Published in final edited form as:
Proteins. 2013 December ; 81(12): 2201–2209. doi:10.1002/prot.24425.

Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20–27

Krishna Praneeth Kilambi[1], **Michael S. Pacella**[2], **Jianqing Xu**[1], **Jason W. Labonte**[1], **Justin R. Porter**[3], **Pravin Muthu**[1], **Kevin Drew**[4], **Daisuke Kuroda**[1], **Ora Schueler-Furman**[5], **Richard Bonneau**[4], and **Jeffrey J. Gray**[1,*]

[1]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland [2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland [3]Thomas C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland [4]Department of Biology, Center for Genomics and Systems Biology, New York University, New York, New York [5]Department of Microbiology and Molecular Genetics, The Hebrew University of Jerusalem, Jerusalem, Israel

Abstract

Rounds 20–27 of the Critical Assessment of PRotein Interactions (CAPRI) provided a testing platform for computational methods designed to address a wide range of challenges. The diverse targets drove the creation of and new combinations of computational tools. In this study, RosettaDock and other novel Rosetta protocols were used to successfully predict four of the 10 blind targets. For example, for DNase domain of Colicin E2–Im2 immunity protein, RosettaDock and RosettaLigand were used to predict the positions of water molecules at the interface, recovering 46% of the native water-mediated contacts. For α-repeat Rep4–Rep2 and g-type lysozyme–PliG inhibitor complexes, homology models were built and standard and pH-sensitive docking algorithms were used to generate structures with interface RMSD values of 3.3 Å and 2.0 Å, respectively. A novel flexible sugar–protein docking protocol was also developed and used for structure prediction of the BT4661–heparin-like saccharide complex, recovering 71% of the native contacts. Challenges remain in the generation of accurate homology models for protein mutants and sampling during global docking. On proteins designed to bind influenza hemagglutinin, only about half of the mutations were identified that affect binding (T55: 54%; T56: 48%). The prediction of the structure of the xylanase complex involving homology modeling and multidomain docking pushed the limits of global conformational sampling and did not result in any successful prediction. The diversity of problems at hand requires computational algorithms to be versatile; the recent additions to the Rosetta suite expand the capabilities to encompass more biologically realistic docking problems.

*Correspondence to: Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218. jgray@jhu.edu.

Additional Supporting Information may be found in the online version of this article.

Keywords

CAPRI; protein interactions; protein docking; binding

INTRODUCTION

Protein interactions are critical for biological function. Structural studies of protein interactions not only aid in deciphering their functions but also provide vital information on putative therapeutic targets. Structural studies of protein interactions often rely on experimental methods such as X-ray crystallography and nuclear magnetic resonance for the determination of protein complexes. These methods are often expensive and time consuming, and structural genomics studies lack the throughput to match the explosion of sequence-based interaction data. Computational prediction of the structures of protein complexes can fill the void and provide an alternative to expedite the investigations.

Since its inception, the Critical Assessment of PRotein Interactions (CAPRI) has served as the premier platform for testing new protein docking methods.[1] CAPRI challenges often involve blind prediction of the structures of protein complexes, given the sequences or the unbound coordinates of the individual partners. Recently, CAPRI has expanded to encompass more varied biomolecular interaction problems. To tackle the challenges in CAPRI rounds 20–27, we used RosettaDock,[2] a Monte Carlo-based docking algorithm, as the basis to develop, test, and refine new protocols.

In previous CAPRI rounds, we highlighted the advantages of the inclusion of experimental biochemical information[3] and the incorporation of backbone flexibility[4] during docking predictions. Availability of biochemical data was the single most important determinant of the docking success, as it drastically shrinks the conformational search space and reduces a global search problem to a local search problem. In CAPRI rounds 20–27, we were able to find and use experimental biochemical information to localize sampling for three targets (Targets 47, 50, and 58). Additionally, algorithms such as EnsembleDock[5] and SnugDock[6] were critical for protein backbone conformational flexibility simulations, leading to several other successful CAPRI predictions. As CAPRI frequently provides only the starting sequences of the protein partners (as opposed to the unbound coordinates), we used homology modeling for constructing the starting structures. Incorporating backbone flexibility with EnsembleDock was a critical strategy to compensate for homology modeling inaccuracies. Over all the rounds, we used EnsembleDock for seven distinct targets.

The latest CAPRI rounds 20–27 presented far more diverse and challenging targets when compared with the previous rounds. The problems included prediction of water molecules at the binding interface after prediction of the complex structure (Target 47), use of experimental SAXS data (Target 58), flexible sugar–protein docking (Target 57), and multidomain assembly (Target 51). In addition, for the first time, CAPRI also requested affinity predictions. For these binding prediction rounds, the assigned tasks included discrimination of designed and natural protein–protein interfaces (Targets 43–45) and the prediction of mutation effects on binding affinities (Targets 55 and 56). The unique amalgam of targets tested the versatility and adaptability of RosettaDock, with each round

demanding new code development and testing. In this article, we outline the methods we used to address each of these challenges and suggest further changes to improve prediction accuracy.

