PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

CASP11 refinement experiments with ROSETTA

Hahnbeom Park,[1,2] Frank DiMaio,[1,2] and David Baker[1,2,3]*

[1] Department of Biochemistry, University of Washington, Seattle, Washington 98195

[2] Institute for Protein Design, University of Washington, Seattle, Washington 98195

[3] Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195

ABSTRACT

We report new Rosetta-based approaches to tackling the major issues that confound protein structure refinement, and the testing of these approaches in the CASP11 experiment. Automated refinement protocols were developed that integrate a range of sampling methods using parallel computation and multiobjective optimization. In CASP11, we used a more aggressive large-scale structure rebuilding approach for poor starting models, and a less aggressive local rebuilding plus core refinement approach for starting models likely to be closer to the native structure. The more incorrectly modeled a structure was predicted to be, the more it was allowed to vary during refinement. The CASP11 experiment revealed strengths and weaknesses of the approaches: the high-resolution strategy incorporating local rebuilding with core refinement consistently improved starting structures, while the low-resolution strategy incorporating the reconstruction of large parts of the structures improved starting models in some cases but often considerably worsened them, largely because of model selection issues. Overall, the results suggest the high-resolution refinement protocol is a promising method orthogonal to other approaches, while the low-resolution refinement method clearly requires further development.

Proteins 2016; 84(Suppl 1):314–322.

Key words: structure prediction; structure refinement; protein loop modeling; protein homology modeling; Monte Carlo simulation.

INTRODUCTION

Since the progress in homology model refinement described in CASP10,[1,2] studies in the area have been in two directions. One direction has aimed to make refinement found in CASP10[3,4] more robust, and the other has searched for progress from alternative sources. This was evident at the CASP11 meeting; many groups adopted molecular dynamics (MD)-based approaches, while others sought to develop distinct approaches. The goal of the latter approaches is to solve challenges that are not well addressed yet, typically "sampling bottleneck" problems, as highlighted by the assessor in the CASP11 conference. There have been clear bottlenecks in general performance for certain starting structures throughout the CASP refinement experiments, in particular, when starting structures have been far off from the native structure, no group has succeeded in improving them substantially.[1] "Discrete search" using Monte Carlo in Rosetta[5] complements the "continuous search" of molecular dynamics (MD) simulation and hence could be useful for tackling those unsolved challenges.

In this report, we first describe new protein structure refinement methods developed within Rosetta, and then describe the strengths and weaknesses of the methods indicated by the CASP11 results.

METHODS

Definition of the refinement space

Our development of new methods was based on the following conceptual understanding of the refinement problem. We define the "refinement space" as the volume

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

Grant sponsor: US National Institutes of Health (to H.P. and D.B.); Grant number: R01GM092802.

*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Seattle, WA. E-mail: dabaker@uw.edu

Received 15 May 2015; Revised 19 July 2015; Accepted 21 July 2015
Published online 23 July 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24862

of phase space that needs to be searched for the refinement of a model structure into the native structure. We consider three orthogonal axes in this space: core-refinement, partial reconstruction, and fold-level reconstruction. Core-refinement is where recent progress has been found from MD-based approaches.[3,4] As we recently described,[6] core-refinement methods primarily improve residues that are nearly correct, but do not substantially improve regions with significant errors. Rosetta sampling methods could be useful for refining along the two other axes and so complement the limitations in core-refinement approaches. When the starting structure is not too far away from the native structure (high-resolution refinement), the native structure can be reached by searching along the first two axes, rebuilding incorrect regions while refining the core. When the core of the protein is incorrectly modeled, full fold-level reconstruction is required to bring an incorrect structure to a roughly correct position, from which core refinement can further improve the model.

This picture of the refinement problem is consistent with the sampling bottlenecks found in core-refinement approaches in CASP experiments. For high-resolution problems, "bottleneck" regions in starting structures (for example, incorrect loops impeding correct core packing) may limit refinement of the core region; consequently, the magnitude of improvement achieved by core-refinement approaches strongly depends on the starting structure. For low-resolution problems, the lack of a properly formed core to serve as a seed for core-refinement similarly limits the magnitude of improvements.

Sampling methods

Our CASP11 protocol applies structural perturbations with a range of magnitudes to an evolving ensemble of conformations. In this section, we first provide a bottom-up explanation, starting from the basic refinement modules. These take as input one or a set of models, and output one or a set of refined models. In Rosetta, operators that take models as inputs and produce new models as output are called "Movers." Movers can readily be combined in series for single processor composite protocols, or in parallel for parallel refinement protocols as described below.

Rosetta_relax[7] alternates between discrete Monte Carlo side-chain optimization and quasi Newton minimization of backbone and side-chain degrees of freedom. Five iterations are carried out: in the first three, minimization is with respect to the internal degrees of freedom; in the last two, with respect to the Cartesian degrees of freedom (x, y, and z). Restraints tether the model to the input structure.

Trajectory_average[6] builds on recent advances in MD-based core-refinement methods but requires considerably less computing power. Multiple independent restrained MD or Monte Carlo minimization (MCM) simulations[7] using the Rosetta implicit solvent energy function[8] are carried out [as movers in Fig. 1(A)] with variations in the energy function and initial side-chain preparation. Structures from those simulation trajectories are collected, filtered, and structurally averaged. The outcome of the protocol is a single refined structure with moderate changes to the starting structure. A detailed description of the method and analysis of its performance is reported in Ref. 6.

