Published in final edited form as:

Proteins. 2017 March ; 85(3): 479–486. doi:10.1002/prot.25168.

Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35

Nicholas A. Marze¹, Jeliasko R. Jeliaskov^{2,5}, Shourya S. Roy Burman¹, Scott E. Boyken^{3,4}, Frank DiMaio^{3,4}, and Jeffrey J. Gray^{1,5,6,7,*}

¹Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland ²T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland

³Department of Biochemistry, University of Washington, Seattle, Washington ⁴Institute for Protein Design, University of Washington, Seattle, Washington ⁵Program in Molecular Biophysics, Johns Hopkins University, Baltimore, Maryland ⁶Institute for NanoBioTechnology, Johns Hopkins University, Baltimore, Maryland ⁷Johns Hopkins School of Medicine, Sidney Kimmel Comprehensive Cancer Center, Baltimore, Maryland

Abstract

The 28th–35th rounds of the Critical Assessment of PRotein Interactions (CAPRI) served as a practical benchmark for our RosettaDock protein–protein docking protocols, highlighting strengths and weaknesses of the approach. We achieved acceptable or better quality models in three out of 11 targets. For the two α -repeat protein–green fluorescent protein (α rep–GFP) complexes, we used a novel ellipsoidal partial-global docking method (Ellipsoidal Dock) to generate models with 2.2 Å/1.5 Å interface RMSD, capturing 49%/42% of the native contacts, for the 7-/5-repeat α rep complexes. For the DNase–immunity protein complex, we used a new predictor of hydrogen-bonding networks, HBNet with Bridging Waters, to place individual water models at the complex interface; models were generated with 1.8 Å interface RMSD and 12% native water contacts recovered. The targets for which RosettaDock failed to create an acceptable model were typically difficult in general, as six had no acceptable models submitted by any CAPRI predictor. The UCH-L5–RPN13 and UCH-L5–INO80G de-ubiquitinating enzyme–inhibitor complexes comprised inhibitors undergoing significant structural changes upon binding, with the partners being highly interwoven in the docked complexes. Our failure to predict the nucleosome–enzyme complex in Target 95 was largely due to tight constraints we placed on our model based on sparse biochemical data suggesting two specific cross-interface interactions, preventing the correct structure from being sampled. While RosettaDock’s three successes show that it is a state-of-the-art docking method, the difficulties with highly flexible and multi-domain complexes highlight the need for better flexible docking and domain-assembly methods.

*Correspondence to: Jeffrey J. Gray, Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland. jgray@jhu.edu.

Nicholas A. Marze and Jeliasko R. Jeliaskov contributed equally to this work.

Keywords

CAPRI; Rosetta; water-mediated interfaces; protein-protein docking; conformational change

INTRODUCTION

Proteins play important roles in cellular structure, metabolic activity, biochemical signaling, and multitudes of other biological functions. A protein's function is determined by its three-dimensional structure, particularly how this structure interacts with other proteins or other biological molecules to form complexes. Consequently, if the structure of protein complexes can be predicted, the nature of their function can likewise be elucidated. Though experimental methods exist to determine protein structure (X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy among others), these are costly, time consuming, and low-throughput. Computational structure prediction is an alternative that can quickly and cheaply generate a structural model of a protein complex.

The Critical Assessment of PRotein Interactions (CAPRI) is a long-running community-wide project that evaluates the performance of state-of-the-art computational protein-protein docking methods.¹ A set of experimentally determined protein complex structures are withheld before publication, and protein docking groups are invited to submit their computational predictions of these structures. These predictions are assessed for accuracy by comparison with the experimentally determined structures. Thus, CAPRI serves as an important benchmark to evaluate the state of the field of computational protein docking, and to reveal remaining challenges. Our group has participated in CAPRI since its inception to evaluate the development of our docking method, RosettaDock.² RosettaDock is, at its core, a Monte-Carlo-based rigid-backbone docking method with side-chain optimization. RosettaDock is extensible, and several ancillary protocols have proven effective in previous CAPRI rounds³; the conformer-selection protocol EnsembleDock⁴ and the flexible-loop induced fit protocol SnugDock⁵ are among the most broadly useful.

The prior set of CAPRI rounds (20–27) highlighted the need for docking tools that could do more than assemble unmodified globular proteins, namely model non-neutral pH, carbohydrates, and water-mediated interfaces.⁶ This trend was continued by CAPRI rounds 28–35, with targets including a DNA/protein nucleosome, protein-peptide complexes, and more water-mediated interfaces. We did not attempt the peptide complexes, knowing that the Rosetta approaches would be predicted by the Furman team, who developed the Rosetta-based FlexPepDock. For the protein complexes, our team developed two new protocols to account for the eccentricities of certain targets in these rounds: (1) Ellipsoidal Dock, to account for oblong proteins such as those in Targets 96/97, and (2) HBNet with Bridging Waters, to predict hydrogen-bonding networks at water-mediated interfaces such as those in Targets 104/105.

