Simultaneous prediction of protein folding and docking at high resolution

Rhiju Das[a,1,2], Ingemar André[a,1], Yang Shen[b], Yibing Wu[c], Alexander Lemak[d], Sonal Bansal[e], Cheryl H. Arrowsmith[d], Thomas Szyperski[c], and David Baker[a,3]

[a]Department of Biochemistry, University of Washington, Seattle WA 98195; [b]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892; [c]Departments of Chemistry and Structural Biology and Northeast Structural Genomics Consortium, State University of New York, Buffalo, NY 14260; [d]Ontario Cancer Institute, Department of Medical Biophysics, and Northeast Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada M5G IL5; and [e]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Saint Louis, MO 63108

Edited by Ken A. Dill, University of California, San Francisco, CA, and approved August 7, 2009 (received for review April 21, 2009)

Interleaved dimers and higher order symmetric oligomers are ubiquitous in biology but present a challenge to de novo structure prediction methodology: The structure adopted by a monomer can be stabilized largely by interactions with other monomers and hence not the lowest energy state of a single chain. Building on the Rosetta framework, we present a general method to simultaneously model the folding and docking of multiple-chain interleaved homo-oligomers. For more than a third of the cases in a benchmark set of interleaved homo-oligomers, the method generates near-native models of large α-helical bundles, interlocking β sandwiches, and interleaved α/β motifs with an accuracy high enough for molecular replacement based phasing. With the incorporation of NMR chemical shift information, accurate models can be obtained consistently for symmetric complexes with as many as 192 total amino acids; a blind prediction was within 1 Å rmsd of the traditionally determined NMR structure, and fit independently collected RDC data equally well. Together, these results show that the Rosetta "fold-and-dock" protocol can produce models of homo-oligomeric complexes with near-atomic-level accuracy and should be useful for crystallographic phasing and the rapid determination of the structures of multimers with limited NMR information.

homo-oligomers | molecular replacement | NMR structure inference | protein structure prediction | symmetry

The majority of expressed proteins function within symmetrical homomeric complexes (1–3). Although a boon for evolving functional diversity (4), this ubiquity of oligomeric structures poses numerous challenges for modern structural biology. The phasing of crystallographic data by molecular replacement and NMR structural inference are complicated by the increasing number of degrees of freedom and spectral degeneracy, respectively, in multimeric systems. The ability to predict de novo the structures of multiple interacting molecular chains would potentially alleviate these problems, allowing unphased diffraction measurements or ambiguously assigned NMR spectra to be resurrected as constraints or as independent validation for *in silico* structural inference.

Accurate modeling of previously unseen homomeric structures has not been demonstrated at high resolution. Significant progress has occurred in modeling the folds of individual monomeric soluble proteins (5–7) and the docking arrangements of predefined monomers (8, 9). However, as proteins interact with other proteins, nucleic acids, or smaller molecules, their lowest free-energy backbone conformations typically shift in response to their partners. These often dramatic structural changes have been a persistent and unsolved issue in blind prediction trials of recent years (8).

In this report, we show how two different methods developed for de novo conformational sampling (for folding and for docking) can be melded into a more general procedure that permits the blind prediction of intertwined complexes of proteins at near-atomic resolution. Some of the test systems involve up to several hundred residues, much larger than previous targets of high-resolution de novo modeling; symmetry plays a crucial role in reducing the number of degrees of freedom that need to be sampled. The new approach is based on the Rosetta framework for molecular modeling (10).

Results and Discussion

Overview of Rosetta "Fold-and-Dock" Protocol. For each molecular complex, we start from fully extended monomer chains in randomly generated symmetric rigid body arrangements (Fig. 1). The first of two stages of automated de novo modeling alternates sets of fragment insertion moves that perturb the backbone conformation of protein monomers (11) with moves that perturb the symmetric docking arrangement of the monomers (12). The conformational energy is determined by the low-resolution energy function previously developed for monomer protein structure prediction (6, 11) with coarse-grained terms for backbone hydrogen bonds and hydrophobic interactions. Cyclic or dihedral symmetry is maintained by cloning the moves to separate monomers, and each move is accepted or rejected based on the standard Metropolis criterion (12). In the second, high-resolution refinement stage, side chains are built into the backbone conformations in all-atom detail, and small, symmetric perturbations are tested in the context of a reasonably accurate, high-resolution energy function that includes van der Waals interactions, the costs of desolvating atoms, and hydrogen bonds (Fig. 1) (10, 13). The lowest energy models are then clustered, and the five most populous clusters are compared with the experimentally determined structures.

We first tested the fold-and-dock protocol on a benchmark set of 27 protein complexes with structures previously determined by high-resolution crystallography. Most of these molecules form intertwined structures in which the number of intermolecular contacts approaches one third of the number of intramolecular contacts (Table 1). We have assumed that the stoichiometry of the crystallized complex is known (as can be rapidly ascertained experimentally by, for example, analytical ultracentrifugation), but tested all possible symmetries available to a fixed number of monomers. For example, tetramer complexes were modeled with both D2 and C4 symmetries.