Normal_mode_sampler samples along the normal modes derived from an anisotropic network model (ANM).[9] One of the ten lowest eigenvalue modes is selected at random, restraints are generated from models perturbed by 0.5 or 1.0 Å Cα RMSD along this direction, and the input model is minimized with respect to the energy supplemented with these restraints.

Segment_remodel rebuilds specified regions either using Rosetta fragment insertion followed by minimization (as in RosettaCM[10]) or kinematic loop closure[11] for internal regions, and fragment insertion[12] for termini.

The Rosetta all atom implicit solvent energy function[8] was used in all of these refinement modules; the *Segment_remodel* and *Normal_mode_sampler* method also use the Rosetta centroid energy function for the initial stages and switch over to an all-atom representation and energy function at the end. Ranking among the models generated by a method that outputs multiple models (*Segment_remodel* and *Normal_mode_sampler*) was done by the normalized sum of Rosetta full-atom energy and GOAP score.[13] *Trajectory_average* outputs a single structurally averaged atomic model so there is no need for selection.

The four sampling methods were integrated into composite sampling protocols using the iterative parallel-computing platform in Rosetta [Fig. 1(A)] adapted from Tyka *et al.*[14] In this three-layer parallel architecture, *Emperor*, *Master*, and *Slave* processors communicate to sample conformations starting through a continuously-evolving pool of structures. For our experiments, this structure pool was maintained at 20 structures. At each iteration, 100 new structures are collected from *Master* processes, added to the current pool, and 20 models are selected using Pareto optimization[15] according to three objective functions: Rosetta energy, GOAP statistical potential,[13] and "structural diversity". The structural diversity of each model was computed as the sum of the *S*-score[16] values to all other pool members.

Segment_remodel and *Rosetta_relax* were combined into a composite "*Rebuild_and_relax*" sampling protocol, where they were repeated for ten iterations in this parallel architecture. The regions to rebuild in *Segment_remodel* are chosen at the beginning of *Rebuild_and_relax* based on the residue fluctuations[17] observed in multiple short independent Rosetta MD simulations (20 ps × 10

