http://pipe.scs.fsu.edu/meta-ppisp.html), a meta web server that is built on raw scores from cons-PPISP, Promate and PINUP through linear regression (Qin and Zhou, 2007a).

Two datasets are used for the assessment. The first, Enz35, consists of 35 proteins in the enzyme/inhibitor category of Docking Benchmark 2.0 (Mintseris *et al.*, 2005), after filtering at 35% sequence identity. The second dataset consists of 25 CAPRI targets (http://capri.ebi.ac.uk/). Only unbound structures are used for interface prediction. For the purpose of assessment, the real interface residues, defined as those with cross-interface contacts <5 Å in the native complex, are mapped to the unbound structure through aligning sequences of the bound and unbound structures. Only surface residues (those with >10% of maximal solvent accessibility) are considered in the assessment.

Figure 1a shows the performance comparison of the six web servers on the Enz35 dataset. Overall, predictions are satisfactory. At a coverage of 50%, the accuracies of cons-PPISP, Promate, PINUP, PPIPRED, SPIDDER and meta-PPISP are 36, 38, 48, 27, 33 and 50%, respectively. According to these accuracies, the ranking of the six web servers is PPI-Pred < SPIDDER < cons-PPISP, Promate < PINUP < meta-PPISP. A possible factor in the performance differences is that Enz35 may have been utilized to different extents in developing the methods.

Figure 1b shows the corresponding comparison on the CAPRI targets. Here the performances of the servers are uniformly worse, likely because 8 of the 25 CAPRI targets are antibody–antigen and other immune system complexes and some of the other complexes are much larger than usually encountered in training datasets. At a coverage of 30%, the accuracies of cons-PPISP, Promate, PINUP, PPIPRED, SPIDDER and meta-PPISP are 26, 26, 28, 23, 25 and 31%, respectively. The ranking is identical to what is found on the Enz35 dataset. The 25 CAPRI targets have a total of 8688 surface residues, of which only 1173, or 14%, are located in the interfaces of the native complexes. Even for the more difficult cases presented by the CAPRI targets, the overall success rate of the prediction methods is twice that of random prediction.

5 APPLICATIONS

The methodologies developed for protein interface prediction and the prediction results have applications in a number of important problems.

5.1 DNA-binding sites

The rationale for predicting protein–protein interfaces is also valid for predicting DNA-binding sites on proteins that interact with DNA. The latter problem has been tackled with some of the methods outlined earlier, such as neural network (Ahmad *et al*., 2004; Tjong and Zhou, 2007) and support vector machine (Kuznetsov *et al*., 2006; Wang and Brown, 2006). The characteristics that distinguishes DNA-binding sites from non-DNA-contacting portions of protein surfaces seem to be stronger than distinguishing characteristics of protein–protein interaction sites (Tjong *et al*., 2007). For example, DNA-binding sites are highly enriched in positively charged Arg and Lys residues and depleted in negatively charged Asp and Glu residues. DNA-binding site prediction thus presents an excellent testing ground for different methodologies.