CAPRI rounds 28–35 also continued to produce challenging globular-protein docking targets. Some of their challenges have already been well documented: proteins that exhibit large conformational changes on binding,^{7,8} partners that must be homology modeled before docking,⁹ and global docking targets lacking a homology complex or specific biochemical

information about the binding site.¹⁰ Two other challenges were more clearly defined by these CAPRI rounds: proteins that become significantly entwined during docking (Targets 98–101), and partners requiring multi-domain assembly before docking (Targets 102/107). The clarification of remaining docking challenges aids the future development of RosettaDock, and the success of our new methods such as Ellipsoidal Dock and HBNet provides a blueprint for future enhancements to the core protocol.

METHODS AND RESULTS

We submitted predictions for 12 targets (13 if Target 107, the re-run of Target 102 is counted separately) across the seven standard CAPRI rounds from 28–35. (The hybrid CASP/CAPRI round 30 is not covered by the scope of this article [see Lensink *et al.*⁹]); we did not submit predictions for any peptide docking targets, as our collaborators in the Furman lab are more adept at the peptide-specific Rosetta methods that we would have used (Schueler-Furman O. FlexPepDock in CAPRI Rounds 28–35. Proteins, Submitted to the CAPRI Special Edition, 2017) We generated two medium- and one acceptable-quality predictions (Table I). Additionally, we achieved one fair-quality water prediction among the two targets that required explicit water predictions. We present our successes first, along with descriptions of the corresponding novel methods used in these predictions. We follow with the targets in which our predictions failed, accompanied by descriptions of the targets' difficulties, and lessons learned in *ex post facto* analysis.

SUCCESSSES

Targets 96/97: α rep–GFP

In Targets 96 and 97, we were challenged to dock GFP to one of two α -repeat (α rep) proteins. The GFP sequence had ~5 point mutations relative to GFP of PDB ID 1JBZ¹¹; we used RosettaDesign^{12,13} to make the appropriate point mutations on the crystal structure. The α rep proteins were highly homologous to the 6-repeat protein of 3LTJ,¹⁴ but with one more (Target 96) and one fewer (Target 97) repeat subunit. To generate the α rep homology models, we spliced in/out a single non-terminal repeat from 3LTJ while maintaining the topological curvature of the template. We then optimized the α rep models, and to account for our uncertainty in the α rep structure we created 30-member docking ensembles using the Rosetta Relax^{15,16} protocol. Though no homologous complex was available in the PDB, the concave face of the α rep protein and the GFP β -barrel exhibited high shape complementarity; as such, we posited that the GFP must dock within the α rep concave face.

The geometric symmetry of GFP's β -barrel necessitated the use of global docking to properly place it in the α rep concave face. A new randomization method, Ellipsoidal Dock (Fig. 1), was used for the global docking of GFP. The standard Rosetta global docking randomizes the Euler angles of the protein partners, implicitly treating them as spheres. When a partner is oblong, like GFP, such randomization inefficiently samples polar regions of ellipsoidal proteins and creates poor contacts in the putative docked complexes [Fig. 1(A,C)]. Ellipsoidal Dock corrects these issues by randomly selecting a point on the ellipsoidal surface approximation of the protein, and aligning the normal vector at this point to the normal vector from the other partner's interface [Fig. 1(B,D)]. The principle of

aligning normal vectors to preserve shape complementarity is similar to that used in ICM-DISCO global docking^{17,18}; however, while ICM-DISCO is built for enumerative searches and uses a poly-hedral surface approximation with one normal vector per face, Ellipsoidal Dock uses a smooth surface with a continuous distribution of normal vectors, appropriate for Rosetta's stochastic Monte Carlo sampling methods.

The Ellipsoidal Dock protein surface approximation uses the standard equation for an ellipsoid:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

where the centroid of the protein's C α atoms defines the origin, and the 1st, 2nd, and 3rd principal components of the C α atom set define the z , x , and y directions, respectively. The parameters c , a , and b are taken as twice the square root of the eigenvalue corresponding with the 1st, 2nd, and 3rd principal component eigenvectors, respectively. The surface area of an ellipsoid does not have a closed-form solution, so approximations are made to sample it evenly. We sample the z -coordinate first from a beta distribution with parameters $\alpha = 1.5$ and $\beta = 1.5$, scaling the distribution over the z -length of the ellipsoid. We also sample the x -coordinate from a beta distribution, but with parameters between 0.5 (used when $\alpha' = \beta'$) and 1.0 (used when $\alpha' \gg \beta'$). Finally, we select the y -coordinate as either the positive or negative y -coordinate corresponding with the chosen x - and z -coordinates.