METHODS AND RESULTS

In the CAPRI prediction of rounds 20–27, we achieved one high, one medium, and two acceptable-quality predictions (Table I). To analyze the target performance, we used several new pertinent metrics (e.g., native water-mediated residue contacts in Target 47) in addition to the standard CAPRI evaluation metrics such as ligand root mean square deviation (Lrmsd), interface root mean square deviation (Irmsd), and fraction of native contacts (f_{nat}).[7]

Because of space limitations, we present details only of the prediction targets that required development of new protocols. Targets addressed using conventional RosettaDock-based protocols are discussed in the Supporting Information. We also discuss here our newly developed pipeline for the scoring challenges.

Target 47: Colicin E2–Im2 immunity protein interface water prediction

In Target 47, the sequence of colicin E2 DNase domain and the unbound crystal structure of the cognate IM2 immunity protein (Protein Data Bank[8] ID: 2WPT: A[9]) were provided. The challenge was to predict not only the structure of the protein complex but also the positions of the water molecules at the binding interface.

As RosettaDock uses a score function with an implicit solvation model,[10] we had to design a new method to predict explicit water positions. We used MODELLER[11] and three crystal structure templates (2WPT: E9-IM2, 1EMV[12]: E9-IM9, and 7CEI[13]: E7-IM7) to build three bound complex homology models of E2-IM2. We further refined each model either by Rosetta local docking refine[2] or RosettaRelax[14,15] (fast relax) to generate nine total homology structures. After superimposing the nine homology models onto their original crystal templates, we used the coordinates of the water molecules in the original crystal templates as starting positions for the water molecules at the E2–IM2 interface. We defined all template water molecules within 3 Å of both protein partners as interface water and represented them by the classical TIP3P model.[16] We then used the RosettaLigand protocol[17] to individually dock each water molecule to the dry homology model starting from its initial position. We then merged all the optimized water positions back into one model. If two water molecules clashed, we inspected the original ligand-docking histories of both waters to find alternative low-energy coordinates without clashes. We conducted a final round of protein side-chain repacking with all the interface water molecules fixed. We ranked the final structures based on their interface scores and submitted nine models with interface water predictions and one model ignoring the water molecules.

The released crystal structure (3U43[18]) revealed 23 water molecules at the interface. For our submitted structure with the best water prediction (Fig. 1), 13 buried water molecules were predicted, capturing 46% of native water-mediated residue contacts (f_{nat}^{wmc}, which is analogous to f_{nat},[7] but measures indirect water-mediated contacts instead of direct receptor–ligand contacts). Overall, six of our models achieved medium-quality ($0.3 < f_{nat}^{wmc} < 0.5$), and

one model was of acceptable quality ($0.1 < f_{nat}^{wmc} < 0.3$), placing us sixth among the 20 participating groups. (The two top-performing groups used explicit solvent MD simulations and recovered more than 60% of the native water-mediated contacts.) Most of the water-mediated contact predictions that failed were due to interface distortions arising from errors in the homology models used for docking. We also did not attempt water placement at the periphery of the complex interface. Solving these two problems would improve water predictions, which might translate to gains in the accuracy of high-resolution docking. However, in this case, the water molecules at the interface did not aid the docking predictions. Our closest submitted structure generated by ignoring the interface water molecules was a high-quality prediction with 0.6 Å Irmsd and 0.82 f_{nat}.

Target 58: g-Type lysozyme–PliG inhibitor SAXS data and pH-sensitive docking

In Target 58, the unbound coordinates of g-type lysozyme (3MGW[19]) and PliG lysozyme inhibitor (4DY3[20]) were given. For the first time in the CAPRI challenge, the experimental SAXS data of the complex was also provided allowing docking algorithms the option to use it to constrain docking orientations.

We first carried out standard global docking without using any known binding site information, leading to a model with a reasonable interface score (−8.54 Rosetta Energy Units). Published biochemical information on the binding site[20] supported the docking orientation. We then performed high-resolution docking refinement on the best model obtained by the global docking run to further improve the relative orientation of two proteins. Finally, we used CRYSOL[21] to evaluate SAXS χ values for the models to quantify the discrepancy between experimental SAXS data and theoretical curves for the models. The SAXS χ values of the top models agreed with the experimental data and improved the confidence of the predictions.

As lysozyme functions in a low pH environment,[22] it served as an ideal target for conducting the first blind tests of our prototype pH-dependent docking algorithm. The Rosetta pH-docking algorithm incorporates dynamic residue protonation states during the side-chain sampling steps in the docking protocol. The favorable protonation states at a given environment pH and docking orientation are determined based on the residue pK_a values evaluated using the Rosetta-pH algorithm.[23] We submitted one structure using pH-docking algorithm at a pH of 6.2 (unbound crystal structure pH of the lysozyme) resulting in a prediction 2.02 Å Irmsd away from the native structure [Fig. 2(A)]. However, the closest structure generated using the standard methods resulted in a slightly better structure with 1.94 Å Irmsd and 0.44 f_{nat}. Overall, our submissions resulted in two medium-quality and one acceptable-quality predictions.