Benchmark of Homo-Oligomeric Proteins Solved by X-Ray Crystallography. Initial tests were carried out on coiled coil structures with simple cyclic geometries as well as more complex dihedral symme-

Author contributions: R.D., I.A., and D.B. designed research; R.D., I.A., Y.S., Y.W., A.L., and S.B. performed research; R.D., I.A., Y.S., Y.W., A.L., S.B., C.H.A., T.S., and D.B. analyzed data; and R.D., I.A., and D.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]R.D. and I.A. contributed equally to this work.

[2]Present address: Departments of Biochemistry and Physics, Stanford University, Stanford, CA 94035.

[3]To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904407106/DCSupplemental.

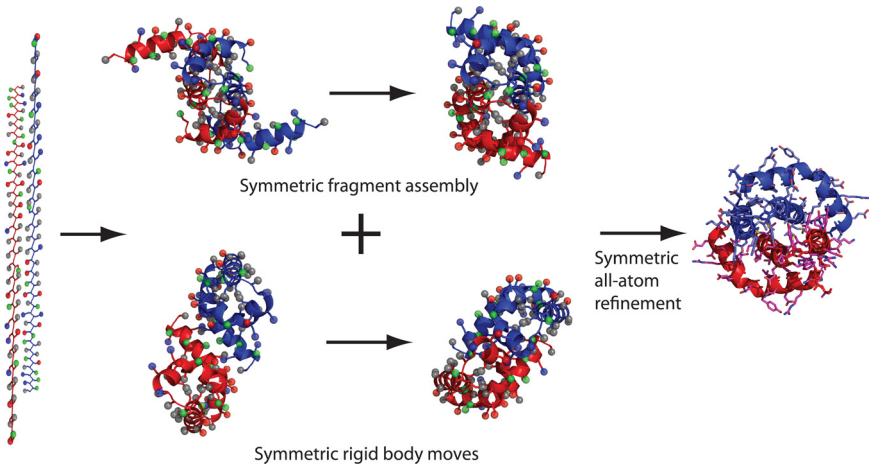

Fig. 1. Overview of the Rosetta fold-and-dock protocol. Starting with extended protein chains, the protein complex is assembled through alternating cycles of fragment insertions, which change the internal coordinates of the monomer, and rigid body perturbations while using a low-resolution representation of the protein. The assembled complexes are then subjected to symmetric all-atom refinement.

tries. These molecules include biologically important complexes such the trimeric coiled coil region of coronin-1 (C3; Fig. 2*B*) and engineered helical proteins with dihedral symmetry (D2; Fig. 2*K*).

In each of these cases, at least one of the five most populous clusters produced by the symmetric folding and docking of multiple chains achieves a backbone accuracy approaching 2 Å or better (C^{α} rmsd; Table 1). Also, in each of these cases, the fine details of hydrophobic side-chain packing are reproduced with near-atomic accuracy (Fig. 2 *A–D*). As a further, blind test of the method, we predicted the structure of a Golgi-associated coiled coil region (14) whose atomic coordinates were released after our modeling trials. The crystal structure and the single Rosetta dimer prediction agree with a C^{α} rmsd of 0.94 Å (Fig. 3*A*). These results demonstrate that the basic principles underlying Crick's "knobs-into-holes"-based manual modeling (15, 16), more constrained conformational searches (17), and other phenomenological rules (18) for the simplest coiled coils can be automatically and quite generally applied to the de novo modeling of complex symmetric helix interaction motifs.

Although helical coiled coils have partly stereotyped conformations, the generality of our method allows it to be applied to symmetric complexes with nonstereotyped folds. Thus, we included in the benchmark set targets with intermolecular beta-strand pairings or helix–loop–helix segments. The cases for which fold-and-dock modeling achieved high-resolution models included not only noncanonical helical bundles (Fig. 2*G*), but also an all-beta protein (Fig. 2*H*) and mixed alpha-beta proteins including the p53 tetramerization motif (Fig. 2*E*). Overall, 11 of the 27 tested molecular complexes gave high-resolution Rosetta models with better than 3-Å backbone accuracy over the full oligomer (Table 1). Further tests demonstrated the importance of simultaneously modeling the folds of chains along with their docking arrangement. In 9 of the 11 cases, the protein fold could not be predicted if the monomer sequence was modeled alone (Table 1).

