Structure
Technical Advance

High-Resolution Comparative Modeling with RosettaCM

Yifan Song,[1,3] Frank DiMaio,[1,3] Ray Yu-Ruei Wang,[1] David Kim,[1] Chris Miles,[1] TJ Brunette,[1] James Thompson,[1] and David Baker[1,2,*]
[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
[2]Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, WA 98195, USA
[3]These authors contributed equally to this work
*Correspondence: dabaker@u.washington.edu
http://dx.doi.org/10.1016/j.str.2013.08.005

SUMMARY

We describe an improved method for comparative modeling, RosettaCM, which optimizes a physically realistic all-atom energy function over the conformational space defined by homologous structures. Given a set of sequence alignments, RosettaCM assembles topologies by recombining aligned segments in Cartesian space and building unaligned regions de novo in torsion space. The junctions between segments are regularized using a loop closure method combining fragment superposition with gradient-based minimization. The energies of the resulting models are optimized by all-atom refinement, and the most representative low-energy model is selected. The CASP10 experiment suggests that RosettaCM yields models with more accurate side-chain and backbone conformations than other methods when the sequence identity to the templates is greater than ~15%.

INTRODUCTION

Protein structures are crucial to understanding biological function, but have been experimentally determined only for a small fraction of known proteins; this fraction continues to decrease as high-throughput sequencing identifies large numbers of protein sequences. Fortunately, structures are now known for at least one representative of most protein families, and comparative modeling methods can be used to generate models of many proteins using these representative structures as starting points (Pieper et al., 2011).

Comparative modeling proceeds in two steps: first, the protein sequence being modeled is aligned to evolutionarily related sequences with known structures; and second, three-dimensional models are built guided by information from these structures. Many excellent methods for comparative modeling have been developed, including the widely used MODELLER program (Eswar et al., 2006; Sali and Blundell, 1993) and, more recently, I-TASSER (Xu et al., 2011) and other methods that explicitly recombine multiple templates.

The Rosetta structure modeling methodology utilizes efficient conformational sampling techniques and a physically realistic all-atom energy function to achieve atomic accuracy in many challenging structural biology problems, including structure determination with sparse experimental data and de novo design of protein structures and interfaces (Fleishman et al., 2011; King et al., 2012; Raman et al., 2010). Previous comparative modeling efforts in Rosetta (Raman et al., 2009; Thompson and Baker, 2011) produced accurate models in some cases but were unable to combine structural information from multiple templates.

Here, we describe RosettaCM, a recently developed comparative modeling method that assembles structures using -integrated torsion space-based and Cartesian space template fragment recombination, loop closure by iterative fragment assembly and Cartesian space minimization, and high-resolution refinement. Results from the CASP10 (Critical Assessment of Techniques for Protein Structure Prediction) blind evaluation of current structure prediction methodology suggest that, given a set of input alignments to templates of known structure, RosettaCM generates models with higher accuracy over all backbone and side-chain atoms than other current methods.

RESULTS AND DISCUSSION

We begin with a brief overview of the RosettaCM protocol; a complete description is provided in the Experimental Procedures. Starting from alignments of the query sequence to templates of known structure, which may be generated using remote homolog detection methods such as PSIBLAST (Altschul et al., 1997) or Hhsearch (Remmert et al., 2012), or using expert knowledge, RosettaCM builds models in three stages as outlined in Figure 1. In the first stage, the query sequence is threaded onto each of the templates, and the resultant threaded partial models are aligned in a single global frame. Full-chain models are then generated by Monte Carlo sampling guided by the Rosetta low-resolution energy function supplemented with distance restraints from the template structures and a penalty for separation in space of residues adjacent in the sequence (Figure S1 available online). Structures are built up using a Rosetta "fold tree" (Das and Baker, 2008); the global position of each segment is represented in Cartesian space, whereas the backbone and side-chain conformation of residues in each segment are represented in torsion space. Two types of Monte Carlo moves are used: first, substitution of the torsion angles from a Rosetta de novo modeling fragment selected from the PDB using local sequence information (Figure 1B); and second, substitution of the coordinates of a template segment (Figure 1C). This recombination of template-derived fragments in