Retrospectively, we examined the correlations between the Rosetta score, Lrmsd, and the experimental SAXS data. Although the SAXS data could filter some decoys that have high RMSD against the native complex, there is no correlation between Lrmsd of the top-scoring 10% of RosettaDock-generated decoys and the SAXS χ values computed by CRYSOL [Fig. 2(B)]. The globular shape of each protein in the complex might be one explanation. Although the SAXS data did not directly aid in ranking the generated top-scoring decoys, they help in decoy enrichment. Among the 10% top-scoring structures, filtering decoys with

the top 10% χ values ($\chi < 1.16$) recovers >50% of the near-native decoys, indicating five-fold enrichment (Supporting Information Fig. S1).

Target 57: Heparin-like saccharide–BT4661 protein docking

For Target 57, the unbound structures of a six-residue heparin-like oligosaccharide and protein BT4661 were provided. The Rosetta modeling suite did not have parameters for the sugar molecules. In addition, Rosetta relies on residue-specific rotamer libraries for side-chain flexibility; however, no such libraries existed for oligosaccharide residues. Fortunately, a few possible approaches for these problems had been recently described in the study by Drew *et al*.[24] Following their work incorporating peptidomimetics in Rosetta, we generated a distinct parameter file for each saccharide residue (defined by a single six-carbon ring plus stereochemistry), modified core Rosetta code to allow minimization of carbohydrate backbone and side-chain torsion angles, and developed a method of sampling discrete low-energy ring conformations. We also defined a library of low-energy side-chain rotamers for each saccharide residue by scanning side-chain torsion angles and performing quantum mechanical calculations on each rotamer [Gaussian software with Hatree–Fock optimization and MP2 energy calculations with a 6-31G(d) basis set]. We used the same technique to generate a library of low-energy oligosaccharide backbone conformers by scanning the ϕ/ψ angles [Fig. 3(A)] for each consecutive heparin residue pair.

We implemented a SugarDock algorithm (flowchart in Supporting Information Fig. S2) based on FlexPep-Dock,[25] a flexible oligomer docking protocol for protein–peptide docking. The docking protocol began with a random rigid-body perturbation, followed by random perturbation of the initial saccharide backbone torsion angles based on quantum mechanical predictions of energy minima. Multiple rounds of small ϕ/ψ perturbations and ring conformation switches [Fig. 3(A)] were performed along with discrete sampling of side-chain rotamers. Following FlexPepDock, we varied the weights on the Van der Waals attractive and repulsive terms gradually down and up, respectively, to sample conformations that are inaccessible due to the high-energy barriers when using standard weights. Finally, the side-chain torsion angles and rigid-body orientation of the oligosaccharide were minimized. We used constraints to prevent the rings from breaking apart during minimization. Our flexible sugar–protein docking algorithm led to an acceptable-quality prediction that recovered 71% of the native contacts [Fig. 3(B)], placing us seventh among 31 participating groups (based on Irmsd).

Targets 43–45: Natural versus designed interface discrimination

Instead of the usual structure prediction challenges, in Targets 43–45, we were asked to develop a metric to discriminate natural interfaces from computationally designed interfaces. Eighty-seven designed and 120 natural complexes were provided for discrimination. We evaluated the following metrics for each protein complex: (1) interface area per residue; (2) number of interface contacts per unit surface area; (3) solvent-accessible surface area (polar and total) of the complex; (4) Van der Waals energy (Rosetta attractive and repulsive Lennard-Jones terms); and (5) solvation penalty (Rosetta LK model[10]). Only the interface area per residue metric (A_{int}/N_{res}) showed a distinction between native and designed complexes, with the designed complexes exhibiting smaller average A_{int}/N_{res} values. The

distributions of all other metrics revealed no significant separation between the native and the designed complexes. The submissions based on the A_{int}/N_{res} metric alone resulted in a 70% AUC on the ROC curve showing true- and false-positive classification (for details, see Ref. 26).

Targets 55 and 56: Hemagglutinin binders affinity predictions

In Targets 55 and 56, we were asked to predict the influence of mutations to each of the 20 natural amino acids at every position in two de novo designed proteins (HB36.4 and HB80.3), by classifying each mutation as deleterious, neutral, or beneficial to binding. We generated predictions in two phases during the challenge. In the first phase, only the crystal structures of the designed proteins bound to hemagglutinin were provided. We first refined the provided structures to relieve clashes and to calculate the wild-type binding score using the Rosetta-Dock score function. We then performed each mutation (every amino acid at every position) and reoptimized the structure by repacking all protein side chains within a 10 Å sphere of the mutation site. Afterward, we reoptimized the rigid body orientation of the designed proteins using gradient-based minimization. Finally, we calculated the binding score ($E_{complex}-\sum_{partners}E_i$) of each mutated complex for analysis. Mutations were classified as beneficial, neutral, or deleterious by selecting binding score cutoffs. The top 2% of mutations were classified as beneficial based on a published frequency of beneficial fitness effects in new mutations.[27] Approximately 20% of the worst-scoring mutations were classified as deleterious, lower than the expected frequency to avoid false positives because of the rigid protein backbone used by our algorithm. The remaining mutations were classified as neutral. For Targets 55 and 56, 54% and 48%, respectively, of all predictions were correct. The percentages of correct predictions for interface positions were slightly lower at 52% and 47%, respectively.