With more than one third of the tested complexes leading to near-atomic resolution models, the rate of success in the benchmark is similar to what is achievable for small monomeric proteins (5, 10). In the cases for which the fold-and-dock method produced high-accuracy models, the extent of convergence of the lowest energy models is a good predictor of model accuracy. For example, there are 10 cases in the benchmark (Table 1), in which at least 40 of the lowest 400 energy models are within 2 Å of each other; nearly all (9 of 10) cases result in high-accuracy models. Perhaps the strongest parallel with previous efforts in modeling biomolecule structures can be drawn from the examples in which the present protocol fails (Table 1). In the majority of these cases (11 out of 16), the energy of the native conformation is lower than the energies achieved with the present level of computational sampling; also, the accuracy of the sampled models is beyond the 2- to 3-Å radius of convergence required to robustly discriminate native-like conformations (8 out of 16 do not sample a single structure better than 3 Å) (Table 1; Fig. S1) (5, 10). Systems with the largest number of degrees of freedom were the most difficult; our protocol failed in all eight cases in which the number of monomer residues exceeded 60. As with the separate problems of monomer structure modeling and docking of prestructured monomers, conformational sampling appears to be the major bottleneck in the simultaneous modeling of the folding and docking of symmetric complexes.

Estimating Phases for Molecular Replacement from High-Resolution Models. The benchmark tests described above suggest that predicting the structure of multimeric proteins from sequence is an attainable goal. We further tested whether this de novo method might have a practical impact on modeling problems commonly encountered in experimental structural biology. One stringent test of the modeling procedure is its ability to provide estimated phases for X-ray crystal diffraction data of the protein of interest via molecular replacement (19, 20). Idealized and simplified multimeric helix assemblies (21, 22) have indeed allowed recent phasing of diffraction datasets, although these efforts have relied on high-throughput screens of hundreds of stereotyped backbone-only conformations. The high-resolution Rosetta models presented in this study offer the possibility of phasing diffraction datasets for nonstereotyped multimeric complexes.

For 24 of the 28 test cases, crystallographic structure factors were publicly available and permitted molecular replacement trials with the Phaser software. Trials with the lowest energy structures derived from de novo modeling of individual chains led to only one case of unambiguous molecular replacement (Table S1). In contrast, the lowest energy models obtained by simultaneous modeling of the fold and docking arrangement gave phasing successes for 10 of these 24 datasets (Table S1), and enabled automated coordinate building for the majority of crystallized residues, using the ARP/wARP software package (23). Thus, as above, it was critical to simultaneously model the folds of chains along with their docking arrangement.

Inference of Oligomeric Protein Structure from Chemical Shift Data. As a final test, we applied the Rosetta fold-and-dock method to a difficult problem in NMR structural inference, the modeling of homo-oligomers using limited data. Experimental collection and assignment of distance constraints necessary for structure calculation remains a time-consuming and laborious task. The process is further complicated for homo-oligomers; spectral symmetry renders intra and intermolecular distance constraints indistinguishable. Specialized experiments have been developed to distinguish

Table 1. Result of the fold-and-dock protocol on a benchmark of protein complexes solved by crystallography and NMR

PDB ID	Modeled residues†	Symmetry	Intermolecular contacts, %	rmsd oligomer prediction, Å‡§	Cluster size¶	Lowest rmsd top 400, Å‡§	rmsd monomer extracted from oligomer, Å§‖	Traditional de novo single subunit rmsd, Å§
				Crystal structure benchmark				
2mlt	26x4	D2	23	7.9	20	1.5	3.1 (1.2)	5.9
2zta	31x2	C2	17	1.6	251	0.6	1.4 (1.0)	10.8
1bdt	44x2	C2	24	12.9	20	0.8	7.0 (5.9)	6.2
2bti	44x2	C2	28	0.9	27	0.8	0.9 (1.5)	4.9
1wrp	47x2	C2	12	11.3	20	2.7	7.2 (7.6)	8.3
1 g6u	48x2	C2	12	1.7	355	0.9	1.5 (1.9)	10.1
1irq	48x2	C2	27	11.9	10	3.3	5.6 (5.2)	10.6
2akf	32x3	C3	20	2.0	343	1.0	1.5 (1.6)	3.6
2odk	50x2	C2	15	12.5	10	3.6	11.8 (6.1)	5.9
2p64	51x2	C2	21	4.8	40	2.6	3.4 (3.6)	6.0
2 h3r	54x2	C2	18	5.3	26	2.1	2.9 (3.3)	10.2
1rop	56x2	C2	14	1.4	40	0.8	1.1 (1.2)	0.8
1igu	59x2	C2	14	14.6	10	5.3	3.7 (6.5)	6.3
1zv1	59x2	C2	11	1.3	29	1.3	0.9 (1.8)	1.2
1 g2z	31x4	D2	30	8.7	70	2.4	3.0 (2.8)	4.9
2fqm	63x2	C2	29	15.6	10	2.4	3.7 (6.3)	5.6
2or1	63x2	C1**	3	8.9	10	2.6	2.6 (1.7)	1.6
1c26	32x4	D2	29	2.7	6	1.8	1.5 (2.8)	5.6
1utx	66x2	C2	8	8.8	10	2.6	2.7 (3.2)	3.4
2dlb	69x2	C2	38	16.2	10	9.1	3.7 (14.1)	10.1
1mby	75x2	C2	15	15.5	10	7.0	3.7 (12.8)	12.6
2nzc	79x2	C2	8	13.5	10	4.5	3.9 (3.8)	2.9
2o1j	43x4	C4	24	1.2	343	1.0	1.8 (0.7)	16.4
1qx8	47x4	D2	30	1.7	56	1.0	1.4 (0.8)	17.8
1fe6	52x4	C4	23	2.4	212	1.9	1.1 (1.9)	20.4
1mz9	45x5	C5	28	1.3	132	1.1	1.7 (1.8)	7.7
1uis	64x6	C6	9	11.6	10	9.8	3.8 (9.9)	2.4
1i8f	71x7	C7	16	11.4	10	10.0	10.1 (7.5)	10.0
				Crystal structure blind prediction††				
3bbp	29x2	C2	15	0.9 (13.7)	—	0.9	1.0 (3.6)	7.9