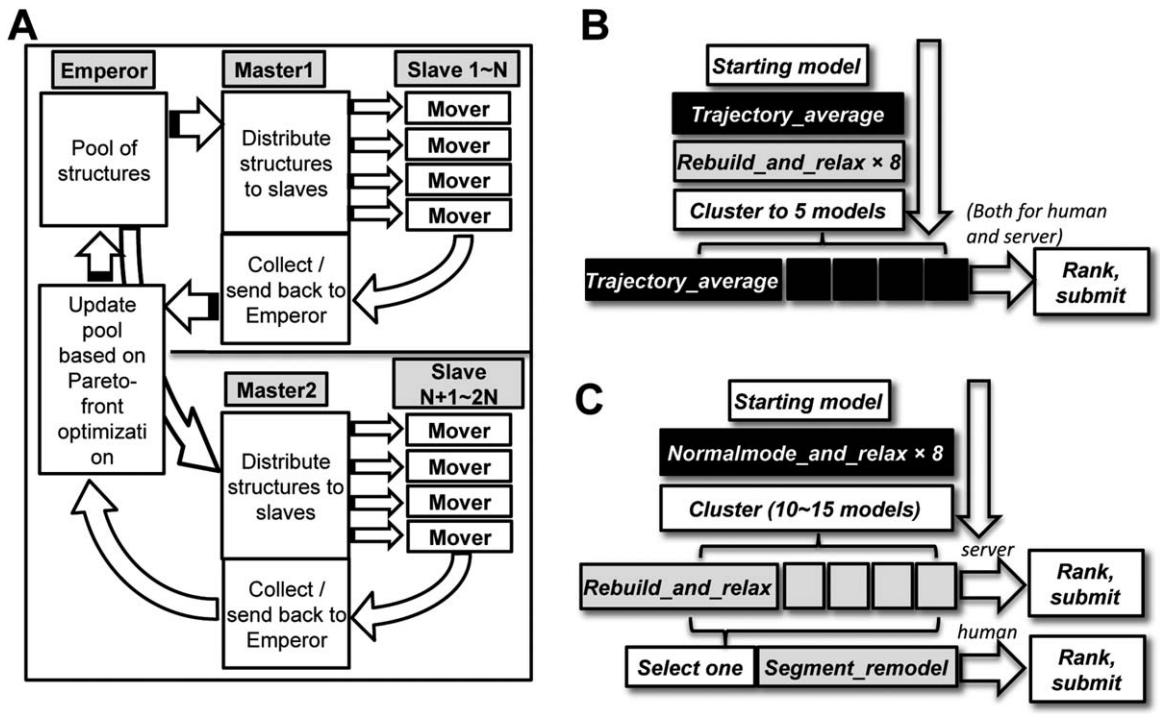

Figure 1

Rosetta parallel-computing refinement protocols used in CASP11. (**A**) General three-layer parallelism algorithm used for *Normalmode_and_relax* and *Rebuild_and_relax*. The roles of CPUs are divided into (i) "Emperor" controlling structural pool, (ii) "Masters" controlling job distribution to "Slaves," and (iii) "Slaves" carrying out actual unit sampling, as originally designed by Tyka *et al.*[14] This architecture can be modified to carry out various modeling tasks such as *Rebuild_and_relax* or *Normalmode_and_relax* by varying the "mover" that perturbs structures. The size of pool was set to 20 in CASP11. (**B**) Strategy for high-resolution targets (starting GDT-HA ≥ 50). The starting model is first fed into *Trajectory_average*[6] to generate a single refined model, followed by *Rebuild_and_relax* on it to generate multiple models with different conformations at selected regions. *Trajectory_average* is applied again separately to each of five models selected from *Rebuild_and_relax* output. (**C**) Strategy for low-resolution targets (starting GDT-HA < 50). An ensemble is generated from the starting model using *Normalmode_and_relax*, clustered into 10 representative structures, and were further refined using by *Rebuild_and_relax*. For human submissions, additional regions (20–50% of the whole structure) were selected and rebuilt by *Segment_remodel* using fold-tree *AbInitio*.[10] Models for both categories of targets were ranked as described in the main text.

trajectories): the most fluctuating regions are rebuilt, along with regions with many buried polar residues or exposed nonpolar residues. Overall, 146 regions (average length of 8.2 residues per region) for 36 targets (4.1 regions per target, the engineered protein from the Baker group (TR769) was skipped) were rebuilt using *Rebuild_and_relax* in the automatic predictions, and an additional 60 regions in the human submissions. Likewise, *Normal_mode_sampler* and *Rosetta_relax* were combined into a composite "*Normalmode_and_relax*" sampling protocol. 10 iterations of *Normal_mode_sampler* in the sampler produces a net change to the starting structure of up to ~3 Å Cα RMSD.

Protocol choice based on target difficulty

One of two overall refinement protocols [Fig. 1(B,C)] was used depending on the starting model GDT-HA to the native structure (iGDT-HA) provided by the CASP11 organizers. We refined 24 targets with iGDT-HA greater than 50.0 using a "high-resolution strategy" [Fig. 1(B), TR769 was skipped] combining *Trajectory_average* and *Rebuild_and_relax* with KIC segment rebuilding. This combination was used because the quality of partial reconstruction strongly relies on the quality of the remaining parts of the model,[18,19] and the success of core-refinement depends on the accuracy of modeling of individual regions. Iteration proceeded as shown in Figure 1(B); multiple models (80 models from 4 independent runs) with varied segments were generated, and a number of cluster centers (usually 5) were selected for the next iteration of core-refinement. Models were ranked as described below.

The remaining 12 targets with iGDT-HA below 50 were refined using a "low-resolution strategy" [Fig. 1(C)] that combines *Rebuild_and_relax* with fragment-based segment rebuilding and *Normalmode_and_relax*. The output structures from five independent runs (overall 100 models) of *Normalmode_and_relax* were clustered and the lowest GOAP[13] score structures within each cluster were selected as cluster representatives. Cluster size varied between 10 and 15 depending on structural

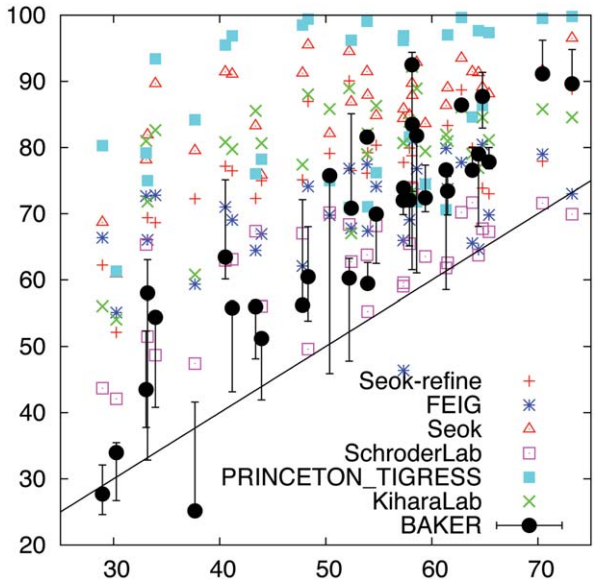

Figure 2

Dependence of structural similarity (GDT-HA$_{\text{starting-to-model}}$) between the submitted models and starting structures (y axis) on target difficulty (*x* axis). Lower GDT-HA$_{\text{starting-to-model}}$ indicates more aggressive submission. Diagonal line shows the structural changes necessary to bring starting structure to the native structure. Submissions by other groups are indicated for comparison.

variation among the sampled models. Finally, each of the cluster representatives was chosen as a starting point for one round of *Rebuild_and_relax* with rebuilding in the predicted unreliable regions (by Rosetta MD simulations, see above). For the human submissions, we tried even more aggressive sampling; regions considered to be incorrectly modeled (roughly 20–50%) were removed, and *Segment_rebuild*—using the RosettaCM[10] fragment based protocol—was run.

For human submissions (Models 2–5) for targets <80 residues and iGDT-HA close to 50 (TR759, TR816, and TR829), we started from a Rosetta *AbInitio*[20] model selected by iGDT-HA and GOAP scores. Because of this, TR816 and TR829—that fell into the high-resolution category for server submissions—followed a strategy closer to that of low-resolution modeling for Models 2–5.

