Scoring Docking Models With Evolutionary Information

Michael Tress,* David de Juan, Osvaldo Graña, Manuel J. Gómez,† Paulino Gómez-Puertas,‡ Jose M. González, Gonzalo López, and Alfonso Valencia
Protein Design Group, CNB-CSIC, Cantoblanco, Madrid, Spain

ABSTRACT **We have developed methods for the extraction of evolutionary information from multiple sequence alignments for use in the study of the evolution of protein interaction networks and in the prediction of protein interaction. For Rounds 3, 4, and 5 of the CAPRI experiment, we used scores derived from the analysis of multiple sequence alignments to submit predictions for 7 of the 12 targets. Our docking models were generated with Hex and GRAMM, but all our predictions were selected using methods based on multiple sequence alignments and on the available experimental evidence. With this approach, we were able to predict acceptable level models for 4 of the targets, and for a fifth target, we located the residues involved in the binding surface. Here we detail our successes and highlight several of the limitations and problems that we faced while dealing with particular docking cases. Proteins 2005;60:275–280.** © 2005 Wiley-Liss, Inc.

Key words: CAPRI; blind test; protein docking; multiple sequence alignments; correlated mutations; tree determinants; conserved regions; interacting surface patches

INTRODUCTION

Protein interactions form the basis of all cellular processes, and the study of protein interactions is fundamental to our understanding of cell systems. Experimental approaches to protein–protein interaction prediction, such as the large-scale application of proteomics methods, are providing overviews of complete protein interaction networks.[1] Obtaining detailed structural information about these complexes is important for understanding the basic biochemical processes, but unfortunately the number of structurally characterized complexes that have been deposited in Protein Data Bank[2] (PDB) is still rather small, something that makes obvious the need for docking algorithms that can reproduce the physical interaction of proteins in protein complexes.

The main interest of our group is the extraction of evolutionary information from multiple sequence alignments and its use for predict interactions within protein interaction networks (see review[3]). Our involvement in the CAPRI experiment is a logical extension of this work.

Since we have been developing methods based on the extraction of evolutionary information for the prediction of protein interactions, we are interested in knowing how these sequence-based methods will function as a means of scoring of protein docking predictions. If evolutionary information can be used to successfully predict interaction partners, perhaps the same information can be employed to help in the prediction of physical regions of interaction between proteins known to interact.

One obvious possibility would be to include residues that may be conserved from an evolutionary point of view. Although it is generally accepted that functionally important regions, such as interaction surfaces, tend to be marked by conserved residues, these residues may only form a small part of interaction surfaces for various reasons.

In some cases, it is possible to observe positions in multiple sequence alignments that are specifically conserved in subfamilies, often called tree determinants.[4] Given a protein family and its subfamilies, tree determinants can point to sequence changes that occurred during family divergence and imply functional changes that have been maintained during evolution. Tree-determinant residues are indicative of the presence of sites of functional specificity and can coincide with specific protein–protein interaction sites that are of functional importance for determining the specificity of interaction.[5–8] In some cases, it has even been possible to use the information derived from these methods to change the binding specificity between proteins by manipulating a few residues in the corresponding interfaces (see review[9]).

Another computational approach that makes use of evolutionary information is based on the concept of inter-protein correlated mutations.[10,11] In the interaction between orthologous pairs of proteins in several species, we can expect to find differences in the way in which mutations affecting the interaction surface in one of the partners have been compensated by mutations in the interaction surface of the other partner. It has been demonstrated that identification of correlated mutations can be used to predict the tendency of pairs of residues to be in physical proximity,[11] and in some cases to predict models of interaction that have been confirmed by experiments.[12]

Grant sponsor: TEMBLOR; Grant number: QLRI-CT-2001-00015. Grant sponsor: Biosapiens; Grant number: LSHG-CT-2003-503265. Grant sponsor: and GeneFun; Grant number: LSHG-CT-2004-503567.

†Present Address: Bioinfo Lab, CAB-INTA, Torrejón de Ardoz, Madrid, Spain

‡Present Address: CBM-CSIC, UAM, Cantoblanco, Madrid, Spain

*Correspondence to: Michael Tress, Protein Design Group, CNB-CSIC, Cantoblanco, Madrid, Spain. E-mail: mtress@cnb.uam.es

Received 14 January 2005; Accepted 3 February 2005
DOI: 10.1002/prot.20570

© 2005 WILEY-LISS, INC.

