

Docking Prediction Using Biological Information, ZDOCK Sampling Technique, and Clustering Guided by the DFIRE Statistical Energy Function

Chi Zhang,^{1†} Song Liu,^{1†} and Yaoqi Zhou^{1,2*}

¹Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, State University of New York at Buffalo, Buffalo, New York

²Department of Macromolecular Science, Key Laboratory of Molecular Engineering of Polymers, Fudan University, China

ABSTRACT We entered the CAPRI experiment during the middle of Round 4 and have submitted predictions for all 6 targets released since then. We used the following procedures for docking prediction: (1) the identification of possible binding region(s) of a target based on known biological information, (2) rigid-body sampling around the binding region(s) by using the docking program ZDOCK, (3) ranking of the sampled complex conformations by employing the DFIRE-based statistical energy function, (4) clustering based on pairwise root-mean-square distance and the DFIRE energy, and (5) manual inspection and relaxation of the side-chain conformations of the top-ranked structures by geometric constraint. Reasonable predictions were made for 4 of the 6 targets. The best fraction of native contacts within the top 10 models are 89.1% for Target 12, 54.3% for Target 13, 29.3% for Target 14, and 94.1% for Target 18. The origin of successes and failures is discussed. *Proteins* 2005;60:314–318.

© 2005 Wiley-Liss, Inc.

Key words: docking; CAPRI; protein–protein interaction; knowledge-based potential

INTRODUCTION

CAPRI provides a challenging opportunity for testing the sampling accuracy of searching algorithms and the quality of scoring functions to rank near-native conformations.^{1,2} Recently, we developed a statistical energy function based on the reference state of distance-scaled, finite, ideal gases (DFIRE) using a structural database of single-chain proteins.³ The initial application of the energy function to protein–protein interactions indicates that it can provide a reasonably accurate prediction of protein–protein (peptide) binding affinity and yield high success rates in discriminating native complex structures from decoys and dimeric interfaces from crystal interfaces.⁴ This spurred our interest to apply it for the selection of near-native conformations in docking prediction. Because we have not yet developed our own sampling methods, we used the well-established ZDOCK program.⁵ The methods used and the results of 6 targets are discussed in detail below.

METHODS

The following steps were used to produce the top 10 models of complex structures for CAPRI targets:

1. Determination of possible binding regions via prior knowledge: For each target, we collected all possible biological and structural information by literature and homology search.^{6,7}
2. Rigid-body sampling around the proposed binding regions: ZDOCK⁵ was used to generate 2000 conformations around each hypothesized binding region by blocking the residues not belonging to the region. For a target without any known information, a complete sampling over the entire surface of target proteins was performed. The default parameters and 10° rotational sampling interval of ZDOCK were used.
3. Scoring by the DFIRE-based energy function: The binding affinities of the conformations generated by ZDOCK were calculated by using the DFIRE-based all-atom statistical energy function.^{3,4} Only interacting residue pairs (that have at least 1 pair of heavy atoms within 4.5 Å of each other) between receptor and ligand were used in binding calculation.⁴
4. Energy-guided clustering: The conformations were ranked by the DFIRE energy. The first cluster contains the structures that are within 5 Å root-mean-square distance (RMSD) from the lowest energy structure. The RMSD value is based on C_α atoms. Only residues belonging to the hypothesized binding region were used to calculate RMSD. For those targets whose binding regions are unknown, all-residue (overall) RMSD was used. The next cluster is based on the clustering around the lowest energy structure from the remaining confor-

[†]These two authors contributed equally to this work.

Grant sponsor: National Institutes of Health; Grant numbers: R01 GM 966049 and R01 GM 068530. Grant sponsor: HHMI (grant to SUNY Buffalo). Grant sponsor: Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Grant sponsor: National Science Foundation of China; Grant number: 20340420391 (to Y. Zhou).

*Correspondence to: Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics and Department of Physiology and Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. E-mail: yqzhou@buffalo.edu

Received 21 December 2004; Accepted 25 January 2005

DOI: 10.1002/prot.20576

TABLE I. Best Model (Based on the Number of Correctly Predicted Interfacial Contacts) and Its rank for Each Target Based on the Evaluation From the CAPRI Evaluation Team

Target	Evaluation criteria							
	Model ^a	Contact ^b	Int-1 ^c	Int-2 ^c	Angle (°) ^d	Distance (Å) ^d	L-rmsd (Å) ^e	I-rmsd (Å) ^e
T10 ^f	—	—	—	—	—	—	—	—
T12	3	49/55	16/19	31/35	7.51	0.45	1.13	0.60
T13	9	38/70	24/25	30/32	21.49	5.69	7.10	2.38
T14	1	46/157	38/68	37/76	15.40	4.61	6.45	1.72
T18	2	64/68	31/33	37/39	21.05	5.20	6.83	2.22
T19	3	2/81	7/25	17/29	103.68	7.16	18.59	10.39

^aThe rank within the top 10 submitted models.

