PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

One contact for every twelve residues allows robust and accurate topology-level protein structure modeling

David E. Kim, Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song, and David Baker*

Department of Biochemistry, University of Washington, Seattle 98195, Washington

ABSTRACT

A number of methods have been described for identifying pairs of contacting residues in protein three-dimensional structures, but it is unclear how many contacts are required for accurate structure modeling. The CASP10 assisted contact experiment provided a blind test of contact guided protein structure modeling. We describe the models generated for these contact guided prediction challenges using the Rosetta structure modeling methodology. For nearly all cases, the submitted models had the correct overall topology, and in some cases, they had near atomic-level accuracy; for example the model of the 384 residue homo-oligomeric tetramer (Tc680o) had only 2.9 Å root-mean-square deviation (RMSD) from the crystal structure. Our results suggest that experimental and bioinformatic methods for obtaining contact information may need to generate only one correct contact for every 12 residues in the protein to allow accurate topology level modeling.

Proteins 2014; 82(Suppl 2):208–218.

Key words: protein structure prediction; rosetta; comparative modeling; homology modeling; *ab initio* prediction; contact prediction.

INTRODUCTION

Predicting the three-dimensional structure of a protein given just the amino acid sequence with atomic-level accuracy has been limited to small (<100 residues), single domain proteins. The ability to consistently predict structures with more complex topologies and structures of larger proteins is currently limited by energy function inaccuracies and to a larger extent, conformational sampling.[1–3] Recent advances in molecular modeling using experimental data such as NMR chemical shifts[4–7] and sparse restraints,[8–13] electron density from diffraction data,[14] and cryoEM[15,16] have shown that even very sparse information can significantly improve modeling. Using such data, models with high-resolution accuracies (<3 Å RMSD) have been generated for larger (>150 residues) and topologically complex proteins.

There has been much recent interest in predicting residue–residue contacts using sequence covariance information,[17–19] and experimental determination of contacting residues using chemical crosslinking followed by mass spectrometry[20] is becoming increasingly powerful. However, there has been little study of how much distance information is required to significantly improve modeling. The contact assisted structure modeling category in the Tenth Critical Assessment of Techniques for Protein Structure Prediction (CASP10) provided a blind test of this critical issue.

In this article, we describe predictions made using Rosetta with contact information provided by the CASP10 organizers. One pair of contacting residues in the native structure, which was not present in the majority of nonassisted server and human predictions, was provided for every 12 residues in each target on average. With this additional information, we were able to model the correct topology (>0.5 TM-score[21]) for all target domains and improve upon the best nonassisted predictions among all predictor groups for 15 of 17 domains with some exceptional high-resolution predictions. Our results suggest even a limited number of accurate contacts can significantly improve structure prediction.

David E. Kim and Frank DiMaio contributed equally to this work.
*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Seattle 98195, Washington. E-mail: dabaker@u.washington.edu
Received 9 April 2013; Revised 12 June 2013; Accepted 21 June 2013
Published online 31 July 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24374

