Predicted protein–protein interaction sites from local sequence information

Yanay Ofran[a,b,c,*], Burkhard Rost[a,b,d]

[a] CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA
[b] Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA
[c] Department of Biomedical Informatics, Columbia University, 630 West 168th Street, New York, NY 10032, USA
[d] North East Structural Genomics Consortium (NESG), Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA

Received 7 March 2003; revised 14 April 2003; accepted 17 April 2003

First published online 13 May 2003

Edited by Robert B. Russell

Abstract Protein–protein interactions are facilitated by a myriad of residue–residue contacts on the interacting proteins. Identifying the site of interaction in the protein is a key for deciphering its functional mechanisms, and is crucial for drug development. Many studies indicate that the compositions of contacting residues are unique. Here, we describe a neural network that identifies protein–protein interfaces from sequence. For the most strongly predicted sites (in 34 of 333 proteins), 94% of the predictions were confirmed experimentally. When 70% of our predictions were right, we correctly predicted at least one interaction site in 20% of the complexes (66/333). These results indicate that the prediction of some interaction sites from sequence alone is possible. Incorporating evolutionary and predicted structural information may improve our method. However, even at this early stage, our tool might already assist wet-lab biology.

Key words: Protein–protein interaction; Neural network; Data mining; Sequence analysis; Protein function; Protein structure; Bioinformatics

1. Introduction

In silico studies of protein–protein interactions pursue two objectives. On the macro level, research focuses on mapping networks of protein interactions [3–9]. On the micro level, the effort concentrates on understanding the mechanisms of interaction and on predicting interaction sites [6,10–16]. With the growth of genome data, an increasing number of computational studies address the first task of delineating all protein networks [17–24]. Very few studies explicitly target the micro level, i.e. focus explicitly on predicting interaction sites from sequence or structure [11,19,25–28]. Identification of interaction sites is critical for a comprehensive understanding of molecular processes, as well as for drug design and quaternary structure prediction. Furthermore, once the residues that interact are identified, it may be simpler to determine with what other protein they interact. Thus, success at the micro level could foster research at the macro level. Gallet et al. have suggested that the hydrophobic moment [29] suffices to predict protein–protein and protein–substrate interaction sites [27]. However, they did not establish the false positive and false negative rates. Pazos and Valencia [19] have introduced an 'in silico two-hybrid system' that attempts both to identify interacting pairs of proteins and to detect the residues that mediate this interaction. Their method is based on the comparison of mutations in pairs of proteins within different genomes; hence, it is somehow confined to sequences that have many available homologues. Other methods use information about three-dimensional (3D) structure to predict interaction sites based on structural features and evolutionary information [28,30].

Why are there so few attempts to predict interaction sites from sequence? Most studies analyze protein–protein interfaces through surface patches that include residues which are non-consecutive in sequence. If the spatial features of surface patches govern the interactions, it would be impossible to predict protein–protein interaction sites from sequence alone. Previously, we showed that residues in interfaces have a significantly different amino acid composition than the rest of the protein [31]. Here, we demonstrate that the majority of interacting residues are clustered in sequence segments of several contacting residues. The combination of these two findings suggests that it might be possible to predict some interaction sites from sequence. To check this hypothesis, we trained neural networks on the largest possible non-redundant dataset of sequence segments for which we have high-resolution data about protein–protein interactions (true positives) and those that are not known to interact externally (true negatives). We found that it is indeed possible to predict many interaction sites from sequence at surprisingly high levels of accuracy.

2. Materials and methods

2.1. Data set

We trained and tested on non-redundant subsets from the Protein Data Bank of experimentally determined 3D structures of proteins (PDB) [1,2]. Interfaces differ between homo- and heteromers, as well as between permanently and transiently interacting peptides [31]. Therefore, we chose to focus on one type of interaction only: the transient interaction between two non-identical chains (protein–protein interactions). We used a data mining procedure [31] to identify complexes of transiently interacting protein chains. 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

*Corresponding author. Fax: (1)-212-305 7932.
E-mail address: ofran@cubic.bioc.columbia.edu (Y. Ofran).

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00456-3

be in a protein–protein interaction if any of its atoms was ≤ 6 Å from any atom of the other protein. As in our previous work [31], we solely applied distance cut-offs to identify contacting residues, i.e. we did not use surface patches. All data are available at http://cubic.bioc.columbia.edu/results/2003/interact_febs/.

