WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

## Prediction Report

# Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home

**Rhiju Das,[1,2] Bin Qian,[1] Srivatsan Raman,[1,3] Robert Vernon,[1] James Thompson,[1,2] Philip Bradley,[1] Sagar Khare,[1] Michael D. Tyka,[1] Divya Bhat,[1,2] Dylan Chivian,[4] David E. Kim,[1] William H. Sheffler,[1,2] Lars Malmström,[1] Andrew M. Wollacott,[1] Chu Wang,[1] Ingemar Andre,[1] and David Baker[1,2,3]***

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

[2] Department of Genome Sciences, University of Washington, Seattle, Washington 98195

[3] Department of Biomolecular Structure and Design, University of Washington, Seattle, Washington 98195

[4] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720

*ABSTRACT*

*We describe predictions made using the Rosetta structure prediction methodology for both template-based modeling and free modeling categories in the Seventh Critical Assessment of Techniques for Protein Structure Prediction. For the first time, aggressive sampling and all-atom refinement could be carried out for the majority of targets, an advance enabled by the Rosetta@home distributed computing network. Template-based modeling predictions using an iterative refinement algorithm improved over the best existing templates for the majority of proteins with less than 200 residues. Free modeling methods gave near-atomic accuracy predictions for several targets under 100 residues from all secondary structure classes. These results indicate that refinement with an all-atom energy function, although computationally expensive, is a powerful method for obtaining accurate structure predictions.*

Proteins 2007; 69(Suppl 8):118–128.
© 2007 Wiley-Liss, Inc.

*Key words: protein structure prediction; fragment insertion; Rosetta; CASP; all-atom refinement; template-based modeling; free modeling.*

## INTRODUCTION

With over 100 domains from structural genomics initiatives and from experimental laboratories, the Seventh Critical Assessment of Techniques for Protein Structure Prediction (CASP7) provided an excellent test of the Rosetta comparative modeling and de novo structure prediction methods as well as the Rosetta all-atom refinement procedure. For all targets, prediction consisted of a search for the lowest energy structure according to the physically realistic Rosetta all-atom energy function, either starting with an extended chain, in the case of free modeling targets, or in the neighborhood of homologous structures, in the comparative modeling case. This large-scale test of computationally intensive all-atom refinement was made possible by the contributions of tens of thousands of individuals participating in the Rosetta@home distributed computing project.

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/
The authors state no conflict of interest.
Grant sponsor: National Institute of General Medical Sciences, National Institutes of Health; Grant number: P20 GM76222; Grant sponsor: Interdisciplinary Training in Genomic Sciences; Grant number: T32 HG00035; Grant sponsors: Howard Hughes Medical Institute, American Leukemia and Lymphoma Society
Rhiju Das and Bin Qian contributed equally to this work.
*Correspondence to: David Baker, Department of Biochemistry, University of Washington, J Wing, Health Sciences Building, Box 357350, Seattle, WA 98195. E-mail: dabaker@u.washington.edu
Received 28 February 2007; Revised 7 May 2007; Accepted 11 May 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21636