In addition to using Ellipsoidal Dock to fully sample the GFP orientation, we also used (1) a larger-than-standard initial translational/rotational perturbation (8 Å translational parameter vs. 3 Å standard) to broaden the search scope along the α rep crevice and (2) Rosetta EnsembleDock to sample different α rep conformers from our docking ensemble. We generated 20,000 candidate structures (decoys) for each target.

Among our 10 submitted models for each target, we achieved one medium-quality structure for Target 96 and two acceptable-quality structures for Target 97. Our highest-quality structure for Target 96 [Fig. 2(A)] had a root-mean-squared displacement of interface atoms (Irmsd) of 1.577 Å and 0.420 fraction of native contacts (f_{nat}), both the second-best among all submissions. This structure was our highest-ranked model. Our highest-quality structure for Target 97 [Fig. 2(B)] had an Irmsd of 2.251 Å and 0.256 f_{nat} .

Targets 104/105: DNase–immunity protein

Targets 104 and 105 presented a dual challenge: first, to predict the complex structure of a DNase (PyoAP41 or PyoS2) with its cognate immunity protein (ImAP41 or ImS2), then to predict the mediating waters and side chains at the protein interface. While the DNase proteins had crystal structures available in the PDB, we had to generate generated homology models for ImAP41 and ImS2. The available homologous structure was colicin Im2 (chain A in PDB ID 3U43¹⁹), a previous CAPRI target (T47) with 50% identity to ImAP41 and 59% identity to ImS2. The starting complexes were then generated by aligning the homology models (ImAP41 or ImS2) and structures (PyoAP41 or PyoS2) to their homolog's position in PDB 3U43. For T104, we then used structural ensembles of PyoAP41 to account

for a flexible loop at the interface and ran an local EnsembleDock to optimize the complex (50,000 decoys). For T105, we ran a local RosettaDock to optimize the complex (20,000 decoys). We used a new method for interface water predictions: HBNet with Bridging Waters (HBNetBW).

We expanded HBNet, a method for designing hydrogen bond networks,²⁰ to include a statistical potential to capture water molecules that form bridging hydrogen bonds between side chains. The two-term potential utilizes the distance between the two protein atoms that hydrogen bond to the water molecule (acceptor or donor polar hydrogen) and the dihedral angle between those two atoms and their base atoms (e.g., the base atom for a carbonyl oxygen acceptor would be the carbon it is double bonded to, and the base atom of the polar hydrogen would be the heavy-atom donor that it is covalently bonded to). We calibrated the potential using interface waters from the Top 8000 dataset²¹ and bicubic spline interpolation; the two-dimensional function that defines the bridging water score is:

$$\text{score}(a_1, a_2, a_3, a_4) = f(\text{distance}(a_2, a_3), \text{dihedral}(a_1, a_2, a_3, a_4)),$$

where a_2 and a_3 are the protein atoms hydrogen bonded to the bridging water, a_1 is the base atom of a_2 , and a_4 is the base atom of a_3 . To identify water positions during HBNet search, if two rotamers have a bridging water score below a specified threshold, they are connected as part of a potential hydrogen bond network and an explicit water molecule is placed at ideal geometry relative to the hydrogen-bonding atoms. We ran HBNetBW on each docked backbone, sampling rotamers of the interface residues to identify the most satisfied networks. There is a substantial energetic penalty associated with burying polar atoms that do not participate in hydrogen bonds (either to solvent or other protein atoms); thus, we hypothesized that using this criterion would be advantageous for discriminating between docked complexes.

For Target 104, all of our models were incorrect. An *ex post facto* analysis revealed that our homology model had the correct complex orientation, and that our docking simulation moved the complex away from that conformation. For Target 105, all four of our submitted models were of medium quality, the best having an Irmsd of 1.757 Å and an 0.481 f_{nat} . Similar to Target 104, however, the unrefined homology model had a more native-like orientation than our docked model. One of our models from Target 105 had a fair-quality water prediction [Fig. 2(C)], with a waters-only f_{nat} of 0.118, indicating that HBNet can be useful even without a perfectly-aligned interface.