PDB ID‡‡	Modeled residues*	Symmetry	Intermolecular contacts, %	rmsd, Å*§§		Lowest rmsd top 400, Å*§§	rmsd monomer, Å*§§	rmsd monomer prediction, Å
				NMR benchmark (chemical-shift-assisted)				
2nwt (gr83)	38x2	C2	26.5	1.9 (9.3)	—	0.9	1.6 (6.0)	6.5
2bzb	48x2	C2	14.4	2.6 (1.7)	—	1.4	2.3 (1.3)	2.1
2js5 (mcr1)	63x2	C2	5.9	7.6 (18.6)	—	3.6	2.6 (2.9)	3.6
2rmm	56x2	C2	7.5	3.2 (1.7)	—	0.8	0.6 (1.3)	1.2
1ns1	73x2	C2	9.5	1.0 (1.4)	—	0.9	0.9 (2.3)	2.5
2b95 (hr2106)	96x2	C2	9.5	1.7 (2.0)	—	1.7	1.7 (1.7)	1.7
				NMR blind predictions				
2k5j (sft1)	44x2	C2	24.3	1.3 (1.1)	—	0.9	1.0 (1.0)	3.9
2k7i (AtT3)	62x2	C2	24.6	5.3¶¶ (8.0)	—	2.7	1.7 (6.5)	6.6

*Only residues in the regular secondary structures are used for rmsd calculation. For definition, *SI Materials and Methods*.

†The C and N termini were trimmed to remove flexible tails. The length refer to the number residues modeled in the simulation.

‡Values for the most accurate of the five models selected through clustering.

§Calculated over Cα atoms over all residues.

¶Number of models in largest cluster, after clustering of 400 lowest energy complexes. This provides a measure of the extent of simulation convergence.

‖Values in parenthesis come from clustering the 400 lowest energy complexes with a single subunit extracted from the complex. The clustering method was identical to the one used for the full oligomer.

**Forms dimer only in the presence of DNA (38). This protein was used as a reference in the benchmark and the results indicate that, in the absence of strong intermolecular interactions, the correct monomeric fold can be found in the folding-and-docking simulation.

††Two competing orientations are observed: parallel and antiparallel. The largest clusters involve antiparallel coiled-coil orientations while biological data dictates the presence of a parallel orientation. The predicted model was the lowest energy parallel coiled-coil.

‡‡NESG target ID in parentheses.

§§Values for the model with lowest energy + chemical shift score. The values in parenthesis are from a clustering analysis using the same method as for the crystal benchmark.

¶¶Convergence was not observed for this blind prediction. Predictions were selected based on clustering of decoys with lowest energy + chemical shift score. The third of the five submitted models was the most accurate and has a rmsd of 3.0 Å (Fig. 3*E*).