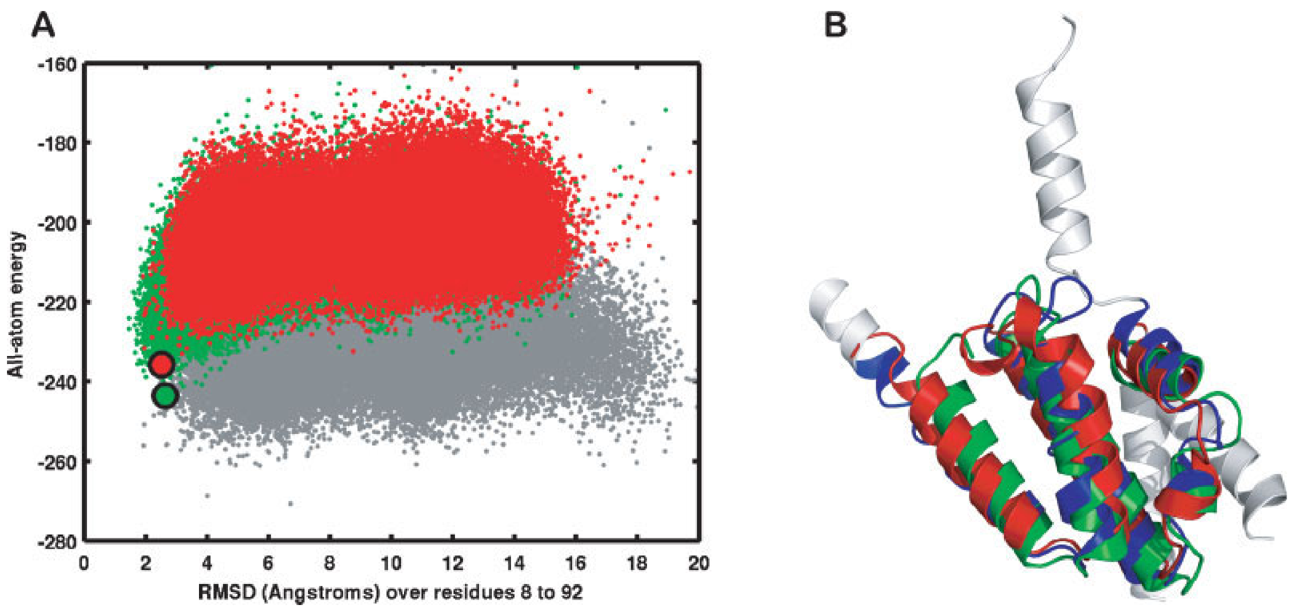

**Figure 1**

*Accurate predictions deriving from enhanced computational power, applied to folding of multiple homologous sequences and to all-atom refinement of hundreds of thousands of conformations.* ***A:*** *All-atom energy versus* $C_\alpha$ *root-mean-squared deviation (RMSD) from the native structure (over regions not making intermolecular contacts in the crystal) for ~60,000 conformations in each of three separate runs for target T0283. Free modeling runs for the target sequence T0283 (gray) did not converge, and the lowest energy conformations did not share the same fold. On the other hand, a homologous sequence (34% sequence identity to target; number 6 out of 7 sequences, drawn from a PFAM alignment[3]) gave outstanding convergence (green), with the 20 lowest energy conformations sharing the same fold (within 2 Å over most of the chain) of each other. Supporting this fold, another homologous sequence (red; 43% sequence identity to target) gave a lowest energy conformation (closed circle) with nearly identical structure.* ***B:*** *The consistency between these folds (red and green) anticipated the excellent agreement with the crystal structure (blue) after mapping the homologous sequences back to the target sequence [see Fig. 4(F)].*

The first success of such a computationally intensive, all-atom approach was in CASP6, in which our free modeling prediction for target T0281 reached an accuracy of better than 2 Å.[1] Following this result, promising in-house benchmarks indicated that one third of free modeling targets with lengths of up to 100 residues could be solved to better than 3 Å resolution with aggressive sampling[2] (see also Fig. 1 for an example from CASP7 target T0283). Further, with the all-atom refinement approach, our comparative modeling protocol appeared to improve upon the best available template in the majority of cases for targets under 200 residues, an advance over prior template-based methods. In this paper, we show that the results of CASP7 were largely consistent with these prior expectations. The in-depth analysis permitted by this large-scale rigorous experiment highlights the strengths of all-atom modeling and illuminates potential strategies for surmounting its current limitations.

# MATERIALS AND METHODS

## Rosetta@home

The distributed computing network Rosetta@home, based on the Berkeley Open Infrastructure Network Computing protocol,[4] went on-line in July, 2005; by the beginning of CASP7 in May, 2006, the network included 140,000 computers, with ~65,000 computers available for use at any given time, yielding an approximate performance of 37 TFlops. The CASP7 schedule allowed an average of ~500,000 CPU-hours to be devoted to predicting each domain.

## Choosing between template-based modeling and free modeling

Results from the BioInfo metaserver[5] were typically inspected 2 days after each target sequence was available. If potential domain parses were suggested by Ginzu and RosettaDom (run as part of the Robetta server[6]),[7] the separate domains were submitted to the 3D-Jury metaserver.[5] Eighty-nine domains with a 3D-Jury score[5] of over 50 were predicted by template-based modeling, and the remaining 32 domains were predicted by free modeling. A breakdown of the targets by our modeling approach is given in Supporting Information Table S1. For several border-line cases with 3D-Jury scores between 40 and 60, both types of modeling were carried out.

## The Rosetta all-atom energy function

The Rosetta all-atom energy function and refinement procedure has been described previously.[8,9] Briefly, the energy function includes terms for van der Waals interactions and for the free energy of solvation in the form of rapidly computed pair-wise approximations first used in molecular dynamics applications.[10,11] However, unlike force fields typically used in molecular dynamics, the free energy function utilizes an orientation-dependent hydrogen-bonding potential instead of a classical electrostatic description with atomic partial charges and backbone and side-chain torsional potentials derived from the Protein Data Bank. Further, bond lengths and angles are kept fixed at ideal values during all-atom refinement. Finally, conformational entropy is assumed to be similar for different compact states and not explicitly taken into account.

## All-atom energy based template selection

In contrast to previous CASP experiments in which we selected the best templates based on the PSI-BLAST or FFAS profile–profile matching score,[6] we carried out an all-atom energy based template selection protocol in CASP7. Starting from models based on up to 30 different templates and alignments obtained from the 3D-Jury server,[5] all-atom refinement[2,8] was carried out, and templates producing the very lowest energy models were identified and used for further modeling. Two to five templates were selected for further modeling for most targets. In cases where the energy differences between refined templates were small, as many as 10 templates were selected.

## Protocol for template-based modeling

Depending on the size of the target and the sequence identity to the template, we used three different template-based modeling strategies. The different size and sequence identity regimes and the method used in each regime are indicated in Figure 2(A) and described in the following paragraphs. The approaches used for each target are listed in Supporting Table S1.