In the second phase of the challenge, we were provided with affinity data in the form of experimental enrichment ratios for a subset of the mutations (~15% of all possible mutations). In this phase, our goal was to use experimental data to tune the RosettaDock score function for improved mutant discrimination. We used generalized linear regression to recalculate weights for the standard terms in the RosettaDock score function to maximize the average binding energy gap between beneficial and deleterious mutations. The resulting score function achieved good discrimination when applied to our initial set of refined mutant complexes. However, after regenerating a second set of mutant complexes, the discrimination was completely lost (Supporting Information Fig. S3). Hence, we did not submit any predictions. These data suggest that the new score function is sensitive to small stochastic changes encountered during Rosetta protocols. More detailed discussion of the methods and results are available in the community publication.[28]

Target 51: Xylanase multidomain assembly

Target 51 was a multidomain assembly problem of xylanase Cthe_2193's six modules: GH5–CBM6–CBM13– Fn3–CBM62–dockerin. After building the complete complex homology model using SWISS-MODEL,[29,30] we docked the domains using a variant of our domain insertion protocol[31] and built the loops using the KIC loop modeling method[32] to connect the domains (see Supporting Information for details). As conformational sampling

even for a single protein interface is challenging, the multiple interface predictions pushed the limits of global docking, and none of the submitted models resulted in a successful prediction.

Standard targets

Some CAPRI targets were approachable with our standard toolset. For these predictions, we employed RosettaDock and its flexible-backbone alternative EnsembleDock. For Target 53, we managed three acceptable-quality predictions, and the best structure had an Irmsd of 3.3 Å and 0.15 f_{nat}. We were unable to achieve any successful prediction for the remaining targets (Targets 46, 48, 49, 50, and 54). Details for these targets are presented in the Supporting Information.

Novel RosettaDock-based scoring protocol

In the CAPRI scoring challenges, participants were provided with a set of protein complexes generated by different algorithms and were tasked with the identification of near-native structures within that set. From the seven targets we attempted, we achieved one high, two medium, and two acceptable predictions, placing us fourth among all the participating groups.

This was our group's first attempt to develop a systematic pipeline for participating in the CAPRI scoring challenge. The pipeline had two major stages: (1) a strategy for leveling the playing field for structures generated by different docking algorithms via a relaxation process, and (2) a new score for identifying the near-native decoys from a set so processed. For each target, after curating input files to achieve equal number of residues and a common sequence, we optimized the structures using a sequence of rotamer-based side-chain packing,[33] rigid-body refinement,[2] relaxation, and minimization[14,15] steps. We then scored the structures using the RosettaDock score function and selected structures with a balance of good interface and total scores after visual inspection of the structures. Using this approach, we were able to achieve a high-, medium-, and acceptable-quality prediction for Targets 47, 49, and 50. For Target 51 involving multiple xylanase domains, we refined the structures after removing connecting loops and domain 5. We manually analyzed the structures and submitted those with good interface and total scores and closable loop distances. The submission achieved an acceptable-quality prediction.

During the later CAPRI scoring rounds, we automated the process by developing an "elliptic score" for discrimination.[34] The elliptic score is based on the observation that near-native structures have both low interface and total scores. Instead of simply using a linear combination (as in Ref. 25), we combined the two scores nonlinearly, increasing the penalty for being very high in only one of the two constituent scores using elliptic level sets in the total score-interface score plane. A geometric depiction of the score derivation (Supporting Information Fig. S7) and a formula are provided in the Supporting Information. We used the new workflow and elliptic score for Target 53 to complete the scoring challenge in an entirely automated fashion. After standard structural refinement, the top-scoring structures yielded a medium-quality prediction, tying for third position among the 11 scoring challenge participants.

DISCUSSION

In the prior CAPRI rounds 13–19,[4] we relied on flexible backbone implementations of the RosettaDock algorithm, EnsembleDock and SnugDock, for predicting protein–protein interfaces. However, in more than half of the CAPRI rounds 20–27, we developed and used new protocols with improvements such as incorporating environmental effects (pH) and nonprotein biomolecules (SugarDock). CAPRI not only helped us to expand the boundaries of RosettaDock but also provided us a guide to future improvements and areas of focus. The powerful object-oriented design of the Rosetta modeling suite[35] proved a major asset as we adapted our protocols to address an array of targets. Here, we outline the steps to improve the new protocols designed for the CAPRI targets and a few possible solutions for the aforementioned challenges.

The recent CAPRI rounds tested the versatility of RosettaDock in tackling diverse docking challenges. As many targets required new protocol development and testing in stringent time frames, there is tremendous opportunity for further improvements in the prototype methods. For example, in interface water prediction (Target 47), we docked each water molecule individually followed by superposition and elimination of clashing water molecules to achieve the final prediction. An advantage of our approach is the speed gain by using a semi-implicit water treatment, similar to other recent work.[36,37] The CAPRI results reveal that additional work is needed to account for the interactions of multiple water molecules through simultaneous multiple ligand docking.