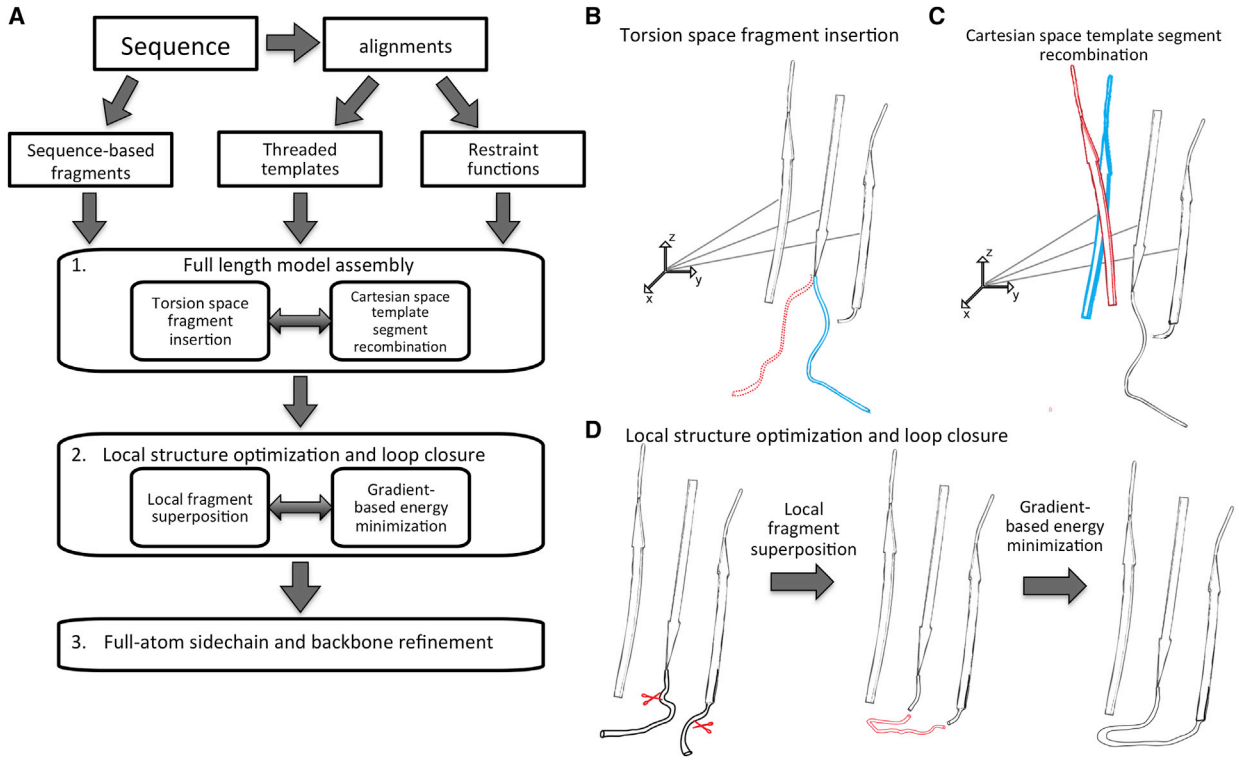

Figure 1. RosettaCM Protocol

(A) Flowchart of the RosettaCM protocol.

(B–D) RosettaCM conformational sampling. See also Figure S1.

(B) Torsion space fragment insertion. Blue indicates before fragment insertion; red, after fragment insertion. Structures are built outward from the origin (small coordinate system) using first the rigid body transforms to the centers of the segments and then the torsion angles from the centers to the end of the segments. Because the effects of torsion angle changes do not propagate beyond segment boundaries, the overall topology is better maintained than in conventional continuous chain torsion space Monte Carlo.

(C) Recombination of template segments in Cartesian space. Blue indicates before and red, after segment replacement.

(D) Local structure optimization and loop closure. First, a fragment is superimposed onto the current pose (red), and second, energy minimization smoothly resolves structural distortions at the fragment junctions.

Cartesian space and Rosetta de novo fragments in torsion space generally converges to the correct topology, but the geometry at segment boundaries is often poor, with clashes, distorted peptide bonds, and poor backbone hydrogen bond geometry.

The second stage improves model geometry and further explores conformational changes away from the starting templates through Monte Carlo sampling with two-step moves (Figure 1D). In the first step, a backbone region is randomly selected and replaced by either a de novo fragment, which spans the region and has N and C termini that can be roughly superimposed on the corresponding residues in the current model, or a template-derived fragment superimposed over all corresponding residues. The de novo fragment substitutions are biased toward regions with poor backbone-bonded geometry, primarily the stage one segment boundaries. In the second step, quasi-Newton minimization is carried out over the entire protein in Cartesian space, using a smoothed version of the Rosetta low-resolution energy function (Rohl et al., 2004), to optimize backbone geometry and hydrogen-bonding interactions. The result of Monte Carlo sampling using these composite fragment superposition and energy minimization moves is smooth and realistic loop closure—facilitated because the loop takeoff and return positions can shift to promote closure—where every local backbone segment is "protein like" (Figures 1D and S1A). Finally, in the third stage, side chains are built on and the structure is optimized by standard Rosetta full-atom refinement using a physically realistic energy function (Tyka et al., 2011).

The balance between the Rosetta energy function, which favors physically realistic conformations, and the template-derived restraint energy functions determines how close the resultant models are to the input template structures. This balance is set by a single overall weight, which we have optimized over a diverse training set as described in the Supplemental Experimental Procedures (Figure S3). In specific applications, the user may wish to alter this parameter and to vary the extent to which each template/alignment contributes to the restraint functions. In the calculations described in the remainder of this paper, the overall weight was set to the value optimal for the training set, and the contributions of each alignment to the restraint functions were weighted based on the alignment likelihood, with close alignments contributing more strongly than weak alignments (see Supplemental Experimental Procedures and Figure S3).