The training of neural networks and other artificial intelligence methods with examples of known complexes and their corresponding alignments is a different avenue for the extraction of information from multiple sequence alignments. Two of the first methods[13,14] trained neural networks with sets of complexes deposited in the PDB and information from the corresponding multiple sequence alignments of the corresponding protein families (i.e., sequence profiles). They represented interaction surfaces as surface patches of neighboring residues. A number of improvements and variants of these methods have been recently proposed.[15,16]

In common docking approaches, the final stage involves the evaluation of the solutions by human experts. We propose to use multiple sequence alignment–based approaches as filters of the docking models at this level, by evaluating the overlap between the interaction surfaces as defined by the docking model using the methods described above—conserved and tree-determinant residues, correlated mutations, and predicted interacting surface patches.

METHODS

General Procedure

The first step in the analysis was to study the literature for a list of biologically relevant residues and docking sites. These may be residues for which there is experimental evidence that they are involved in the docking, or equally important residues for which there is biological evidence against their involvement in the docking surface.

Evolutionarily related sequences were collected for each of the interacting chains with BLAST.[17] CLUSTALW[18] was used to construct multiple sequence alignments. In some cases, the alignment required manual intervention or the removal of some of the loosely related sequences, followed by additional alignments with CLUSTALW or T-COFFEE.[19] Furthermore, where possible, we took care to avoid including sequences that had not conserved the particular interaction under study.

From the final multiple sequence alignments we calculated the evolutionary information following the methods described below.

Conserved Regions

For CAPRI Rounds 4 and 5 we included a method that calculates stable regions in multiple sequence alignments.[20,21] The method uses PSI-BLAST profiles and a smoothing function to evaluate residue positions in query-profile alignments.

Tree Determinants

Tree-determinant residues were calculated using the sequence space and mutational behavior procedures as described in del Sol Mesa et al.[4] These methods both extract family specific residues but rely on different principles. The mutational behavior method detects sequence positions based on their parallel evolution with respect to the whole sequence. This is calculated as a rank-order correlation coefficient for the whole multiple sequence alignment:

$$r = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2} \sqrt{\sum_j (S_j - \bar{S})^2}},$$

where R_i and S_i are the rank-order values of the matrix elements belonging to the protein change matrix and position change matrix, respectively. $\bar{R}$ and $\bar{S}$ are their average values.

In the case of the sequence space method, residues are detected by performing a principal component analysis of the alignment. The points representing the protein sequences in this space are automatically clustered into protein families, and the vectors describing these clusters are used to assign residues to each of the clusters.

For the mutational behavior method, residues were selected as tree-determinant if the scores were higher than 0.9, while for the sequence space method, residues were selected if the residues were completely conserved at least in the protein cluster containing the target sequence, if and only if this position was not conserved among all the protein clusters.

Correlated Mutations

The results were calculated as in Gobel et al.[10] and Pazos et al.[11] using an alignment constructed by concatenating sequences from the same organism that are closely homologous to the corresponding target interacting proteins. For this method to work, at least 6 concatenated sequences are required. The Spearman correlation coefficient was calculated as shown above for the mutational behavior method but using actual values instead of ranks. Pairs with correlation coefficients higher than 0.9 were selected. The application of the correlated mutation method is restricted to targets in which the 2 proteins have coevolved in various species. For this reason, the docking of antigen–antibody and many enzyme–inhibitor complexes is beyond the scope of application of correlated mutations.

For those cases without enough sequences from the same organisms, intrasequence correlated mutations were calculated separately for each of the interacting proteins, and those corresponding to exposed residues were used as low-quality indicators of potential interacting sites.

Prediction of Interaction Patches

The method described by Fariselli et al.[14] was used to predict surface interacting patches. This method is based in the ability of a supervised neural network system to capture those features of interaction patches after an adequate training procedure. This training was performed with multiple sequence alignment–derived information for a set of nonredundant patches of surface residues. After the training, the neural network is able to provide predictions for individual residues with an estimated 73% accuracy. For this work, reliability scores ≥ 0.5 were considered informative.

Scoring Models

The various lists of interesting residues were filtered by cross-checking with the experimental evidence and by structural criteria (solvent accessibility).

For each case, a broad list of approximately 10,000 rough models were obtained by running GRAMM[22] with very permissive parameters. This set of decoys was sorted by taking into account the list of interesting residues. At this stage, docking models that did not fit to the experimental information were rejected. The best 10 GRAMM models were additionally expanded using the program Hex[23] in order to improve "soft" GRAMM results. In most cases, we submitted models from both GRAMM and Hex for comparison purposes. However, we found that there was a tendency for the rough models produced by GRAMM to have too many clashes for the experiment.