After the CAPRI blind challenge, we ran HBNetBW on the revealed crystal structures for Targets 104 and 105 and the closest homology model to each. We removed water molecules from the structures. We then relaxed (cycles of minimization and side-chain repacking) the structures using Rosetta. Next, we ran HBNetBW using identical parameters to those during analysis of submitted docked complexes. In regions of the interface where the backbone was close to that of the crystal structure, the native side-chain hydrogen bond networks were largely recapitulated, and a couple of the bridging water molecules were placed in agreement with interface waters in the crystal structure; for example, running HBNetBW on the T105 homology model generated a network with a bridging water molecule between Tyr640,

Tyr55, and His34 that is in very close agreement to the experimental crystal structure. However, many false-positive networks and water placements were also generated—multiple networks are identified for each fixed-backbone decoy, making it challenging to choose which networks and water placements to keep and which to discard. Ranking networks according to satisfaction and connectivity led to success in designed protein-only networks²⁰; however, as used here, these metrics are only as reliable as the bridging water identification and placement, and our results suggest that there is significant room for improvement to both.

FAILURES

Target 95: nucleosome–Bmi1/Ring1b–UbcH5c

The challenge in Target 95 was to dock the ubiquitinating enzyme complex Bmi1/Ring1b–UbcH5c to a nucleosome. The unbound forms of both partners were available, in 3RPG²² and 3LZ0²³ for the enzyme complex and the nucleosome, respectively. The scientists who solved the unbound enzyme crystal structure (3RPG) hypothesized a binding mode for the complex of interest predicated on two structural constraints: (1) a 2 Å distance constraint between the ubiquitin donor residue (UbcH5c, Cys85) and the ubiquitin acceptor residue(s) (nucleosome H2A, Lys119/Lys118), and (2) an ambiguous interaction constraint between four enzyme complex residues (Bmi1, Lys62/Lys64, Ring1b, Arg97/Arg98, all implicated through mutagenesis experiments as DNA-binding) and the nucleosome DNA.²² We used this binding mode as one starting structure for a local RosettaDock simulation (10,000 decoys). We also identified two other binding modes meeting the Bradley constraints [Fig. 2(D)] and launched local RosettaDock simulations from each of these starting structures (10,000 decoys each). Constraint (1) was enforced with a flat harmonic score function penalty during all docking runs, albeit at a looser minimum penalized distance of 15 Å, while constraint (2) was used as a post-filter to remove any structures without at least one key residue contacting the nucleosome DNA.

All of our submitted models were incorrect (closest Irmsd: 11.4 Å). Examination of the complex crystal structure (4R8P) showed that constraint (2) was not preserved. While the constraint assumed that the patch of four positively charged amino acids would contact the DNA directly, they in fact make salt bridges with the protein core of the nucleosome [Fig. 2(E)]. As a result, our post-filter constraint prevented us from finding the correct binding mode.

Targets 98–101: UCH-L5(±Ub)–[RPN13 or INO80G]

Targets 98–101 provided a combinatorial docking challenge which asked us to dock deubiquitinating enzyme UCH-L5, with or without its conjugate ubiquitin (Ub), to either of two inhibitors, RPN13 or INO80G. Unbound structures of UCH-L5, RPN13, and Ub were available (3IHR, 2KQZ, and 1UBQ, respectively). We homology modeled INO80G by threading from PDB structure 2KQZ, loop-building, and refining in Rosetta. Additionally, we built a homology UCH-L5–Ub complex by aligning the two proteins to PDB structure 4IG7. Using the FloppyTail protocol²⁴ we modeled the tails of RPN13, which are unresolved in 2KQZ, and the homologous regions of INO80G. We found no biochemical

data or homology complexes that clearly identified a binding site, necessitating a global docking search. Due to the uncertainty in the monomer structures, we ran EnsembleDock with 30-member ensembles (generated by relaxing our top homology models). 20,000 decoys were generated for each target.

These targets were quite difficult: across all four targets, no CAPRI group submitted a model of acceptable-quality or better. Comparison of the complex binding mode to the unbound structures revealed that RPN13 undergoes a significant conformational shift upon binding, in which a helical bundle hinges open to bind around a helical element from UCH-L5, which itself undergoes a substantial kinking upon binding [Fig. 2(F)]. Though INO80G has no unbound structure to compare with its bound forms, the inhibitor is similarly entwined with the UCH-L5 helix. This binding mode is doubly difficult to predict. Firstly, predicting conformational change upon binding has been observed to be difficult in previous CAPRI challenges, particularly when the change is so large. Secondly, the degree of structural entwinement between the two partners requires a hybrid folding/docking algorithm to predict correctly: the bound forms of RPN13 and INO80G would have high energies in solution due to their open hydrophobic pocket, and even if these forms could be predicted, due to the high degree of entwinement they would be almost impossible to dock by rigid-body methods.