The methods are presented in order of increasingly aggressive sampling. For targets with high sequence identity to the closest templates, less aggressive sampling was carried out, as the starting model was likely to be relatively close to the template. Less aggressive sampling was also used for larger targets for which the greatly increased

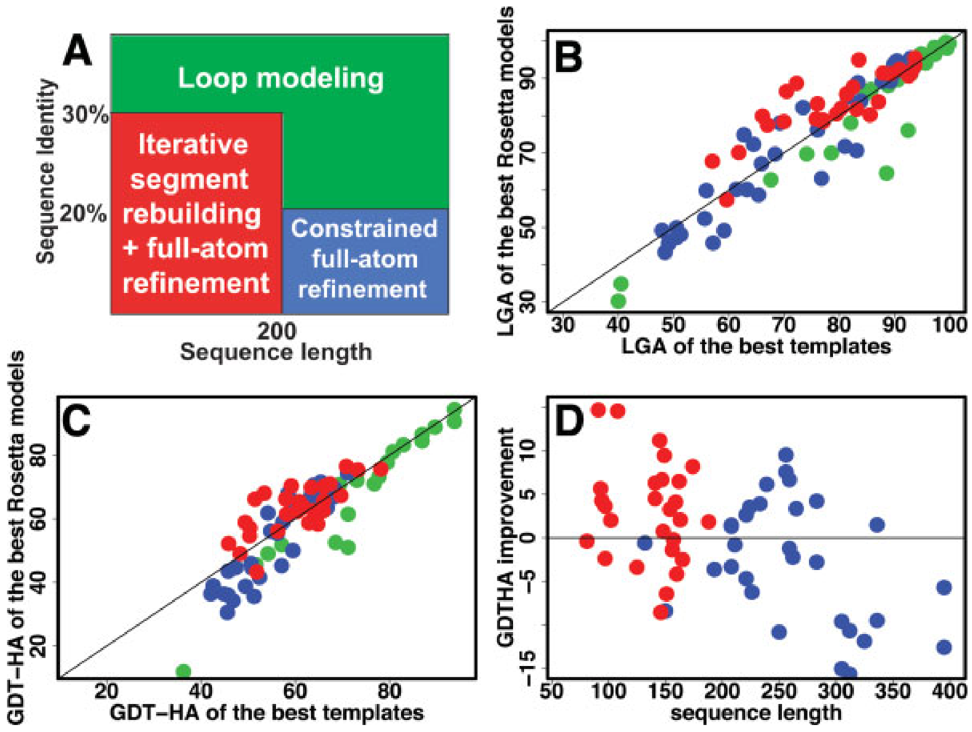

**Figure 2**

*Summary of comparative modeling results.* ***A:*** *Categorization of targets by length and template sequence identity used to determine the modeling protocol.* ***B, C:*** *LGA (Local/Global Alignment)[12] (B) and GDT-HA[12] in structurally alignable regions (GDT-HA) (C) of the best of our submitted models versus those of the best templates in PDB.* ***D:*** *GDT-HA improvement in structurally alignable regions correlates with the target size. The data points in Panels B, C, and D are colored according to the color scheme in Panel A. The structurally alignable regions were defined by comparing the native structure to the best template with the 3DPAIR structure alignment program.[13]*

size of the search space considerably reduced the probability of locating native-like folds at high resolution. We thus chose to focus extensive sampling on proteins less than 200 amino acids as the most effective way of allocating the available computing power.

*Class I. Loop modeling: targets with sequence identity to the closest template greater than 30% and targets longer than 200 residues with 20–30% sequence identity to the closest template.* The target sequence was threaded onto the best template backbone, and regions containing insertions or deletions as well as regions with low local sequence similarity relative to the template were built using an improved version of the Rosetta loop modeling protocol,[14] which incorporates loop closure by cyclic coordinate descent[15] followed by gradient based energy minimization. Side-chains were modeled using a combinatorial search through an extended version of the Dunbrack backbone-dependent rotamer library supplemented with side-chain conformations from the template using Monte Carlo sampling.[16] Regions with atomic clashes after repacking were identified, and gradient based minimization of the Rosetta all-atom energy was used to refine their backbone and side-chain torsion angles.

*Class II. Loop modeling with constrained all-atom refinement: targets longer than 200 residues with template sequence identity below 20%.* Following the application of protocol (I) to many alternative alignments to each of the selected templates,[17] full-chain, all-atom refinement was carried out using template-based consensus $C_\alpha$–$C_\alpha$ distance constraints derived from the 3D-Jury template-based models with the lowest Rosetta all-atom energies; quadratic penalties were imposed for distances greater than the largest distance observed in the low energy templates by more than one standard deviation of the $C_\alpha$–$C_\alpha$ distance distribution.[18]

*Class III. Iterative segment rebuilding and all-atom refinement: targets shorter than 200 residues with template sequence identity below 30%.* Protocol (I) was carried out for each target, and the structurally diverse regions in low-energy models were identified and rebuilt, followed by all-atom refinement of the entire model without $C_\alpha$–$C_\alpha$ distance constraints. This process was repeated for 10 iterations, selecting a diverse subset of the lowest energy models at each stage for input to the next round of refinement.