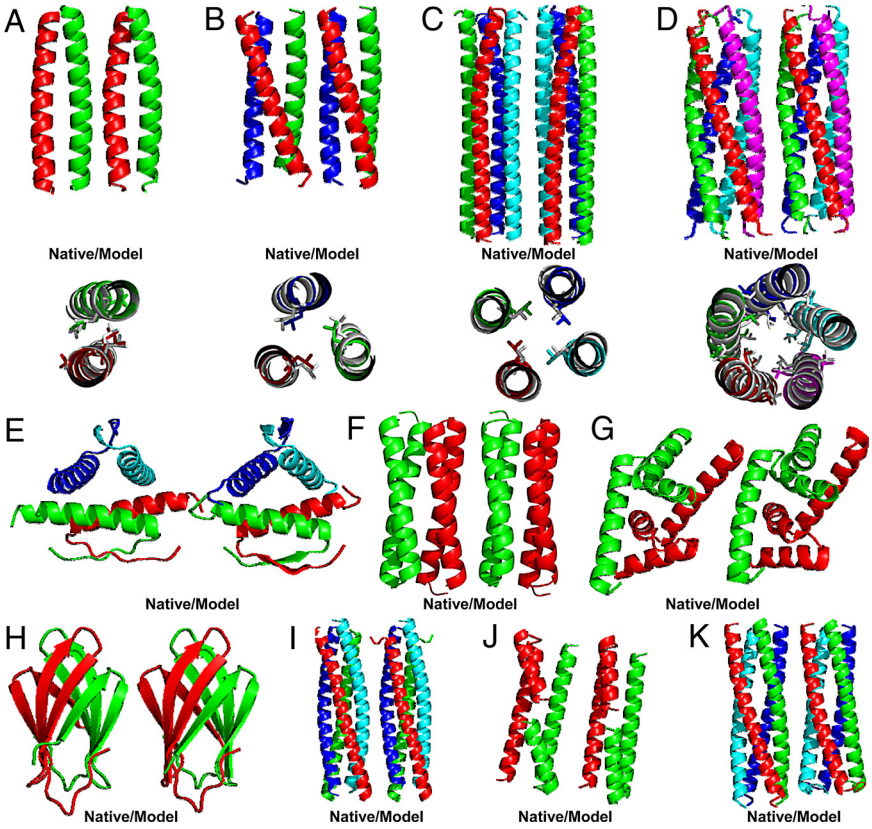

Fig. 2. High-resolution structure prediction using the fold-and-dock protocol. (*A–I*) Comparison between native crystal structure (*Left*) and most accurate of the five models selected by clustering of the low energy models (*Right*); PDB ID, 2zta (*A*); 2akf (*B*); 1fe6 (*C*); 1mz9 (*D*); 1c26 (*E*); 1rop (*F*); 1zv1 (*G*); 2bti (*H*); 2o1j (*I*); 1g6u (*J*); 1qx8 (*K*). For the cyclic coiled-coil structures (*A–D*), the packing of core leucine side chains is illustrated.

the two types of data, but these techniques require the preparation of mixed isotopically labeled samples, and typically yield only a few intermolecular constraints. To address this problem, we built on recent findings that NMR structure determination of monomers can be greatly accelerated by coupling de novo modeling with backbone chemical shift data without further NOE, scalar coupling, residual dipolar coupling (RDC), or isotope-edited measurements (24, 25). The structure of a dimeric complex was recently predicted from models of monomeric components generated using chemical shift fragments (26). We brought together the chemical-shift-derived fragment method with our symmetric folding-and-docking method to model eight complexes under investigation by the Northeast Structural Genomics consortium (NESG; Table 1). The set contains proteins with all-α, α/β, as well as all-β secondary structure in the range of 44 to 96 residues per monomer. Models were ranked based on a sum of the all-atom energy and a chemical shift score, which measures how well chemical shifts estimated from the models agree with the experimental values (25). The single model with lowest value of this combined score is taken as the prediction (Table 1 and Fig. 3; Fig. S2).

In six of the eight cases, the lowest energy Rosetta model had near-atomic accuracy (<3-Å C^{α} rmsd) (Fig. 3 *B–D*; Fig. S2). The overall rate of success (75%) is higher than the crystal structure benchmark described above (41%), due to enhanced conformational sampling with higher quality chemical-shift-derived fragments. The most remarkable result is seen for the protein HR2106, a mixed α/β protein with a total of 192 residues, resulting in a 1.7-Å model (Fig. 3*D*). This topologically complex protein is significantly larger than previous reported successes of chemical-shift-aided protein modeling (24, 25).

Blind Prediction of NMR Structures. For two of the NMR targets, blind predictions were made before NMR structures were released. First, significant convergence was not observed in the modeling of 2k7i, a complex of two 62-residue chains, suggesting that blind prediction would be inaccurate (Fig. S2). Five models were instead selected based on a combination of the all-atom energy and the chemical shift score. The third of these models is quite accurate, giving an rmsd of 3.0 Å (2.7–3.3 Å over the 20 conformers) over secondary structure elements compared with the model obtained with the complete NMR data (Fig. 3*E*). We investigated whether this quite good model could be distinguished using experimentally measured residual dipolar coupling (RDC) data, which provide orientational information useful for validating NMR structures (27). The correlation between measured and back-calculated residual dipolar couplings (the quality(Q)-factor (28)) was determined for each of the five models, and the third model indeed fit the data the best (Q-factor of 1.06, 1.41, 0.55, 0.73, and 1.03, respectively, for these five models). Thus, additional data can make possible accurate structure determinations when convergence is too poor for the correct structure to be unambigously identified (Fig. S3*A*).

The second blind NMR test case, for a slightly smaller complex, produced even more striking results. Modeling runs for the protein 2k5j, a complex of two 44-residue (number of structured residues) chains, gave outstanding convergence, with 52 of the 200 lowest energy models within 1.0-Å C^{α} rmsd of each other, suggesting that the modeling had reached near-atomic accuracy. Indeed, when compared with 20 NMR conformers determined subsequently and independently using a full suite of 2136 NOE constraints (Fig. 3*F*), the lowest energy Rosetta model achieved an accuracy of 0.79–1.20 Å over all modeled residues, and 0.79–1.09 Å over the secondary structure elements. Despite making use of far less data, the Rosetta model achieves equally good agreement with RDC data compared with the model obtained with the traditional NMR method (Q-factors of 0.36 for the lowest 10 energy models versus 0.38 for 10 first conformers, respectively) (Fig. 3*G*; Fig. S3*B*). The RDC comparison suggests that even higher accuracy might be achievable if RDC data are used for fold-and-dock refinement rather than just validation (29); Rosetta models with quality factors as low as 0.24 are