Target 102 (107): HxuA–hemopexin

Target 102 challenged us to assemble and dock the multi-domain protein HxuA to the heme storage protein hemopexin. The challenge was repeated in Target 107 with the unbound structure of HxuA provided. The unbound structure of hemopexin was suggested in both challenges as PDB ID 1QHU²⁵. In our attempt to assemble HxuA, we used the Robetta server to predict the individual domains, then we used ClusPro^{26–29} to dock the domains together one-by-one, with Rosetta CCD^{30,31} used to close the linking regions. Robetta predicted four domains, with domain 1 matching PDB ID (4I84³²), domain 4 being a similar β -solenoid shape, and domains 2 and 3 being small helical and sheet linking domains, respectively. The unbound structure of HxuA provided in Target 107 revealed a few key errors in our assembled structure: (1) domain 2, predicted by Robetta as helical, is entirely a beta-sheet, (2) HxuA is not made up of distinct domains, but rather is a single extended β -solenoid, and (3) we mistakenly inserted domain 4 at the N-terminus, effectively inverting the entire domain. No binding site was identified through either homology or biochemical data, so we ran a Rosetta global dock, using Ellipsoidal Dock to account for the elongated HxuA β -solenoid. We produced 10,000 decoys for both Target 102 and Target 107.

In both targets, all of our submitted models were incorrect. The failures in Target 102 can largely be attributed to our incorrect model of HxuA. The failures in Target 107 are less easily attributed. A component of the failure is likely the size of the complex. HxuA is 884 residues; to fully sample the protein with a Rosetta global dock, we would generally produce between 100,000 and 1,000,000 decoys; however, the time constraints of the CAPRI competition limited us to 10,000 decoys. Perhaps more critically, though, the binding conformation is mediated by a 23-residue loop on HxuA that undergoes significant remodeling during binding, inserting into the heme-binding site of hemopexin. As our

simulation did not account for any backbone flexibility, this loop was completely unavailable to the binding site in hemopexin.

Target 103: UBE2Z–FAT10

Target 103 presented the challenge of predicting the complex of UBE2Z, a ubiquitin-conjugating (E2) enzyme, and FAT10, a diubiquitin analogue. We used Modeller³³ to generate homology models for UBE2Z and FAT10 (from PDB structures 3CEG³⁴ and 4KSL,³⁵ respectively) for use in our docking ensembles. We ran a 50,000 decoy, high-perturbation local EnsembleDock from a putative binding conformation based on homology to other E2 ubiquitinating enzymes. During docking, we imposed a 16 Å distance constraint between the C-terminal residue of FAT10 and Cys-160 on UBE2Z, with the latter posited as the FAT10 carrier site. We achieved no models of acceptable or better quality, either in the evaluation of the full complex, or in the separate evaluations of the binding sites of either the C-terminal or N-terminal FAT10 domains.

Target 59: Rps28b–Edc3

In Target 59, we were challenged to dock ribosomal protein Rps28b to mRNA decapping enzyme Edc3. 20-member unbound NMR ensembles were provided for each partner: 1NE3³⁶ for Rps28b, and 4A53³⁷ for Edc3. We ran 10,000 local EnsembleDock decoys from each of six putative binding sites manually identified by examining the solvent-accessible faces of Rps28b in the context of the ribosome and Edc3 in the context of homologous hexamers. We achieved no acceptable or better structures in our ten predicted structures, but we did achieve one acceptable structure in our 100 uploaded structures.

DISCUSSION

Our group's CAPRI performance reveals strengths and limitations in our docking abilities. We achieved a successful prediction in 3 of 12 targets. Compared to the community as a whole, however, our performance is not atypical, as six targets did not elicit a single successful prediction from any team. We did not participate in the peptide docking targets, but our collaborators in the Furman lab did (Schueler-Furman O. FlexPepDock in CAPRI Rounds 28–35. Proteins, Submitted). When the results of our submissions (3*/2**) are adjoined with those from the Furman lab (3*/3**/1***) on different targets, Rosetta docking approaches (6*/4**/1***) had acceptable or better predictions in six targets and medium or better predictions in four targets, which would position the combined ranking somewhere in the top ten of all predictors.³⁸ Other CAPRI predictors included Rosetta refinement in their approaches (Baker, Bradley, Guerois, and so forth). In fact, Guerois, who incorporated Rosetta-based refinement as the final refinement and discrimination stage in their pipeline, predicted 9/18 targets correctly. Furthermore, for targets with at least one correct prediction, the Rosetta-based approaches yielded models closest to native. Thus, Rosetta remains a state-of-the-art computational tool that can successfully predict a diverse set of protein complexes.