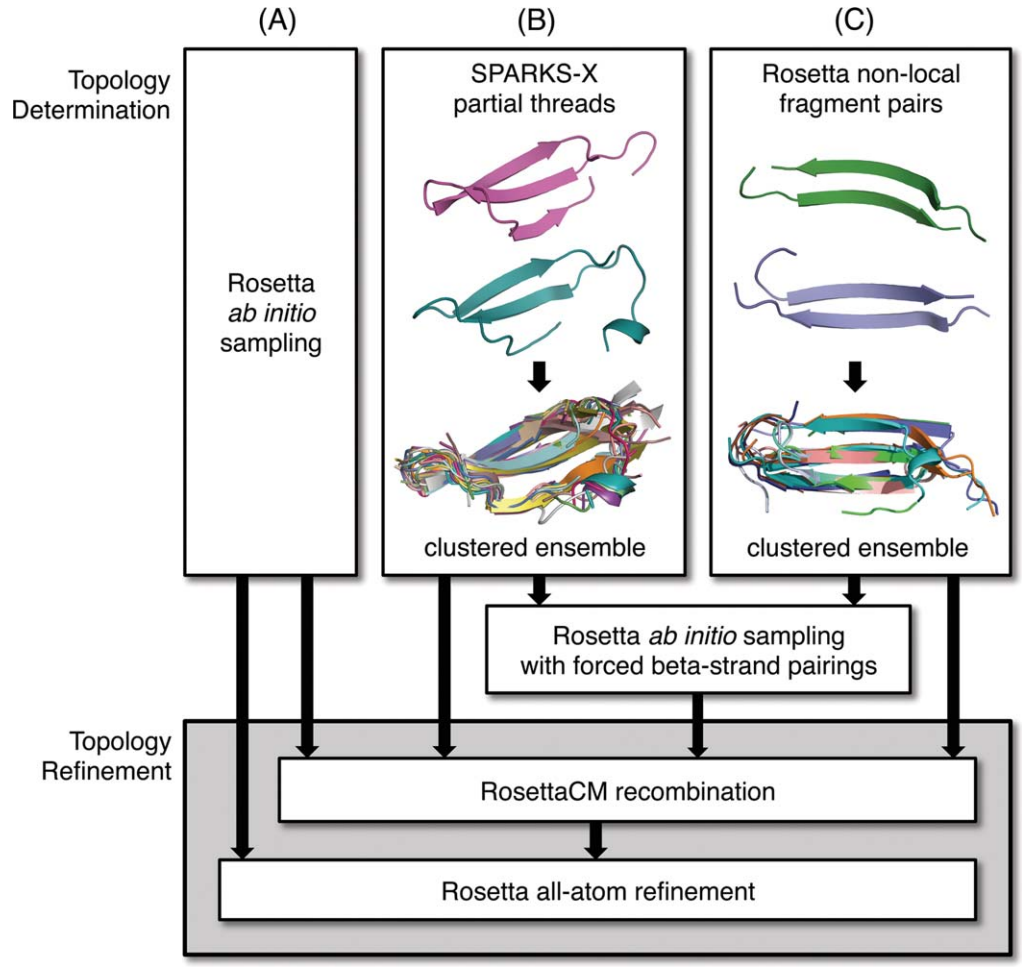

Figure 1

Schematic representation of the protocol used for contact assisted structure prediction. The protocol consists of two stages. In the first stage (Topology Determination), topologies and partial structures with satisfied contacts and good secondary structure were obtained from (**A**) Rosetta *ab initio* sampling methods with constraints, (**B**) partial threaded models from the top 1000 SPARKS-X alignments, and (**C**) Rosetta non-local fragment pairs. For some targets, beta-strand pairings were predicted from either (B) or (C) and enforced using the Rosetta broken chain fold-tree structure prediction method. In the second stage (Topology Refinement), the models and partial structures predicted from the first stage were used as input for the RosettaCM recombination protocol to remodel regions where contacts were not satisfied from the first stage and to sample full-length topologies. All models were optimized using Rosetta all-atom refinement.

MATERIALS AND METHODS

General strategy

We used a two-stage approach to generate models from contact information as outlined in Figure 1. In the first stage, alternative topologies are sampled and the lowest energy topology compatible with the contact information is selected. In the second stage, the fit to the contact information and the energy are further optimized by sampling alternative structures with the selected topology.

Topology determination

Topologies satisfying most to all contacts were generated using three different approaches. In the first approach [Fig. 1(A)], topologies were predicted using Rosetta *ab initio* structure prediction methods with constraints, and were supplemented with similar structures identified using TM-align[22] against server models and PDB templates. In the second approach [Fig. 1(B)], we predicted topologies from clustered ensembles of partial threaded models from SPARKS-X alignments.[23] Backbone segments that contained residues from unsatisfied contacts were removed before clustering. In the third approach [Fig. 1(C)], topologies were predicted from clustered ensembles of nonlocal Rosetta fragment pairs, which were derived from contacting pairs of local 5–20 mer fragments from the same PDB chain. Fragments were selected using the Rosetta fragment picker[24] with an additional score term favoring fragments with

satisfied contacts. Representative partial structures were generated by averaging the backbone coordinates of the clustered ensembles. For some targets, these partial structures were used to predict beta-strand pairings, which were enforced during sampling using the Rosetta broken-chain fold-tree modeling protocol[25] as described below.

Topology refinement

While the first stage generally converged to a topology satisfying the majority of constraints, the resulting models and partial structures often still had unsatisfied contacts or missing contact residues. The RosettaCM[26] recombination protocol was used to resolve these issues; it efficiently samples alternative nonlocal and local structures while maintaining the overall topology of the starting structure of the input models. Nonlocal segments are sampled by recombining segments from globally superimposed input structures in Cartesian space and local segments are sampled by fragment replacement in torsion space in the context of the global topology. Backbone segments that contained residues from unsatisfied contacts were removed before being used as input for RosettaCM, forcing these segments to be remodeled. The input was also supplemented with models and partial structures whose loop positions (as defined by DSSP[27]) were removed to increase loop sampling diversity. Full-length models produced by hybridization were subjected to Rosetta all-atom refinement.[28–30]

Atom pair distance constraints

To guide sampling, a simple atom pair distance constraint function previously developed for experimental restraints[10] was added to the standard Rosetta energy for both low-resolution sampling and all-atom refinement. The constraint energy is a function of the distance between Cβ atoms (Cα for glycine) $f(x)$:

$$f(x)=\begin{cases}(x-lb)^2 & \text{for} \quad x < lb \\ 0 & \text{for} \quad lb \le x \le ub \\ (x-ub)^2 & \text{for} \quad ub \le x \le ub+rswitch \\ x-ub-rswitch+rswitch^2 & \text{for} \quad x > ub > rswitch\end{cases}$$

where *lb* is a lower bound, *ub* is an upper bound, and *rswitch* is a constant of 0.5. Since assisted contacts were defined as residue pairs with Cβ (Cα for glycine) distances within 8 Å in the native structure, we used an upper bound of 8 Å and a lower bound of 1.5 Å. For noncontacts, pairs of residues that were not in contact in the native structure, we set the upper bound to 99 Å and the lower to 8 Å.

Achieving an optimal balance between the constraint energy and the standard Rosetta energy is critical for satisfying contacts while sampling protein-like topologies. We determined the optimal balance empirically by carrying out preliminary sampling with a range of constraint weights. If the weight is too low, few contacts are satisfied, while if the weight is too high, contacts are satisfied but models are irregular and lack secondary structure because the constraint energy overwhelms the physical chemistry implicit in the energy function. The constraint weight was gradually ramped up to the optimal value found in the preliminary calculations over the course of the trajectories. Constraints between residues close along the sequence reached full strength before those between residues distant along the sequence (as in the Rosetta NMR modeling protocol[28]) to avoid trapping in local minima of the constraint function in the early stages of each trajectory.

Rosetta *ab initio* structure prediction methods

Three Rosetta *ab initio* structure prediction methods were used with constraints and have been described in previous work.[25,31,32] All simulations were run on the distributed computing network, Rosetta@home, which enabled rigorous conformational sampling of 20,000–900,000 models per target and Rosetta all-atom refinement for each simulation. In this section, we give a brief overview of each method.