The amount of computer resources used for the automated pipeline was generally less than 1 day per target using 20 cores (<500 core hours). Scaled to a 200-residue protein, the high-resolution strategy, with 6 runs of *Trajectory_average* and 8 runs of *Rebuild_and_relax* [Fig. 1(B)], takes 320 core hours; the low-resolution strategy, with 5 runs of *Normalmode_and_relax* and 10–15 runs of *Rebuild_and_relax* [Fig. 1(C)], takes 250 core hours. Additional computations carried out for human predictions varied significantly based on target, but were generally on the order of hundreds of core hours.

Model selection and ranking

For human model submissions, we considered the extent of change required for complete refinement (to bring the starting model to the native). In Figure 2, the structural changes to the starting models (GDT-HA$_{\text{model-to-starting}}$) are shown for our and other top-ranked groups' submissions as functions of iGDT-HA (GDT-HA of starting structures). Lower GDT-HA$_{\text{model-to-starting}}$ implies less similarity between submitted and starting structure and less conservative refinement. For our CASP11 submissions, we selected models where GDT-HA$_{\text{model-to-starting}}$ was close to iGDT-HA. As shown in the figure, our submissions matched this condition reasonably, while most of the other top-groups made more conservative submissions except for group 396 (SchröderLab).

Ranking of the five submitted models used GOAP score. For human submissions of low-resolution targets, structural changes to starting models were also considered and Model 1 was selected from among the more conservative models.

RESULTS

Overall results strongly depend on strategy

In Figure 3(A), GDT-HA model quality values are compared between the starting models and our automated (bottom) or human (top) submissions. Throughout the article, we describe results for our human submissions (group BAKER) unless specified. It is clear in both cases that our high-resolution strategy (colored red in the figure) brought overall improvements to both Model 1 and the best model among our five submissions (top5). On the other hand, the performance of the low-resolution strategy (colored blue in the figure) is questionable, at least for model selection, where there is a large gap between the overall quality of Model 1 and top5 submissions. A similar trend is found by *Z*-score analysis [Fig. 3(B)]. The panel at the bottom shows the *Z*-score difference between Model 1 and top5; the large decrease in *Z*-scores for low-resolution targets is due to model selection failures. The large difference between the performances of the high- and low-resolution strategies is not surprising as the methods are quite different.

Table I summarizes performance according to four different measures: GDT-HA, SphereGrinder,[21,22] Cα RMSD, and MolProbity.[23] In the accumulated *Z*-scores in GDT-HA for Model 1 submissions (against other groups submissions, setting minimum *Z*-score = −2.0), positive contributions by 24 high-resolution targets (+15.1, fourth among the groups) are offset by the negative contributions from 12 low-resolution targets (−11.6, third worst). Similarly, using Cα RMSD, BAKER group