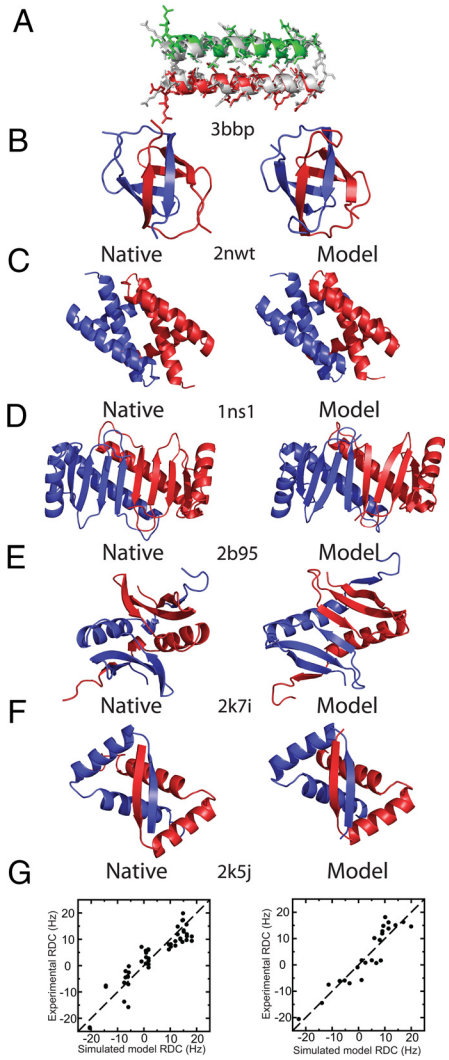

Fig. 3. High-resolution structure prediction using the fold-and-dock protocol. (*A*) Blind prediction of the structure of a Golgi-associated coiled coil region (PDB ID 3bbp). Red and green crystal structure, white blind prediction. (*B–E*) High-resolution structure prediction using the fold-and-dock protocol and chemical shift data. (*B–D*) Comparison between experimentally determined NMR structure (*Left*) and lowest energy Rosetta model (*Right*) for 2nwt (*B*), 1ns1 (*C*), and 2b95 (*D*). (*E*) Blind prediction of 2k7i. Experimentally determined structure (*Left*) and best blind prediction (*Right*). (*F*) Blind prediction of 2k5j. Experimentally determined structure (*Left*) and best blind prediction (*Right*). (*G*) Comparison between experimentally determined RDCs and simulated values for the first conformer in the NMR-ensemble of 2k5j refined without using NOE-data (*Left*) and best Rosetta model (*Right*).

present in the generated pool of models (Fig. S3*B*). Based on these and previous results, chemical-shift-based Rosetta has become an integral part of the NESG high-throughput structure determination pipeline for multimers (30). A third blind prediction of a dimeric structure was validated as this paper went to press. The predicted (lowest energy) structure, a non-intertwined monomer, had a backbone rmsd of 0.8 Å to the subsequently determined crystal structure. This illustrates that the method can produce excellent models for non-intertwined as well as intertwined oligomers–whether the lowest energy structure is intertwined or not is a function of the amino acid sequence, and need not be specified in advance.

Conclusion

We have demonstrated that simultaneous prediction of the fold and docking arrangement for protein complexes with many different symmetries and modest monomer sizes is feasible at high resolution. Further applications of the presented method can be envisioned in conjunction with experimental techniques beyond crystallography and NMR spectroscopy. First, large symmetric complexes are common targets of cryoelectron microscopy. Attaining high-resolution models from these low-resolution data may be possible by incorporating density maps into the conformational search. Second, oligomeric membrane complexes, which have been difficult to characterize experimentally, are attractive targets for the method. Many physiologically important channels and transporters are in a suitable size range for the protocol described here, and chemical cross-linking and accessibility data may provide rapid validation of high-resolution models. Last, the techniques described here, when coupled to existing design algorithms, provide a computational foundation for rationally engineering proteins that self-assemble into symmetric containers, fibrils, and channels of new and useful function.

Materials and Methods

The computational protocol developed in this work builds on previously described methods for de novo (5) and symmetrical protein assembly (12) structure prediction in Rosetta.