The combined set of GRAMM and Hex solutions was ranked by the proximity of interesting residues in the docking solutions. This evaluation is done using the Xd formula, as described by Pazos et al.,[11] which gives an estimate of the global proximity of the predicted residues to the interaction surface in the form of the weighted harmonic distances.

Positive Xd indicates cases for which the population of predicted residues is closer together relative to the entire population of residues. The correct docking models should have higher values of Xd than the incorrect decoys, since we are expecting the predicted residues to be closer together.

RESULTS

Target 9—LicT Homodimer

LicT is a transcriptional antiterminator protein that regulates expression of *Bacillus subtilis* operons involved in β-glucosides metabolism.

Each monomer in the homodimer contains an RNA-binding domain (CAT) and 2 phosphoenolpyruvate:sugar phosphotransferase system (PTS) regulation domains (PRD). These are phosphorylated at conserved histidines when the substrate is available. In the LicT activated form, PRD1 H100 and H159 are dephosphorylated, while in PRD2, H207 and H269 are phosphorylated.

In this case, the initial docking models were generated with Hex. The exploration space was restricted to a maximum rotation angle of 15° and 4 Å root-mean-square deviation (RMSD) from the starting orientation. We derived conserved positions and tree determinants from the 23 sequences in the corresponding family alignments, the accessible residues from Dictionary of Protein Secondary Structure (DSSP), and sites of potential protein–protein interactions from the neural network. Selection of the final models was done according to the Xd values of accessible, conserved, and tree-determinant residues and the visual inspection of the models (see Fig. 1).

The real structure reveals that only PRD1 domains interact. The PRD2 domains are flipped, so the conserved histidines are completely accessible to phosphorylating enzymes. This intramonomer structural change between phosphorylated and dephosphorylated states would be difficult to predict even with flexible docking.

For the best model, about 35% of interface residues were correctly predicted (55% if only PRD1 domains are taken into account). The model with the lowest local RMSD had the highest Xd, indicating that, in this case, the sequence information was informative for the selection of the best docking solution in our set of decoys. The native docking solution had equally high Xd [see Fig. 1(C)], suggesting that Xd would have been useful in selecting the native confirmation if it had been present in our starting set.

Targets 11 and 12—Dockerin and Cohesin

Cohesin domains are part of the cellulolytic scaffoldin complex in the bacterium *Clostridium thermocellum*. There are 9 cohesin domains in each scaffoldin complex. The cohesin is responsible for binding the dockerin domain. The cohesin–dockerin binding is a crucial part of formation of the complex. The dockerin domain is part of the cellulytic enzyme structure and is homologous to the EF-hand calcium-binding loop.

For the docking of Target 11, the cohesin structure was unbound and the dockerin had to be generated as a homology model from a PDB structural template that had 50% sequence identity. However, for the small dockerin chain, there was experimental evidence to suggest that the interacting residues were to be found in both parallel helices. We were not able to extract much sequence-based information for this chain, because dockerin is essentially 2 short, 20-residue repeats where residue signals are masked by the strong signals from the calcium-binding residues and the structural residues integral to holding the dockerin together.

Despite this, 8 of our predictions were acceptably close to the correct orientation (see Fig. 2), with our fourth model having a *fnat* (fraction of predicted contacts over native) of 0.26.

However, we were distracted by experimental information that implicated residues 11, 12, 45, and 46 in the binding. As a result, we chose models where all these residues were close to the interacting surface and these models limited the orientation of the bound dockerin. In fact only 45 and 46 are in contact in the native structure, though the internal symmetry of dockerin suggests that it might conceivably bind in 2 orientations.

Our predictions did improve when the bound dockerin was introduced in Target 12, although the model that we submitted directly from GRAMM that had an 0.44 *fnat* and 2.06 Å local RMSD had a few too many clashes.

Target 18—Xylanase and Inhibitor

Target 18 was a complex between xylanase from *Aspergillus niger* and an inhibitor, *Triticum aestivum* xylanase inhibitor-I (TAXI). While the binding of the inhibitor might have come about from evolutionary pressures, it was not be possible to say the same for the xylanase. Fortunately the experimental evidence suggested that the binding was competitive, so we knew that TAXI must somehow block the binding site.

