SUCCESSFUL DISCRIMINATION OF PROTEIN INTERACTIONS

Carlos J. Camacho and David W. Gatchell

Department of Biomedical Engineering, Boston University,

44 Cummington Street, Boston, MA 02215

Short title: Prediction of protein interactions

Keywords: Docking, Desolvation, clusters, CAPRI, decoys, SmoothDock

Corresponding Author: Carlos J. Camacho

Department of Biomedical Engineering

Boston University

44 Cummington Street

Boston, MA 02215

E-mail: ccamacho@bu.edu

Phone/Fax: (617) 353-4842 ; (617) 353-6766

ABSTRACT

We present results from the prediction of protein complexes associated with the first Critical Assessment of PRediction of Interactions (CAPRI) experiment. Our algorithm, *SmoothDock*, comprises four steps: first, we perform rigid-body docking using the program DOT, keeping the top 20,000 structures as ranked by surface complementarity; second, we re-rank these structures according to a free energy estimate that includes both desolvation and electrostatics and retain the top 2,000 complexes; third, we cluster the filtered complexes using a pairwise RMS deviation criterion; finally, the twenty-five largest clusters are subject to a smooth docking discrimination algorithm where van der Waals forces are taken into account. We predicted targets 1, 6 and 7 with RMS deviations of 9.5, 2.4 and 2.6 Å, respectively. More importantly, from the perspective of biological applications, our approach consistently ranked the correct model first (i.e., with highest confidence). For target 5 we identified the binding region but not the correct orientation. Although we were able to find reasonable clusters for all targets, low affinity complexes (K_d < nM) were harder to discriminate. For 4 out of 7 targets, the top models predicted by our automated procedure were among the best community-wide predictions.

INTRODUCTION

Finding physical interactions between proteins involved in common cellular functions is one of the most important problems in biology. The first international Critical Assessment of Prediction of Interactions (CAPRI) experiment[1] was designed to evaluate current computational approaches that address this critical problem. Most methods used involved protein docking algorithms whose goals are to obtain models for the bound complex from the coordinates of the component molecules.

Current docking methods evaluate a vast number of docked conformations by simple functions that measure single component correlation functions[2,3]. However, in addition to near-native states, these methods produce many false positives, i.e., structures with good scores but high root-mean-square-deviations (RMSDs). Using scoring functions that better account for the chemical affinity between the individual molecules[2] and refining the interacting surfaces[4,5], conformations with RMSDs less than 10 Å are generally found within the top tens of structures, though the highest ranked complexes are often far from the native structure. Therefore, perhaps the most difficult challenge left in protein docking is the ability to discriminate native-like structures from these remaining false positives.

Discrimination among decoys of protein complexes is very difficult given the high sensitivity of the scoring functions to small side chain and backbone displacements. In Camacho and Vajda[6], we described a novel method that yields a more physically meaningful free energy ranking of decoys. Specifically, the decoys are first clustered using a pairwise RMSD criterion[7]. Then, each cluster is minimized according with a free energy target function that attempts to mimic the driving forces of

the binding process at different length scales[6,7]. In a test set of eight independently crystallized receptor/ligand structures this method was able to refine complexes that were around 10 Å away from the native complex to 2 Å RMSD by optimizing the free energies around each cluster[7]. This multi-cluster refinement procedure allows us to compare not individual decoys but the average free energies (see Eq. 1) of the optimized clusters. We have argued that this average free energy should be a better estimate of the potential of mean force, and therefore a better discriminator of the binding sites with highest affinity[7].

Here, we report our results on the first blind experiment of prediction of protein interactions (CAPRI). We obtained very good results for four of the seven targets, obtaining not one, but at least two top community-wide models within the five submissions allowed for each target. Our automated platform involves four steps: (a) rigid-body docking; (b) filtering decoys; (c) clustering decoys; and, (d) refinement and discrimination of native-like clusters.

MATERIALS AND METHODS

The scheme used to predict complexes for round 1 (3 targets) and round 2 (4 targets) of the CAPRI experiment is outlined below. In what follows, we briefly describe the four steps of our docking algorithm *SmoothDock*. A more elaborate description of the methodology is published in another article in this issue (see, e.g., Ref. 7).

Step 1: Rigid-body docking using the Fast-Fourier Transform (FFT) based program DOT [8,9] was performed for each receptor/ligand target. The output of this program was the top 20,000 receptor/ligand complexes sampled by the DOT program and ranked according to surface complementarity. Any experimental constraint on the binding area was also imposed here.