2.2. *Prediction method*

We trained standard feed-forward neural networks with back-propagation and momentum term [32,33] on windows of nine residues consecutive in sequence. A window was defined as an interaction site, if the central residue was in contact with a residue in another protein. This yielded a set with 59 559 true positives. We trained on two thirds of the data and tested on the remaining one third. Then, we rotated around, such that each protein was once used for testing, i.e. we actually trained three different versions of all networks. No chain in the test set had an HSSP value smaller than 2 to any chain in the training set [34]. (Note: the HSSP curve [34,35] relates alignment length to pairwise sequence identity in order to establish levels of significant sequence identity; for alignments longer than 250 residues, an HSSP distance +2 implies 22% identical residues.) The neural network has one hidden layer with 189 input, 300 hidden, and two output units (interaction site or not). Next, we filtered the raw network predictions. Our analysis of protein interfaces suggested that most interacting residues have other interacting residues in their sequence neighborhood (Fig. 1). Therefore, we eliminated all isolated raw predictions, i.e. those with fewer than four predicted residues within a window of six adjacent residues (three on either side).

2.3. *Measuring accuracy*

We evaluated the performance of our method by its accuracy (number of correctly predicted protein–protein (p-p) sites/number of predicted p-p sites), and coverage (number of correctly predicted p-p sites/number of observed p-p sites). Note that all estimates were derived for the test set, and that there was no significant pairwise sequence similarity between any protein in the test and training sets that could have enabled homology-based predictions [6,16].

2.4. *Random and simple predictions*

To obtain the expected coverage and accuracy at random we shuffled the predictions and randomly assigned them to the residues in the test set. This process accounts for any size effect that can be caused by the number of predictions. Furthermore, it enabled us to find a specific expectation for each scaling of the prediction. Note that we generated different random models for different values of the ROC curve (Fig. 2). While the random reshuffling established to which extent our predictions could have been reproduced by random, we also tested a simple prediction method to establish that the performance of our method was non-trivial. In particular, we predicted all hydrophobic residues [29] predicted to be exposed by PROFacc [36–38] to be interaction sites.

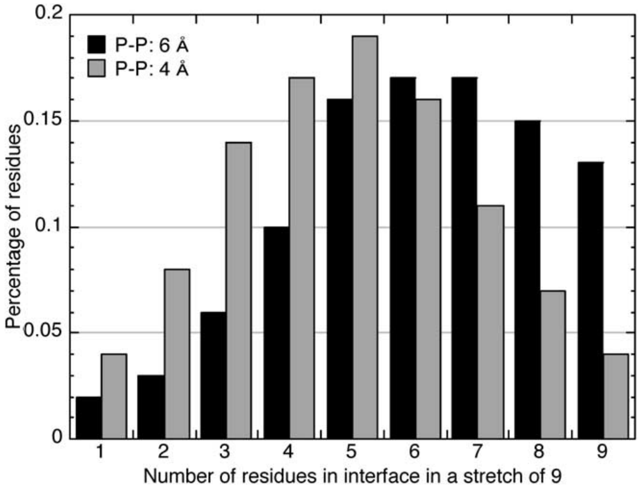

Fig. 1. Distribution of interacting residues in segments of nine consecutive residues. We employed two different distance thresholds to consider a residue involved in protein–protein interfaces, namely when the closest atom pair between two residues in different proteins was closer than 4 (gray) or 6 (black) Å. Although the distribution for the less permissive 4 Å cut-off is moved slightly to shorter segments, both distributions clearly demonstrate that most interface residues have other contacting residues in their sequence neighborhood.