^bThe number of correctly predicted residue–residue contacts versus that of the crystal structure.

^cThe number of correctly predicted interface residues of smaller partner (Int-1) and larger partner (Int-2).

^dThe θ angle and translation distance needed to superimpose the predicted and crystal structure.

^eThe overall RMSD (L-rmsd) and interface RMSD (I-rmsd).

^fThe 2 models we submitted for T10 have no correctly identified contacts.

mations. This procedure was repeated until the top 15 clusters are obtained. For each cluster, the conformation with the lowest DFIRE-energy is used as the representative structure for further analysis.

- Manual investigation: All 15 representative structures were inspected individually. Structures were discarded or reranked if they are not consistent with prior knowledge or similar to the conformation with lower binding energy. The latter was to increase the structural diversity for the top 10 predicted structures. The number of atomic clashes was also reduced by using a simple geometric constraint for side-chains before a predicted model was submitted.

RESULTS

The overall performance of our best predictions for each target is summarized in Table I. We provided close solutions for T12, T13, and T18, with more than 50% fraction of native contacts correctly predicted. The best submitted model for T14 has a 1.7 Å interface RMSD, although it only correctly predicted 46 of 157 contacts. The predictions for T10 and T19 were unsuccessful.

Target 10 [Trimeric Form of the Tick-Borne Encephalitis Virus (TBEV) Envelope Protein]

For this symmetric trimer, we assumed each subunit to be parallel to virus surface and have equal contact area with the virus surface based on limited literature information available.⁸ We thus manually arranged these subunits in a symmetric starlike or trianglelike form, and adjusted the distance of each subunit to the center to maximize the contact area between the subunits. These assumptions led to predictions that are far from the native complex structure in which each subunit is perpendicular to the virus surface.⁹

Target 12 (Unbound Cohesin–Bound Dockerin)

Early studies suggested that the binding region between cohesin and dockerin is well conserved through hydrophobic interaction of surface residues.^{10,11} Thus, we constrained rigid-body sampling (step 1) around the solvent-exposed area of helices 1 and 3 of dockerin and the

solvent-exposed area of β -strands 3, 5, 6, and 8 (including loops 2-3, 4-5, 5-6, and 6-7) of cohesin.

After we score and cluster the 2000 generated conformations, the second cluster was manually reranked as the third model, so that the reranked second model is structurally more different from the first model. This was done despite the fact that the DFIRE energy of the representative structure in the second cluster is almost identical to that of the first cluster. The difference was less than 0.02%. The third model (originally the second cluster) was shown to be a high-quality prediction with an interface RMSD of 0.6 Å, and 49 of 55 residue contacts correctly predicted compared to the native¹² [Fig. 1(A)]. Although the first model has essentially the same energy score as the third model, the former contains only 6 of 55 correct contacts and has an interface RMSD of 7.37 Å. The interface patches, however, were mostly predicted for this model (58% and 71% of dockerin and cohesin, respectively). In addition to model 1 and model 3, model 8 has more than 23% fraction of native contacts and an interface RMSD of 2.97 Å.

Target 13 [Unbound Surface Antigen 1 (SAG1)–Bound Antibody]

The publication of SAG1 crystal structure revealed a homodimeric configuration with the glycosylphosphatidylinositol (GPI)-anchored site located in the D2 domain.¹³ However, polymorphism (a possible indicator of immune pressure) was detected only in the D2 domain (membrane proximal part) rather than in the D1 domain (membrane distal part) of SAG1 alleles.¹³ Thus, we did not restrict the sampling to the membrane distal portion of SAG1, the region that is spatially more favorable to accommodate the antibody. In addition, it is not clear if SAG1 exists in a homodimeric form when it binds to the FAB. Thus, the dimeric interface was also not blocked from sampling. Only the complementarity determining region (CDR) of the antibody was used as the prior knowledge in subsequent sampling, scoring, and clustering steps.