Simulation of sugar molecules is challenging because of their high conformational flexibility. The Rosetta modeling suite provides a potentially fast and accurate platform for modeling sugars as it uses fast sampling of discrete libraries of pregenerated low-energy conformers. For Target 57, our algorithm incorporated several sugar flexibility moves including sampling of the side-chain and backbone torsional angles and ring conformation moves, leading to an acceptable-quality prediction. Rigorous benchmarking and generalizing the tool to accommodate the vast chemical diversity of oligosaccharide residues will be needed to deploy the protocol widely for the many important structural biology problems involving sugars.[38]

Traditionally, Rosetta has not included the effects of environmental conditions such as temperature, salt, or pH. We recently incorporated tools to account for pH by dynamically sampling side-chain protonation states.[23] Protonation and deprotonation is critical in docking as protonation state change contributes to the binding energy in approximately half of the protein complex interfaces of Docking Benchmark 4.0.[39] Hence, we developed the Rosetta pH-sensitive docking algorithm to sample residue protonation states dynamically during simulations. Target 58 was the ideal case for the successful first blind test of our prototype pH-sensitive docking algorithm. However, for Target 58, even standard docking produced comparable results indicating room for further improvements.

Besides new code development, two familiar factors made the recent CAPRI rounds especially demanding. First, global docking predictions were unreliable, which has been a recurring theme for us in several past CAPRI rounds.[3,4,40] We used global docking to

successfully predict a medium-quality structure for the lysozyme–inhibitor complex (Target 58). However, global docking limitations hindered our performance on the other targets. To test the reasons for the docking failures, we selected the crystal structures of Target 46 (3Q87[41]) and Target 50 (3R2X[42]) and used RosettaDock global and local docking runs to generate decoys. For Target 46, the energy funnel plot generated from the global docking decoys showed very limited sampling within 5 Å from the native structure [Supporting Information Fig. S4(A)]. In addition, the decoys generated from native crystal structure refinement scored much lower than the global docking decoys. These observations clearly implicate sampling limitations of global RosettaDock. For Target 50, local RosettaDock produced a pronounced energy funnel [Supporting Information Fig. S4(B)], but global docking failed, again showing the challenge of sampling a larger conformational space. For Target 49, we were able to achieve an acceptable-quality prediction during the scoring challenge, but failed during the predictions, further highlighting global sampling issues. Biochemical information can be used to reduce the search space, but finding reliable experimental data can be difficult if the targets are not well studied. As recent studies show that templates for all possible structural complexes are already available,[43] using protein complex databases[44,45] to reduce the conformational search space might be a viable solution. In addition, other approaches such as the replica exchange docking algorithm,[46] starting docking simulations from decoy clusters based on inter-residue contacts[47] or pairwise RMSD,[48] and using faster FFT-based docking algorithms[49,50] are possible options to improve the sampling during global docking.

Second, targets from recent rounds often involved homology modeling and binding affinity predictions, two tasks that are not the primary focus of Rosetta-Dock, which was designed to predict accurate complex predictions from unbound crystal coordinates. Seven of the targets in CAPRI rounds 20–27 required homology modeling for at least one of the protein partners. Previously, we demonstrated using SnugDock (antibody–antigen complexes) that an ensemble of homology models can compensate for homology modeling errors.[6] For Target 53, using EnsembleDock with homology models helped us to achieve an acceptable-quality prediction. Unfortunately, none of the other three submissions generated using EnsembleDock were successful. In addition, Targets 43–45 and 55 and 56, which involved binding affinity predictions, were especially challenging. The uninspiring performance for these targets implicates deficiencies in the Rosetta score function. The score function was tuned to maximize the energy gap between near-native and nonnative structures. Calibrating the score function using an objective function with constraints on both binding energy errors (compared with experimental values) and RMSD of an ensemble of structures (including homology models) from the native crystal structure might improve the prediction accuracy.

Recent work has expanded the horizon of structural biological simulations to cellular-level interactomes,[44,51–54] indicating the need for docking algorithms to adapt and face the current challenges. Our recent additions to RosettaDock are a small step expanding the capabilities beyond idealized protein–protein docking; however, a giant leap is required to move toward the ultimate goal of biocomputational algorithms: a live, whole-cell biomolecular interaction simulation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the CAPRI organizers, specifically Dr. Sameer Velankar and Dr. Marc Lensink for the communications answering the target-related queries and for evaluating all the final targets, and Dr. Julia Koehler Leman for feedback on the manuscript. We thank the crystallographers who offered their complexes as the CAPRI targets. We also thank all the developers of the Rosetta molecular modeling suite (www.rosettacommons.org), which was used as the basis for all the studies.

Grant sponsor: U.S. Department of Health and Human Services National Institute of General Medical Sciences; grant number: R01-GM078221; Grant sponsor: National Science Foundation; grant number: CBET-0846324; Grant sponsor: European Commission European Research Council; grant number: 310873; Grant sponsor: National Science Foundation; grant number: CHE-1151554; Grant sponsor: National Science Foundation; grant number: IOS-1126971; Grant sponsor: U.S. Department of Health and Human Services National Cancer Institute; grant number: U54-CA143907; Grant sponsor: U.S. Department of Health and Human Services National Institute of General Medical Sciences; grant number: T32-GM88118; Grant sponsor: Israel Science Foundation; grant number: 319/11; Grant sponsor: USA-Israel Binational Science Foundation; grant number: 2009418.