A long-standing question in the structure prediction field is the extent to which comparative models improve over the available template structures. Many widely used comparative modeling packages and servers produce models that cover the entire sequence of a protein, whereas the available templates in general do not; hence, comparative models generally have more residues superimposable on the actual structure than the original template. Less trivial are improvements in the aligned regions, which require shifts away from the starting template coordinates. To assess the extent to which RosettaCM improves models beyond the best-available template over a large and unbiased set of structures, we participated in the CAMEO project (Continuous Automated Model EvaluatiOn, http://www.cameo3d.org/) in which recently solved structures deposited in the PDB but not yet publicly released are made available to prediction servers; all models must then be submitted prior to the public release data. Analysis of statistics collected between May 1, 2012 and March 31, 2013 by the CAMEO experiment showed that RosettaCM consistently improves over the available templates in the aligned regions (Figure S2A).

To compare RosettaCM to the earlier Rosetta "rebuild and refine" protocol (LoopRelax), a benchmark set was selected from CAMEO to cover different ranges of modeling difficulties (Table S2). RosettaCM differs from the earlier protocol both in the explicit use of multiple templates and in the loop closure/structure optimization protocol. To separate out these effects, we first compared the methods using a single template for each case. As shown in Figure S2, using a combination of fragment insertion and Cartesian-space minimization of the Rosetta low-resolution energy function improves over the cyclic coordinate descent method used in the earlier protocol; this is likely because this approach allows readjustments promoting loop closure over the whole backbone. Further improvements are observed (Figure S2) when multiple input templates are used in modeling compared to just using the top-ranked (in the sequence-based search of the PDB) template; the explicit template recombination in RosettaCM is a considerable advantage when different parts of the query sequence are better modeled by different templates. The improvements over the earlier rebuild and refine protocol are primarily for intermediate-difficulty targets (Figure S2G).

It is not trivial to accurately assess the performance of a structure modeling method relative to methods developed by other groups. Even if the structure modeling software is available, there are generally a number of settings, and a nonexpert may not run the calculations in an optimal way. For this reason, to evaluate the strengths and limitations of RosettaCM, we analyze its performance in the CASP10 structure prediction experiment. In CASP10, the RosettaServer ran the RosettaCM protocol starting from templates and alignments identified by Hhsearch (Remmert et al., 2012), SPARKS-X (Zhou and Zhou, 2004), and RaptorX (Peng and Xu, 2009). As noted above, the accuracy of comparative models depends not only on the quality of the model-building approach but also on the input templates and alignments. To evaluate the model-building approach in RosettaCM independent of template recognition and alignment generation, we focused on the subset of closer homology targets for which most methods used the same templates and alignments. We used the structural similarity (as measured by the GDT) between the first models submitted in the CASP10 experiment by the RosettaServer and the state of the art HHpredA (a widely used public server) (Söding et al., 2005) and ZhangServer (the top-performing server in CASP10) as a measure of the extent of convergence on templates and alignments. The 63 domains for which the average GDT between the first models was 70% or above were selected for detailed analysis (Table S1) to reduce the impact of differences in template selection.

To compare the performance of the methods, we utilized statistics computed and made publicly available by the CASP10 (http://predictioncenter.org/casp10/) organizers (http://predictioncenter.org/casp10/index.cgi) and the Zhang lab (http://zhanglab.ccmb.med.umich.edu/casp10/). The accuracy of the modeled protein backbone was assessed using the GDT (Zemla et al., 1999), the accuracy of the side-chain placements by the GDC-SC (Zemla, 2003), and the accuracy of the polar interactions by the fraction of recovered native hydrogen bonds (see Supplemental Experimental Procedures for detailed descriptions of these metrics). According to all three metrics, on the 63 targets for which template selection and alignment generation were straightforward, the RosettaServer models were better than those of other servers both on average and in having the most top models (Figure 2). Overall stereochemical quality—as reported by the MolProbity score (Davis et al., 2007)—was also highest for the RosettaServer models. On the complete set of 127 domains, RosettaServer had the most top models (Figure 2, left), although the performance of the Zhang server was considerably better according to the standard CASP sum of *Z* scores metric (Tramontano et al., 2001) because the RosettaServer did quite poorly on several targets due to errors in template identification and domain parsing.

What is the origin of the improved model-building performance evident in Figure 2? To build a good model, a comparative modeling method should (1) improve over the closest template in the aligned regions, and (2) properly reconstruct the loops and other regions not present in the templates. The per-residue changes in model accuracy relative to the closest available templates for RosettaCM and several other top methods are compared in Figure 3B. Most residues are already quite close to the correct positions in the starting templates, and hence, most frequently, the deviations are close to zero. A subset of residues is in significantly different positions in the starting template and the actual structure, and for these residues, modeling methods can make substantial improvements. For this subset, RosettaCM produced the largest number of improvements over the target set as indicated by the greater number of decreases in deviation of more than 1.5 Å (Fig 3B, left most bars). Of residues that are improved by over 1.5 Å, 27% are on a helix, 3% are on a strand, and 70% are either on a loop structure or at the junction between a loop and a helix or strand.