### Free modeling (Classes IV and V)

The remaining targets were predicted by free modeling and fall into Classes IV or V, depending on whether assessors later classified them as template-based modeling or free modeling targets. The protocol for free modeling consisted of the low resolution Rosetta de novo structure prediction method followed by all-atom refinement. As in CASP6, the first step was the generation of a large pool of conformations by Rosetta fragment assembly,[19] guided by a low resolution energy function that favors hydrophobic burial and β-strand pairing. Fragments were picked to match secondary structure predictions from PSI-PRED,[20] JUFO,[21] SAM,[22] and PROF.[23] We again attempted to ensure diversity in this set of conformations by folding multiple homologs for each sequence (see, e.g., Fig. 1) by forcing the exploration of different secondary structures through manually imposed torsional “bar-codes” (RD, PB, DK, D. Baker, unpublished results), and by seeding simulations with long-range β sheet pairings.[24] Further, a new term was introduced to reproduce pair-wise distance correlations between side-chain interaction centers (Supporting Figure S1A). Finally, the number of fragment insertions was increased by 10-fold, leading to better annealing of β strands; concomitantly, terms involving the burial of loops and helices were down-weighted to prevent over-convergence into the false minima of the low-resolution energy function, which can be quite inaccurate (Supporting Figure S1B).

The second step of the free modeling protocol was the same full-chain, all-atom refinement procedure used in template-based modeling, described above. A similar two-step protocol was carried out previously for CASP6 target T0281[1] and a recently published benchmark.[2] In CASP7, however, all conformations were all-atom refined, rather than just cluster centers. Further, predictions for sequence homologs were mapped back to the target sequence (using the loop modeling plus all-atom refinement protocol described above) after all-atom-refinement with the homologous sequence, rather than before the refinement.[1] Modifications to the free modeling protocol to enforce symmetry or to resample folds near promising conformations are described in the text.

### Choice of submitted predictions

Submitted predictions were drawn from the lowest all-atom energy conformations (typically the best 100–1000 out of $10^5$–$10^6$ conformations), with clustering and human judgment (see below) used to choose a final set of five reasonably diverse submissions. For some submissions, additional packing (see Supporting Figure S2) or exposed surface area[25] (see below) metrics were used to filter the low energy set.

## RESULTS

### Different levels of success for different target classes

We present our results in the CASP7 experiment grouped by the approaches used to make each prediction. For the three template-based modeling approaches, the choice of method was based on the length of the target and the sequence similarity to the closest template: the more distant the sequence similarity to potential templates and the shorter the length of the protein, the more aggressive the all-atom refinement. For free modeling targets, all-atom refinement was carried out for all predictions.

A reviewer of this paper questioned the utility of computationally intensive all-atom refinement for improving protein structure prediction by pointing out that the excellent predictions of the Zhang group were made using a reduced model (with a $C_\alpha$ atom and an interaction center per residue) yet had similar quality, as assessed by Global-Distance Test (GDT)[12] scores. To address this important question, for each group of targets we compare the GDT-HA (High Accuracy) *Z*-score distributions for the best of our five submitted models to those of the best of the submitted models for each of five other representative top groups. The following analysis of our successes and failures in each of these classes gives insight into which classes benefit the most and which benefit least from all-atom refinement.

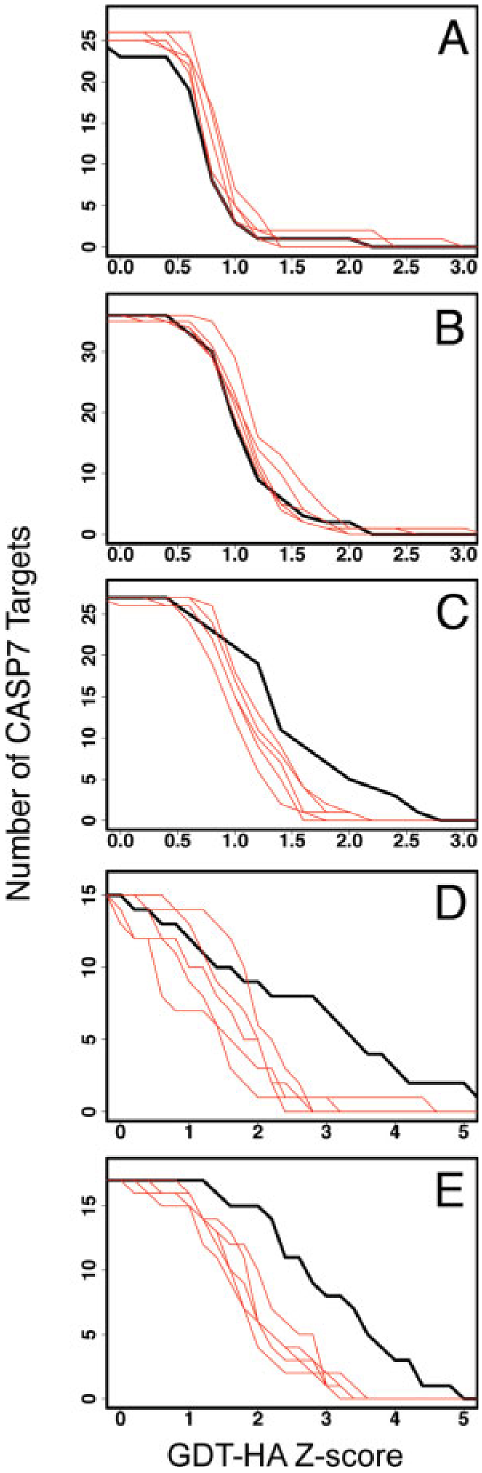

**Figure 3**

*Comparison of Z-scores (based on GDT-HA) of Rosetta models to Z-scores of the top CASP7 groups (as selected by assessors) highlights limitations and successes of the Rosetta all-atom refinement method for different categories of targets.* ***A:*** *Class I, proteins with high sequence-identity to existing templates, predicted by loop-modeling.* ***B:*** *Class II, proteins with low-sequence-identity (<20%) templates longer than 200 residues, predicted by loop-modeling and ?all-atom refinement.* ***C:*** *Class III, proteins with low-sequence-identity (<30%) templates less than 200 residues, predicted by iterative segment rebuilding coupled to all-atom refinement.* ***D:*** *Class IV, proteins predicted by free modeling that were later classified as template-based modeling targets.* ***E:*** *Class V, free modeling targets.*

