

ISIS: interaction sites identified from sequence

Yanay Ofra^{1,2,*} and Burkhard Rost^{1,2}

¹CUBIC & North-East Structural Genomics Consortium, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA and ²Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St Nicholas Avenue, Rm 801, New York, NY 10032, USA

ABSTRACT

Motivation: Large-scale experiments reveal pairs of interacting proteins but leave the residues involved in the interactions unknown. These interface residues are essential for understanding the mechanism of interaction and are often desired drug targets. Reliable identification of residues that reside in protein–protein interface typically requires analysis of protein structure. Therefore, for the vast majority of proteins, for which there is no high-resolution structure, there is no effective way of identifying interface residues.

Results: Here we present a machine learning-based method that identifies interacting residues from sequence alone. Although the method is developed using transient protein–protein interfaces from complexes of experimentally known 3D structures, it never explicitly uses 3D information. Instead, we combine predicted structural features with evolutionary information. The strongest predictions of the method reached over 90% accuracy in a cross-validation experiment. Our results suggest that despite the significant diversity in the nature of protein–protein interactions, they all share common basic principles and that these principles are identifiable from sequence alone.

Contact: yanay.ofra@columbia.edu

1 INTRODUCTION

Many interactions but few interaction sites known. Large-scale experiments contribute substantially to unraveling the map of all protein–protein interactions in cells (Gavin *et al.*, 2002; Giot *et al.*, 2003; Ho *et al.*, 2002; Li *et al.*, 2004; Uetz *et al.*, 2000). However, they do not capture the residues that are involved in these interactions. Such information is a key to understanding protein function in detail. Given an experimental three-dimensional (3D) structure it is possible to identify the interface residues. However, for over 97% of all experimentally characterized pairs of interacting proteins, we do not have such high-resolution experimental information. Several studies have gone a step further and showed that even a 3D structure of unbound proteins could suffice for the identification of interface residues (Fariselli *et al.*, 2002; Fernandez-Recio *et al.*, 2005; Jones and Thornton, 1997; Neuvirth *et al.*, 2004). However, for the vast majority of known proteins there is no experimental 3D structure, and for most of them even a high-resolution model is not available. A computational-method that reliably identifies interface residues from sequence could, therefore, be extremely valuable.

Interfaces successfully and specifically predicted. Developing a method that predicts interface residues requires the identification of the biophysical features that enable the interaction. Different studies

have offered different, occasionally contradicting, accounts on the biophysical nature of interface residues. Thus, it has often been assumed that there are no common denominators to interface residues. This assumption has recently been challenged by several large-scale analyses which demonstrated that based on the composition of interacting residues, we can clearly distinguish between different types of interfaces (Jones and Thornton, 1996; Ofra and Rost, 2003a). In particular it has been shown that residue–residue contacts inside a globular domain differ from those between domains; that the interfaces in homo-oligomers differ from those in hetero-oligomers and that interfaces between permanently interacting chains (obligomers) differ from those between transiently interacting ones (oligomers). In fact, interfaces between transiently interacting proteins differ so substantially in their amino acid composition from all other interfaces that predictions from simple sequence features are in principle possible (Ofra and Rost, 2003b). Several studies have recently corroborated our initial hypotheses (Koike and Takagi, 2004; Res *et al.*, 2005; Wang *et al.*, 2005).

Better information–better prediction. We also suggested that using additional information such as that provided in multiple sequence alignment should help (Ofra and Rost, 2003b). This part of our original hypothesis was verified (Res *et al.*, 2005). Several other characteristics may be common to many protein interfaces and thus may enhance predictions. For instance, interface residues likely to be accessible to solvent in the unbound state. We also expect details about secondary structure to be relevant for protein interactions. Therefore, in this study we used evolutionary profiles along with predictions of solvent accessibility and secondary structure (Rost, 2004) to predict whether a residue is likely to be part of a protein–protein interface. Combined with the unique amino acid composition of protein–protein interfaces, these features characterize interface residues and differentiate them from the rest of the protein. We employed sequence analysis and structure prediction tools to elicit these features from sequence. Then, we used them as input for a combination of machine learning algorithms to predict which residues in a sequence are spatially located in protein–protein interface. The data that we used for training and testing were collected from high-resolution 3D structures of protein–protein complexes deposited in the PDB, i.e. Protein Data Bank (Berman *et al.*, 2000).