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. [PubMed: 12784359]
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. Chaudhury S, Sircar A, Sivasubramanian A, Berrondo M, Gray JJ. Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6–12. Proteins. 2007; 69:793–800. [PubMed: 17894347]
4. 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]
5. 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]
6. 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]
7. Mendez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins. 2005; 60:150–169. [PubMed: 15981261]
8. 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. [PubMed: 10592235]
9. Meenan NAG, Sharma A, Fleishman SJ, MacDonald CJ, Morel B, Boetzel R, Moore GR, Baker D, Kleanthous C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci USA. 2010; 107:10080–10085. [PubMed: 20479265]
10. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999; 35:133–152. [PubMed: 10223287]
11. Eswar, N.; Webb, B.; Marti-Renom, MA.; Madhusudhan, MS.; Eramian, D.; Shen, M.; Pieper, U.; Sali, A. Current protocols in protein science. Wiley; 2001. Comparative protein structure modeling using MODELLER. DOI: 10.1002/0471140864.ps0209s50
12. Kühlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C. Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. J Mol Biol. 2000; 301:1163–1178. [PubMed: 10966813]

13. Ko T-P, Liao C-C, Ku W-Y, Chak K-F, Yuan HS. The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. Structure. 1999; 7:91–102. [PubMed: 10368275]

14. Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science. 2005; 309:1868–1871. [PubMed: 16166519]

15. Misura KMS, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins. 2005; 59:15–29. [PubMed: 15690346]

16. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998; 102:3586–3616.

17. Meiler J, Baker D. RosettaLigand: protein–small molecule docking with full side-chain flexibility. Proteins. 2006; 65:538–548. [PubMed: 16972285]

18. 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]

19. Kyomuhendo P, Myrnes B, Brandsdal B-O, Smalås AO, Nilsen IW, Helland R. Thermodynamics and structure of a salmon cold active goose-type lysozyme. Comp Biochem Physiol B Biochem Mol Biol. 2010; 156:254–263. [PubMed: 20398783]

20. Leysen S, Vanderkelen L, Van Asten K, Vanheuverzwijn S, Theuwis V, Michiels CW, Strelkov SV. Structural characterization of the PliG lysozyme inhibitor family. J Struct Biol. 2012; 180:235–242. [PubMed: 22634186]

21. Svergun D, Barberato C, Koch MHJ. CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr. 1995; 28:768–773.

22. Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry. 1990; 29:2403–2408. [PubMed: 2337607]

23. Kilambi KP, Gray JJ. Rapid calculation of protein pK_a values using Rosetta. Biophys J. 2012; 103:587–595. [PubMed: 22947875]

24. Drew K, Renfrew PD, Craven TW, Butterfoss GL, Chou F-C, Lyskov S, Bullock BN, Watkins A, Labonte JW, Pacella M, Kilambi KP, Leaver-Fay A, Kuhlman B, Gray JJ, Bradley P, Kirshenbaum K, Arora PS, Das R, Bonneau R. Adding diverse noncanonical backbones to Rosetta: enabling peptidomimetic design. PLoS One. 2013; 8:e67051. [PubMed: 23869206]

25. Raveh B, London N, Zimmerman L, Schueler-Furman O. Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS One. 2011; 6:e18934. [PubMed: 21572516]

26. Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, Zhou H-X, Mitchell JC, Demerdash ONA, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko J, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovalı ŞK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Pérez-Cano L, Pons C, Fernández-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AMJJ, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodríguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang S-Y, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol. 2011; 414:289–302. [PubMed: 22001016]

27. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007; 8:610–618. [PubMed: 17637733]

28. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JPGLM, Trellet M, Bonvin AMJJ, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Kilambi KP, Gray JJ, Popov P, Grudinin S, Esquivel-Rodríguez J, Kihara D, Zhao N, Korkin D, Zhu X, Demerdash

ONA, Mitchell JC, Kanamori E, Tsuchiya Y, Nakamura H, Lee H, Park H, Seok C, Sarmiento J, Liang S, Teraguchi S, Standley DM, Shimoyama H, Terashi G, Takeda-Shitaka M, Iwadate M, Umeyama H, Beglov D, Hall DR, Kozakov D, Vajda S, Pierce BG, Hwang H, Vreven T, Weng Z, Huang Y, Li H, Yang X, Ji X, Liu S, Xiao Y, Zacharias M, Qin S, Zhou H-X, Huang S-Y, Zou X, Velankar S, Janin J, Wodak SJ, Baker D. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. Proteins. 2013 in press.

29. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 2006; 22:195–201. [PubMed: 16301204]
30. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL repository and associated resources. Nucleic Acids Res. 2009; 37(Database Issue):D387–D392. [PubMed: 18931379]
31. Berrondo M, Ostermeier M, Gray JJ. Structure prediction of domain insertion proteins from structures of individual domains. Structure. 2008; 16:513–527. [PubMed: 18400174]
32. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. 2009; 6:551–552. [PubMed: 19644455]
33. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997; 6:1661–1681. [PubMed: 9260279]
34. Porter, J. Honors Research Paper. Johns Hopkins University; Baltimore, MD: 2012. The selection of near-native decoys in protein-protein docking by score combinations using the Rosetta all-atom force field.
35. Leaver-Fay A, Tyka M, Lewis S, Lange O, Thompson J, Jacak R, Kaufman K, Renfrew D, Smith C, Sheffler W, Davis I, Cooper S, Treuille A, Mandell D, Richter F, Ban Y-EA, Fleishman S, Corn J, Kim D, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek J, Karanicolas J, Das R, Meiler J, Kortemme T, Gray J, 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]
36. Fennell CJ, Kehoe CW, Dill KA. Modeling aqueous solvation with semi-explicit assembly. Proc Natl Acad Sci USA. 2011; 108:3234–3239. [PubMed: 21300905]
37. Lee MS, Salsbury FR, Olson MA. An efficient hybrid explicit/implicit solvent method for biomolecular simulations. J Comput Chem. 2004; 25:1967–1978. [PubMed: 15470756]
38. Wormald MR, Petrescu AJ, Pao Y-L, Glithero A, Elliott T, Dwek RA. Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. Chem Rev. 2002; 102:371–386. [PubMed: 11841247]
39. Hwang H, Vreven T, Janin J, Weng Z. Protein–protein docking benchmark version 4.0. Proteins. 2010; 78:3111–3114. [PubMed: 20806234]
40. Daily MD, Masica D, Sivasubramanian A, Somarouthu S, Gray JJ. CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock. Proteins. 2005; 60:181–186. [PubMed: 15981262]
41. Liger D, Mora L, Lazar N, Figaro S, Henri J, Scrima N, Buckingham RH, van Tilbeurgh H, Heurgué-Hamard V, Graille M. Mechanism of activation of methyltransferases involved in translation by the Trm112 “hub” protein. Nucleic Acids Res. 2011; 39:6249–6259. [PubMed: 21478168]
42. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science. 2011; 332:816–821. [PubMed: 21566186]
43. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA. 2012; 109:9438–9441. [PubMed: 22645367]
44. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013; 10:47–53. [PubMed: 23399932]
45. Meyer MJ, Das J, Wang X, Yu H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics. 2013; 29:1577–1579. [PubMed: 23599502]

46. Zhang Z, Lange OF. Replica exchange improves sampling in low-resolution docking stage of RosettaDock. PLoS One. 2013; 8:e72096. [PubMed: 24009670]
47. Oliva R, Vangone A, Cavallo L. Ranking multiple docking solutions based on the conservation of inter-residue contacts. Proteins. 2013; 81:1571–1584. [PubMed: 23609916]
48. London N, Schueler-Furman O. FunHunt: model selection based on energy landscape characteristics. Biochem Soc Trans. 2008; 36:1418–1421. [PubMed: 19021567]
49. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. Proteins. 2006; 65:392–406. [PubMed: 16933295]
50. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. Proteins. 2007; 69:511–520. [PubMed: 17623839]
51. Vakser IA. Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol. 2013; 23:198–205. [PubMed: 23294579]
52. Frembgen-Kesner T, Elcock A. Computer simulations of the bacterial cytoplasm. Biophys Rev. 2013; 5:109–119. [PubMed: 23914257]
53. Roberts E, Stone JE, Luthey-Schulten Z. Lattice microbes: high-performance stochastic simulation method for the reaction-diffusion master equation. J Comput Chem. 2013; 34:245–255. [PubMed: 23007888]
54. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012; 490:556–560. [PubMed: 23023127]

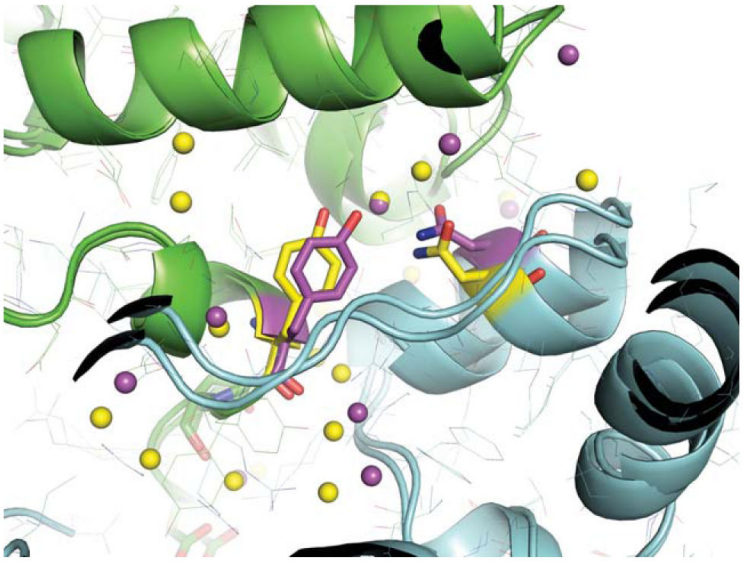

Figure 1.
Target 47: Water-mediated residue contacts at the interface of colicin E2 DNase domain–Im2 immunity protein complex. The predicted contacts in the model (magenta) are compared with those in the crystal structure (yellow). In the best model, 13 buried water molecules were predicted, capturing 46% of native water-mediated residue contacts.