### *Target Class I. Loop modeling*

The first class of targets are those modeled without all atom refinement of the entire protein chain. This set consists of targets with sequence identity to the closest template greater than 30% or sequence identity of 20–30% and length greater than 200 residues. As described in the methods, these targets were predicted using loop modeling and side-chain refinement but not full-chain all-atom refinement.

The GDT-HA *Z*-score distribution for our predictions for this class of targets are clearly worse than those of other top performing groups [Fig. 3(A)], and the differences are exacerbated if only the first submitted model is considered (not shown). In addition, our models are usually worse than the best templates as indicated by scores such as the Local/Global Alignment score [LGA; Fig. 2(B)] and GDT-HA over aligned regions [Fig. 2(C)]. On one hand, this poor performance partly reflects failure to select the closest templates as well to utilize evolutionary information available from multiple templates for loops and for core elements, as was innovatively carried out by other groups. On the other hand, the remodeled loops were quite accurate in a number of cases (Supporting Table S2); an example (T0315 residues 142–149) is shown Figure 4(A).

### *Target Class II. Loop modeling with constrained all-atom refinement*

The second class of targets consists of those proteins modeled with a single round of constrained all-atom refinement. This set consists of targets with sequence

DOI 10.1002/prot

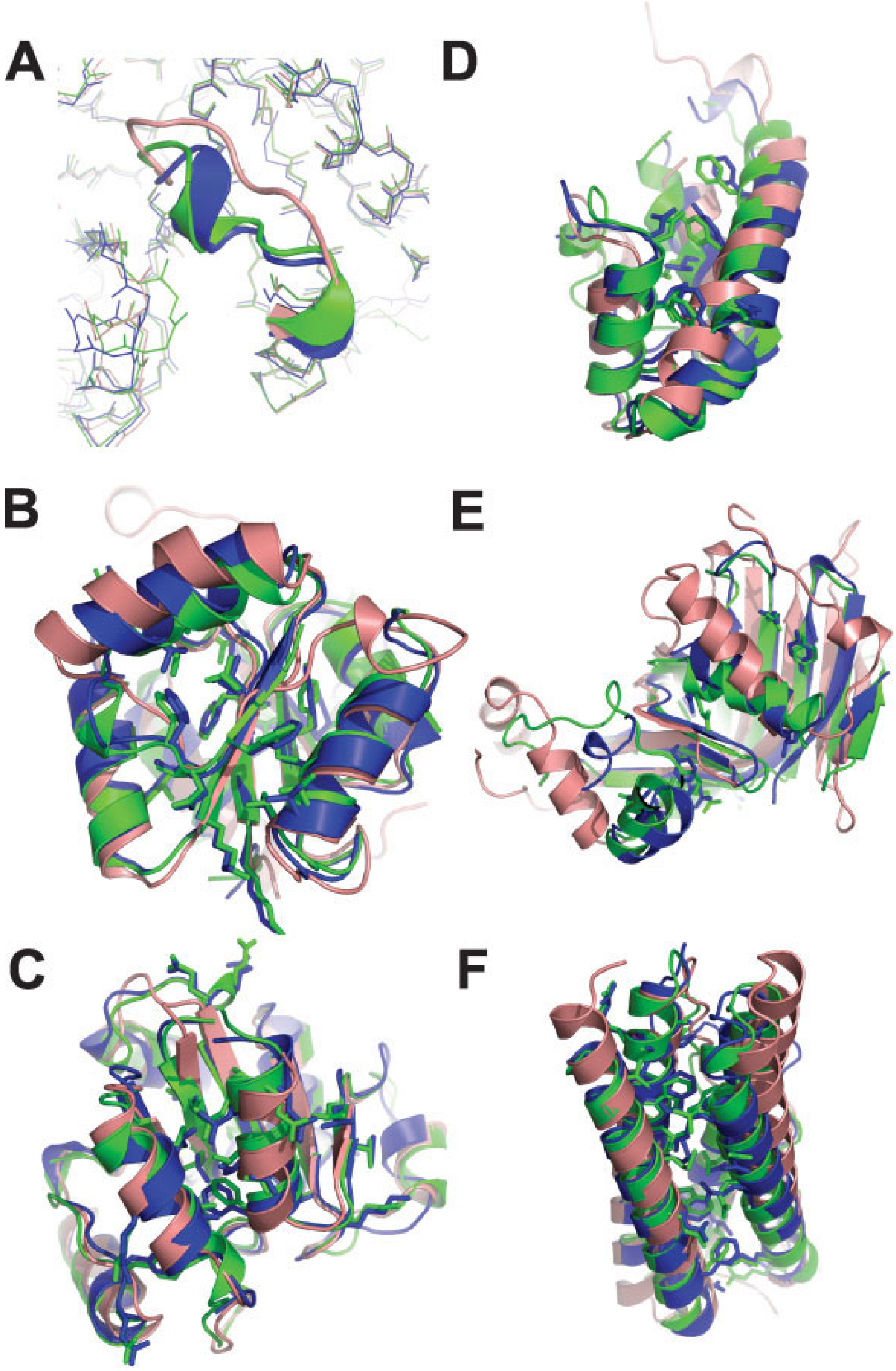

**Figure 4**

*Examples of successful template-based predictions. For each target, the crystal structure is shown in blue, the best of our submitted models in green, and the best template in pink. Selected core side-chains are shown as sticks in blue (crystal structure) or green (prediction).* ***A****: Loop modeling of target T0315 (residues 142–149 shown as cartoon),* ***B****: T0341 domain 1,* ***C****: T0329 domain 1,* ***D****: T0330 domain 2,* ***E****: T0331,* ***F****: T0385.*

identity to the template below 20% and lengths greater than 200 amino acids. As described in the methods, during the prediction season we considered it unlikely that we would be able to sample sufficiently well to locate the global minimum for longer proteins, and hence invested less computational effort into these targets.

The GDT-HA *Z*-score distributions for this class of targets are also generally worse than those of other top groups [Fig. 3(B)]. In addition, our models are worse than the best templates in most cases as indicated by LGA [Fig. 2(B)] and GDT-HA [Fig. 2(C)] scores. Aggressive sampling is necessary to produce significant improvements over the starting template, but if sampling is insufficient or the energy function is in error, refinement can degrade rather than improve the starting model (see also below, "What went wrong?"). Two predictions in which the constrained all-atom refinement protocol improved upon the templates are illustrated in Figure 4(B,C). These predictions display increased accuracy in backbone conformations relative to the best templates as well as accurately packed core side-chains.