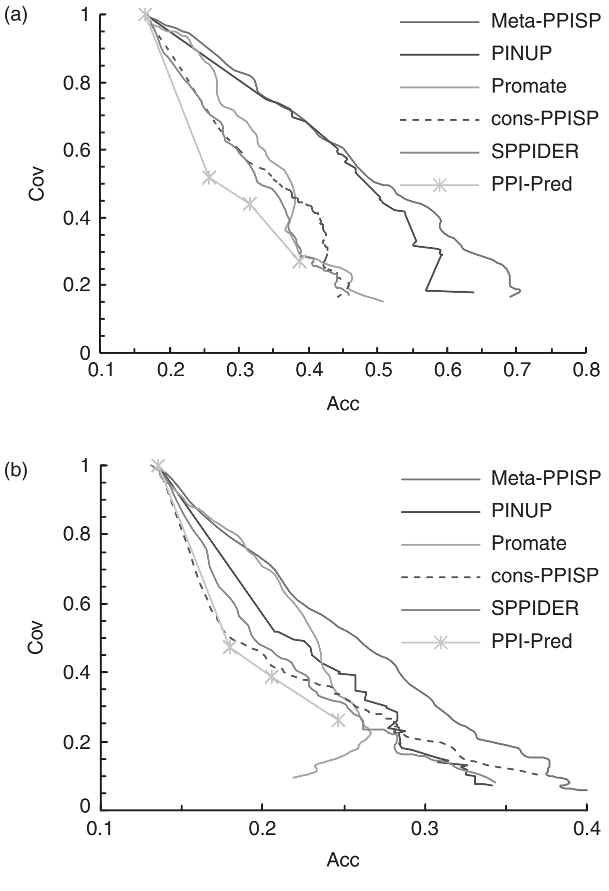

Fig. 1. Comparison of performances among six web servers on (**a**) Enz35 proteins and (**b**) CAPRI targets. Results for cons-PPISP, Promate, PINUP and meta-PPISP are reported previously (Qin and Zhou, 2007a). The two lists of proteins are found at http://pipe.scs.fsu.edu/meta-ppisp_Results.pdf.

5.2 Functional sites

Protein interaction sites have often been used almost interchangeably with functional sites. As seen earlier, in prediction of interaction sites, structural and physical attributes have been used. In prediction of functional sites, emphasis is placed on evolutionary relations, e.g. as presented by phylogenetic trees (Landau *et al*., 2005; Res *et al*., 2005). It is likely that better use of evolutionary relations will improve interface prediction, while incorporation of structural and physical properties will help functional site prediction.

Disease-causing mutations often occur in protein interfaces (Brautigam *et al*., 2006; Eudes *et al*., 2005; Stroud *et al*., 2006; Toomes *et al*., 2004; Zhou, 2004). Interface prediction can thus directly contribute to the understanding of disease mechanisms and help the design of therapeutic agents.

5.3 Docking

Docking refers to the procedure by which the structure of a protein complex is built from the unbound structures of the subunits. Conformational changes, both local and large-scale, present enormous complications. The vast search space for potential poses and the difficulty in ranking docked poses are two major problems. Interface prediction can help in both problems (Qin and Zhou, 2007b). Predicted interface residues can help limit the initial search; this is referred to as front-end use. Alternatively, they can assist scoring the docked poses; this is referred to as back-end use. The roles of interface prediction in docking can be viewed as similar to those of secondary structure prediction in structure prediction for single proteins or protein domains. Corresponding to the prediction of secondary structures from local sequences, interface residues are predicted without reference to the partner protein(s).

Both front-end use and back-end use have been made of interface prediction in docking studies (Chelliah *et al*., 2006; Heuser *et al*., 2005; Tjong *et al*., 2007; Tress *et al*., 2005; van Dijk *et al*., 2005). The advantage of front-end use is that the search space is narrowed at the beginning; the drawback is that docking may be misled by inherently inaccurate interface prediction. On the other hand, after an unbiased search, back-end use allows predicted interface residues to be combined with other scores (such as those based on interaction energy) to rank docked poses, and is thus more tolerant of inaccuracy. The confidence in the predicted interface residues is a factor in deciding whether to make a front-end or back-end use in docking.

As an illustration, Figure 2 presents a comparison of rankings of near-native poses by ZDOCK (Chen *et al.*, 2003) and by simulated interface prediction on 20 CAPRI targets. Simulated interface prediction entails randomly substituting a fraction of the real interface residues by an equal number of non-interface residues. The assumed interface residues are then taken as the benchmark and its coverages by the interfaces in 2000 docked poses, also generated by ZDOCK, are used to rank the poses. Even with 60% substitution (corresponding to 40% coverage and accuracy), the simulated interface prediction can rank near-native poses better than the original ZDOCK scores.