The new Rosetta methods we tested during these CAPRI rounds worked especially well, leading to all three of our group's success cases. Ellipsoidal Dock's ability to appropriately

adapt our search to the oblong shape of GFP led to our successes in predicting Targets 96 and 97, which were both difficult targets for the community. We were the top predictor group for Target 96, and in the top-five for Target 97. HBNetBW, even its early stage of development, was able to achieve a fair-quality interface water prediction for Target 105 despite errors in the docked partners. However, like in previous CAPRI experiments, all of our successes were small protein complexes with little flexibility upon binding and with clues about the native binding sites, either by homology complex or obvious shape complementarity.

These CAPRI rounds reveal the shortcomings of our docking methods and the remaining challenges for the docking community as a whole. Large and multi-domain targets remain quite challenging, even when they are otherwise tractable docking challenges. Target 95, a 1639 residue complex that is nearly rigid upon binding and has an abundance of biochemical data restraining the complex, was only successfully predicted by three predictor groups, with only one medium-quality model between them. Target 102, a 1,098 residue complex that also exhibits a full loop remodeling at the active site upon binding, did not elicit a single acceptable-quality or better model from any predictor, even when the full unbound structure of the larger partner was provided. To allow prediction of large complexes, future global docking methods must be able to sample the resultant large conformational space more efficiently.

The latter large target is also indicative of the other key remaining challenge, large conformational changes during binding, which will often confound all existing docking methods. Targets 98–101 are all small complexes in which the structure of the binding residues of the larger partner is predictable by homology. Not one target, even T98 where unbound structures of both partners were provided, had a single correct model predicted. While this difficulty can be attributed to the conformational changes upon binding, it can also be attributed to the severe entwinement of the two partners in the bound state. This entwinement requires a more sophisticated set of docking methods where flexibility and docking orientation are sampled concurrently, as opposed to existing methods such as Rosetta's EnsembleDock, which largely separates the sampling of the flexibility and the docked conformation. The revelation of these CAPRI rounds that there are two distinct challenges in flexible docking will provide insight into future development of flexible docking methods.

References

1. Vajda S, Vakser IA, Sternberg MJE, Janin J. Modeling of protein interactions in genomes. *Proteins*. 2002; 47:444–446. [PubMed: 12001222]
2. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*. 2003; 331:281–299. [PubMed: 12875852]
3. Sircar A, Chaudhury S, Kilambi KP, Berrondo M, Gray JJ. A generalized approach to sampling backbone conformations with Rosetta-Dock for CAPRI rounds 13–19. *Proteins*. 2010; 78:3115–3123. [PubMed: 20535822]
4. Chaudhury S, Gray JJ. Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *J Mol Biol*. 2008; 381:1068–1087. [PubMed: 18640688]

5. Sircar A, Gray JJ. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol*. 2010; 6:e1000644. [PubMed: 20098500]
6. Kilambi KP, Pacella MS, Xu J, Labonte JW, Porter JR, Muthu P, Drew K, Kuroda D, Schueler-Furman O, Bonneau R, Gray JJ. Extending Rosetta-Dock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20–27. *Proteins*. 2013; 81:2201–2209. [PubMed: 24123494]
7. Bonvin AM. Flexible protein-protein docking. *Curr Opin Struct Biol*. 2006; 16:194–200. [PubMed: 16488145]
8. Kuroda D, Gray JJ. Pushing the backbone in protein-protein docking. *Structure*. 2016 In Press.
9. Lensink MF, Velankar S, Kryshchuk A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Rie Lee G, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RAG, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrmann TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JPGLM, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond ASJ, Visscher K, Kastiris PL, Bonvin AMJJ, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jiménez-García B, Moal IH, Fernández-Recio J, Joungh JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Marze N, Kuroda D, Roy Burman SS, Gray JJ, Chermak E, Cavallo L, Oliva R, Tovchigrechko A, Wodak SJ. Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. *Proteins*. 2016; 84:323–348. [PubMed: 27122118]
10. Moreira IS, Fernandes PA, Ramos MJ. Protein-protein docking dealing with the unknown. *J Comput Chem*. 2010; 31:317–342. [PubMed: 19462412]
11. Hanson GT, McAnaney TB, Park ES, Rendell MEP, Yarbrough DK, Chu S, Xi L, Boxer SG, Montrose MH, Remington SJ. Green Fluorescent Protein Variants as Ratiometric Dual Emission pH Sensors. 1. Structural Characterization and Preliminary Application. *Biochemistry*. 2002; 41:15477–15488. Available at <<http://pubs.acs.org/doi/abs/10.1021/bi026609p>>. [PubMed: 12501176]
12. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci*. 2000; 97:10383–10388. [PubMed: 10984534]
13. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YEA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popov Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. 2011; 487:545–574. [PubMed: 21187238]
14. Urvoas A, Guellouz A, Valerio-Lepiniec M, Graille M, Durand D, Desravines DC, van Tilbeurgh H, Desmadril M, Minard P. Design, production and molecular structure of a new family of artificial alpha-helical repeat proteins (α Rep) based on thermostable heat-like repeats. *J Mol Biol*. 2010; 404:307–327. [PubMed: 20887736]
15. Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005; 309:1868–1871. [PubMed: 16166519]
16. Misura KMS, Baker D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins Struct Funct Genet*. 2005; 59:15–29. [PubMed: 15690346]
17. Fernández-Recio J, Totrov M, Abagyan R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins*. 2003; 52:113–117. [PubMed: 12784376]
18. Fernández-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol*. 2004; 335:843–865. [PubMed: 14687579]
19. Wojdyla JA, Fleishman SJ, Baker D, Kleanthous C. Structure of the ultra-high-affinity colicin E2 DNase–Im2 complex. *J Mol Biol*. 2012; 417:79–94. [PubMed: 22306467]
20. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, Sankaran B, Seelig G, Zwart PH, Baker D. De novo design of protein homo-oligomers