Only model 9 among the 10 predictions was a reasonable prediction compared to the native.¹⁴ The model correctly identified 38 of 70 contacts and almost all the residues in

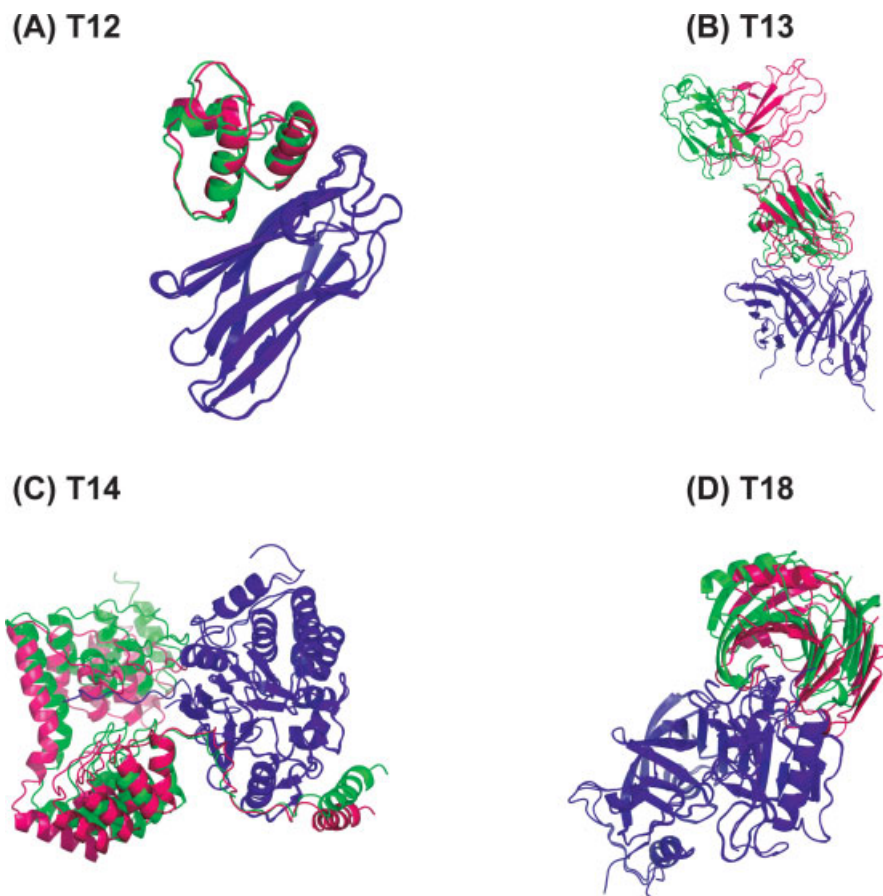


Fig. 1. (A) T12: Cohesin (in blue) versus dockerin (crystal structure in green and model 3 in red). (B) T13: Truncated FAB (in blue) versus SAG1 (crystal structure in green and model 9 in red). There is significant conformational movement for the C-terminal domain of SAG1 after binding. (C) T14: PP1ffi (in blue) versus MYPT1 (crystal structure in green and model 1 in red). (D) T18: TAXIffi (in blue) versus *A. niger* xylanase (crystal structure in green and model 2 in red).

the interface of both ligand and receptor (93.8% and 96.0%, respectively) [(Fig. 1(B)]. The rank of this model based on the DFIRE energy is 26th in the 2000 docked conformations. The clustering of the 2000 conformations promoted its rank to the 12th cluster. After manual refinement, the model was finally ranked as 9th by removing 3 other models that are either similar to the better ranked models or docked to the region around GPI-anchored site—the same region that may be occupied by the membrane.

Target 14 [Unbound Protein Phosphatase-1 (PP1ffi)–Bound Myosin Phosphatase–Targeting Subunit (MYPT1)]

There is a rich body of biological information available for PP1.¹⁵ The crystal structure of PP1 γ -RVxF peptide is also available.¹⁶ Based on the binding interface between RVxF peptide and PP1, we used only residues 28–41 of MYPT1 (RVxF motif region), as well as residues 241–300 of PP1 δ (the hydrophobic channel of β -sandwich) for docking and further analysis.

Model 1 [Fig. 1(C)] was chosen from 2000 conformations because of its similarity to the complex structure of RVxF peptide and PP1, its tight fit between docking partners,

and its relatively low DFIRE energy (ranked 160th in 2000 conformations). Model 2, on the other hand, was created purely based on the structural alignment to the RVxF–PP1 complex.¹⁶ The two models have interface RMSDs of 1.72 Å and 3.65 Å, respectively. We also attempted to select and refine other models into more compact conformations, but none of them turned out to be correct predictions as compared to the native,¹⁷ although all of them (3rd–10th submissions) are representative structures with lower DFIRE energy. Manual intervention led to 2 acceptable models. However, strict use of available information might have prohibited us from sampling higher quality models. This is reflected in low false positives (2.6% and 9.5% for ligand and receptor, respectively) but high false-negative native contacts (51.3% and 44.1% for ligand and receptor, respectively).