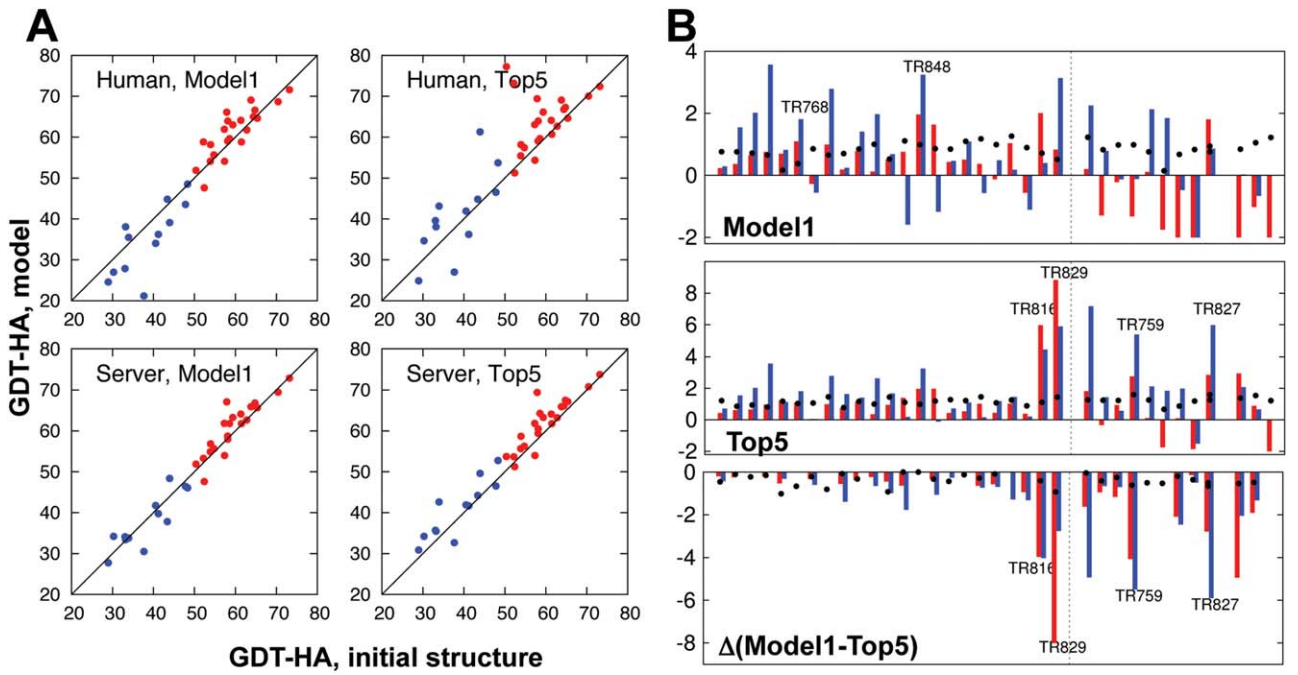

Figure 3

Overall results in CASP11 refinement category by Baker group. (**A**) GDT-HA comparison between starting structures and refined models for human (upper) and server (lower) submissions. Results for the models submitted as Model 1 are plotted at left, and for the best among five models (top5) at right. Dots above diagonal lines are those targets improved over starting structures. The high-resolution strategy was used for the red targets, and the low-resolution strategy, for the blue targets. (**B**) Per target *Z*-score for Model 1 (upper), best among five (top5, middle), and the difference between the two (lower). Red and blue bars represent *Z*-scores in GDT-HA and SphereGrinder,[21,22] respectively. Black dots show *Z*-scores in MolProbity.[23] Positive values are improvements for all the quality measures. Dotted lines at the center of each panel show the border between targets that underwent high-resolution and low-resolution refinement strategies.

did the best for high-resolution targets (summed *Z*-score +26.4; second rank SHORTLE with +18.0), but second worst in low-resolution targets; SphereGrinder shows similar trends: best in high-resolution (summed *Z*-score by +23.3; second rank Seok-refine +19.5) but mediocre in low-resolution. If *Z*-scores less than 0 are not penalized (as in CASP10[1]), the statistics are somewhat different (Supporting Information Table S1). Finally, MolProbity was improved for all the targets regardless of strategy.

In the following, we analyze the results separately for the targets subjected to the high- and low-resolution strategies.

High-resolution targets: significant improvements from the reconstruction of unreliable regions

The aims of our experiment for high-resolution targets were (a) to determine to what extent the implicit solvent simulation implemented in Rosetta can capture recent

Table I
Improvements Over Starting Models for Human Submissions

Target category	Num. Targets		GDT-HA (%)	Sphere Grinder (%)	Cα RMSD (Å)	MolProbity
iGDT-HA ≥ 50[a]	24[b]	Model 1	+1.0 (+15.1)	+3.8 (+23.3)	−0.50 (+26.4)	−1.1 (+19.7)
		Top5	+3.9 (+35.0)	+7.2 (+46.5)	−0.84 (+45.4)	−1.4 (+27.3)
iGDT-HA < 50[c]	12	Model 1	−3.9 (−11.6)	−1.1 (+2.3)	+1.31 (−6.6)	−1.3 (+10.0)
		Top5	+2.0 (+6.4)	+4.3 (+21.4)	+0.27 (+1.5)	−1.6 (+15.0)
Overall	36	Model 1	−0.6 (+3.5)	+2.2 (+25.6)	+0.10 (+19.8)	−1.2 (+29.7)
		Top5	+3.3 (+41.4)	+6.3 (+67.9)	−0.47 (+46.9)	−1.5 (+42.3)

Average improvements are given using four different measures, and the sum of *Z*-scores relative to other group predictions are in parentheses. Summed *Z*-scores are calculated using −2.0 for the minimum value, and positive values are better for all measures. Analysis with minimum *Z*-score = 0.0 is provided in Supporting Information **Table S1**.

[a]Targets with starting GDT-HA higher than or equal to 50.0, to which high-resolution strategy is applied.

[b]TR769 is not submitted as the target protein is engineered from Baker group.

[c]Targets with starting GDT-HA lower than 50.0, to which low-resolution strategy is applied.

advances in core-refinement and (b) if there is synergy between partial reconstruction and core-refinement methods. To address these questions, we consider two distinct quality metrics: GDT-HA and SphereGrinder (or Cα RMSD). As pointed out by the assessor, each quality measure reports on different aspects of protein model quality: GDT-HA is most sensitive to precise placement of protein core atoms, reporting on the performance of the core-refinement method, while SphereGrinder (and Cα RMSD) is most sensitive to repositioning of substantially incorrect regions, and therefore reports more on the performance of the segment reconstruction method.

For the Model 1 submissions, the average GDT-HA improvement for high-resolution targets is less (+1.0) than that of the methods based on explicit water MD simulations (Group 288 (Feig) +2.9, Group 396 (Schröder) +1.5). Also the fraction of targets that show an improvement in GDT-HA (54%) is relatively small compared to that of other top-groups; a conservative core-refinement approach by Group 106 (Tigreless_Princeton)[24] yielded a similar mean improvement in GDT-HA (+1.1) but with the highest fraction of models improved among all groups (88%). A control experiment running only our core-refinement method (*Trajectory_average*) shows similar mean improvement but with an increased fraction of success (77%). This suggests that running aggressive reconstruction (approximately five regions per target) on top of the outcome of *Trajectory_average* can sometimes improve GDT-HA but also can hurt, and adds noise to the consistency of improvements brought by *Trajectory_average*. Focusing on GDT-HA only, aggressive reconstruction efforts might be regarded as not very successful.