One further problem that came up was that while there were enough sequences to make calculations for sequence-

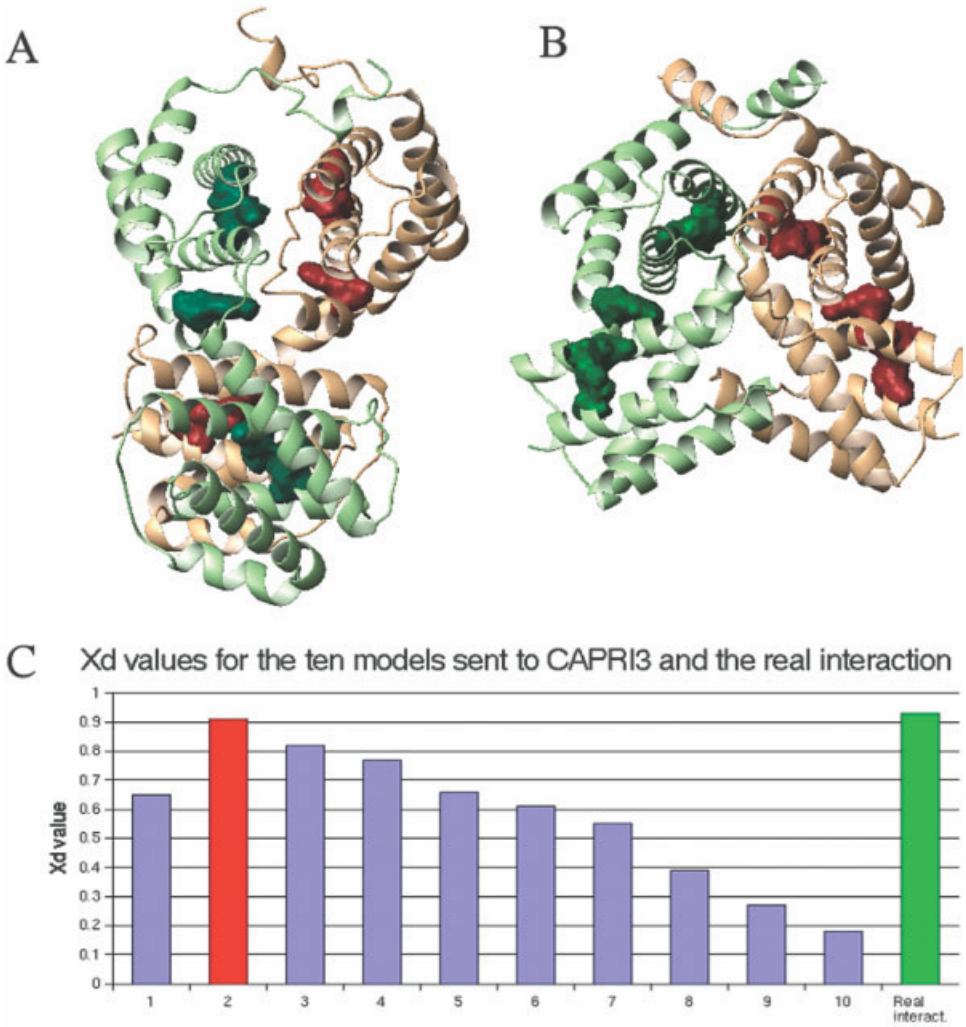

Fig. 1. Model 2 for Target 9. (**A**) Our best-ranked model showing conserved histidines in spacefill. (**B**) The native LicT homodimer structure showing the position of the conserved histidines. (**C**) A graph showing Xd values for the 10 models that we sent compared to the Xd for the real interaction (in green). The red bar corresponds to the model with the best *fnat*.

based residues for TAXI, the sequences we found were heavily biased toward the evolutionarily related aspartic proteases: There were only 4 sequences that probably functioned as inhibitors. Nevertheless, we were able to predict correctly that 3 loops on the surface of TAXI were the most likely to be involved in the binding of the xylanase (see Fig. 3).

While we correctly predicted the binding surface of TAXI, however, we were not able to orientate the enzyme correctly in the 3 predictions submitted, although one of the models did have the best overall RMSD.

CONCLUSIONS

Our methods, while limited by the requirement of a common evolutionary history to a subset of targets, are able to predict interacting regions in docking targets some of the time. However, although we have used multiple sequence alignment–based evaluation schemes to predict a high proportion of "acceptable" docking models, we were not able to use exclusively sequence-based methods to predict models classed as "good" and higher. In part this was due to factors beyond our control. For example, for the targets in CAPRI Round 3, we did not have a sufficiently high model sample space, and for Targets 10 and 14, our rough docking strategy was not able to provide docking solutions that fit the necessary experimental criteria.