Target 18 [Unbound *Aspergillus niger* Xylanase–Bound *Triticum aestivum* Xylanase Inhibitor (TAXI)]

There was no biological information about the structure of TAXI. On the contrary, the crystal structure¹⁸ and site-directed mutagenesis¹⁹ have suggested that the pos-

sible binding modes of xylanase are located around the region involving the “thumb” hairpin loop (for XIP-I inhibition) and cleft (for substrate binding). However, unaware of whether TAXI and XIP-I (both are xylanase inhibitors with unrelated structure) share similar inhibition mechanism (and thus share similar binding mode in xylanase), we chose not to block any residues for subsequent docking analysis. Actually, most of the low DFIRE energy conformations generated are involved around the thumb portion (consistent with the explanation that inhibition occurs by preventing substrate from accessing the catalytic cleft¹⁹). Seven of the 10 predictions we submitted correctly identified more than 55% of interface patch in xylanase, and most of them are high in rank (the top 6).

Our second model correctly predicted 64 of 68 native contacts and more than 93% interface residues of both docking partners, with an interface RMSD 2.22 Å from the native²⁰ [Fig. 1(D)]. Its DFIRE energy was 17th of the 2000 and ranked 7th after energy guided-clustering. We submitted it as the second model after manual investigation, because the interface between xylanase and TAXI in this conformation is well packed.

Target 19 (Unbound Ovine Prion–Bound Antibody)

We found 3 known epitopes for the unbound ovine prion between residues 121 and 230: residues 132–156,^{21,22} 163–171,²³ and 220–231.²² Thus, we performed 4 rounds of docking sampling: CDRs versus epitope I, CDRs versus epitope II, CDRs versus epitope III, and CDRs versus the entire surface of prion. Two models were selected from CDR versus each epitope and 4 from CDRs versus the entire surface of prion. These selections were based on a total of 8000 conformations. However, there was no near-native structure in these 8000 conformations [interface RMSD (I_{rmsd}) < 6 Å]. This is in fact because FAB binds to a new epitope (residues 188–199).²⁴ The sampling based on the entire prion surface was also not successful. The best model submitted has only 2 of 81 native contacts.²⁴ Given the small size of the prion (only 110 residues), a fifth round of sampling in the region between epitopes II and III (i.e., residues 175–215) might improve the sampling result. The use of NMR structure for prion also led to the poorer sampling (see Discussion section).

DISCUSSION

As indicated in Table I, the method used here yields a reasonable answer within the top 10 models for 4 of 6 targets. Two poorly predicted cases are due to the false assumption about the possible binding region and the unsuccessful strategy to generate near-native conformations. To various extents, the prior knowledge was useful for all 4 targets. The ZDOCK sampling method with default setting is shown to be adequate for most targets: The exception for Target 19 is largely due to the use of prion model from NMR structure. Results from Weng’s group on the same target indicate that building a homology model of prion from a crystal structure yielded a substantially more accurate sampling by ZDOCK.²⁵

The DFIRE energy function used here generally ranks near-native structures among the top 30 of 2000. For

example, the best predictions we made for T12, T13, and T18 are ranked 2nd, 26th, and 17th in energy, respectively. The energy-guided clustering can further improve the ranking of the best near-native structures from 26th to 12th for T13, and 17th to 7th for T18. We found that the cluster size for model 3 of T12 is the smallest among the top 10 predictions, but the cluster size for model 9 of T13 is the largest. Thus, a combination of size-based and energy-based ranking, with flexible cluster radius, might be useful to further improve the prediction.

The performance of the DFIRE-based statistical energy function in detecting near-native structures is far from perfect. This is in sharp contrast to the fact that the DFIRE energy function can give a highly accurate prediction of binding affinity (a correlation coefficient of 0.87 and an RMSD of 1.76 kcal/mol with 69 experimental data points⁴). One possible reason is that many near-native conformations generated by ZDOCK have a large number of atomic clashes, which may have degraded the performance of DFIRE somewhat. Close examination indicates that DFIRE only gives good ranks for those near-native structures with a small number of atomic clashes. Thus, application of DFIRE to RDOCK decoys²⁶ that remove atomic clashes may further improve the results. Another possible reason is because statistical potentials are trained on native structures only. Moreover, pairwise interaction in the absence of explicit solvation makes highly compact, globular-shaped structures more favorable. Developing an accurate and efficient energy function for docking prediction continues to be a challenging task in the foreseeable future.