Step 2: Following the procedure detailed elsewhere[4,7], for each complex we computed the effective desolvation and electrostatic binding affinity between receptor and ligand. We then filtered the 500 best desolvation energy[10] and 1,500 best electrostatic energy[4] complexes for a total of 2,000 complex candidates.

Step 3: We clustered the filtered complexes using a pairwise RMS deviation criterion, and retained the twenty-five clusters with the highest number of neighbors[7]. For targets 1-3, the complexes were clustered using an all C_α RMSD criterion and a 10 Å cutoff, whereas for targets 4-7, we utilized a C_α binding site RMSD criterion and a cutoff radius of 7 Å. All clustering was done in a hierarchical manner such that no overlaps occurred between distinct clusters.

Step 4: Using 10 representative structures from each cluster, the smooth docking algorithm described in Camacho et al.[6] was used to optimize our free energy function around each cluster. We submitted the top ranked complexes from those clusters that converged to the lowest free energies as estimated by Equation 1:

$$\Delta G = E_{elec} + E_{desolv} + E_{vdw} \qquad (1)$$

RESULTS

A summary of our top predictions is given in Table 1. For targets 1, 6 and 7, we obtained low RMSD structures with respect to the co-crystallized complex structures. For target 5, our top three predictions were near the binding site (see Table 1). In what follows, we discuss in detail the application of the *SmoothDock* algorithm to each target.

Target 1: Hexameric Hpr kinase/phosphotase and phosphocarrier Hpr[11].

From the literature provided for this target[11], we learned that Hpr kinase/phosphotase (HprK/P) catalyzes the ATP-dependent phosphorylation of Ser46 in Hpr. HprK/P also contains the characteristic P-loop nucleotide binding domain[11] at the interface between two monomers. This observation allowed us to concentrate on 1/6 of the total surface area of the receptor. Namely, the rigid-body docking in Step 1 of the method was performed using only the solvent accessible surface area of two of the six chains of the hexamer, chains A and B of HprK/P. Furthermore, in Step 2, we filtered out all the hits that overlapped with missing chains C, D, E and F of HprK/P.

The fifth largest cluster center obtained from the top 2,000 free energy structures had a RMSD with respect to the native structure of 5.86 Å. Figure 1 shows the free energy minimization of the top five clusters. Our multi-cluster refinement procedure (Step 4) successfully refined cluster 5 (solid line in Fig. 2) to the lowest free energy as measured by Eq. 1. However, upon refinement the cluster center moved away from the native structure to around 9 Å RMSD. The marked increase in the RMSD of this structure was caused by the significant differences between the target and the crystal structures of HprK/P, mainly a misoriented helix and a missing loop on the binding surface

of the receptor. Model 2 was also a very reasonable model with a final RMSD of 11.5 Å, and Models 3 and 4 had approximately 10% of the correct contacts at the interface and RMSDs of ~17 Å RMSD from the native state. The fifth model (Fig. 1) was not submitted because its cluster had an average free energy almost 5 kcal/mol higher than the average free energy of the fourth cluster.

Targets 2 and 3: Viral capsid VP6 domain from Bovine rotavirus and Fab antibody[12], and X31 Flu hemagglutinin and Fab HC63[13].

The sizes of these receptors, more than 1,100 residues each, presented huge challenges to our method, which has been developed and tested on proteins typically consisting of no more than 200 residues[6,7]. At the time of Round 1 of CAPRI, our only choice was to chop the receptors into three domains – top (binding site), middle, and bottom – and run the *SmoothDock* algorithm for each of the three domains separately. We filtered out all hits that overlapped with missing parts of the receptor. The best clusters obtained after Step 3 ranked second when docking to the top domains, with RMSDs of 13.5 and 7.36 Å for Targets 2 and 3, respectively. However, after discrimination the free energy estimate of these clusters was 8 kcal/mol or higher than the top clusters. Hence, none of these clusters met our selection criterion.

For Target 2, two clusters converged to free energies much lower than the other clusters, whereas for Target 3 we found only one cluster with significantly lower free energy than the rest. We submitted two models from each of these clusters, i.e., four models for Target 2 and two models for Target 3. It is interesting to note that the complexes selected by the blind search were found to have better estimated energies than the crystal structures themselves (Table 2). Almost every measure ranks the free energies of the predicted models lower than those of the crystal structures.

The apparent failures of our method are partially rationalized by the large cavities observed at the interfaces of the complex crystal structures. These cavities are most likely filled with structural water molecules that, for the most part, are neglected by our empirical free energies. As shown in Table 2, almost every free energy estimate fails to correctly discriminate the native structure indicating that the origin of the binding affinity for these complexes is not yet well understood.

Targets 4-6: Three camelid VHH domains in complex with porcine pancreatic alpha-amylase[14].