The vast majority of our predictions was correct for proteins for which we had high-resolution complexes available (test set). Note that none of these test proteins had any significant sequence similarity to any of the proteins used for development (Methods). Although our prediction does not rely on knowing 3D structures such knowledge does improve its performance (data not shown). By default, we use the sequence to predict some structural features

*To whom correspondence should be addressed.

and then use these features for the prediction. Three-dimensional structure was used, however, in developing the method, in order to determine which residues are in the interface for the purpose of forming the testing and training set.

2 METHODS

2.1 Datasets

Dataset of known interactions. For training and testing, we used non-redundant subsets (below) from PDB (Berman *et al.*, 2002; Bernstein *et al.*, 1977). Here, we focused on one interaction type, namely the transient interaction between two non-identical chains of two different proteins. We used a data-mining procedure (Ofran and Rost, 2003a) to differentiate between complexes of transiently interacting proteins and other interactions. Applied to the non-redundant PDB, this procedure yielded 1134 chains in 333 complexes; there were 59 559 contacting residues. A residue was defined to be in a protein–protein interaction if any of its atoms was within 6 Å of any atom in the other protein.

Aligning proteins. First, we aligned all proteins in our dataset with MaxHom (Sander and Schneider, 1991; Schneider and Sander, 1996) against SWISS-PROT (Bairoch and Apweiler, 2000). Then, we built PSI-BLAST (Altschul *et al.*, 1997) profiles using a filtered version of all currently known sequences with three iterations (D. Przybylski and B. Rost, manuscript in preparation). We used these PSI-BLAST profiles both as input to the PROFphd series of methods predicting secondary structure and solvent accessibility (Rost, 2002b) and to the method described here that predicted residues in protein–protein interactions.

Scores for measuring sequence similarity. The simplest way to measure sequence similarity is percentage pairwise sequence identity (PIDE), i.e. the percentage of residues identical between two proteins divided by residues aligned (not counting gaps). The second measure we used was given by the statistical expectation values as reported by BLAST (E-VAL). The third scoring scheme that we used was the HVAL, i.e. the distance from the Sander–Schneider curve (Rost, 1999; Sander and Schneider, 1991):

$$\text{HVAL} = \text{PID} - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32\{1+\exp^{-L/1000}\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad (1)$$

where L was the number of residues aligned between two proteins, PIDE the percentage of pairwise identical residues. An HVAL of 0 defines the line, above which (almost) no two naturally evolved proteins differ grossly in their 3D structures. To illustrate the curve for alignment lengths around 100 residues, 33% pairwise sequence identity suffices to infer structure, above 250 residues 21% is significant and below 11 residues even 100% identity is not enough to infer structural similarity. Although derived to describe structural similarity, HVAL also distinguishes well between proteins of similar and dissimilar function (Nair and Rost, 2002; Rost, 2002a; Rost *et al.*, 2003).

Non-redundant subsets. In order to reduce the bias from too similar sequences in the database, we built sequence-unique subsets for all types of proteins under consideration. ‘Sequence-unique’ was defined such that no pair in the set had an HVAL > 2 [Equation (1)]. Given an all-against-all pairwise alignment for the biased set, we simply used a greedy search to find the largest subset that fulfilled the above condition. Note that while this level suffices to infer some coarse-grained structural similarities in the cores of two proteins, it usually does not suffice to map similar surfaces, and it clearly does not suffice for the homology-inference of protein–protein interactions. A comprehensive analysis of conservation of interactions indicates that when sequence identity is lower than 80% interactions are not conserved.