Standard *ab initio*

The standard Rosetta *ab initio* structure prediction method[31] was used for the majority of targets. Conformational sampling is carried out using a Monte Carlo fragment replacement strategy guided by a low-resolution energy function that favors protein-like features. Bond angles and bond lengths are kept fixed, and side-chains are represented by a single "centroid" interaction center; the only degrees of freedom are the backbone phi, psi, and omega torsion angles. Conformational sampling proceeds, starting from an extended chain, by random replacement of backbone torsion angles with torsion angles from fragments with similar local sequence selected from PDB templates using the Rosetta fragment picker. Variable fragment lengths of 3–19 residues were used as previously described.[33]

Broken chain fold-tree *ab initio*

Topologies with long-range beta-strand pairings are difficult to sample using the standard fragment assembly strategy because the precise geometry of long-range beta-strand pairings is difficult to achieve through random backbone torsion angle moves. Because of this, we used the Rosetta broken chain *ab initio* structure prediction method[25] for targets whose PSIPRED secondary structure predictions[34] suggested mostly beta topologies. Using this method, beta-strand pairings that are predicted *a priori* can be enforced: alternative local structures

are sampled in torsion space, while different beta-strand pairing geometries are explored by explicitly sampling rigid-body transformations. The protein chain is represented by a fold-tree—a directed, acyclic, connected graph composed of continuous peptide segments linked by rigid body transformations. For each beta-strand pairing, a chain break is made with a bias towards positions with higher predicted loop frequency in the intervening segment to prevent cyclic connections. Starting from extended continuous segments and beta-strand pairing connections, conformational sampling is carried out by Monte Carlo replacement of fragments and rigid body transformations taken from a library of beta-pairing geometries from known structures. An additional "chain-break" term with a weight that gradually increases throughout the simulation is also used to favor closing of chain breaks. Three-dimensional coordinates are constructed from the backbone torsion angles and rigid body transformations by traversing the fold-tree.

Ab initio fold and dock

The Rosetta symmetric fold and dock protocol[32] starts with extended backbone conformations of each subunit and a randomized symmetric configuration with no atomic-contacts between subunits. Comformational sampling proceeds by Monte Carlo symmetric fragment replacement, supplemented with two types of symmetric rigid body subunit perturbations for every 10 fragment moves, (1) random rotation and translation, and (2) translation along the symmetry axis into atomic contact. For symmetric oligomers, Rosetta all-atom refinement includes random rigid body perturbations in addition to the small backbone moves sampled in the standard refinement protocol. The symmetric coordinate system[35] maintains symmetry in both low-resolution sampling and high resolution refinement by explicitly sampling only the symmetric degrees of freedom.

Model selection

Models with the most satisfied contacts, and protein-like topologies with good secondary structure content based on visual inspection and from the contribution of backbone hydrogen bonding terms to the all-atom energy were selected from the lowest 5% energy population. For some targets with mostly beta secondary structure, models in the top 15% population with highest contact order[36] were considered. Final model ranking was carried out using a tighter constraint bound of 4 Å in addition to the Rosetta all-atom energy and visual inspection.

RESULTS

Figures 2 and 3 provide an overall qualitative view of the submitted contact assisted Rosetta models. The native structure is shown on the left, our best contact assisted submitted model in the middle, and the best submitted nonassisted model among all predictor groups on the right. It is evident that the contact assisted predictions are significantly better than the nonassisted predictions, and that the majority of contact assisted predictions have the correct overall topology.

Topology level accuracy

Accurate structure prediction may provide functional insight for proteins with unknown structure and function[37] through sequence-independent structure–structure comparisons of predicted models against the PDB; proteins with similar folds may have related biological functions. We used the TM-score[21] quality metric for determining whether our contact assisted predictions had a sufficient level of accuracy for topology classification. On the basis of a consensus definition of SCOP[38] and CATH,[39] >99.9% of proteins are not in the same fold when the TM-score = 0.4, but when the TM-score = 0.6, >90% are in the same fold; hence TM-scores greater than 0.5 indicate that the overall fold is very likely correctly modeled. Our model 1 contact assisted predictions all have TM scores to the native structure above 0.5, with an average TM-score of 0.64, suggesting they all have the correct fold. For five of the targets, the TM-score of the best nonassisted prediction among all predictor groups is below 0.5 (Tc658-D1, Tc684-D2, Tc719-D6, Tc734, and Tc735-D2); our submitted contact assisted models for these targets have TM-scores greater than 0.6.

Comparison to nonassisted predictions

For each target, we compared our best submitted contact assisted prediction with the best nonassisted prediction among all predictor groups, and our model 1 contact assisted prediction with the best model 1 nonassisted prediction among all predictor groups using GDT-TS[40] quality scores provided by the CASP10 automated evaluation (http://predictioncenter.org/casp10/results.cgi). The GDT-TS quality measure is calculated by averaging the percentage of equivalent residue pairs that are placed within the distance of 1, 2, 4, and 8 Å from the minimum RMSD superposition of the predicted and native structures. Our best submitted models had higher GDT-TS scores than the best nonassisted predictions for 15 of the 17 target domains [Fig. 4(A)] with a mean difference of 13.5 GDT-TS. Similarly, 14 of the 17 target domains had improved model 1 predictions as compared to the best nonassisted model 1 predictions [Fig. 4(B)] with a mean difference of 11.9 GDT-TS.

There is clearly a consistent improvement over most of the best nonassisted predictions; however, a portion of the large-scale improvement may be due to extra