It was found that for 8 of the 20 CAPRI targets, cons-PPISP predictions can improve the rankings of near-native poses (Qin and Zhou, 2007b). In these successful cases, true positives were obtained on both partners forming the native complex. For the failed cases, no true positives were obtained on at least one of the two partners. The message seems to be that a few true positives on both partners, when not overwhelmed by false positives, can help the ranking of near-native poses.

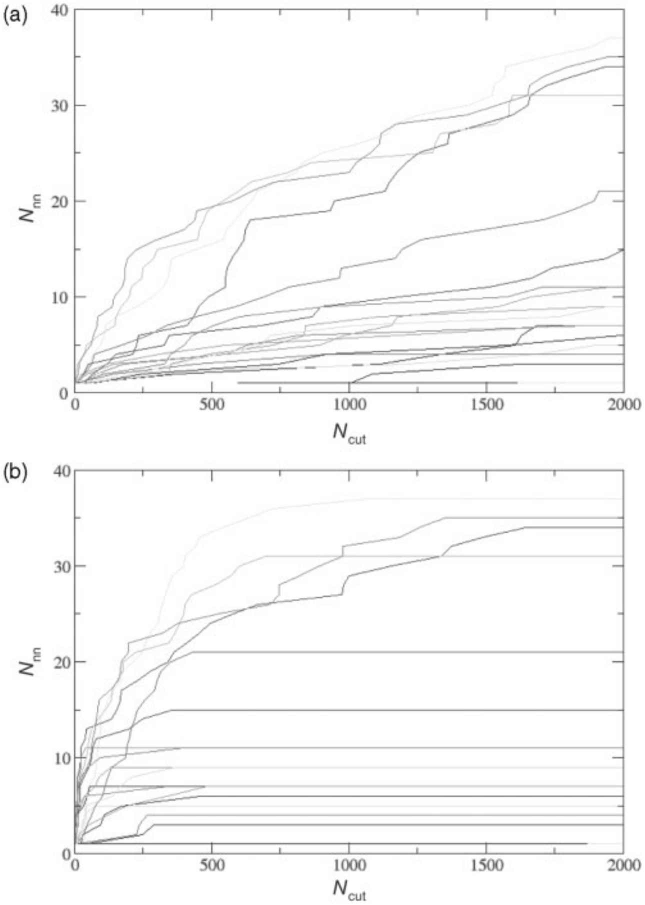

Fig. 2. Number of near-native poses within the first N_{cut} of poses, according to (**a**) ZDOCK scores and (**b**) simulated interface prediction with 40% coverage and accuracy. Near-native poses are those with <10 Å root-mean-square-deviations (L_rmsd) from the native complex. Each of the 20 CAPRI targets is represented by a line with a distinct color. A point with coordinates (N_{cut}, N_{nn}) means that, among the first N_{cut} poses with the highest scores, there are N_{nn} near-native poses. An analogous comparison between ZDOCK and interface prediction by cons-PPISP is found in Qin and Zhou (2007b).

6 CHALLENGES

In the past few years, there have been intense methodological developments in interface prediction. Predictions are now satisfactory for complex-forming proteins that are well represented in the Protein Data Bank (e.g. small to medium enzymes and inhibitors), but less so for under-represented ones (e.g. large proteins). It seems clear that the performance of existing methods will improve as the Protein Data Bank is enriched with structures of protein complexes in the coming years. Meanwhile, a number of challenging problems are vying for the attention from method developers.