Examples of the improvements are shown in Figure 4. In Figures 4A–4C, the difference in model quality relative to the best template is shown along the linear sequence for the RosettaServer model and for several other top servers. The RosettaServer models show pronounced dips below the x axis, indicating improvement relative to the best template. The structural comparisons in the lower insets illustrate structural

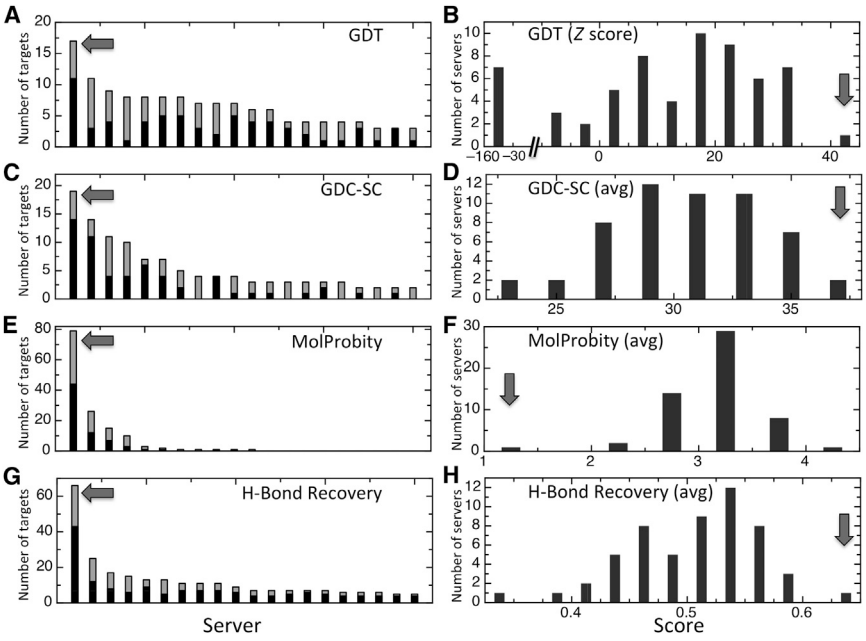

Figure 2. RosettaCM Performance in CASP10

For each CASP10 target, performance statistics were downloaded from the CASP10 website and used to rank the servers based on (A and B) global structural similarity, as measured by the GDT-TS metric (Zemla et al., 1999), (C and D) accuracy of side-chain placement, as measured by the GDC-SC metric (Keedy et al., 2009), (E and F) stereochemical quality, as assessed by the MolProbity score (Davis et al., 2007), and (G and H) the fraction of native hydrogen bonds (http://zhanglab.ccmb.med.umich.edu/casp10/). (A), (C), (E), and (G) indicate, for each of the four metrics, the number of targets for which each server produced the best-scoring model; servers are ordered on the x axis based on this number. The counts for the 63 easier target subset are shown in black, and those for the rest of the targets in gray. The arrow indicates the RosettaCM result. (B), (D), (F), and (H) are histograms of the sum or average of each of the four scores over the 63 easier target subset (sum of GDT-TS *Z* scores is in B; average GDC-SC, MolProbity score, and fraction of native hydrogen bonds are in D, F, and H). The y axis is the number of servers in the total score interval on the x axis. Arrows indicate the RosettaCM score interval. Models with better stereochemistry have lower MolProbity scores. Seven servers with summed GDT-TS *Z* score < −30 (B) were excluded from the GDC-SC, MolProbity, and native hydrogen bonds summaries because evaluations of side-chain and physical properties of the models are only meaningful when the global structure is correct. The CASP10 targets the average GDT between the first model by the three servers (RosettaServer, HHpredA, and ZhangServer), and the templates used by RosettaCM are listed in Table S1. Additional analysis of the 63 easy targets based on sequence identity between the target and the closest template is shown in Figure S4.

changes taking place during modeling for the regions indicated by the red arrows. The most-often observed scenario was improvement in loop regions (Figure 4A). Concerted improvements in secondary structure placement and loop geometry were also often observed (Figures 4B and 4C).

Loop region improvement is illustrated by T0667. In target T0667, there is a deletion in the residue 161–163 loop in the closest template (2WTM). There is another template with a loop of the same length (1ISP), but the conformation is quite different (2 Å over the three loop residues; Figure 4D). The RosettaCM model is much closer to the native structure (0.9 Å over the loop region; Figure 4G) compared to the other server models (Figure 4J). The improvement in loop modeling lowers the rmsd for the residues indicated by the arrow in Figure 4A. The improvement in model accuracy comes from combining fragments from the lower-ranked template, and energy minimization after the fragment is superimposed.