There were substantial conformational changes associated with complex formation for several targets, while with Targets 11 and 12, we were deceived by experimental evidence that in the end did not fit with the orientation of the native complex. The solutions for these complexes emphasized the variability and complexity of the biology associated with the CAPRI docking targets and showed that care must be taken even when interpreting experimentally derived information.

For the remaining target (Target 18), we were able to correctly predict the binding surface of the inhibitor, but without experimental evidence, we were forced to make

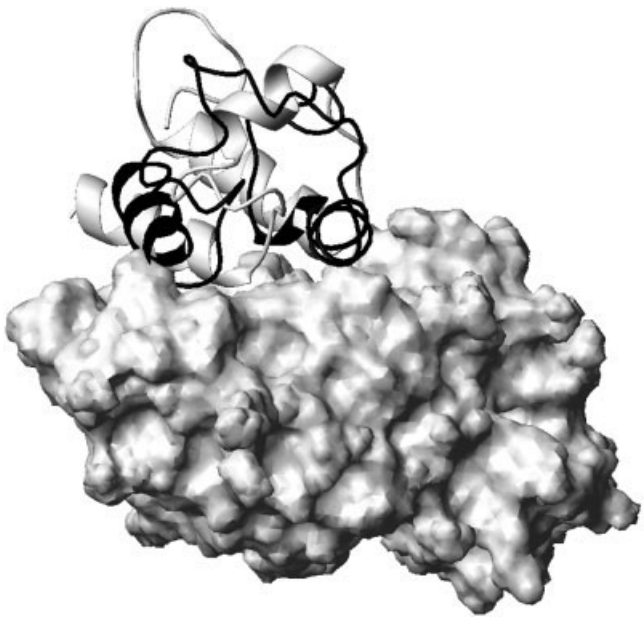

Fig. 2. Target 11. Superposition of model with the native structure. The cohesin chain is below in spacefill, while the relative orientation of the model (black) and native (white) dockerins are shown above.

Fig. 3. Native structure for Target 18 (pdb code: 1T6G). Native structure of the xylanase (in black)–TAXI (gray) complex with the residues predicted to be interacting in the TAXI protein shown in spacefill.

guesses as to which part of the enzyme active site the inhibitor would block.

Despite the difficulties we encountered, the results obtained were in agreement with the scope of our approach, which was focused on extracting valuable docking models from a set of low-quality decoys.

It is clear that our overall approach can be improved. One of the more obvious drawbacks of our approaches is the limited size of the sample space we are working with. As docking methodologies become more widely available, it will become possible to employ a range of docking methods in order to broaden the sample space available to us. In addition, the continued growth of the sequence databases ought to improve the number and distribution of sequences available for each target, something that can only increase the capacity of our sequence-based approaches.

REFERENCES

1. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM. Protein interaction networks from yeast to human. Curr Opin Struct Biol 2004;14:292–299.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
3. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.
4. del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. J Mol Biol 2003;326: 1289–1302.
5. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
6. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol 1995;2:171–178.
7. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
8. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 2002;316:139–154.
9. López-Romero P, Gómez MJ, Gómez-Puertas P, Valencia A. Prediction of functional sites in proteins by evolutionary methods. In: Kamp RM, Calvete JJ, Choli-Papadopoulou, T, editors. Methods in proteome and protein analysis. Berlin/Heidelberg: Springer-Verlag; 2004. p 319–336.
10. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18: 309–317.
11. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein–protein interaction. J Mol Biol 1997;271:511–523.
12. Carettoni D, Gómez-Puertas P, Yim L, Mingorance J, Massidda O, Vicente M, Valencia A, Domenici E, Anderluzzi D. Phage-display and correlated mutations identify an essential region of subdomain 1C involved in homodimerization of *Escherichia coli* FtsA. Proteins 2003;50:192–206.
13. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
14. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
15. Yan C, Dobbs D, Honavar V. A two stage classifier for identification of protein–protein interface residues. Bioinformatics 2004;20: 371–378.
16. Bradford JR, Westhead, DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–1494.
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller

W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
18. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
19. Notredame C, Higgins D, Heringa J. T-Coffee: A novel method for multiple sequence alignments. J Mol Biol 2000;302;205–217.
20. Tress ML, Jones DT, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 2003;330: 705–718.
21. Tress ML, Graña O, Valencia A. SQUARE—determining reliable regions in sequence alignments. Bioinformatics 2004;20: 974–975.
22. Vakser I. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins 1997;Suppl 1:226–230.
23. Ritchie DW, Kemp GJL. Protein docking using spherical polar Fourier correlations. Proteins 2000;39:178–194.