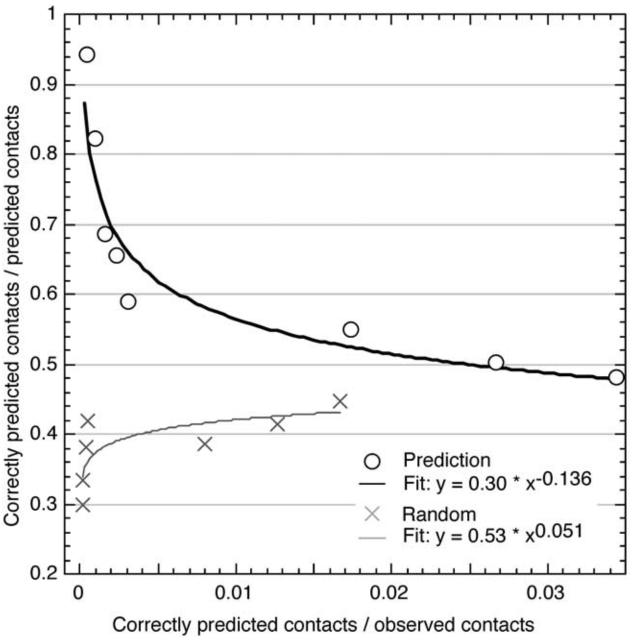

Fig. 2. Significant improvement over random. The random results were obtained as follows. The predictions of the network were scrambled and assigned randomly to the residues in the test set. Then the filtering stage was applied to these ‘predictions’, to reveal any size effect that might result from the distributions of the contacts and the predictions. The number of correctly predicted contacts/number of predicted contacts (accuracy, *y*-axis) represents the fraction of correct positive predictions; the *x*-axis (number of correctly predicted/number of observed contacts) represents the fraction of interacting residues that were correctly predicted as a percentage of all known interactions. The random predictions never reached levels of coverage > 2%, and its accuracy hovered around 0.4. Our method had substantially better accuracy for any level of coverage. Note the accuracy drops significantly if we force the system to detect more than 0.5–1% of all the observed contacts. However, at a level at which we detect at least one interaction site in each protein, 70% of the predictions are correct.

3. Results

3.1. *Local sequence segments contribute substantially to protein–protein interactions*

Are interaction sites formed by residues consecutive in sequence? We found that more than 98% of the protein–protein contacts had at least one interacting residue in their local sequence vicinity, i.e. within four residues N- or C-terminal; 80% had five or more contacts in their neighborhood (black bars in Fig. 1). When applying a more restrictive distance threshold for contacts (≤ 4 Å), we found slightly fewer residues in the sequence neighborhood (gray bars in Fig. 1). However, the majority of nine-mers still contained five or more contacting residues. Combined with the observation that interacting residues tend to have unique compositions [31], this finding suggested that interaction sites are detectable from sequence alone.

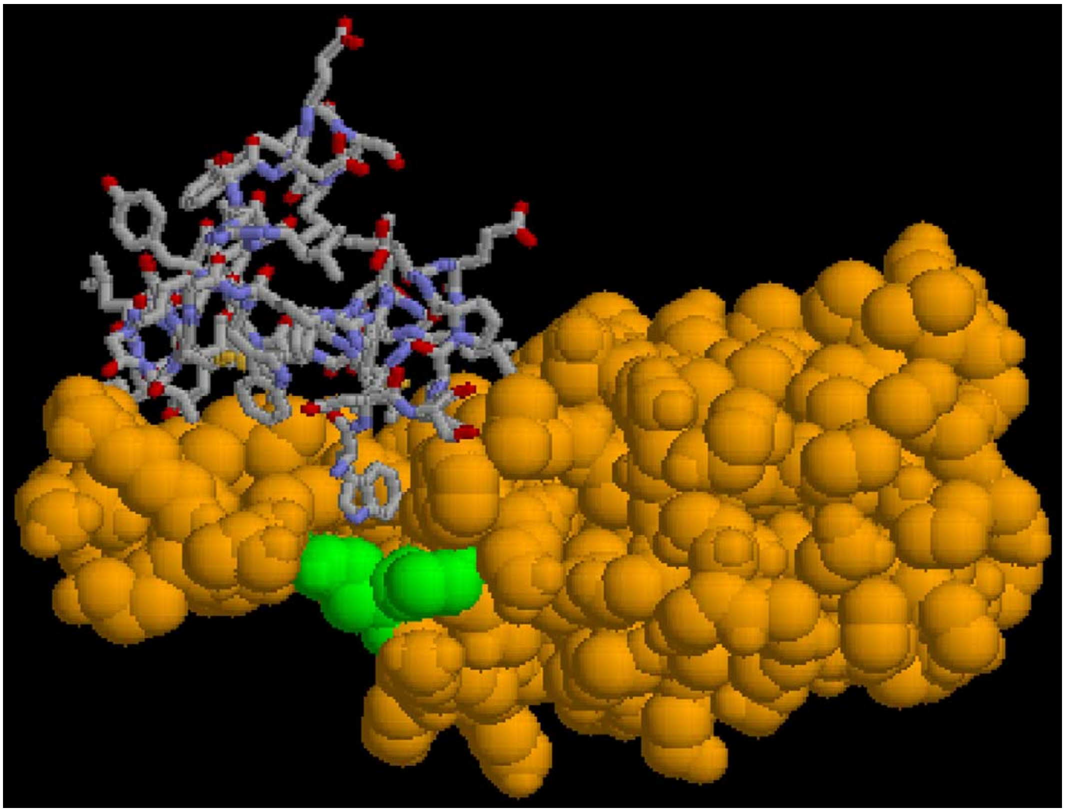

Fig. 3. Example for prediction mapped onto 3D structure. When scaled for highest accuracy (94%), our method correctly identified some contacts in 28 chains; one of these is presented here. The method identified two residues (green) in the ubiquitin ligase skp1–skp2 complex [39]. Both of the predictions are part of a pocket that accommodates the Trp109 in SKP-2 F-box protein. Note that there were no wrong predictions in this complex at the given threshold for the prediction strength.