Concerted backbone repositioning is illustrated by T0702 and T0685. In T0702, a nonconservative glycine-to-histidine substitution at position 5 results in a side-chain-side-chain hydrogen bond with Asn 59 that does not exist in the template, which is associated with a helix shift and loop structure change relative to the closest template, 2RCY (Figure 4E). The RosettaCM model recapitulates this hydrogen bond, and the associated helix shift and loop changes (Figure 4H). The other server models do not reproduce the hydrogen bond or the backbone structural changes (Figure 4K). These changes together improve the rmsd in the region indicated by the arrow in Figure 4B. Similarly in T0685, the interhelix interaction between a Phe and Tyr in the top template used by all three servers, 2C2A, is changed to Ala and Gly in the target, which causes two helices to collapse toward each other (Figure 4F). RosettaCM is able to model this change as well as the loop connecting the helix accurately (Figure 4I). In comparison, other methods either stayed close to the template structure or modeled the helix shift but not the conformational changes in the loop region (Figure 4L).

RosettaCM is freely available to academic users as part of the Rosetta software suite. As detailed in Supplemental Experimental Procedures, the user provides—in addition to the protein sequence—a set of template structures and sequence alignments to these structures. Available experimental data—electron density maps, NMR data (chemical shifts, RDCs, and NOEs), and X-ray diffraction data—can be input into RosettaCM to supplement homologous structure information. RosettaCM is also available through the ROBETTA server, which uses Hhsearch (Remmert et al., 2012), SPARKS-X (Zhou and Zhou, 2004), and RaptorX (Peng and Xu, 2009) to generate the input alignments. It is clear that improved results can be obtained using more sensitive remote homolog detection and sequence alignment methods, and methods' developers working in these areas should be able to use RosettaCM to build improved models. In particular, the superb remote homolog detection by the Zhang group based on structural similarity with de novo models should greatly improve modeling of proteins based on very distantly related targets.

The run time of RosettaCM is determined by the number of independent trajectories carried out. A single modeling trajectory for a 200-residue protein takes about 10 minutes, and—for sequences with greater than 25% sequence identity to a protein of known structure—only five to ten trajectories are necessary

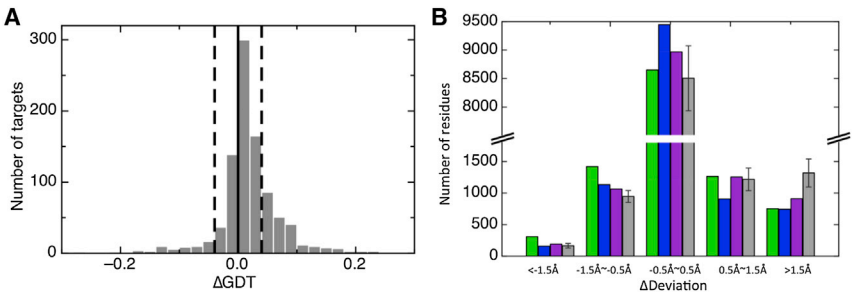

Figure 3. RosettaCM Improves Model Accuracy in the Aligned Regions Relative to Starting Template Structures

(A) Distribution of ΔGDT for 847 targets used in CAMEO benchmark test, where ΔGDT is the difference in GDT between RosettaCM models and the top-ranked template calculated over the aligned region (positive values are improvements). A scatter plot comparing GDT of RosettaCM models over the aligned region and the top-ranked templates is shown in Figure S2.

(B) Histogram of per-residue changes in model accuracy relative to the closest available templates over the 63-target subset from CASP10. Per-residue deviation data for each target were obtained from the CASP10 web page. The x-axis (Δdeviation), is calculated as distance(model) — distance(template), where distance(model) and distance(template) are pre-residue distance between the model (or the template) and the native structure. Negative values indicate improvements over the closest template. Numbers of residues in different deviation ranges are shown for RosettaServer (green), HHpredA (blue), ZhangServer (magenta), and the average of the rest of the top 20 servers (gray). The SDs of the rest of the top 20 servers are shown as error bars. Comparisons to all the servers are shown in Figure S4.

for accurate modeling (see Supplemental Experimental Procedures). Hence, RosettaCM could be used in conjunction with servers such as HHsearch, which produce accurate alignments using robust statistics with very little wait time.

EXPERIMENTAL PROCEDURES

RosettaCM Protocol

The workflow in the RosettaCM modeling protocol is outlined in Figure 1. The inputs to RosettaCM are alignments of the sequence of the protein of interest to proteins of known structures, and standard Rosetta de novo modeling fragment sets to model the unaligned regions and to explore deviations from the templates in the aligned regions. The alignments to proteins of known structure can be generated using remote homolog detection programs such as PSIBLAST (Altschul et al., 1990), Hhsearch (Remmert et al., 2012), SPARKS-X (Zhou and Zhou, 2004), and RaptorX (Peng and Xu, 2009), or using expert knowledge of the protein family and any available experimental information. The user can provide an optional file specifying the weight to be given to each alignment during modeling ("weights" file); if no weights file is provided, the input alignment file should be ordered such that the most confident alignments are first, i.e., RosettaCM assumes in the absence of a provided weights file the decrease in alignment accuracy from the top-ranked model to the nth ranked model observed for HHsearch alignments for a large set of proteins. Rosetta de novo fragment files can be generated using the Rosetta program or ROBETTA server as described elsewhere.