- with modular hydrogen-bond network-mediated specificity. *Science*. 2016; 352:680–687. [PubMed: 27151862]
21. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr Sect D Biol Crystallogr*. 2010; 66:12–21. [PubMed: 20057044]
 22. Bentley ML, Corn JE, Dong KC, Phung Q, Cheung TK, Cochran AG. Recognition of UbcH5c and the nucleosome by the Bmi1/Ring1b ubiquitin ligase complex. *EMBO J*. 2011; 30:3285–3297. [PubMed: 21772249]
 23. Vasudevan D, Chua EYD, Davey CA. Crystal structures of nucleosome core particles containing the ‘601’ strong positioning sequence. *J Mol Biol*. 2010; 403:1–10. [PubMed: 20800598]
 24. Kleiger G, Saha A, Lewis S, Kuhlman B, Deshaies RJ. Rapid E2–E3 assembly and disassembly enable processive ubiquitylation of cullin-RING ubiquitin ligase substrates. *Cell*. 2009; 139:957–968. [PubMed: 19945379]
 25. Baker EN, Paoli M, Anderson BF, Baker HM, Morgan WT, Smith A. Crystal structure of hemopexin reveals a novel high-affinity heme site formed between two [[beta]]-propeller domains. *Nat Struct Biol*. 1999; 6:926–931. [PubMed: 10504726]
 26. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res*. 2004; 32:W96–W99. [PubMed: 15215358]
 27. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*. 2004; 20:45–50. [PubMed: 14693807]
 28. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins Struct Funct Genet*. 2006; 65:392–406. [PubMed: 16933295]
 29. Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR, Vajda S. How good is automated protein docking? *Proteins*. 2013; 81:2159–2166. [PubMed: 23996272]
 30. Canutescu AA, Dunbrack RL. Cyclic coordinate descent: a robotics algorithm for protein loop closure. *Protein Sci*. 2003; 12:963–972. [PubMed: 12717019]
 31. Wang C, Bradley P, Baker D. Protein–protein docking with backbone flexibility. *J Mol Biol*. 2007; 373:503–519. [PubMed: 17825317]
 32. Baelen S, Dewitte F, Clantin B, Villeret V. Structure of the secretion domain of HxUA from *Haemophilus influenzae*. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2013; 69:1322–1327.
 33. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinform*. 2016; 54:5.6.1–5.6.37.
 34. Sheng Y, Hong JH, Doherty R, Srikumar T, Shloush J, Avvakumov GV, Walker JR, Xue S, Neculai D, Wan JW, Kim SK, Arrowsmith CH, Raught B, Dhe-Paganon S. A human ubiquitin conjugating enzyme (E2)-HECT E3 ligase structure-function screen. *Mol Cell Proteomics*. 2012; 11:329–341. [PubMed: 22496338]
 35. Rivkin E, Almeida SM, Ceccarelli DF, Juang Y-C, MacLean TA, Srikumar T, Huang H, Dunham WH, Fukumura R, Xie G, Gondo Y, Raught B, Gingras A-C, Sicheri F, Cordes SP. The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis. *Nature*. 2013; 498:318–324. [PubMed: 23708998]
 36. Wu B, Yee A, Pineda-Lucena A, Semesi A, Ramelot TA, Cort JR, Jung J-W, Edwards A, Lee W, Kennedy M, Arrowsmith CH. Solution structure of ribosomal protein S28E from *Methanobacterium thermoautotrophicum*. *Protein Sci*. 2003; 12:2831–2837. [PubMed: 14627743]
 37. Fromm SA, Truffault V, Kamenz J, Braun JE, Hoffmann NA, Izaurralde E, Sprangers R. The structural basis of Edc3- and Scd6-mediated activation of the Dcp1:Dcp2 mRNA decapping complex. *EMBO J*. 2012; 31:279–290. [PubMed: 22085934]
 38. Lensink MF, Velankar S, Wodak SJ. Modeling protein-protein and protein-peptide complexes: CAPRI. 6.