3.2. *Predictions substantially better than random and simple method*

The first stage raw network predictions were characterized by a high coverage (correctly predicted/observed) at the cost of low accuracy (correctly predicted/predicted): only 20–42% of the predictions were correct (compared to 14–38% that are expected at random), but approximately 30% of the contacting residues were found (22% expected at random). The filtering eliminated many of the wrong predictions. For all levels of prediction strength, our filtered predictions were clearly above random (Fig. 2). We previously found that single residue frequencies contain rather weak preferences for protein–protein interactions. It was then not surprising that a neural network trained on single residues did not outperform the random prediction markedly (data not shown). Interestingly, the simple method that predicted all exposed hydrophobic residues as interaction sites also did not perform much better than random (data not shown).

3.3. *Extreme point: very high specificity for few strong predictions*

When we calibrated our system to the point of its strongest predictions, 94% of the predicted protein–protein interaction sites were correct, i.e. constituted observed contacts according to our definition. At this accuracy, we successfully identified 58 sites from 28 chains in 20 complexes. In these 28 chains, all these strong predictions were correct. The random model yielded 0 correct predictions for the same cut-off. At 70% accuracy, we identified 197 sites (expected at random: 12 residues) from 95 chains in 66 complexes; in 81 of these chains, all predictions were correct.

3.4. *Safely identifying few interaction sites can assist experiments*

Although most interactions involve many residues, single point mutations usually suffice to disrupt the interaction. Therefore, correctly identifying even a single residue in the interface may make it possible to experimentally manipulate that interaction. For instance, the key for interaction in a complex of ubiquitin protein ligase and its substrate recognition component is the insertion of W109 on the F-box protein [39] (gray in Fig. 3) into a pocket of the SKP1 protein (orange). Our method correctly predicted two residues in SKP1 (green), both were in the inner side of the pocket.

4. Discussion and conclusions

4.1. *Interaction sites are mapped to local sequence segments*

The vast majority of residues in protein–protein interfaces were clustered in consecutive, local sequence neighborhoods (Fig. 1). Clearly, there are many long-range effects that influence the interactions and are not expected to be detectable from sequence, yet our results demonstrate that at least in some interaction sites, the local sequence signal suffices for prediction from sequence.

4.2. *Really true negative?*

One of the main problems in assessing predictions of protein–protein interactions is the difficulty of determining the false positive rate. Even if we have high-resolution information about a complex involving an interaction between proteins A and B, it is impossible to rule out that there are additional interaction sites on the surface of A, e.g. for an interaction with another protein C. Therefore, the few existing methods for the prediction of interaction sites refrain from elaborating on their false positive rate. Yet, there is little value in a prediction method that does not assess its false positive rate. A good approximation for the false positive rate of a method can be obtained from the accuracy vs. coverage curves for known complexes. As a rule of thumb, a method that predicts many residues that are not experimentally confirmed is likely to have a high false positive rate. We adopted a rather radical perspective in which we considered everything that was not observed in the 3D structure of the complex as true negatives. Thus, our estimates for accuracy were extremely conservative or pessimistic. Another way of assessing performance is through comparison with random predictions [40]. This can be done by scrambling the predictions and reassigning them randomly to the data points. If the predictions are accurate and have a low false positive rate, we expect them to be substantially better than random. In contrast to the simple prediction method (all exposed hydrophobic residues predicted as interaction sites), our filtered predictions were very impressive by this comparison (Fig. 2).

4.3. Simple method already useful

To compare our predictions to other methods, we implemented the previously published hydrophobic moment predictor [27]. On our data set that method did not outperform our simple prediction method, i.e. 12% of the predictions were correct when 11% of the observed contacts were retrieved. The low coverage at the residue level notwithstanding, analysis of the results at the protein level suggested that even this rather simple method is already useful. The results are encouraging also because of the prospects for the future. Previous works [19,27] laid the conceptual foundations for predicting interaction sites from sequence using evolutionary information or biophysical features. Combining the local sequence signal used by the method introduced here with evolutionary information, energy considerations and structural features may yield progress for this hard problem.