The reconstruction method appears more promising by other measures, however. In SphereGrinder and Cα RMSD, the average magnitude of improvements outperforms that of the core-refinement approach: +3.8/−0.50 Å (SphereGrinder/Cα RMSD) by BAKER and +2.7/−0.44 Å by BAKER_RefineServer, compared to *Trajectory_average* only (+0.8/−0.02 Å; for Cα RMSD, negative values are improvements). These results are better than those of other groups who mainly employed core-refinement approaches, group 288 (Feig, +0.6/+0.14 Å) and 333 (Kihara, +1.3/−0.06 Å). The fraction of models improved by these measures is also higher than those improved by GDT-HA, with 71% improved under each metric. This is consistent with our view that partial reconstruction is on a different axis than core refinement, and also with the assessor's view that model quality improvement from the reconstruction of incorrect regions is less well captured by GDT-HA than by metrics such as SphereGrinder or RMSD.

Analyzing representative examples provides further evidence that the structural improvements result from the combination of the core refinement and segment reconstruction protocols. TR848 is one of the highlights with significant improvements in an unreliable region in the starting structure [Fig. 4(A)]; our Model 1 improves in SphereGrinder and Cα RMSD by 14.1% and 1.1 Å, respectively (best among all groups), but is not the best by GDT-HA improvement (5.1%). Model 1 by group 288 (Schröder), who applied explicit water MD simulation, shows the largest improvement in GDT-HA at 6.0%, but very modest improvement using SphereGrinder (1.1%), and a slight worsening using Cα RMSD (+0.2 Å). Deviation between these quality metrics stems from the difference in regions each group mostly refined; our model improved the significantly wrong part, while the model of group 288 improved the rest of the structure. We can expect that combining two approaches would have brought even more significant improvements to this target. For TR768 [Fig. 4(B)], *AbInitio* rebuilding successfully reconstructed two missing repeats in a LRR (leucine-rich repeat) protein. The reconstructed region turned out to interact with the remainder of the protein through bridging waters (red spheres in the figure), highlighting the importance of explicit waters that were not considered in our approach.

Overall, the results suggest that our idea for high-resolution refinement was at least partially successful in CASP11. Reconstruction of incorrect regions significantly improved SphereGrinder and Cα RMSD (but not the GDT-HA), supporting the idea that segment reconstruction can complement continuous MD or MC sampling methodologies. On the other hand, the core-refinement method we employed—based on implicit solvent simulations—did not perform as well as explicit water MD simulations used by other groups; a current research problem is whether the effect of explicit water molecules can be introduced into otherwise implicit solvent simulations to get the accuracy benefit of explicit water simulation with reduced computational cost.

Low-resolution targets: selection remains challenging

For the targets where we used a more aggressive low-resolution protocol, the overall quality of our Model 1 submissions was inferior by all the measures to those made using our high-resolution protocol, as well as to those of other groups. GDT-HA and SphereGrinder worsen by −3.9 and −1.1 on average, respectively (Table I), and the net *Z*-score for the 12 low-resolution targets is −11.6 (−0.97 per target) and +2.3 for GDT-HA and SphereGrinder, clearly worse than the +15.1 and +23.3 for the 24 high-resolution targets (Table I). This poor performance in Model 1 submissions is primarily due to poor model selection; for the best of the five submitted models, there was an improvement over starting the structures in GDT-HA by +2.0 and SphereGrinder by +4.3 on average. Failures in model selection have two causes. First, the score function used for selection is