ACKNOWLEDGMENTS

We are grateful to the organizers, evaluators, and experimental contributors of the CAPRI challenge for their enormous efforts, and to Dr. Zhiping Weng for helpful discussions and for providing us the ZDOCK program.

REFERENCES

1. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* 2003;52:2–9.
2. Méndez R, Leplae R, Maria LD, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins* 2003;52:51–67.
3. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002;11:2714–2726.
4. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* 2004;56:93–101.
5. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* 2003;52:80–87.
6. Altschul SF, Madden TL, Schäffer 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.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne P. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
8. Kuhn RJ, Rossmann MG. When it’s better to lie low. *Nature* 1995;375:275–276.
9. Bressanelli S, Stiasny K, Allison SL, Stura EA, Duquerroy SD, Lescar J, Heinz FX, Rey FA. Structure of a flavivirus envelope

- glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J* 2004;23:728–738.
10. Shimon LJ, Bayer EA, Morag E, Lamed R, Yaron S, Shoham Y, Frolov F. A cohesin domain from *Clostridium thermocellum*: the crystal structure provides new insights into cellulosome assembly. *Structure (Camb)* 1997;5:381–390.
 11. Lytle BL, Volkman BF, Westler WM, Heckman MP, Wu JHD. Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium binding domain. *J Mol Biol* 2001;307:745–753.
 12. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romao MJ, Fontes CM. Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. *Proc Natl Acad Sci USA* 2003;100:13809–13814.
 13. He XL, Grigg ME, Boothroyd JC, Garcia KC. Structure of the immunodominant surface antigen from the *Toxoplasma gondii* SRS superfamily. *Nat Struct Biol* 2002;9:606–611.
 14. Graille M, Stura E, Bossus M, Muller BH, Letourneur O, Battail-Poirot N, Sibai G, Gauthier M, Rolland D, Le Du MH, Ducancel F. Structure of the immunodominant epitope displaced by the surface antigen 1 (SAG1) of *Toxoplasma gondii* complexed to a monoclonal antibody. 2005. Submitted for publication.
 15. Cohen PTW. Protein phosphatase 1—targeted in many directions. *J Cell Sci* 2002;115:241–256.
 16. Eglöf MP, Johnson DF, Moorhead G, Cohen PTW, Cohen P, Barford D. Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. *EMBO J* 1997;16:1876–1887.
 17. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. Structural basis of protein phosphatase 1 regulation. *Nature* 2004;429:780–784.
 18. Kregel U, Dijkstra BW. Three-dimensional structure of endo-1,4-betaxylanase I from *Aspergillus niger*: molecular basis for its low pH optimum. *J Mol Biol* 1996;263:70–78.
 19. Tahir TA, Berrin JG, Flatman R, Roussel A, Roepstorff P, Williamson G, Juge N. Specific characterization of substrate and inhibitor binding sites of a glycosyl hydrolase family 11 xylanase from *Aspergillus niger*. *J Biol Chem* 2002;277:44035–44043.
 20. Sansen S, De Ranter CJ, Gebruers K, Kristof Brijns K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of *Aspergillus niger* xylanase by *Triticum aestivum* xylanase inhibitor-I. *J Biol Chem* 2004;279:36022–36028.
 21. Williamson RA, Peretz D, Pinilla C, Ball H, Bastidas RB, Rozensteyn R, Houghten RA, Prusiner SB, Burton DR. Mapping the prion protein using recombinant antibodies. *J Virol* 1998;72:9413–9418.
 22. Peretz D, Williamson RA, Kaneko K, Vergara J, Leclerc E, Schmitt-Ulms G, Mehlhorn IR, Legname G, Wormald MR, Rudd PM, Dwek RA, Burton DR, Prusiner SB. Antibodies inhibit prion propagation and clear cell cultures of prion infectivity. *Nature* 2001;412:739–743.
 23. Brun A, Castilla J, Ramirez MA, Prager K, Parra B, Salguero FJ, Shiveral D, Sanchez C, Sanchez-Vizcaino JM, Douglas A, Torres JM. Proteinase K enhanced immunoreactivity of the prion protein-specific monoclonal antibody 2A11. *Neurosci Res* 2004;48:75–83.
 24. Eghiaian F, Grosclaude J, Lesceu S, Debey P, Doublet B, Treguer E, Rezaei H, Knossow M. Insight into the PrPC→PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. *Proc Natl Acad Sci USA* 2004;101:10254–10259.
 25. Wiehe K, Pierce B, Mintseris J, Tong WW, Anderson R, Chen R, Weng Z. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins* 2005;60:207–213.
 26. Li L, Chen R, Weng Z. RDOCK: refinement of rigid-body protein docking predictions. *Proteins* 2003;53:693–707.