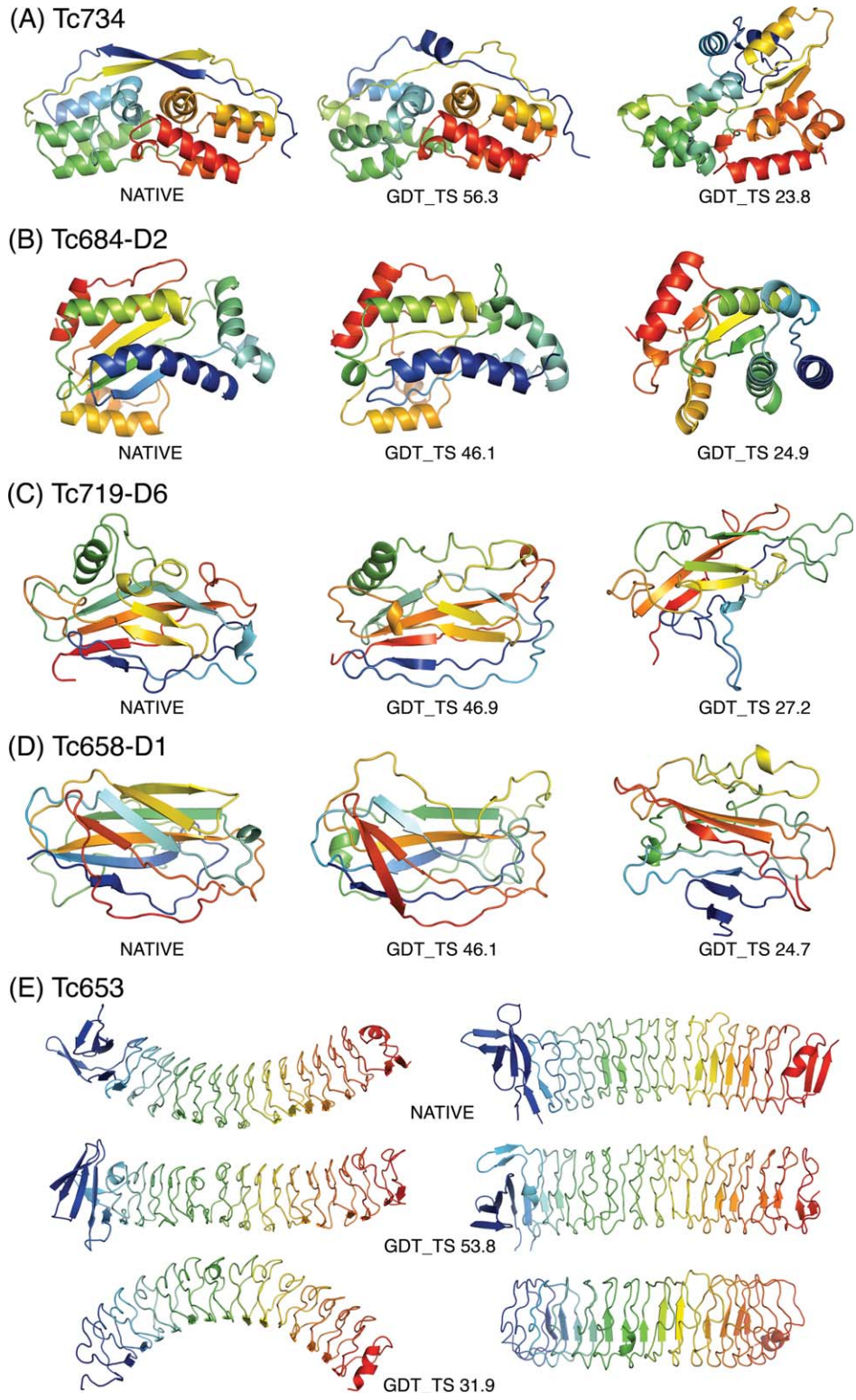

Figure 2

Contact assisted predictions significantly improved over the best nonassisted predictions. The native structure is on the left, our best submitted model in the middle, and the best nonassisted prediction among all predictor groups on the right. (**A**) Tc734, (**B**) Tc684-D2, (**C**) Tc719-D6, (**D**) Tc658-D1, and (**E**) Tc653 (native is on top, our model is in the middle, and the best nonassisted prediction among all predictor groups is on the bottom; orthogonal views are shown on the left and right; the best nonassisted prediction has typical LRR-like curvature, which is opposite to the atypical curvature of the native).