Camelids produced functional antibodies devoid of light chains and CH1 domains[14]. Targets 4 and 5 bound outside the catalytic site with almost no inhibition of the amylase activity. Surprisingly, a large number of framework residues are involved in the interactions of two of the VHHs with amylase. This unexpected behavior adversely affected our predictions for these targets since we disregarded clusters whose primary interaction sites did not involve at least one of the Complementary Determining Regions (CDRs).

Target 4: For this target the only clusters that involved CDRs were near the catalytic site. Consequently, none of our predictions were close to the crystal structure. In particular, we had a cluster center 12.5 Å RMSD away from the complex that was deemed not viable and therefore discarded before refinement (Step 4 of the algorithm). Given the low affinity found for this complex (K_d = 230 nM[14]), it is unlikely that our free energy estimate would have distinguished this complex from noise.

Target 5: The top three predictions for this target were close to the binding region. Model 1 had one receptor/ligand correct contact (the most of any model ranked first by its respective predictors). Model 2, with an RMSD of 26 Å with respect to the crystal structure, buried the highest number of

residues involved in the bound complex – a total of 35, 22 (out of 29) receptor and 13 (out of 25) ligand, residues. Finally, Model 3 had 7 (out of 64) correct contacts (there was only one model with more contacts, i.e., 10). Although these models failed to have the correct orientation, they ranked among the best predicted models for this target. Finally, we should note that the affinity of this complex was rather poor, around 24 nM[14].

Target 6: For this target, *SmoothDock* worked very well as the affinity of this complex was found to be in the nM range[14], i.e., we observed a large energy gap that led to a better signal-to-noise ratio (see Fig. 1). Thus, the energetic discrimination of the best complexes was straightforward. After clustering (Step 3), the center of the sixth cluster was found to have an RMSD of 7.24 Å from the native structure. Refinement (Step 4) improved the ranking of this cluster to No. 1 (Fig. 1) with a final RMSD of 2.4 Å. The second best cluster resulted in a 6.6 Å RMSD prediction. However, given the similarity of this prediction with our top-ranked structure, we submitted this structure as our 4th best model (see also Table 1).

Targets 7: T-Cell receptor (TCR) ß-chain in complex with Streptococcal pyrogenic exotoxin A[15].

In retrospect, Target 7 was perhaps the most difficult complex to predict. Indeed, the best cluster center found in the blind search had an RMSD of 19 Å with respect to the crystal. In general, we have found that clusters that are further than 15 Å from the native structure are not discriminated well by *SmoothDock*. However, since we found a close homologue, 1SBB[16], for this cluster in the Protein Data Bank[17], we added this specific structure as its own cluster for refinement. We submitted this *ad hoc* optimized structure as our first submission, obtaining the best community-wide RMSD. The best cluster did improve somewhat, and after refinement, we submitted the

structure closest to the homologue, i.e., Model 2 with an RMSD of 8.3 Å. Based on our free energy estimate, Eq. 1, these two models were not ranked higher than Models 3-5, but we biased our submissions toward models similar to the 1SBB homologue.

Although a completely blind prediction would have failed for this complex, it is fair to say that an unbiased manual intervention could have nevertheless resulted in a reasonable prediction. Indeed, our models selected by free energy alone (Models 3-5) had some red flags of their own. For example, the two largest contributors to the binding free energy on Model 3 are the N-terminal residues ASP1 and ASP3. Thus, given the high mobility associated with the protein termini it would have been reasonable to disregard this complex altogether. Model 4 binds to the membrane bound substrate of the TCR, hence it was also an unlikely candidate. Finally, Model 5 does not involve the CDRs of the T-cell receptor in its binding at all. At this point, we have not hard wired these types of constraints in our automated technique.

CONCLUSIONS: Lessons from CAPRI

A natural question to ask is, "Are automated docking methods more accurate than procedures using manual intervention?" We found that the only benefit from human intervention is to implement *known* biochemical constraints that might be available for a given target, e.g., the restrictions imposed by the P-loop binding domain for HprK/P in Target 1[11], and the binding interface of the homologue 1SBB for Target 7[16]. Indeed, the one target for which we arbitrarily biased the search, Target 4, resulted in a complete failure to predict even a single near-native complex.

A disappointing result from CAPRI was that, despite finding complexes with both good energies and shape complementarities (Table 2), we fail to predict near-native complexes for Targets 2 and 3. We do not yet understand these observations, though one explanation may be the shortcomings of quantitative estimates of the binding free energy. In retrospect, we conclude that the best strategy to predict protein interactions is an *unbiased* (other than biochemical constraints) search and discrimination of protein complexes.