### Target Class III. Iterative segment rebuilding and all-atom refinement

The third class of targets are those modeled using iterative segment rebuilding coupled with all-atom refinement. This set consists of proteins of less than 200 amino acids for which an evolutionarily related protein of known structure was clearly detectable, but with less than 30% sequence identity. During the prediction season we considered this class of targets to be the most likely to benefit from aggressive refinement as the sequence dissimilarity from the template suggested structural changes, and the relatively small size made the search problem more tractable.

The more extensive sampling coupled with the smaller size of the search space made the all-atom refinement significantly more successful for this class of targets [Fig. 3(C)] than for Class II targets [Fig. 3(B)]: the all-atom refinement produced many more models with *Z*-scores greater than 1.5 than did methods not using all-atom refinement. Among the 27 domains for which we carried out this protocol, 21 had better LGA scores than the best template [Fig. 2(B)] and 18 had better GDT-HA scores over the aligned regions than the best template [Fig. 2(C)]. Several targets (T0330_D2, T0331, T0380, T0368, T0357) achieved the largest improvements among all the CASP7 submissions; most of the remaining targets showed improvements in some regions of the structure. Examples of these improvements are illustrated in the superpositions of the crystal structure, the best template, and our best predictions in Figure 4(D–F). T0330 domain 2 [Fig. 4(D)] is noteworthy because the model is better than the best template throughout the entire structure, and the core side-chain packing is very similar to that observed in the native structure. In T0331 [Fig. 4(E)], the outer helix and long hairpin loops inherited from the best template (1ty9.pdb) were rebuilt and refined towards the native conformation successfully. In T0385 [Fig. 4(F)], a long helix reoriented during refinement to achieve better packing in the core of the four-helix bundle, making the model significantly better than the best template.

### Target Classes IV and V. Free modeling

We predicted the remaining 32 models in CASP7 by free modeling. We have subdivided these targets into Classes IV and V, depending on whether assessors later classified them as template-based modeling or free modeling problems, respectively. For the 15 domains in Class IV, the fact that the potential templates were actually identified by few or no predictors and the relatively high accuracy of our predictions [Fig. 3(D)] indicated that free modeling was a reasonable choice. The final class of targets (Class V) were later classified as having no structurally similar templates, and we modeled these 17 domains without exception using our free modeling methodology. As with the Class IV targets, many of the models were quite good in a relative sense [Fig. 3(E)].

Unlike previous CASP experiments, there were very few cases in which human intuition or additional information could be applied during the free modeling predictions. Human intuition played a major role in only one target, T0299. Inspection of the secondary structure prediction and domain boundary predictions suggested an approximately symmetric conformation with two ferredoxin-like folds (cf. T0272 in CASP6[1]); assembling models of the two domains[26] led to a prediction with an accuracy of 5.1 Å over 180 residues [Fig. 5(A)].