- Large-scale conformational changes. As alluded to earlier, interface prediction methods can overcome the complications arising from local-conformational changes. However, large-scale conformational changes, e.g. those involving domain–domain rearrangements, may prove detrimental. For example, interface residues in the native complex may have been scattered in the unbound structures, and thus could be eliminated during the clustering process. It appears that interface prediction and methods that predict domain rearrangements (e.g. based on hinge motion) can help each other in reaching their respective goals.
- One protein, many partners. One protein can interact with many partner proteins and form interfaces involving different parts of its surface. Figure 3 presents the binding patches on the SUMO-conjugating enzyme UBC9 (Reverter and Lima, 2005; Walker *et al.*, 2007). In cases like these, it is possible that the different binding patches are predicted but biochemical data have to be relied upon to identify which is for which partner protein

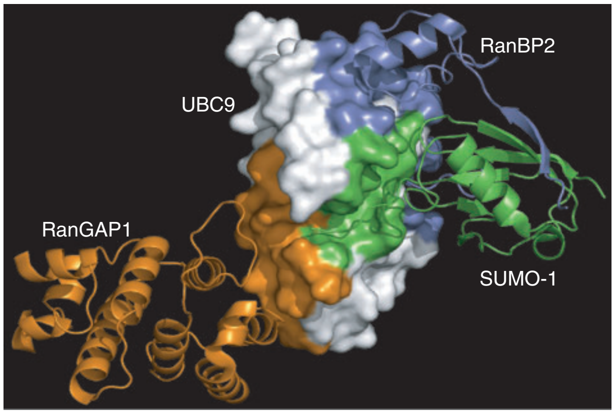

Fig. 3. The interactions of UBC9 with its many partners. The picture is generated using PDB entry 1z5s. The binding patch for RanGAP1 is identical to that for another UBC9 target, E2-25K.

(Qin and Zhou, 2007b). Methods that can predict correlations between interface residues of two partner proteins will be very useful here. For a protein that binds both another protein and DNA, either at the same patch or at two distinct patches, an interesting question is whether methods for predicting protein–protein interfaces and methods for predicting DNA-binding sites can correctly identify the same patch in the former case and the distinct patches in the latter. Some promising results in this regard were found in a recent study (Tjong *et al*., 2007).

- Multi-component complexes. Many protein functions involve supracomplexes formed by more than two components. Sometimes the supracomplex can be viewed as formed by adding one component at a time. In such cases, one can predict the interfaces sequentially (Tjong *et al*., 2007). It remains to be seen how general this strategy is; other strategies are also worth exploring.

Interface prediction has become more and more successful. In combination with other computational tools and experimental data, interface prediction will continue to improve molecular understanding of protein interactions and functions.

ACKNOWLEDGEMENT

This work was supported by NIH grant GM058187.

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al*. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

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.

Bradford,J.R. *et al*. (2006) Insights into protein–protein interfaces using a Bayesian network prediction method. *J. Mol. Biol*., **362**, 365–386.

Brautigam,C.A. *et al*. (2006) Structural insight into interactions between dihydrolipoamide dehydrogenase (E3) and E3 binding protein of human pyruvate dehydrogenase complex. *Structure*, **14**, 611–621.

Burgoyne,N.J. and Jackson,R.M. (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, **22**, 1335–1342.

Chelliah,V. *et al*. (2006) Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *J. Mol. Biol*., **357**, 1669–1682.

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.

Chen,R. *et al*. (2003) ZDOCK: an Initial-stage protein-docking algorithm. *Proteins*, **52**, 80–87.

Chung,J.L. *et al*. (2006) Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins*, **62**, 630–640.

Cole,C. and Warwicker,J. (2002) Side-chain conformational entropy at protein–protein interfaces. *Protein Sci*., **11**, 2860–2870.

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

Crowley,P.B. and Golovin,A. (2005) Cation-π interactions in protein–protein interfaces. *Proteins*, **59**, 231–239.

de Vries,S.J. *et al*. (2006) WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins*, **63**, 479–489.

Eudes,R. *et al*. (2005) Nucleotide binding domains of human CFTR: a structural classification of critical residues and disease-causing mutations. *Cell Mol. Life Sci*., **62**, 2112–2123.

Fariselli,P. *et al*. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem*., **269**, 1356–1361.

Fernandez-Recio,J. *et al*. (2004) Identification of protein–protein interaction sites from docking energy landscapes. *J. Mol. Biol*., **335**, 843–865.