Acknowledgements: Thanks to Rana Samuels (Columbia) for her invaluable comments on the manuscript. Thanks to Jinfeng Liu (Columbia) for computer assistance and to Kaz Wrzeszczynski (Columbia) for crucial insights. Thanks also to the anonymous referee and to Rob Russell (EMBL) for helpful comments. The work was supported by Grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

[1] Bernstein, F.C. et al. (1977) J. Mol. Biol. 112, 535–542.
[2] Berman, H.M., Westbrook, J., Feng, Z., Gillliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucleic Acids Res. 28, 235–242.
[3] Kanehisa, M. and Goto, S. (2000) Nucleic Acids Res. 28, 27–30.
[4] Overbeek, R. et al. (2000) Nucleic Acids Res. 28, 123–125.
[5] Rzhetsky, A. et al. (2000) Bioinformatics 16, 1120–1128.
[6] Aloy, P. and Russell, R.B. (2002) Proc. Natl. Acad. Sci. USA 99, 5896–5901.
[7] Karp, P.D., Paley, S. and Romero, P. (2002) Bioinformatics 18, S225–S232.
[8] Valencia, A. and Pazos, F. (2002) Curr. Opin. Struct. Biol. 12, 368–373.
[9] Bock, J.R. and Gough, D.A. (2003) Bioinformatics 19, 125–134.
[10] Jernigan, R.L. and Bahar, I. (1996) Curr. Opin. Struct. Biol. 6, 195–209.
[11] Jones, S. and Thornton, J.M. (1996) Proc. Natl. Acad. Sci. USA 93, 13–20.
[12] Teichmann, S.A., Murzin, A.G. and Chothia, C. (2001) Curr. Opin. Struct. Biol. 11, 354–363.
[13] Thornton, J.M. (2001) Science 292, 2095–2097.
[14] Valdar, W.S. and Thornton, J.M. (2001) Proteins 42, 108–124.
[15] Sheinerman, F.B. and Honig, B. (2002) J. Mol. Biol. 318, 161–177.
[16] Aloy, P. and Russell, R.B. (2003) Bioinformatics 19, 161–162.
[17] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Science 285, 751–753.
[18] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) Nature 402, 83–86.
[19] Pazos, F. and Valencia, A. (2002) Proteins 47, 219–227.
[20] Pazos, F. and Valencia, A. (2001) Protein Eng. 14, 609–614.
[21] Sprinzak, E. and Margalit, H. (2001) J. Mol. Biol. 311, 681–692.
[22] Rzhetsky, A. and Gomez, S.M. (2001) Bioinformatics 17, 988–996.
[23] Gomez, S.M. and Rzhetsky, A. (2002) Pacific Symposium on Biocomputing, pp. 413–424.
[24] Lu, L., Lu, H. and Skolnick, J. (2002) Proteins 49, 350–364.
[25] Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (1997) J. Mol. Biol. 271, 511–523.
[26] Jones, S. and Thornton, J.M. (1997) J. Mol. Biol. 272, 133–143.
[27] Gallet, X., Charloteaux, B., Thomas, A. and Brasseur, R. (2000) J. Mol. Biol. 302, 917–926.
[28] Fariselli, P., Pazos, F., Valencia, A. and Casadio, R. (2002) Eur. J. Biochem. 269, 1356–1361.
[29] Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1982) Nature 299, 371–374.
[30] Armon, A., Graur, D. and Ben-Tal, N. (2001) J. Mol. Biol. 307, 447–463.
[31] Ofran, Y. and Rost, B. (2003) J. Mol. Biol. 325, 377–387.
[32] Bohr, H., Bohr, J., Brunak, S., Fredholm, H., Lautrup, B. and Petersen, S.B. (1990) FEBS Lett. 261, 43–46.
[33] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[34] Rost, B. (1999) Protein Eng. 12, 85–94.
[35] Sander, C. and Schneider, R. (1991) Proteins 9, 56–68.
[36] Rost, B. and Sander, C. (1994) Proteins 20, 216–226.
[37] Rost, B. (1996) Methods Enzymol. 266, 525–539.
[38] Rost, B. (2001) J. Struct. Biol. 134, 204–218.
[39] Schulman, B.A. et al. (2000) Nature 408, 381–386.
[40] Goebel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Proteins 18, 309–317.