2.2 Training neural networks

First level prediction. We trained standard feed-forward neural networks with back-propagation and momentum term (Bohr *et al.*, 1988, 1990; Qian and Sejnowski, 1988; Rost and Sander, 1993) on windows of nine consecutive residues. A window was defined as positive, if the central residue had

any atom that was within 6 Å of any atom in a different protein. This yielded a set with 59 559 positive samples. We trained on two-thirds of the data and tested it on the remaining one-third.

Second level refinement filter. Next, we filtered the raw network predictions. Our analysis of protein interfaces at the sequence level suggested that most interacting residues have other interacting residues in their sequence neighborhood (Ofran and Rost, 2003b). Therefore, we eliminated predictions with fewer than seven raw predictions within ten adjacent residues (five on either side).

2.3 Evaluation of performance

Measuring accuracy. We evaluated the performance of our method by its accuracy (number of correctly predicted interface residues/number of predicted interface residues), and coverage (number of correctly predicted interface residues/number of observed interface residues). We also computed the total two-state accuracy (number of correct predictions/number of residues). Note that all estimates were derived for the test set too distant for homology-based predictions (Aloy and Russell, 2002).

Random prediction. To obtain the expected coverage and accuracy at random we reshuffled the predictions in the following way: each protein was represented by two strings of the same length, one representing its sequence and the other representing the predictions (‘P’ for an interacting residue, ‘-’ for a non-interacting residue). Then, we split the prediction string into half and assigned the predictions of the first half of the sequence to the second and vice versa. This process accounted for any size effect that could be caused by the number of predictions and for any effect caused by the heterogeneous distribution of contacting residues along the sequence. Furthermore, it enabled us to find a specific expectation for each scaling of the prediction. We generated different random models for different values of the ROC-like curve (Figure 1). Our background model captured how random our predictions were rather than how well we could predict interface residues at random.

Estimates for accuracy. We divided our dataset of non-redundant complexes into three parts; one we used for training the neural networks, one for deciding when to stop training (cross-training set) and the last to estimate the performance (test or validation set). We rotated around, such that each protein was once used for testing, i.e. we actually trained three different versions of all networks. All our estimates were valid for the testing set in which we used no information about structure to fit any parameter.

3 RESULTS AND DISCUSSION

Assumption: not observed = not existing. One important question for the evaluation of performance is how false positives are treated, i.e. residues that are predicted to be in transient protein–protein interfaces but have not been observed, or more extremely, happen not to be part of the dataset that we chose. Residues that are not observed to physically interact in any particular complex might still interact with other proteins. In fact, many proteins are observed in different interactions often using different interaction sites for different targets (Gavin *et al.*, 2002; Giot *et al.*, 2003; Li *et al.*, 2004; Uetz *et al.*, 2000). Nevertheless, we considered any residue not observed in the given complexes as negatives. This solution was conservative in the sense that it clearly underestimated our performance, at least for the major scores that we reported, namely the accuracy in predicting interaction residues.

Significant improvement in performance. Our first finding was that the raw neural network output was significantly better than random. Since few residues in protein–protein interfaces are isolated (Ofran and Rost, 2003b), we filtered the raw network output by simply omitting isolated predictions. This second step considerably improved the performance of our method (Figure 1). Using different

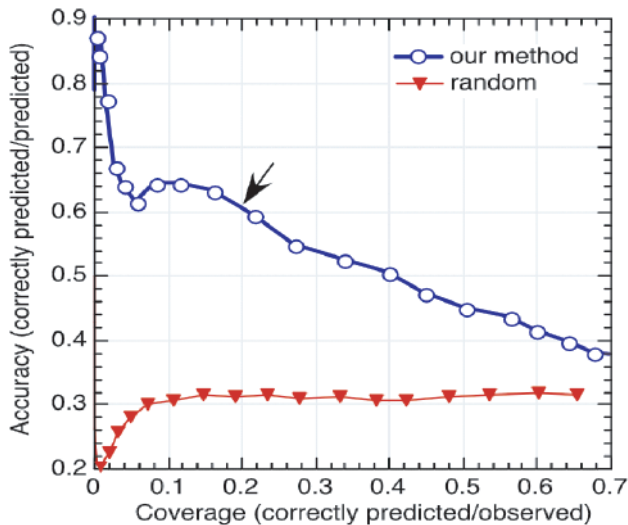


Fig. 1. Accuracy versus coverage for our prediction method (circles) and a random prediction (triangles) using PDB interfaces as gold standard. The data were compiled for a set of proteins that was not used for developing the method. The stronger the confidence in our prediction, the higher the accuracy and the lower the coverage, i.e. when we select the strongest predictions, most of these are right. Around 0.61 accuracy (arrow), our method correctly predicted at least one residue in most of the proteins in our dataset.