Friedrich,T. *et al*. (2006) Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics*, **22**, 2851–2857.

Heuser,P. *et al*. (2005) Refinement of unbound protein docking studies using biological knowledge. *Proteins*, **61**, 1059–1067.

Hoskins,J. *et al*. (2006) An algorithm for predicting protein–protein interaction sites: abnormally exposed amino acid residues and secondary structure elements. *Protein Sci*., **15**, 1017–1029.

Jones,S. and Thornton,J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol*., **272**, 133–143.

Koike,A. and Takagi,T. (2004) Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel*., **17**, 165–173.

Kufareva,I. *et al*. (2007) PIER: protein interface recognition for structural proteomics. *Proteins*, **67**, 400–417.

Kuznetsov,I.B. *et al*. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, **64**, 19–27.

Landau,M. *et al*. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*., **33**, W299–W302.

Li,J.J. *et al*. (2006) Identifying protein–protein interfacial residues in heterocomplexes using residue conservation scores. *Int. J. Biol. Macromol*., **38**, 241–247.

Li,M.H. *et al*. (2007) Protein–protein interaction site prediction based on conditional random fields. *Bioinformatics*, **23**, 597–604.

Liang,S. *et al*. (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res*., **34**, 3698–3707.

Lichtarge,O. *et al*. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol*., **257**, 342–358.

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

Murakami,Y. and Jones,S. (2006) SHARP2: protein–protein interaction predictions using patch analysis. *Bioinformatics*, **22**, 1794–1795.

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) Predicted protein–protein interaction sites from local sequence information. *FEBS Lett*., **544**, 236–239.

Porollo,A. and Meller,J. (2007) Prediction-based fingerprints of protein–protein interactions. *Proteins*, **66**, 630–645.

Qin,S.B. and Zhou,H.-X. (2007a) meta-PPISP: a meta web server for protein–protein interaction site prediction. *Bioinformatics*, submitted for publication.

Qin,S.B. and Zhou,H.-X. (2007b) A holistic approach to protein docking. *Proteins*, submitted for publication..

Res,I. *et al*. (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Reverter,D. and Lima,C.D. (2005) Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex. *Nature*, **435**, 687–692.

Sen,T.Z. *et al*. (2004) Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics*, **5**, 205.

Stroud,J.C. *et al*. (2006) Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14**, 159–166.

Tjong,H. and Zhou,H.-X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res*., **35**, 1465–1477.

Tjong,H. *et al*. (2007) PI^{2}PE: protein interface/interior prediction engine. *Nucleic Acids Res*., in press.

Toomes,C. *et al*. (2004) Mutations in *LRP5* or *FZD4* underlie the common familial exudative vitreoretinopathy locus on chromosome 11q. *Am. J. Hum. Genet*., **74**, 721–730.

Tress,M. *et al*. (2005) Scoring docking models with evolutionary information. *Proteins*, **60**, 275–280.

van Dijk,A.D.J. *et al*. (2005) Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins*, **60**, 232–238.

Walker,J.R. *et al*. (2007) A novel and unexpected complex between the SUMO-1-conjugating enzyme UBC9 and the ubiquitin-conjugating enzyme E2-25 kDa, to be published.

Wang,B. *et al*. (2006a) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett*., **580**, 380–384.

Wang,B. *et al*. (2006b) Inferring protein–protein interacting sites using residue conservation and evolutionary information. *Protein Pept*. *Lett*., **13**, 999–1005.

Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*., **34**, W243–W248.

Yan,C.H. *et al*. (2004) Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. *Neural Comput*. *Appl*., **13**, 123–129.

Zhang,Y. *et al*. (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl Acad. Sci. USA*, **103**, 2605–2610.

Zhou,H.-X. (2004) Improving the understanding of human genetic diseases through predictions of protein structures and protein–protein interaction sites. *Curr*. *Med*. *Chem*., **11**, 539–549.

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