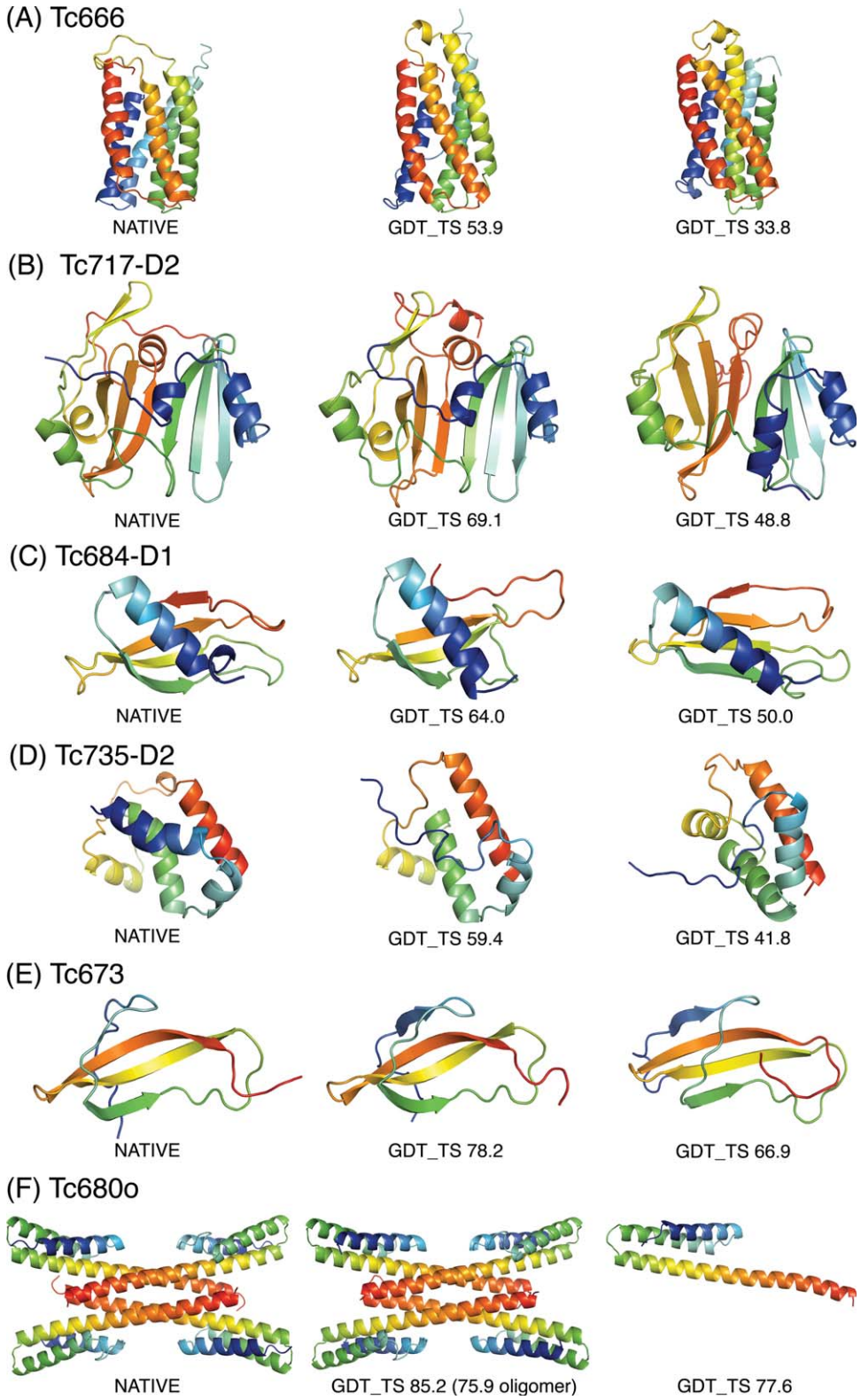

Figure 3

Contact assisted predictions with topology-level accuracy similar to best nonassisted predictions. In each panel the native structure is on the left, our best submitted model in the middle, and the best nonassisted prediction among all predictor groups on the right. For Tc680o, since nonassisted tetramer predictions were not made, the best single chain nonassisted prediction among all predictor groups is shown. (**A**) Tc666, (**B**) Tc717-D2, (**C**) Tc684-D1, (**D**) Tc735-D2, (**E**) Tc673, and (**F**) Tc680o.

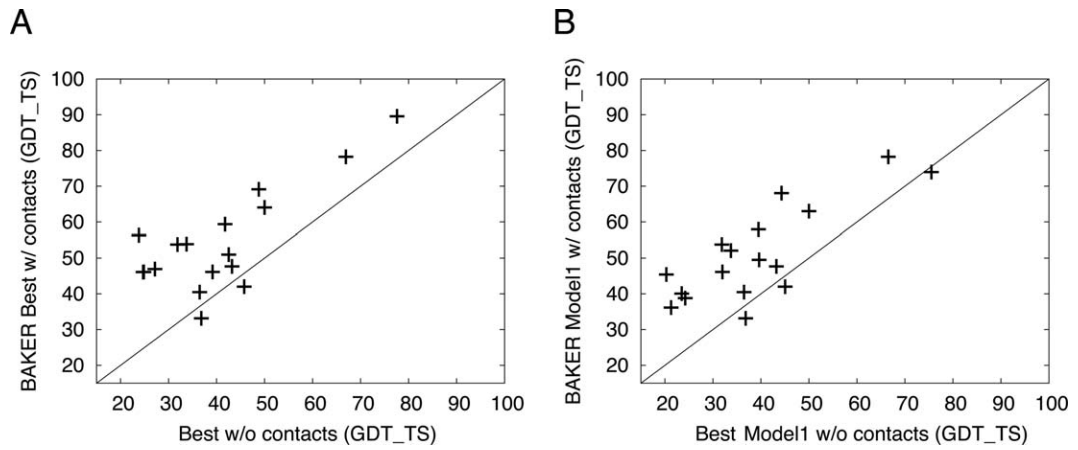

Figure 4

Contact assisted predictions are significantly improved over most nonassisted predictions based on GDT-TS scores. (**A**) Best BAKER contact assisted predictions versus the best nonassisted predictions among all predictor groups for the 17 target domains (the GDT-TS scores for Tc658-D1 and Tc684-D2 are close enough to appear as the same point). (**B**) Model 1 BAKER contact assisted predictions versus the best model 1 nonassisted predictions among all predictor groups for the 17 target domains.

information provided beyond contacts such as domain boundaries, terminal residues absent in the native structures, and oligomeric state (Tc680o). Nonassisted predictors were not given information about these features, which if predicted accurately, would possibly lead to improved structure predictions. Extra information beyond contacts was not provided for 5 full-length contact assisted targets (Tc649, Tc653, Tc666, Tc678, and Tc734). For four of these targets, our best submitted predictions had higher GDT-TS scores than the best nonassisted predictions with a mean difference of 15.9 GDT-TS, and similarly, our model 1 predictions had improved GDT-TS scores as compared to the best model 1 nonassisted predictions with a mean difference of 14.3 GDT-TS.

Model quality

An overview of the methods used for our best predictions for each target ordered by the difference in GDT-TS (ΔGDT-TS) as compared to the best nonassisted predictions among all groups is provided in Table I. This section will focus on the targets for which *Z*-scores calculated from the distribution of GDT-TS scores of the BEST assisted predictions from each group are greater than 3.0 (Tc734, Tc719-D6, Tc653, and Tc717-D2), in addition to our successful quaternary structure prediction for Tc680o. These targets provide good examples of the methods used for all targets.

Target Tc734 (model 4 GDT-TS, 56.3, and GDT-TS *Z*-score among BEST assisted predictions, 3.25) was predicted using Rosetta *ab initio* and RosettaCM methods [Fig. 2(A)]. Internal nonoverlapping sequence similarity[41] and the lowest energy models from Rosetta *ab initio* sampling suggested symmetric domains (residues 30–111 and 135–216) so we modeled the domains separately. The lowest energy models of each domain converged to a similar fold with a cluster radius of less than 2 Å RMSD. Cluster representatives of both domains were recombined with full-length models using RosettaCM. The best domain input models had 74.1 GDT-TS and 3.0 Å RMSD for the first domain and 89.1 GDT-TS and 1.4 Å RMSD for the second domain. RosettaCM improved upon the best full-length input model from 40.7 to 56.3 GDT-TS and 6.0 to 4.3 Å RMSD. The best prediction had the most contacts satisfied among our submitted models with only one close but unsatisfied contact. The domains and their relative orientation were modeled quite well but a nonlocal beta-strand pair, which twists like an overhand knot, was incorrectly modeled likely because of inaccurate secondary structure prediction.