RosettaCM builds models from these inputs as described in the following paragraphs. The RosettaCM script provided in the Supplemental Experimental Procedures carries out all of the steps.

Probabilistic distance restraints are generated from the weighted input alignments as described previously (Thompson and Baker, 2011). For short gaps, the contribution of alignments lacking a particular pair of residues to these distance constraints is the background distance distribution (see Thompson and Baker, 2011). If there is a gap longer than 50 residues in one template, then the contribution of this template to the gapped residues is excluded, and the contribution of the rest of the templates is renormalized to avoid blurring out the restraints in domains that are only represented in a subset of the alignments. Models are then assembled and optimized in three stages. In the first stage, complete chain models are built up by recombining fragments from the aligned template structures and de novo fragments representing the unaligned regions. In the second stage, deviations from the templates are explored, and gaps in the models are closed using a combination of fragment superposition and Cartesian space minimization. In the third stage, side-chain and backbone conformations are optimized using Rosetta full-atom refinement.

Stage 1

Global Superposition

A stochastic procedure is used to select a template, which is then used to generate a global superposition of the aligned portions of the templates. Because the global alignment most consistent with the actual structure is unknown in advance, this is done independently for each model generated to sample different possible global superpositions. First, for each alignment, the sequence of interest is threaded onto the corresponding template structures to generate a set of partial threads. One of the partial threads is randomly selected as the base model for the superposition with probability given by the user specified or default weight assigned to the alignment as described above. For each of the remaining partial threads, the coordinates are transformed to minimize the rmsd with the base thread over the residues they have in common. Partial threads with no residues in common with the base model are eliminated. If a partial thread is parsed into multiple domains (using DDOMAIN; Zhou et al., 2007), each domain is superimposed independently, resulting in global orientation between domains similar to that in the base model, with structural variations within domains from the different partial threads.

Template Fragment Generation

To allow recombination of structural elements present in the global superposition, each partial thread is broken up into segments corresponding to secondary structure elements. Secondary structure is first assigned using DSSP (Kabsch and Sander, 1983), and continuous helices of at least six residues or strands of at least three residues are added to a fragment list. The interconnecting loops are split and joined to the connected helix or strand segment. Secondary structure segments separated by less than three loop residues are grouped into the same segment so kinked helices and tight β hairpins are treated as a single rigid segment.

Fragment Recombination

Full-chain models are generated by recombining the template-derived segments with Rosetta de novo fragments that cover the regions not represented in the templates. Unaligned regions are split in half, and each half is associated with the adjacent aligned region. Structures are generated from the template and Rosetta de novo fragments according to a Rosetta fold tree: the global position of each segment is represented in Cartesian space, whereas the backbone and side-chain conformations of each segment are represented in torsion space. A Monte Carlo trajectory is carried out with two types of moves. (1) Substitution of the backbone torsion angles of a randomly selected Rosetta de novo fragment for the current torsion angles of these residues, and regeneration of coordinates according to the fold tree (Figure 1B). The segmentation of the protein limits propagation of torsion angle changes to the closest end of the segment. (2) Substitution of the coordinates of a randomly selected partial thread-derived fragment (in the global frame) for the current model coordinates of these residues (Figure 1C).

The scoring function used in the Monte Carlo trajectory is a linear combination of the Rosetta low-resolution (centroid) energy function, which favors compact structures with buried hydrophobic residues and paired β strands, the template-derived restraint functions described above, and a chain break term that penalizes large distances between residues adjacent in the sequence that can arise at fold tree boundaries (the middle of unaligned regions). As in the Rosetta de novo structure prediction protocol (Rohl et al., 2004), these terms are gradually phased in. At the beginning of the trajectory,