confidence thresholds (picking different points in Figure 1) it is possible to increase accuracy (true positives/all positives) at the expense of coverage (true positives/true positives + false negatives). At $\sim 61\%$ accuracy, we found at least one correct interface residue in over 90% of the proteins in our test set. Defining interacting residues based on 3D structures provided enough data to train neural networks and to estimate sustained performance.

Clean comparison to other method impossible. Many methods for the prediction of interaction sites have been introduced (Armon *et al.*, 2001; Fariselli *et al.*, 2002; Fernandez-Recio *et al.*, 2005; Jones and Thornton, 1997; Koike and Takagi, 2004; Neuvirth *et al.*, 2004; Pazos and Valencia, 2002; Res *et al.*, 2005; Wang *et al.*, 2005; Wodak and Mendez, 2004). We wanted to compare the performance of our method to that of the methods that rely exclusively on sequence. Since developers used different datasets, we could not compare our method directly with the performance reported in the literature. Furthermore, not all methods are publicly available and those that are available often use different definitions for interaction sites. Some of the methods attempt to predict a very specific type of interface, while others attempt to identify all functionally important residues. Therefore, to enable a meaningful comparison, we implemented two methods that represent two approaches: (1) using only sequence as input and (2) using a combination of sequence and evolutionary knowledge. We trained these methods on the same dataset and compared their performance with that of our new method.

Predicted structure improves performance. What we could explicitly assess was the importance of combining predicted structural features with evolutionary and other information. Our hypothesis in this study was that this sort combination would significantly improve our ability to distinguish between residues that reside in protein–protein interfaces and residues that do not.

Correlations between such features and binding are typically so subtle that we cannot use simple linear statistics to predict them. Therefore we used artificial neural networks for this task. These networks implicitly yet reliably identified common denominators of protein–protein interfaces. Thereby, we developed a method that predicts residues in protein–protein interfaces for uncharacterized sequences. Analyzing performance on a large scale suggested that at a level of accuracy that corresponds to $\sim 60\text{--}70\%$ of our positive predictions (interacting residues) and to $\sim 98\%$ of our negative predictions (non-interacting) we can identify more than 10% of the residues found in the interface. Note that Figure 1 reports the positive accuracy and coverage, namely, the performance on interface residues in the test set. It does not report the negative values (accuracy and coverage for predicting residues that are not in the interface). The total two-state accuracy of our method (namely the total number of correctly predicted residues, both positive and negative, over the total number of residues) is 0.68.

Many proteins have no homologues. In the past we have shown that amino acid sequence alone suffices to identify some interface residues (Ofra and Rost, 2003b). Several studies elaborated on this notion and demonstrated how adding evolutionary information can improve these predictions. However, one problem of the evolution-based methods is their inability to deal with proteins that have very few or non-known sequence homologues. The percentage of sequences that have no detectable homologue is estimated to be around 30 (Fischer and Eisenberg, 1999). Hence, the applicability of methods that rely solely on homology is limited, particularly when it comes to proteomic-scale analysis. About 5% of the sequences in our test set did not have any sequence homolog in publicly available databases. Conservation-based methods would not be able to analyze these sequences. However, the average total two states accuracy of ISIS for these proteins was 0.59—lower than the accuracy for proteins with an elaborate evolutionary profile but still high compared with other methods (Table 1).