Description of Symmetry. The computational setup used previously in Rosetta to dock already folded monomeric structures into symmetric homo-oligomers has been described previously (12); in this previous work the internal backbone coordinates of each monomer were held fixed. This approach works for non-interleaved homo-oligomers but in general will fail for interleaved structures as the structure of the monomer depends strongly on interactions with other subunits. Here, we have retained the basic framework for modeling symmetric structures while adding in full backbone flexibility to allow monomers to fold up in the context of symmetrically related copies of themselves. During the simulations, backbone, side chain, and rigid body degrees of freedom are all varied. Backbone symmetry is enforced by maintaining identical bond angles, bond lengths, and torsion angles in symmetry-related partners in the system. Side chains are placed using two different methods; combinatorial and one-at-a-time noncombinatorial optimization (31, 32) using a backbone-dependent rotamer library (33). Rotamers are simultaneously inserted into all symmetry related partners of the system, and the energy of the full symmetrical oligomer is evaluated (12, 32). The most straightforward manner to describe rigid body symmetry is to maintain a global reference coordinate system and use symmetry transforms to find the coordinates of symmetry-related partners in the system. This system leads to difficulties in describing complex symmetries and to perform gradient based minimization. Instead, in Rosetta, each symmetry-related partner maintains its own reference coordinate systems that are in themselves related by symmetry. In this setup, the symmetry-related partners have identical coordinates, but in their own reference frames. During the simulation protocol, only rigid body degrees of freedom that maintain symmetry are allowed to vary. The nature of these perturbations depends on the type of symmetry that is simulated. In this work, two types of symmetries are used, cyclic and dihedral symmetry; complete descriptions of the samples degrees of freedom are given in Fig. S4.

Subsystems. For larger oligomers, it is unnecessary to explicitly simulate the whole systems, because a smaller subsystem can describe all of the interactions in the system if interactions at very long distances are ignored. The minimal number of monomers that explicitly needs to be simulated is equal to the number of different interaction surfaces between subunits in the system. To avoid edge effects, the smallest subsystem is a single monomer with all directly neighboring subunits present, and the total energy is calculated as a multiple of the energy of the central monomer.

Energy Function. The energy function in the low-resolution search is a linear combination of mostly knowledge-based terms modeling residue-environment and residue-residue interactions, secondary structure packing, chain density, and excluded volume (34). The high-resolution energy function is composed of a Lennard–Jones potential to model side-chain packing, an orientation-dependent hydrogen bond term parameterized from quantum mechanics (35) and analysis of high-resolution structures (36), the Lazaridis–Karplus implicit solvation model (37), pair-interaction terms modeling long range electrostatics, a side-chain torsional potential derived from the Dunbrack backbone-dependent rotamer library (33), and backbone torsional potential dependent on secondary structure

and amino acid type. Detailed description of the energy terms can be found in Rohl et al. (13) and in supporting information in Kuhlman et al. (38).

Conformational Search Protocol. The initial state of the simulated system is generated with random rigid body degrees of freedom and the subunit backbones in extended conformations. The protocol starts with a low-resolution symmetrical fragment insertion where torsion angles of randomly selected 3 or 9 residue fragments in the protein chains are replaced with torsion angles from nonhomologous proteins with known structure (6); 40,000 fragment insertions attempts are made. At every 1 in 10 fragment insertions, the rigid body of the protein chains are adjusted using two type of moves: random perturbations (translation/rotation) and sliding of subunits into atomic contact (translation). In the random perturbation step, the subunits are randomly and symmetrically rotated or translated up to 5° or 0.5 Å, respectively. In the sliding step, the subunits are translated into atomic contact along the symmetry axis of the system. In the case of dihedral symmetry, the subunits are consecutively translated into atomic contact along the different symmetry axis in a randomly selected order. Fragment and rigid body moves are accepted based on the Metropolis criteria.

The models produced by the low-resolution protocol are energy minimized using an all-atom Monte Carlo Minimization protocol as previously described (5). A number of different symmetrical moves are used in the refinement: small random perturbation of backbone torsion angles, one-at-a-time (noncombinatorial) rotamer optimization using a backbone-dependent rotamer library (33), gradient-based minimization of the backbone, and side chain degrees of freedom and insertion of nonperturbing three or nine residue fragments. In addition to these moves, random rigid body perturbations were also incorporated into the protocol.

The 2×10^4 to 5×10^6 models were generated using Rosetta@home, and the 400 lowest energy models were clustered based on rms similarity to select the best models. As in previous work (6), modeling runs that did not pass 50th percentile energy cuts half-way and three-quarters of the way through the simulation were terminated. A clustering threshold of 2 Å was used, and five models representing the center of the five largest clusters were selected as predictions. NMR models, which can make use of additional limited experimental information, were selected as described below.

To compare the energies of models generated by the de novo protocol with native structures, the native structures were subjected to all-atom refinement. The native structures were first idealized by replacing the bond angles and bond lengths with the ''ideal'' values used in the de novo protocol (using a quasi-Newton optimization to adjust the backbone torsion angles to minimize the overall pertubation to the 3D structure) to facilitate comparison.

NMR Structure Inference. For NMR targets, the standard protein fragment selection procedure in Rosetta was replaced with the chemical shift filtered fragment selection used in CS-Rosetta (25). In the CS-Rosetta method, the software SPARTA (39) is used to predict the backbone and C^{β} chemical shifts of a database protein structures for which high-resolution crystal structures are available. Three and nine residue fragments are selected from this structural database based on similarity between predicted chemical shifts and experimentally determined chemical shifts of the target protein (for further details, see ref. 25). The standard fold-and-dock protocol is then used to generate high-resolution models of homo-oligomers.