Additional information from external sources was useful in only two further cases, T0300 and T0319. The target description for T0300 indicated that the protein was a dimer, and the sequence gave a strong signal for a parallel coiled-coil at the N-terminus.[27,28] After enforcing an asparagine–asparagine pairing geometry,[29] fragment insertions were carried out symmetrically on each of two 102-residue chains, followed by filtering for intermolecular pairing of each molecule's lone β-strand, and then symmetric all-atom refinement. The agreement of the resulting models to the crystal structure did not reach atomic accuracy, but was nonetheless excellent, given that the protein crystallized not as the physiological dimer but as a tetramer [Fig. 5(B)]. In the case of T0319, the homology of the N-terminal and C-terminal β-stranded regions to a zinc-binding domain was previously hypothesized in the literature by Bujnicki and coworkers[30] based on a sensitive fold recognition study. After a long-range β-strand pairing was enforced[24] to maintain the zinc binding site, all-atom refinement led to excellent predictions for both the β domain and the "inserted" α-helical domain [Fig. 5(C)].

Because of the scarcity of additional information on the targets, CASP7 provided an excellent test of automatic structure prediction with all-atom refinement. Aggressive sampling with the all-atom energy function led to several outstanding predictions in CASP7 for pro-

DOI 10.1002/prot

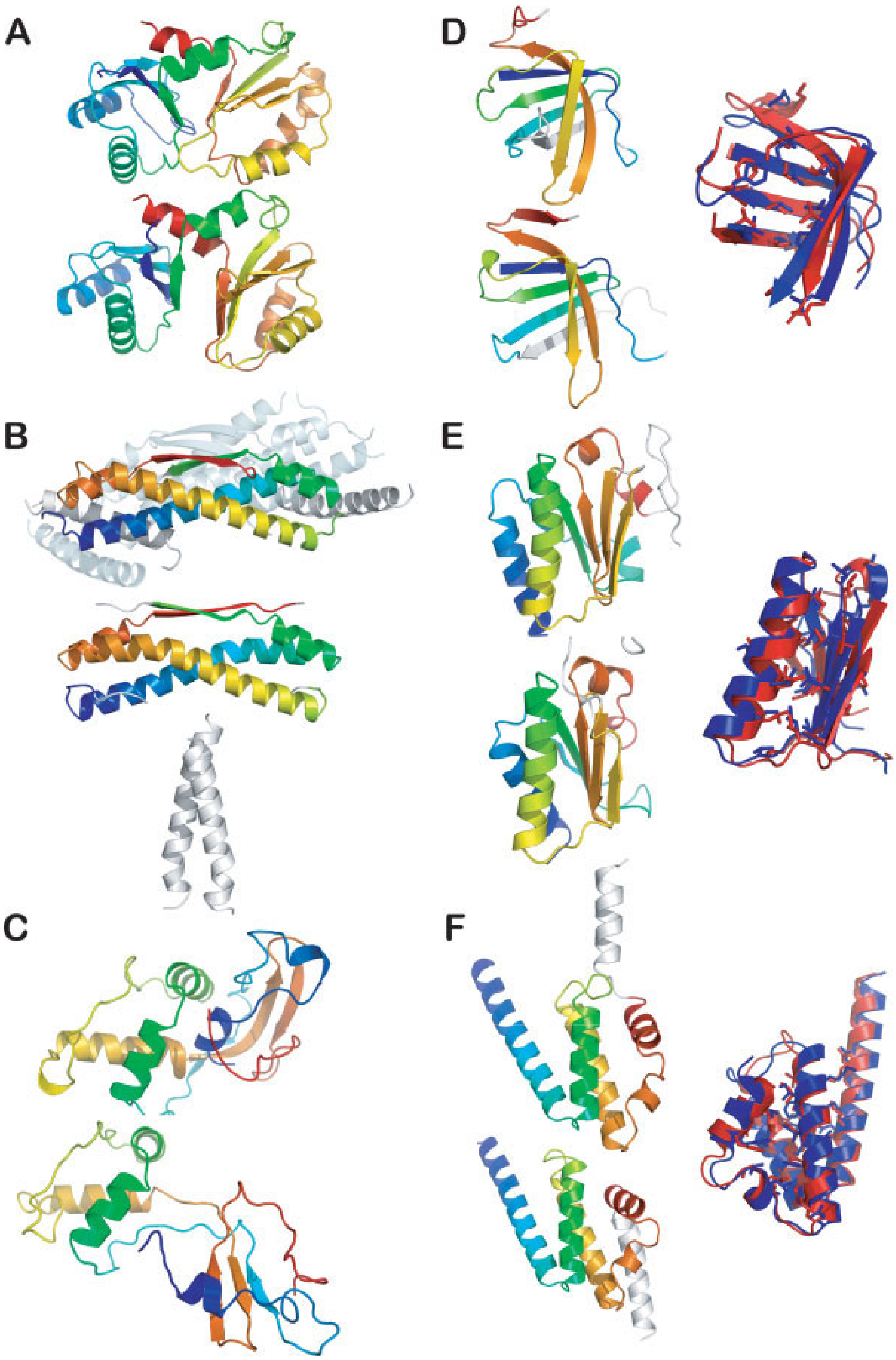

**Figure 5**

*Examples of successful free modeling predictions. For each panel, the crystal structure of a CASP7 target is shown at the top of the panel, and the best of our five submitted models at the bottom. Panels* ***D–F*** *display additional superimpositions of the native state (blue) on the model (red) over regions that align to better than 2 Å and core side-chains shown as sticks.* ***A****: T0299; rainbow coloring highlights 180 residues of Model 1 that agree with the crystal structure within 5.1 Å.* ***B****: T0300; rainbow coloring highlights agreement of 111 residues of dimer model 5 with the native structure within 3 Å. The additional two molecules that form a tetramer in the crystal structure via coiled-coils are shown as transparent, cyan cartoons.* ***C****: T0319; model 1 gave agreement of 4.0 Å over 67 residues, in the* α*-helical domain. D: T0316 domain 3; model 1 gave 72 residues aligned within 2.9 Å. E: T0354; model 2 gave 90 residues aligned within 3.3 Å. F: T0283; model 3 gave 90 residues aligned within 1.4 Å. Note that the targets shown in Panels A and D–F were classified as template-based-modeling problems by assessors, but our predictions were based on free modeling (see text).*