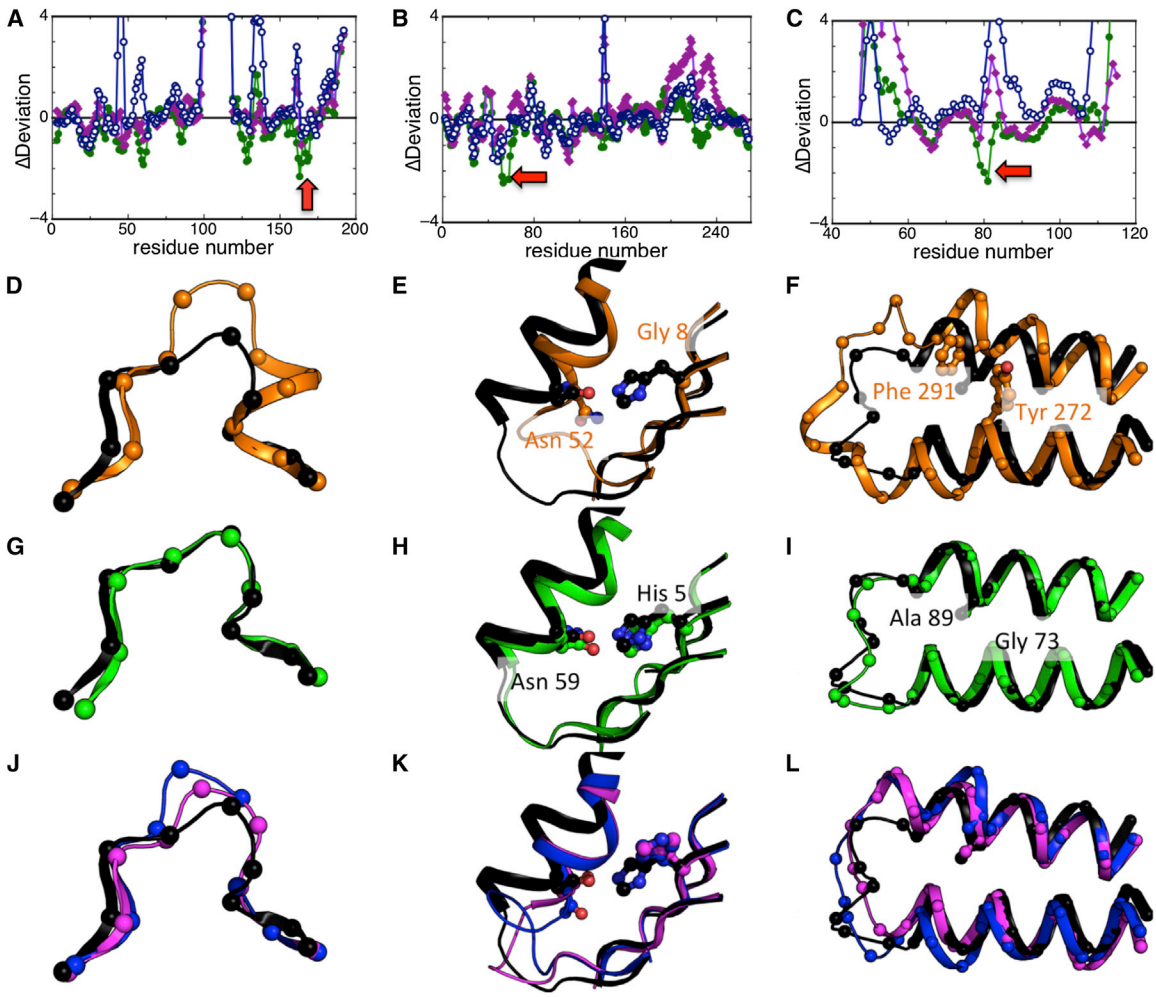

Figure 4. Examples of Improvements over Starting Templates in CASP10

(A)–(C) Per-residue change in accuracy relative to the best template of RosettaCM (green), HHpred (blue), and ZhangServer (magenta) for T0667 (left), and T0702 (middle), and T0685 (right). Values less than zero indicate regions in which the submitted model is closer to the true structure than the best template. Results are shown for first submitted models. The structural comparisons in (D)–(L) are over the region with the largest improvements over the templates indicated by the red arrow in (A)–(C). (D–L) The native structures are in black, the best template is in orange (D–F), and models from RosettaServer are in green (G–I). HHpredA and ZhangServer models are in blue and magenta for comparison (J–L). Orange labels, aligned template residue identities; black labels, the target residue identities.

only excluded volume interactions are considered, then secondary structure pairing and hydrophobic burial, and then the remaining terms. The chain break term between residues in separate branches of the fold tree but adjacent in the linear sequence is $|r-r_0|$, where r is the distance between the bonded atoms, and r_0 is the idealized bond length between the atoms. This term is set to zero at the first half of the Monte Carlo trajectory and linearly ramped up to full strength in the second half. This allows large structural changes to be sampled while still favoring separations small enough so that the gaps are closable in the second phase.

The total number of steps in these first-stage Monte Carlo trajectories was set to 10,000, with 5,000 attempts at inserting template fragments and 5,000 for de novo fragment insertions. Both the total numbers of Monte Carlo steps and the ratio between two types of fragment insertions are adjustable. This first stage takes about 1 min for a 150-residue protein. The lowest-energy structure sampled during the trajectory is passed on to the second stage.

Stage 2

The models generated in stage 1 contain all residues and generally have the correct overall topology but are suboptimal in two ways: first, the aligned regions are often very close to one of the input template structures; and second, the backbone geometry at the junctions between fold tree branches is often quite distorted. To sample further from the input template structures, and to close the loops, a Monte Carlo trajectory using a two-step move is carried out. The first step consists of random selection of a de novo or template-based fragment, and substitution into the current conformation of the coordinates of the superimposed fragment. In the de novo fragment case, the N- and C-terminal residues of the fragment are superimposed on the corresponding residues of the current conformation (Figure 1D), and the fragment insertions are biased toward regions in which the backbone is most distorted as assessed by the local bond length and bond angle energies. In the template fragment case, the superposition is over all residues in the fragment, not just the termini. Following the fragment insertion, Cartesian space quasi-Newton (BFGS) minimization is carried out using a differentiable version of the Rosetta centroid energy function described in the next paragraph, the template-derived restraints, and explicit bond length, bond angle, and improper torsion energy terms in place of the relatively weak chain break term used in stage 1 (Figure 1D).