The most important lesson from the first CAPRI experiment has been the validation of our automated prediction of protein interactions algorithm *SmoothDock*. For 4 of the 7 targets, we produced some of the best predictions community-wide. Interestingly, for all these targets we had more than one good prediction ranked in our top five models. More importantly, from a biological and experimental perspective, for three of these targets we ranked our best prediction first (with highest confidence).

ACKNOWLEDGEMENTS

We thank Dr. Larry Brown III for his insights on the geometric constraints of the phosphorelation site which we used to restrict the binding area in Target 1; Mike Silberstein for building some models of the missing loop in Target 1; and, David Bergstein for performing molecular dynamic simulations of the phosphocarrier. We are especially grateful to Dr. J. Fernández-Recio for computing the ICM energies for our models in Table II. This research has been supported by NIH GM61867-01.

Figure and Table Captions

Fig. 1. Free energy refinement and discrimination of Target 1 and Target 6. The average total free energy (Eq.1) of the top 10 best ranked complexes is plotted as a function of the number of sampled structures, for the top clusters that converged to the lowest free energies in Step 4 of the *SmoothDock* algorithm. We note that at the beginning of the plot, vdW interactions are almost negligible, thus it is possible for the total free energy (which includes the vdW energy) to increase at the beginning of the refinement process. The solid lines correspond to the clusters ranked No. 1. The dotted lines correspond to the best ranked clusters not submitted. For Target 6, the long-dashed line corresponds to the cluster ranked No. 2, though it was submitted as model 4.

Table I. [a]Number of the model predicted by the *SmoothDock* algorithm. A value of one indicates that the model was our best *a priori* prediction for that target. [b]The best RMSD of all top submissions, i.e., models submitted 1st by each of their respective predictors, and the best RMSD of all submissions regardless of the order in which they were submitted. [c]The highest number of correct contacts predicted for all top submissions and for all submissions regardless of the order in which they were submitted. [d]Rankings of our best predictions relative to the top submissions (top row), and to all of the predicted complexes (bottom row). Values are shown for RMSD and correct contacts. [e]RMSDs were not calculated if no near-native complexes were predicted. [f]No good models were found for this complex. [g]Our second submission for this target correctly identified 22/29 and 13/25 of the ligand and receptor binding site residues.

Table II. [a]Internal energy as calculated using CHARMM 19 parameters[18]. [b]Poisson-Boltzmann Equation as calculated by CONGEN[19]. [c]Analytical Continuum Electrostatics potential[20]. [d]Internal Coordinate Mechanics[5]. [e]Binding energies were calculated for site HL on the native receptor. [f]Binding energies were calculated for site HL on the native receptor.

References

1. Janin, J, Welcome to CAPRI: A critical assessment of predicted interactions. Proteins 2002;47:257-257.

2. Smith GR, Sternberg MJE. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28-35.

3. Camacho CJ, Vajda S. Protein-protein association kinetics and protein docking. Curr Opin Struct Biol 2002;12:36-40.

4. Camacho CJ, Gatchell DW, Kimura SR, and Vajda S. Scoring docked conformations generated by rigid-body protein-protein docking. Proteins 2000;40:525-537.

5. Fernández-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci 2002;11:280-291.

6. Camacho CJ, Vajda S. Protein docking along smooth association pathways. Proc Natl Acad Sci USA 2001;98:10636-10641.

7. Gatchell D, Vajda S, Camacho CJ. Sampling, Clustering, Refinement and Discrimination of Protein Interactions using SmoothDock. To be Submitted.

8. Ten Eyck LF, Mandell J, Roberts VA, Pique ME. Surveying molecular interactions with DOT. In: Hayes A, Simmons M, editors. Proceedings of the 1995 ACM/IEEE Supercomputing Conference. New York: ACM Press, 1995.

9. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, Vakser I. Molecular surface recognition: determinination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195-2199.

10. Zhang C, Vasmatzis G, Cornette JL. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707-726.

11. Fieulaine S, Morera S, Poncet S, Mijakovic I, Galinier A, Janin J, Deutscher J, Nessler S. X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. Proc Natl Acad Sci USA 2002;99:13437-13441.

12. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M. An antibody that prevents the hemagglutinin low pH fusgenic transition. Virology 2002;294:70-74.

13. Vaney MC, Rey F (to be published).

14. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic α-amylase. Inhibition and versatility of binding topology. J Biol Chem 2002;277:23645-23650.