Better two-state performance. Res *et al.* (2005) assessed the performance of different methods based on their total two-state accuracy. Following their footsteps we benchmarked the total two-state accuracy of our method compared with the two other predictors we implemented: one that is based only on sequence and the other that uses evolutionary information. Our results for the latter two methods were virtually identical to those reported by Res *et al.* ISIS, which incorporates predicted structural features, surpasses the sequence or evolutionary-based methods even in the lack of extensive evolutionary profile.

It is important to note that our training set is limited to complexes found in PDB. Thus, the prediction might incur any bias that PDB has. In particular, a significant fraction of the interactions in the cell involves membrane proteins, which are underrepresented in PDB. It is, therefore, hard to assess the performance of the method on membrane proteins. Thus, one could expect that ISIS will perform adequately on the extracellular and intracellular segments of trans-membrane proteins. It is rather unlikely, though, that ISIS could predict successfully interaction sites that are embedded in the membrane.

In summary, we show that even when there is no experimental 3D structure, structural analysis is essential for successful prediction of interaction sites. Better utilization of structure prediction may enhance these predictions even further. These results show, once again, that a combination of all relevant features improves the

Table 1. Two-state accuracy depending on input information used

Method	Total two-state accuracy
Prediction based on sequence alone	0.58
Prediction based on evolutionary conservation	0.6
ISIS	0.68

ISIS is the method that uses all input features described in the Methods section.

performance of a prediction method. More importantly, by not relying on a single feature, ISIS ascertains that good predictions will be available even for proteins that have no known homologues.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Columbia) for computer assistance and to Guy Yachdav (Columbia) for setting up the ISIS Internet server. Special thanks also to Lawrence Shapiro, Wayne Hendrickson, Barry Honig, David Hirsh and Oliver Hobert (all Columbia) for helpful discussions. The work of Y.O. and B.R. was supported by the grants RO1-GM63029-01 and R01-GM64633-01 from the National Institutes of Health (NIH). Last, not least, thanks to all those who maintain excellent databases and to all experimentalists who enabled this work by making their data publicly available.

REFERENCES

- Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
- Altschul,S. *et al.* (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Armon,A. *et al.* (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.
- Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Berman,H.M. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
- Bernstein,F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
- Bohr,H. *et al.* (1988) Protein secondary structure and homology by neural networks. *FEBS Lett.*, **241**, 223–228.
- Bohr,H. *et al.* (1990) A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett.*, **261**, 43–46.
- 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.* (2005) Optimal docking area: a new method for predicting protein–protein interaction sites. *Proteins*, **58**, 134–143.
- Fischer,D. and Eisenberg,D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.
- Gavin,A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
- Giot,L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
- Ho,Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
- Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
- 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.
- Li,S. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
- Mika,S. and Rost,B. (2006) (in press).
- Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Prot. Sci.*, **11**, 2836–2847.
- 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. (2003a) Analysing six types of protein–protein interfaces. *J. Mol. Biol.*, **325**, 377–387.
- Ofran,Y. and Rost,B. (2003b) Predicted protein–protein interaction sites from local sequence information. *FEBS Lett.*, **544**, 236–239.
- Pazos,F. and Valencia,A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.
- Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.
- Res,I. *et al.* (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.
- Rost,B. (1999) Twilight zone of protein sequence alignments. *Prot. Eng.*, **12**, 85–94.
- Rost,B. (2002a) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
- Rost,B. (2002b) Prediction in 1D: secondary structure, membrane helices, and accessibility. In Bourne,P. and Weissig,H. (eds), *Structural Bioinformatics*. John Wiley, pp. 559–588.
- Rost,B. (2004) How to use protein 1D structure predicted by PROFphd. *Meth. Mol. Biol.*
- Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
- Rost,B. *et al.* (2003) Automatic prediction of protein function. *Cell Mol. Life Sci.*, **60**, 2637–2650.
- Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
- Schneider,R. and Sander,C. (1996) The HSSP database of protein structure–sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.
- Uetz,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
- Wang,B. *et al.* (2005) Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.*, **580**, 380–384.
- Wodak,S.J. and Mendez,R. (2004) Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.*, **14**, 242–249.