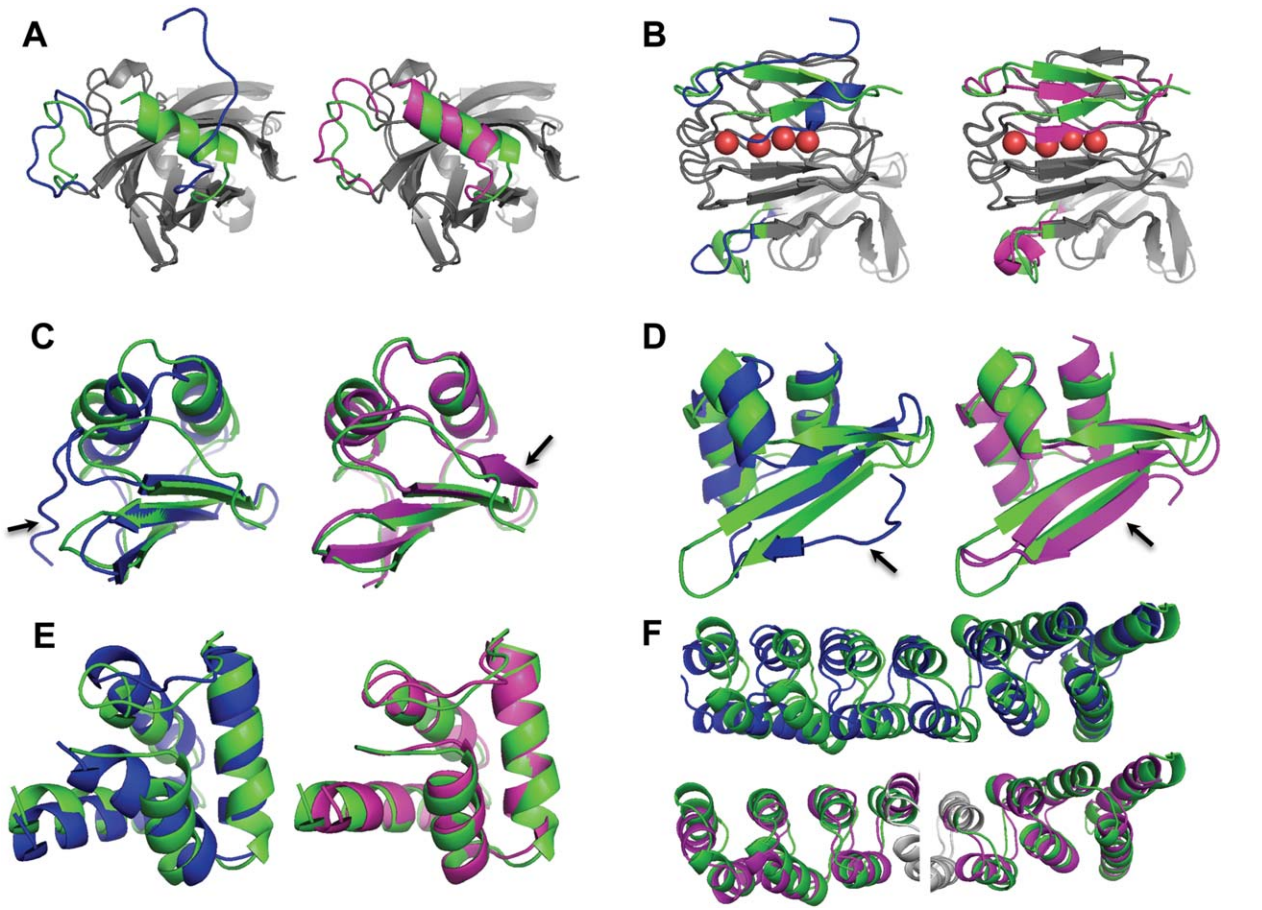

Figure 4

Examples of submitted models with large improvements. TR848, TR768, TR829, TR759, TR816, and TR827 (from A to F). In all panels, native, starting, and refined structures are colored in green, blue, and magenta, respectively. Regions most improved are shown by arrows. (**A** and **B**) Targets to which high-resolution strategy was applied; regions reconstructed are highlighted by colors (otherwise gray). (A) TR848_1 and (B) TR768_1 are improved by 5.1/14.1% and 4.7/5.6% in GDT-HA/SphereGrinder, respectively. In the reconstructed part of TR768_1, b-strands shift to occupy space filled by water molecules (red spheres) in the native structure. Refinement by fully reconstructing starting structures: (**C**) TR829_2, (**D**) TR759_3, and (**E**) TR816_2. Improvements in GDT-HA/SphereGrinder are (D) 25.8/43.3%, (E) 16.9/26.6%, and (**F**) 22.4/18.4%, respectively. Cα RMSD changes are (C) 6.2–1.2 Å, (D) 4.2–2.1 Å, and (E) 2.5–1.2 Å. (F) TR827_4 for which low-resolution strategy brought major improvements: 8/12% improvements in GDT-HA/SphereGrinder, respectively. Superpositions of model to the native structure over the two halves of the protein are shown.

clearly imperfect. Second, Model 1 for low-resolution targets was selected using a conservative strategy (see Methods). As an example, for TR816 and TR829, Model 2 had best score among the five models and also very significant improvement over the starting structure.

Significant improvements were obtained for targets for which large modifications were made to the starting structures: TR829 Model 2 [67 residues and $\text{GDT-HA}_{\text{model-to-starting}} = 45.9$, Fig. 4(C)] shows improvements in GDT-HA and SphereGrinder by 25.8% (from 51.1 to 76.9) and 43.3% (54.5 to 97.8), respectively, and in Cα RMSD from 6.2 to 1.2 Å. To our knowledge, this amount of improvement is the largest among all submissions not only in CASP11 (best GDT-HA improvement from others by +18.5, TR765 Model 2 by group 288 (Feig)) but also in all CASPs so far (best GDT-HA improvement by +19.3 in previous CASPs, TR462 Model 5 by group 470 (jacobson) in CASP8). The dramatic improvements in Cα RMSD and SphereGrinder come from the reconstruction of the whole structure including the N-terminus (arrow in the panel). Significant improvements were also obtained in other cases where total structure rebuilding was used: for example TR816 Model 2 [Fig. 4(D), 68 residues and $\text{GDT-HA}_{\text{model-to-starting}} = 52.6$] and TR759 Model 3 [Fig. 4(E), 62 residues and $\text{GDT-HA}_{\text{model-to-starting}} = 49.2$]. On the other hand, larger targets—for which the starting structure was only partially rebuilt—show less improvement. An exception is TR827, which could be refined through making hinge motions (such as normal mode motions) to the core region of starting structure [Fig. 4(F), 193 residues and $\text{GDT-HA}_{\text{model-to-starting}} = 41.3$].