15. Sundberg EJ, Hongmin L, Liera AS, McCormick JK, Tormo J, Schlievert PM, Karjalainen K, Mariuzza RA. Structures of two streptococcal superantigens bound to TCR β chains reveal diversity in the architecture of T cell signaling complexes. Structure 2002;10:687-699.

16. Li H, Llera A, Tsuchiya D, Ysern X, Schlievert PM, Karjalainen K, Mariuzza RA. Three-dimensional structure of the complex between a T cell receptor beta chain and the superantigen staphylococcal enterotoxin B. Immunity 1998;9:807-816.

17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235-242.

18. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187-217.

19. Bruccoleri RE. Application of systematic conformational search to protein modeling. Molecular Simulations 1993;10:151-174.

20. Schaefer, M. Karplus, M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem 1996;100:1578-1599.

Table I. Comparison of CAPRI Predictions

ID	Receptor	Ligand	Our Submissions			Community Submissions			Rankings[d]
			Rank[a]	RMSD (Å)	Correct contacts	Best RMSD[b] (top/all)	Correct contacts[c] (top/all)		
1	1JB1	1SPH	1	9.5	11/52	9.5	12/52		1st/2nd
			2	11.5	8/52	7.5	17/52		3rd/3rd
2	1QHD	Bound Fab	1	74.75	0/52	6.3	20/52		Lower 50%
			2	37.01	2/52	2.3	50/52		in all cases
3	2VIU	Fab HC63	1	57.16	0/63	---[e]	6/63		Lower 50%
						4.6	45/63		in all cases
4	1PIF	Ig VH Domain 1	1	54.53	0/58	---	0/58		n/a[f]
			4	38.07	0/58	---	1/58		
5	1PIF	Ig VH Domain 2	1	35.92	1/64	---	1/64		---/1st
			2[g]	26.59	0/64	---	10/64		---/2nd
			3	32.39	7/64				
6	1PIF	Ig VH Domain 3	1	2.42	54/65	2.42	54/65		1st/1st
			4	6.64	34/65	0.7	60/65		6th/3rd
7	1BEC	1B1Z	1	2.62	29/37	2.62	31/37		1st/3rd
			2	8.36	20/37	2.62	31/37		1st/3rd

[a]Number of the model predicted by the *SmoothDock* algorithm. A value of one indicates that the model was our best *a priori* prediction for that target. [b]The best RMSD of all top submissions, i.e., models submitted 1st by each of their respective predictors, and the best RMSD of all submissions regardless of the order in which they were submitted. [c]The highest number of correct contacts predicted for all top submissions and for all submissions regardless of the order in which they were submitted. [d]Rankings of our best predictions relative to the top submissions (top row), and to all of the predicted complexes (bottom row). Values are shown for RMSD and correct contacts. [e]RMSDs were not calculated if no near-native complexes were predicted. [f]No good models were found for this complex. [g]Our second submission for this target correctly identified 22/29 and 13/25 of the ligand and receptor binding site residues.

Table II. A Comparison of Free Energies for Targets 2 and 3

ID	Model	C_a RMSD (Å)	vdW (kcal/mol)	ACP+Elec+vdW (kcal/mol)	Int[a] (kcal/mol)	PBE[b] (kcal/mol)	ACE[c] (kcal/mol)	ICM[d] (kcal/mol)
2[e]	Native	n/a	-12845.40	*-21680.89*	4902.44	-19.58	-36445.85	-30.4
	Oriented	0.68	-12995.80	*-21924.63*	4823.65	-1.93	-36520.48	-37.1
	1	74.75	-13005.10	*-21936.30*	4829.52	-11.56	-36496.30	-45.8
	2	37.01	-13006.70	*-21931.37*	4824.70	-2.10	-36507.10	-42.8
	3	71.57	-12994.50	*-21922.13*	4824.38	-14.75	-36501.60	-46.8
	4	40.1	-13006.20	*-21937.14*	4824.38	6.01	-36577.80	-31.1
3[f]	Native	n/a	-13641.10	*-23121.51*	5273.17	25.96	-41131.07	-15.7
	Oriented	1.05	-13714.70	*-23270.25*	5211.72	11.24	-41701.64	-27.2
	1	57.16	-13798.70	*-23465.76*	5215.01	5.29	-41785.63	-16.8
	2	63.01	-13820.50	*-23487.16*	5225.04	-18.08	-41786.75	-36.4

[a]Internal energy as calculated using CHARMM 19 parameters[18]. [b]Poisson-Boltzmann Equation as calculated by CONGEN[19]. [c]Analytical Continuum Electrostatics potential[20]. [d]Internal Coordinate Mechanics[5]. [e]Binding energies were calculated for site HL on the native receptor. [f]Binding energies were calculated for site HL on the native receptor.