Figure 1.

Global randomization in (A) standard Rosetta and (B) Ellipsoidal Dock. The standard global randomization pulls the docking partners apart, randomizes the Euler angles of the implicit sphere circumscribing one protein partner, and pulls the partners back into contact. Ellipsoidal Dock calculates the normal vectors from the surface of one partner at the starting point of contact, as well as at a random point on the surface, and superimposes the two normal vectors. The distributions of 200 oblong candidate complexes as generated using standard randomization (C) and Ellipsoidal Dock (D) are shown. The red spheres represent the center of mass of the antibody in each candidate complex.

Figure 2.

A & B: Our best-quality models for Targets 96 & 97 (red), superimposed with their complex crystal structures (blue). (A) shows our medium-quality model for Target 96; (B) shows the better of our two acceptable-quality models for Target 97. **C:** Our best medium-quality model for Target 105, superimposed with the crystal structure. Our model is colored in red/purple shades, while the crystal structure is colored in blue shades. **D:** Two of the three binding modes used in our docking simulations for Target 95. The mode implicated by Bradley et al. is shown in orange, while our first novel mode is shown in magenta. The six residues implicated as DNA-binding are colored in dark blue; in all our modes, these residues contact DNA directly. **E:** The native binding mode for Target 95, shown in green. The six residues implicated in DNA-binding are again colored in dark blue; none of these residues contact DNA directly, instead making salt bridges with positively charged residue, colored in red. **F:** The bound UCH-L5–RPN13 complex (grey and pink), with the unbound RPN13 superimposed on top (blue). Upon binding, the RPN13 helical bundle hinges open to accommodate the UCH-L5 C-terminal helix.

Table I

Summary of Our Results in CAPRI Rounds 28–35 for Targets We Predicted

Target	Complex	Type	Metrics of best model				Quality	Method	Search scale
			f_{nat}	Lrmsd	Irmsd				
59	Rps28b–Edc3	U–U	0.00	29.7	9.9	–	–	EnsembleDock	Local
95	Nucleosome–Bmi1/Ring1b–UbcH5c	U–U	0.00	23.2	11.4	–	–	Constrained	Local
96	α rep7–GFP	H–U	0.42	4.0	1.6	**	**	Ellipsoidal Dock	Partial global
97	α rep5–GFP	H–U	0.26	6.4	2.3	*	*	Ellipsoidal Dock	Partial global
98	UCH-L5–RPN13	U–U	0.01	24.6	9.9	–	–	EnsembleDock	Global
99	UCH-L5–Ub–RPN13	H–U	0.00	20.5	7.1	–	–	EnsembleDock	Global
100	UCH-L5–Ub–INO80G	H–H	0.05	22.3	7.4	–	–	EnsembleDock	Global
101	UCH-L5–INO80G	U–H	0.00	38.1	18.3	–	–	EnsembleDock	Global
103	Ube27–FAT10	H–H	0.01	42.5	14.5	–	–	Constrained EnsembleDock	Local
104	DNase–ImAP41	H–H	0.07	13.4	5.7	–/bad ^a	–/bad ^a	Standard/HBNetBW ^a	Local
105	DNase–ImS2	H–H	0.48	4.1	1.8	**/fair ^a	**/fair ^a	Standard/HBNetBW ^a	Local
102/107	HxuA–Hemopexin	U–U	0.00	27.4	17.0	–	–	Standard	Global

Target, the CAPRI target ID; Complex, the docking partners; Type, model for each docking partner: homology model (H) or unbound structure (U); f_{nat} , fraction of native contacts recovered; Lrmsd, ligand RMSD in Å; Irmsd, interface RMSD in Å; Quality, the CAPRI rating of the best predicted model: high (***), medium (**), acceptable (*), or incorrect (–); Method, the Rosetta docking protocol used to generate predictions; Search scale, breadth of the docking search.

^aInterface water prediction ratings.