Target Tc719-D6 (model 2 GDT-TS, 46.9, and GDT-TS *Z*-score among BEST assisted predictions, 3.73) was predicted using a variety of methods [Fig. 2(C)]. Since, the secondary structure prediction suggested a beta-topology, we used the broken chain fold-tree sampling protocol. Three antiparallel beta-strand pairs (residues 604 and 673, 611 and 694, and 604 and 699) were predicted from clustered ensembles of nonlocal Rosetta fragments. One pair was incorrectly predicted, 604 and 673. The predicted pairs were randomly enforced during sampling. Four templates (2jtyA, 2p52A, 2vn6A, and 3gfuA) were also identified among SPARKS-X alignment templates from a TM-align search against the selected models. The selected models and partial threaded templates were recombined using RosettaCM, which improved upon the best full-length input model from 41.9 to 46.9

Table I
Methods Used to Model CASP10 Targets

Target	Len	Contacts	ΔGDT-TS	Standard *ab initio*	RosettaCM	SPARKS-X partial threads	Fold-tree *ab initio*	Nonlocal fragments	Look-back search	Fold and dock	Refine w/membrane score
Tc734	216	20	32.5	X	X						
Tc653	414	12[a]	21.9	X	X						
Tc658-D1	166	16	21.4				X	X			
Tc684-D2	168	18	21.3	X							
Tc717-D2	166	15	20.3	X	X						
Tc666	195	14	20.1	X							X
Tc719-D6	163	13	19.8		X		X	X	X		
Tc735-D2	88	7	17.6	X							
Tc684-D1	73	8	14.0	X							
Tc680	96 (384[b])	3 (6[c])	12.0							X	
Tc673	62	5	11.3	X							
Tc678	161	12	8.4	X							
Tc735-D1	233	28	7.0		X	X					
Tc676	173	17	4.5	X							
Tc705-D2	344	34	4.0		X	X			X		
Tc649	210	16	−3.7		X	X					
Tc691	141	15	−3.7		X	X	X				

Targets are ordered by the difference in GDT-TS (ΔGDT-TS) between the best contact assisted submitted model and the best nonassisted prediction among all groups.
[a]Noncontacts were provided for Tc653.
[b]Homo-oligomeric tetramer length.
[c]Interchain contacts.

GDT-TS and 6.6 to 5.8 Å RMSD. The best model was our only submitted model that had all contacts satisfied for this target.

Target Tc653 (model 1 GDT-TS, 53.8, and GDT-TS *Z*-score among BEST assisted predictions, 3.47) was modeled using Rosetta *ab initio* [Fig. 2(E)]. Twelve noncontacts were provided along with information stating that the best server predictions had the majority of local contacts satisfied but the gross shape substantially deviated from the native. Server predictions generally agreed upon a leucine rich repeat (LRR) like structure. Because of its large size (414 residues), we parsed the sequence into 13 overlapping segments with at least 1–2 LRR repeat unit lengths of overlap and modeled the individual segments. Each segment converged with similar LRR like structure with cluster radii less than 2 Å RMSD. Models were selected among cluster representatives and full-length models were constructed by superimposing overlapping segments, and further refined using RosettaCM. The overall shape of our submitted predictions was substantially different than the best server and human models. Rather than having a typical LRR curved shape [Fig. 2(E); bottom] with a regular beta-sheet on the interior part of the curve, our models were straightened with a slight opposite curvature and a broken up beta-sheet on the exterior part of the curve [Fig. 2(E); middle]. Our models were significantly closer to the native structure, which has an opposite curvature when compared to typical LRR structures and a broken up beta-sheet on the exterior [Fig. 2(E); top]. The termini were not modeled correctly but a large part of the repeating structure was modeled quite well with 1.6 Å RMSD over 198 residues in our best submitted model.

Target Tc717-D2 (model 5 GDT-TS, 69.1, and GDT-TS *Z*-score among BEST assisted predictions, 3.02) was predicted using Rosetta *ab initio* and RosettaCM methods [Fig. 3(B)]. The lowest energy models from Rosetta *ab initio* sampling converged to a common fold with a cluster radius under 4.0 Å RMSD. The C-terminal loop and helix, and a short helix and lone beta-hairpin in the region around residues 132–160 varied in structure whereas the main beta-sheet and the N-terminus, which packs against the sheet, were relatively well converged with satisfied contacts. RosettaCM was used to remodel the regions that did not converge. Although the final predictions had more satisfied contacts on average than the original models selected from *ab initio* sampling, 3 out of 21 of the original models were slightly closer to native as compared to our best prediction. Consistent with the convergence observed in the N-terminus and beta-sheet, our best submitted model had atomic-level accuracy in the region where two short N-terminal helices pack against a locally formed four-stranded antiparallel beta-sheet [Fig. 5(B)]. The RMSD is 1.1 Å over 61 residues.