A single model is predicted for each target protein by selecting the model with lowest Rosetta energy adjusted with a term measuring the similarity between the predicted chemical shift of the model and the experimentally determined shifts (25).

Availability. The code is freely available to academic users at www.rosettacommons.org in release Rosetta 2.3.0. For command line, see *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Keith Laidig and Darwin Alonso for flawless administration of computational resources; the users of Rosetta@home for donating their computer time; developers in the Rosetta community for contributions to our shared codebase; and Suzanne Pfeffer for informing us about the GCC185 coiled-coil sequence before the release of its crystal structure. This work was supported by the Howard Hughes Medical Institute (to D.B.), the Knut and Alice Wallenberg Foundation (to I.A.), the Jane Coffin Childs Foundation and a Burroughs-Wellcome Career Award at the Scientific Interface (to R.D.), the Intramural Research Program of the National Institutes of Health National Institute of Diabetes and Digestive and Kidney Diseases (to Y.S.), and the National Institutes of Health Protein Structure Initiative Grant U54-GM074958 (to T.S. and C.H.A.). C.H.A. holds a Canada Research Chair in Structural Proteomics.

1. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 29:105–153.
2. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput Biol* 2:e155.
3. Andre I, Strauss CE, Kaplan DB, Bradley P, Baker D (2008) Emergence of symmetry in homooligomeric biological assemblies. *Proc Natl Acad Sci USA* 105:16148–16152.
4. Kuriyan J, Eisenberg D (2007) The origin of protein interactions and allostery in colocalization. *Nature* 450:983–990.
5. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
6. Das R, et al. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 69:118–128.
7. Jauch R, Yeo HC, Kolatkar PR, Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins* 69:57–67.
8. Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* 69:704–718.
9. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in modeling of protein structures and interactions. *Science* 310:638–642.
10. Das R, Baker D (2008) Macromolecular modeling with Rosetta. *Annu Rev Biochem* 77:363–382.
11. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209–225.
12. Andre I, Bradley P, Wang C, Baker D (2007) Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci USA* 104:17656–17661.
13. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. *Methods Enzymol* 383:66–93.
14. Burguete AS, Fenn TD, Brunger AT, Pfeffer SR (2008) Rab and Arl GTPase family members cooperate in the localization of the golgin GCC185. *Cell* 132:286–298.
15. Crick FHC (1953) The packing of alpha-helices: simple coiled coils. *Acta Crystallogr* 6:689–697.
16. O'Shea EK, Klemm JD, Kim PS, Alber T (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* 254:539–544.
17. Nilges M, Brunger AT (1993) Successful prediction of the coiled coil geometry of the GCN4 leucine zipper domain by simulated annealing: comparison to the X-ray structure. *Proteins* 15:133–146.
18. Kammerer RA, et al. (2005) A conserved trimerization motif controls the topology of short coiled coils. *Proc Natl Acad Sci USA* 102:13891–13896.
19. Qian B, et al. (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature* 450:259–264.
20. Petsko GA (2000) The grail problem. *Genome Biol* 1:Comment 002.
21. Strop P, Brzustowicz MR, Brunger AT (2007) Ab initio molecular-replacement phasing for symmetric helical membrane proteins. *Acta Crystallogr D* 63:188–196.
22. Howard RJ, Clark KA, Holton JM, Minor DL, Jr (2007) Structural insight into KCNQ (Kv7) channel assembly and channelopathy. *Neuron* 53:663–675.
23. Langer G, Cohen SX, Lamzin VS, Perrakis A (2008) Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3:1171–1179.
24. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* 104:9615–9620.
25. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.
26. Montalvao RW, Cavalli A, Salvatella X, Blundell TL, Vendruscolo M (2008) Structure determination of protein-protein complexes using NMR chemical shifts: Case of an endonuclease colicin-immunity protein complex. *J Am Chem Soc* 130:15990–15996.
27. Lipsitz RS, Tjandra N (2004) Residual dipolar couplings in NMR structure analysis. *Annu Rev Biophys Biomol Struct* 33:387–413.
28. Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 120:6836–6837.
29. Rohl CA, Baker D (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* 124:2723–2729.
30. Wu Y, et al., Protocol for NMR-based Structural Genomics, High-quality protein structures solved in high-throughput. Twenty-third International Conference on Magnetic Resonance in Biological Systems, August 24–29, 2008, San Diego, CA.
31. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* 97:10383–10388.
32. Thompson MJ, et al. (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* 103:4074–4078.
33. Dunbrack RL, Jr, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 6:1661–1681.
34. Simons KT, et al. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34:82–95.
35. Morozov AV, Kortemme T, Tsemekhman K, Baker D (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci USA* 101:6946–6951.
36. Kortemme T, Morozov AV, Baker D (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* 326:1239–1259.
37. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Proteins* 35:133–152.
38. Kuhlman B, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302:1364–1368.
39. Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302.