Given our poor performance in selecting Model 1 during CASP11, we investigated after CASP11 whether an automated selection scheme could have done significantly better. In terms of average model quality, in retrospect automatic Model 1 selection using the normalized sum of Rosetta energy plus GOAP score would have considerably outperformed our manually selected Model 1 submission: this would have resulted in improvement in GDT-HA by +1.8 and summed *Z*-score in GDT-HA by +15.8 over all 36 targets, far better than our actual Model 1 submissions (Table I, average GDT-HA −0.6 and summed *Z*-score +3.5). The combination of the two metrics performs better than each individually; using only Rosetta energy or GOAP score for selection yields average GDT-HA improvement by +0.2 or 0.0 only, respectively, compared to +1.8 when combined. The improvement of the automated selection over human inspection mainly comes from successful model selection on a small number of difficult targets (TR829, TR816, and TR827) for which conservative Model 1 selection failed to pick models with dramatic improvements. Concluding that automatic model selection is better than human inspection based selection is likely over interpreting these results, but it is clear that conservative selection methods are not likely to perform well when the starting conformation is far from the native structure.

Although the CASP11 version of our low-resolution strategy was not very successful, the experiment suggests first, that our understanding of the low-resolution refinement problem is reasonable, and second, clear directions for improvement: extending the large-scale structure rebuilding approach used successfully for the three small targets (TR829, TR816, and TR759) to larger proteins and improving model selection.

DISCUSSION

In succeeding CASP refinement experiments, strict high-resolution model quality metrics (such as GDT-HA looking at only Model 1) have been used for assessing all targets in different difficulty ranges.[1,25,26] Due to the difficulty of improving GDT-HA for low-resolution targets conservative approaches became more attractive. Indeed, in CASP11 the overall submissions by the participants for low-resolution targets were very conservative (Fig. 2) and narrow in quality distributions (average standard deviation in GDT-HA values for Model 1 submissions for all groups is only 2.9). To spur progress in the field, it may be useful to use metrics for hard refinement problems that reflect correct topology more than higher resolution detail which is hard to achieve starting far from the native structure.

Overall, the CASP11 experiment revealed both strengths and weaknesses of our Rosetta based refinement protocols. The results of the experiment will be very useful in developing a next generation of refinement methods.

ACKNOWLEDGMENTS

We thank to Dr. Per Greisen, Sergey Ovchinnikov, David E. Kim, and Ray Wang in University of Washington for their technical helps during CASP. We also thank to Dr. TJ Brunette in University of Washington and Dr. Chaok Seok in Seoul National University for helpful discussions.

REFERENCES

1. Nugent T, Cozzetto D, Jones DT. Evaluation of predictions in the CASP10 model refinement category. Proteins 2014;82: 98–111.
2. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round X. Proteins 2014;82: 1-6
3. Mirjalili V, Feig M. Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. J Comput Chem Theory 2012;9:1294–1303.
4. Mirjalili V, Noyes K, Feig M. Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins 2014;82: 196–207.
5. Leaver-Fay A, Tyka MD, Lewis SM, Lange OF, Thopson J, Jacak R, Kaufman K, Renfrew PD, Smith Colin A, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YEA, Fleishman S, Corn J, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek J, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D Bradley P. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2014;487:545–574.
6. Park H, DiMaio F, Baker D. The origin of consistent protein structure refinement from structural averaging. Structure 2015;23:1123–1128.
7. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci 2014;23:47–55.
8. O'Meara MJ, Leaver-Fay A, Tyka M, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, Kuhlman B. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 2015;11: 609–622.
9. Eyal E, Yang LW, Bahar I. Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics 2006;22:2619–2627.
10. Song Y, DiMaio F, Wang RYR, Kim DE, Miles C, Brunette TJ, Thompson J, Baker D. High resolution comparative modeling with RosettaCM. Structure 2013;21: 1735–1742.
11. Stein A, Kortemme T. Improvements to robotics-inspired conformational sampling in rosetta. PLoS One 2013;8: e63090
12. DiMaio F, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits TC, Cheng Y, Baker D. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. Nature Methods 2015;12:361–365.
13. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101: 2043–2052.
14. 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.
15. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 2002;6:182–197.
16. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci. 2009;15:900–913.

17. Park H, Seok C. (2012;). Refinement of unreliable local regions in template-based protein models. Proteins, 80: 1974–1986.
18. Park H, Lee GR, Heo L, Seok C. Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. PLoS ONE 2014;9:e113811
19. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
20. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
21. Lukasiak P, Antczak M, Ratajczak T, Szachniuk M, Blazewicz J. Quality assessment methodologies in analysis of structural models. In: Proceedings of the 25th European conference on operational research, Vilnius, Lithuania, 8–11 July; 2012. p 80.
22. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82:7–13.
23. Chen VB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Laura W, Richardson, Jane S Richardson David C. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D 2010;66:12–21.
24. Khoury GA, Tamamis P, Pinnaduwage N, Smadbeck J, Kieslich CA, Floudas CA. Princeton_TIGRESS: protein geometry refinement using simulations and support vector machines. Proteins 2013;82: 794–814.
25. MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA. Assessment of the protein-structure refinement category in CASP8. Proteins 2009;77: 66–80.
26. MacCallum JL, Pérez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. Proteins 2011;79: 74–90.