Tc680o (model 5 GDT-TS, 75.9, and GDT-TS *Z*-score among BEST assisted predictions, 2.21) was modeled using the Rosetta *ab initio* fold and dock protocol.[32] Participants were asked to model the quaternary structure knowing that the target was a tetramer and that there were no contacts below 7.5 Å in the pairs of chains A:C and B:D. Three intrachain contacts were provided

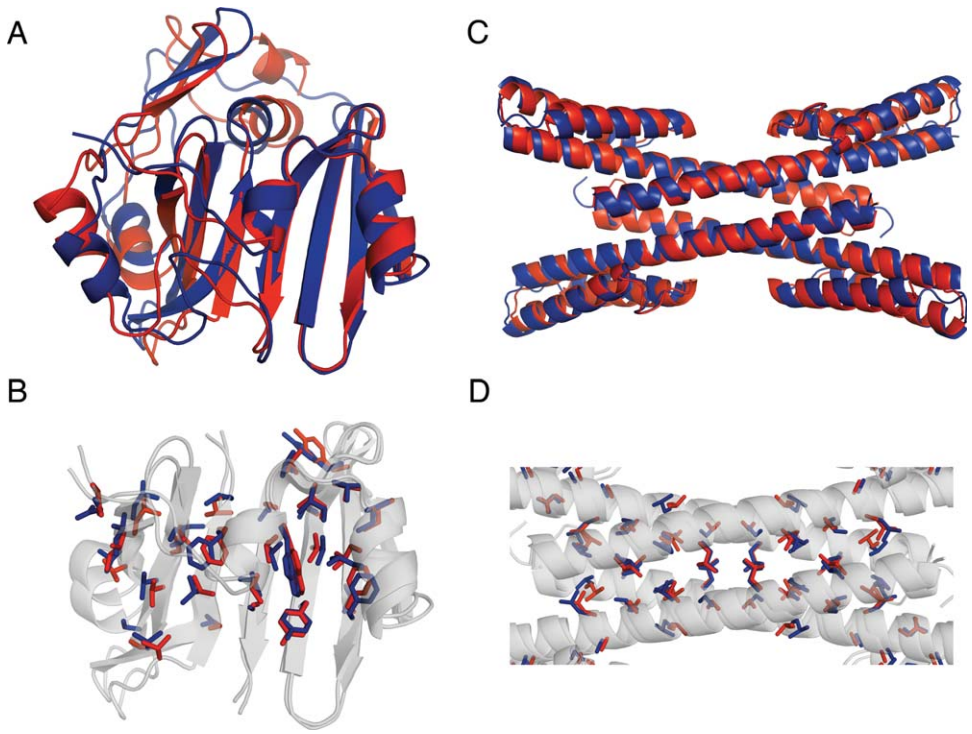

Figure 5

Examples of predictions with near atomic-level accuracy. The core side chains of the submitted model (red) and native (blue) are highlighted. (**A**) Tc719, (**B**) the converged section of Tc719, (**C**) the symmetric homo-oligomeric tetramer, Tc680o, and (**D**) the interface of Tc680o.

along with six interchain contacts between A:B and B:C pairs of chains. Given this information, models with D2 dihedral symmetry were generated using the Rosetta symmetric modeling framework.[35] The lowest energy models converged to a common quaternary fold with structural variation mainly within the short helices that are not involved in the interface between chains. Final models were selected from cluster representatives with most to all contacts between pairs of chains A:C and B:D greater than 7.5 Å. Our best submitted oligomeric model had near atomic-level accuracy particularly at the interface between chains [Fig. 5(D)], and has a global and single chain RMSD of 2.9 and 2.2 Å, respectively. Our best single chain submitted prediction, model 3, has an RMSD of 1.5 Å.

There were a number of other noteworthy predictions. Tc684-D2 [Fig. 2(B)], Tc658-D1 [Fig. 2(D)], and Tc666 [Fig. 3(A)] are longer than 160 residues and have a mean GDT-TS improvement of 20.9 over the best nonassisted predictions among all groups. Interestingly, Tc666 was predicted reasonably well (GDT-TS 53.9 and 4.9 Å RMSD) using *ab initio* sampling despite being a membrane protein. As a last step, we refined our models using the Rosetta membrane energy function[42] with a high constraint weight but minimal structural changes to the models were made.

For targets less than 100 residues in length (Tc684-D1 [Fig. 3(C)], Tc735-D2 [Fig. 3(D)], and Tc673 [Fig. 3(E)]), the mean improvement in GDT-TS over the best nonassisted predictions among all groups is 14.3. Tc684-D1 (model 5 GDT-TS, 64.0, and GDT-TS Z-score among BEST assisted predictions, 1.89) has a global RMSD of 4.3 Å. Deviations from the native structure are mostly in two hairpin loops with the highest B-factors. Excluding these regions gives an RMSD of 2.4 Å over 51 residues. Tc735-D2 (model 3 GDT-TS, 59.4, and GDT-TS *Z*-score among BEST assisted predictions, 2.43) has a global RMSD of 5.3 and 1.5 Å RMSD over 45 residues. The secondary structure prediction was incorrect for the N-terminal helix, which is kinked with a proline residue in the native structure. Our submitted predictions had loop conformations in place of the helix. The loops and small helix covering residues 303–335, which have relatively high crystallographic B-factors, also deviate from the native structure. Tc673 (model 1 GDT-TS, 78.2, and GDT-TS *Z*-score among BEST assisted predictions, 2.55) has a global RMSD of 2.5 and 1.2 Å RMSD over 42 residues. Since the native structure is a dimer, deviations are likely due to interactions between chains, which were not considered when modeling the monomer.

What went wrong

In several cases the local structure in the models was incorrect; unlike the chemical shift information used in

CS-Rosetta,[5,6] the contact information does not directly improve sequence based fragment selection. Inaccurate secondary structure prediction was an issue for a number of targets (e.g., Tc666, Tc734, and Tc735-D2) and many of the native structures had a significant amount of loop conformation. Despite this, topologies were successfully predicted likely because there was enough contact information to overcome errors in local structure. Significant structural deviations mostly occurred in regions with incorrect secondary structure prediction and loop regions. Supplementing experimental contact information with information on local secondary structure, for example from NMR chemical shifts or MS-HD exchange, may help resolve this problem.