The differentiable centroid energy function makes use of smooth reparameterizations of the centroid pair and environment terms, which enforce pair distributions and nonpolar burial, respectively, and the C_β and cenpack terms,

which enforce native-like core packing (Rohl et al., 2004). The smooth reparameterizations fit mixtures of Gaussians to empirically derived distributions; the relatively small number of Gaussians needed to fit these distributions (generally two to four for each pair distance distribution) offers a significant reduction in parameters versus the previous table-based parameterization. Neighbor counts are sigmoid smoothed. The resulting continuously differentiable energy function allows minimization with centroid energies, which allows optimization of backbone hydrogen bond and covalent-bonded geometry, without requiring the expensive rotamer optimization calculations needed to accurately compute all-atom energies.

As described previously, each move in the second-stage Monte Carlo trajectories involves fragment insertion by superposition followed by full backbone minimization. The total number of attempted moves is 1,500, with 1,000 template fragment insertions and 500 de novo fragment insertions. This second stage takes 5 min for a 150-residue protein. The lowest-energy structure sampled during the trajectory is passed on to the final full-atom refinement phase.

Stage 3

The low-energy structures resulting from the stage 2 trajectories have near-ideal backbone geometry, but side chains are not explicitly represented. In stage 3, the Rosetta Monte Carlo combinatorial side-chain optimization method is used to build on side chains, and the recently developed Rosetta "FastRelax" protocol is used to iteratively refine the side-chain and backbone conformations (Tyka et al., 2011). Annealing is carried out by ramping up and down the strength of the repulsive interactions and, at each iteration, repacking the side chains and subjecting the whole structure to quasi-Newton optimization of the side-chain and backbone coordinates first in internal coordinates and then in Cartesian coordinates. The Rosetta full-atom energy function supplemented with the alignment-derived restraint function is used in all calculations with the weight on the repulsive interactions varied as described above.

Model Selection

Final models (which may be generated from different seed alignments) are collected, and the best 10% of the models by energy is identified. These structures are then clustered, and the center of the largest cluster (where each model is weighted such that low-energy models have highest weight; Xiang et al., 2002) is selected as the top model. In cases such as CASP where multiple models are desirable, additional models are identified by repeating the clustering process after the 10% of the models closest to the selected model is removed.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results, Supplemental Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.str.2013.08.005.

ACKNOWLEDGMENTS

We thank Andriy Kryshtafovych for providing CASP10 analysis and helpful discussions. Research reported in this publication was supported by NIGMS of the National Institutes of Health under award number R01GM092802-02.

Received: March 15, 2013
Revised: July 28, 2013
Accepted: August 2, 2013
Published: September 12, 2013

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Das, R., and Baker, D. (2008). Macromolecular modeling with rosetta. Annu. Rev. Biochem. *77*, 363–382.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 3rd, Snoeyink, J., Richardson, J.S., and Richardson, D.C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. *35*(Web Server issue), W375–W383.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr. Protoc. Bioinformatics *Chapter 5*, Unit 5 6.

Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.M., Wilson, I.A., and Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science *332*, 816–821.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Keedy, D.A., Williams, C.J., Headd, J.J., Arendall, W.B., 3rd, Chen, V.B., Kapral, G.J., Gillespie, R.A., Block, J.N., Zemla, A., Richardson, D.C., and Richardson, J.S. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins *77*(*Suppl 9*), 29–49.

King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., André, I., Gonen, T., Yeates, T.O., and Baker, D. (2012). Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science *336*, 1171–1174.

Peng, J., and Xu, J. (2009). Boosting protein threading accuracy. Res. Comput. Mol. Biol. *5541*, 31–45.

Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. *39*(Database issue), D465–D474.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins *77*(*Suppl 9*), 89–99.

Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., et al. (2010). NMR structure determination for larger proteins using backbone-only data. Science *327*, 1014–1018.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*, 173–175.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*(Web Server issue), W244–W248.

Thompson, J., and Baker, D. (2011). Incorporation of evolutionary information into Rosetta comparative modeling. Proteins *79*, 2380–2388.

Tramontano, A., Leplae, R., and Morea, V. (2001). Analysis and assessment of comparative modeling predictions in CASP4. Proteins (*Suppl 5*), 22–38.

Tyka, M.D., Keedy, D.A., André, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607–618.

Xiang, Z., Soto, C.S., and Honig, B. (2002). Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc. Natl. Acad. Sci. USA *99*, 7432–7437.

Xu, D., Zhang, J., Roy, A., and Zhang, Y. (2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab

initio folding and FG-MD-based structure refinement. Proteins *79*(*Suppl 10*), 147–160.

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. *31*, 3370–3374.

Zemla, A., Venclovas, C., Moult, J., and Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. Proteins (*Suppl 3*), 22–29.

Zhou, H., and Zhou, Y. (2004). Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins *55*, 1005–1013.

Zhou, H., Xue, B., and Zhou, Y. (2007). DDOMAIN: dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci. *16*, 947–955.