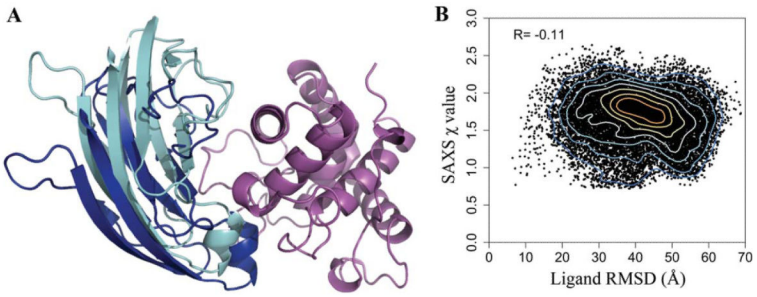

Figure 2.

A: Medium-quality prediction of the g-type lysozyme in complex with the PliG inhibitor. The orientation of the inhibitor generated using pH-sensitive docking at an environment pH of 6.2 (cyan) is compared with the native crystal structure (blue) after superimposing the coordinates of the lysozyme. **B**: Ligand RMSD (Lrmsd) versus SAXS χ value evaluated using CRYSOL for top-scoring 10% of predicted complex structures.

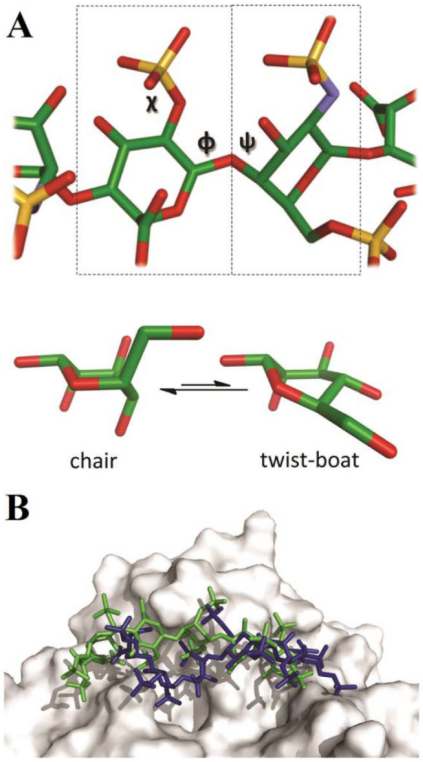

Figure 3.

Target 57: **A**: Detail of the sugar model, outlining the degrees of freedom manipulated by our algorithm: the backbone φ/ψ angles, side-chain χ angles, and ring conformations (two representative conformations shown of the eight low-energy conformations in the library). **B**: Acceptable-quality prediction (green) of the heparin-like oligosaccharide in complex with the protein BT4661 is compared with the crystal ture (blue). The model recovered 71% of the native sugar–protein residue–residue contacts.

Table I

RosettaDock Performance in CAPRI Rounds 20–27

Target No.	Target	Special	Method	Search scale	Type	Prediction rating	Metrics of the best prediction model			Scoring rating
							Irmsd	Lrmsd	f_{nat}	
T46	Mtq2–Trm112		EnsembleDock	Global	H-H	–	5.95	20.53	0.03	–
T47	E2–Im2	Interface water	EnsembleDock/ RosettaLigand	Local	H-U	***/++	0.65	1.34	0.82	***
T48	T4moC–[Hydroxylase + T4moD]		EnsembleDock	Global	U-U	–	5.73	15.05	0	–
T49	T4moC–Hydroxylase		EnsembleDock	Global	U-U	–	5.01	14.8	0	*
T50	HB36.3–Hemagglutinin		RosettaDock	Global HB36.3	H-B	–	8.03	16.99	0.04	**
							12.66	34.94	0.09	
							11.82	24.74	0.13	
T51	Xylanase domains	Multidomain	RosettaDock	Global	H-U	–	17.48	44.03	0.07	*
T53	Rep4–Rep2		EnsembleDock	Global	H-U	*	3.34	7.02	0.15	**
T54	Rep16–Neocarzinostatin		EnsembleDock	Global	H-U	–	12	37.79	0.06	–
T57	BT4661–Heparin	Sugar	SugarDock	Global/Patches	H-U	*	2.69	6.5	0.71	na
T58	Lysozyme–PliG inhibitor	SAXS	pH-dock/EnsembleDock	Local	U-U	**	1.94	7.85	0.44	na

Target, the docking partners; Special, nonconventional additional challenges for each target; Methods, the protocol used to solve the target; Search scale, the level of sampling for both the docking partners; Type, the docking partner models used: homology models (H), unbound structure (U), or bound structure (B); Prediction rating, the CAPRI rating of the best predicted model; the interface RMSD (Irmsd), ligand RMSD (Lrmsd), and fraction of recovered native contacts (f_{nat}) of the best pre dicted model are also shown; Scoring rating, the CAPRI rating of the best model in the scoring rounds. High, medium, and acceptable quality predictions are indicated by ***, **, and *, respectively. Targets for which scoring rounds were not conducted are denoted as not applicable (na).