Our best predictions for two targets, Tc649 and Tc691, had lower GDT-TS scores than the best nonassisted predictions among all groups, and also had the lowest TM-scores among our best submitted predictions for all contact assisted targets. For Tc649, the failure to generate models with satisfied contacts resulted from the use of incorrect partial threads. For Tc691, we used the broken-chain fold-tree protocol, but among 6 predicted strand pairings, only 1 was correct and 2 were shifted by 1 residue. Despite having mostly incorrect pairings, reasonable topologies with close to satisfied contacts were modeled and refined using RosettaCM. Our best prediction had the correct fold with 3.0 Å RMSD over 80 residues but there were substantial deviations at the termini and beta-hairpin loops, which were largely unstructured in our model. These regions are involved in significant interactions between chains in the native structure; the biological unit is a homo-oligomeric tetramer, and this is likely why this target was difficult to model accurately using template-free sampling methods.

DISCUSSION

The results presented here suggest one correct residue–residue contact for every 12 residues in a protein may be sufficient to accurately model overall protein topology, provided that the contacts are mainly nonlocal and broadly distributed, similar to the contacts selected by the CASP10 organizers. There is a trade off in both experimental methods such as chemical crosslinking followed by mass spectrometry, and bioinformatics methods based on residue covariance and other properties, between the number of predictions made and the average prediction accuracy. Our results suggest the development of experimental and bioinformatics methods for residue–residue contact determination should perhaps focus on accurately predicting this relatively small number of contacts. The combination of methods for accurately generating contact information with the structure modeling methods described here could be very powerful in rapidly determining accurate fold level protein models.

It may be possible to further improve the accuracy of the models by using recently developed iterative refinement methods such as RASREC[12] appropriately constrained using the contact information. Our results also suggest that supplementing contact information with local structure information, for example NMR chemical shifts or hydrogen-deuterium exchange data from MS or NMR would produce considerable increases in model accuracy.

The collection of Rosetta modeling methods used in this work was largely developed for solving structures using experimental data. For example, the symmetric fold and dock protocol used for Tc680o was used in conjunction with experimental data to solve the structure of the Type III secretion system needle,[43] and the *ab initio* modeling with constraints protocol used for many of the targets was developed for NMR structure determination.[28] The classic protein structure prediction problem remains extremely challenging because of the vast size of conformational space and the inaccuracy of current potentials; we believe that focus on the more tractable but still challenging problem of structure determination from sparse experimental data, which greatly reduces the search problem, is a constructive way to make progress in this fundamental research area.

ACKNOWLEDGEMENTS

The authors thank Rosetta@home participants for providing the computing resources necessary for this work, and Keith Laidig and Darwin Alonso for developing the computational and network infrastructure. They also thank the CASP10 organizers and the contributing structural biologists who provided targets.

REFERENCES

1. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
2. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D. Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 2011;405:607–618.
3. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–1735.
4. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci U S A 2007;104:9615–9620.
5. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 2008;105:4685–4690.
6. Shen Y, Vernon R, Baker D, Bax A. De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 2009;43:63–78.
7. Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 2008;36(Web Server issue):W496–W502.

8. Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. TOUCHSTONEX: protein structure prediction with sparse NMR data. Proteins 2003;53:290–306.
9. Qu Y, Guo JT, Olman V, Xu Y. Protein structure prediction using sparse dipolar coupling data. Nucleic Acids Res 2004;32:551–561.
10. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR structure determination for larger proteins using backbone-only data. Science 2010;327:1014–1018.
11. Hirst SJ, Alexander N, McHaourab HS, Meiler J. RosettaEPR: an integrated tool for protein structure determination from sparse EPR data. J Struct Biol 2011;173:506–514.
12. Lange OF, Baker D. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins 2012;80:884–895.
13. Schmitz C, Vernon R, Otting G, Baker D, Huber T. Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol 2012;416:668–677.
14. DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, Das D, Vorobiev SM, Iwai H, Pokkuluri PR, Baker D. Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 2011;473:540–543.
15. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D. Refinement of protein structures into low-resolution density maps using rosetta. J Mol Biol 2009;392:181–190.
16. Chan KY, Trabuco LG, Schreiner E, Schulten K. Cryo-electron microscopy modeling by the molecular dynamics flexible fitting method. Biopolymers 2012;97:678–686.
17. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PloS one 2011;6:e28766.
18. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 2011;108:E1293–E1301.
19. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–190.
20. Singh P, Panchaud A, Goodlett DR. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. Anal Chem 2010;82:2636–2642.
21. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 2010;26:889–895.
22. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 2005;33:2302–2309.
23. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27:2076–2082.
24. Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D. Generalized fragment picking in Rosetta: design, protocols and applications. PloS one 2011;6:e23294.
25. Bradley P, Baker D. Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 2006;65:922–929.
26. Song Y, DiMaio F, Wang RY, Kim DE, Miles C, Brunette T, Thompson J, Baker D. High resolution comparative modeling with RosettaCM. Structure, in press.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
29. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 2007;69(Suppl 8):118–128.
30. Tyka MD, Jung K, Baker D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. J Comput Chem 2012;33:2483–2491.
31. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
32. Das R, Andre I, Shen Y, Wu Y, Lemak A, Bansal S, Arrowsmith CH, Szyperski T, Baker D. Simultaneous prediction of protein folding and docking at high resolution. Proc Natl Acad Sci U S A 2009;106:18978–18983.
33. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 2009;77(Suppl 9):89–99.
34. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
35. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I. Modeling symmetric macromolecular structures in Rosetta3. PloS one 2011;6:e20450.
36. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
37. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.
38. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36(Database issue):D419–D425.
39. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res 2011;39(Database issue):D420–D426.
40. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
41. George RA, Heringa J. The REPRO server: finding protein internal sequence repeats through the Web. Trends Biochem Sci 2000;25:515–517.
42. Barth P, Schonbrun J, Baker D. Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci U S A 2007;104:15682–15687.
43. Loquet A, Sgourakis NG, Gupta R, Giller K, Riedel D, Goosmann C, Griesinger C, Kolbe M, Baker D, Becker S, Lange A. Atomic model of the type III secretion system needle. Nature 2012;486:276–279.