teins with well-defined secondary structure. The lowest all-atom energy conformations for the small protein targets T0335 and T0350 converged quite well; compared with the NMR-derived structures, the resulting models were accurate to within 3 Å over most of the structure. Some submissions (e.g., T0300, T0319, T0304, and T0356 domain 1), while not reaching such high resolution, agreed well with the crystal structures over subdomains [see, e.g., Fig. 5(B,C), and Supporting Figure S3]. For T0304, in particular, a new protocol carrying out conformational space annealing with backbone constraints allowed efficient resampling of promising conformations, leading to lower energy models and better predictions (Supporting Figure S3). Overall, the use of all-atom refinement appears to have been the major reason for the dominance of Rosetta@home predictions over models from the Robetta server, which carried out an identical fragment insertion protocol but not the subsequent all-atom refinement.

The most striking cases in which the high-resolution free modeling methodology outshone the standard low-resolution fragment insertion procedure are shown in Figure 5(D–F). The prediction for all-β domain 3 of T0316 [2.9 Å over 72 residues; Fig. 5(D)] gave the correct global strand arrangement and nearly perfect strand register. The model for the α + β protein T0354 [1.8 Å over 77 residues; Fig. 5(E)] displays correct strand register in the β sheet and accurate helix-sheet packing. Finally, the model for all-α protein T0283 [1.4 Å over 90 residues; Fig. 5(F)] correctly placed all the helices except for one making contact with a crystal neighbor. In each of these cases, core side-chains in our best predictions superimpose well with the native structure (Fig. 5). Interestingly, despite being classified as template-based modeling problems by assessors, we modeled these three cases starting with extended chains. These results indicate that free modeling with all-atom refinement is capable of generating accurate blind predictions for proteins of each secondary structure class.

## WHAT WENT WRONG?

The performance of Rosetta in CASP7 offers informative lessons about the power and the limitations of all-atom refinement in structure prediction. In particular, our experiences exposed the necessity of using information beyond an all-atom energy function for accurate predictions; incorrect assumptions of energy-based refinement; structure expansion due to side-chain modeling; and factors that make conformational space too large to achieve reasonable all-atom sampling.

### Ignoring evolutionary information

For target Classes I and II, we failed to make use of evolutionary information present in multiple templates, and the resulting models were clearly inferior to those of many other groups. Building models based on a single template is clearly not an optimal approach in this era in which large amounts of relevant evolutionary information in the form of related structures frequently may be available.

Further, among Class IV targets, there were several targets for which excellent templates were available but for which the free modeling protocol was applied (e.g., T0285, T0306, T0347_D1, T0349, and T0356_D2). The resulting predictions were significantly less accurate than the best templates in the Protein Data Bank. Indeed, during the prediction season, template-based models gave lower Rosetta all-atom energies and turned out to be more accurate than the free modeling predictions in the few cases where both methodologies were used for submissions (e.g., T0363, T0373, and T0383). Use of consensus templates from the CASP7 server predictions would have easily ameliorated our ignorance of good templates for most of these cases.

### Incorrect assumptions of energy-based refinement

The fundamental assumption of energy-based all-atom refinement is that the native state is the global free energy minimum of the modeled system. For nearly all CASP7 targets, the system was assumed to be a monomer. Thus it is not surprising to observe poor accuracy for segments of the proteins contacting binding partners or neighbors in crystals (e.g., T0312, T0363, T0368 in template-based modeling; T0300, T0309 in free modeling). For long extended regions that lack interactions with the rest of the monomeric protein, it is essentially impossible for the physical energy guided process to work well. In lieu of simulating full oligomeric complexes, protein structural family derived scoring functions like those pioneered by Zhang *et al.*,[31] may complement the Rosetta all-atom physics-based terms and allow better modeling of binding interfaces, particularly for template-based problems. For free modeling, prediction of native structure will likely be difficult unless accurate information regarding the number of interacting monomers and potential interactions are given as input to the modeling.

### Structure expansion due to side-chain modeling

During the prediction season, we noticed that all-atom side-chain modeling tended to produce an expansion of template-based modeling predictions, causing a general degradation of models unless extensive sampling led to side-chain conformers that could be well-packed. For example, such expansion of the starting template occurred for the earlier targets in template-based modeling Class II, T0289, T0293, and T0298. After this problem was noticed, we employed $C_\alpha$–$C_\alpha$ constraints for subsequent targets during the initial all-atom refinement, but as shown in the comparison in Figure 3(B), the size of the search space made degradation of models a more likely outcome than improvement.

Side-chain-induced expansion was also observed in free modeling targets, especially in all-α-helical targets, such as T0307, T0382, and T0361 (Supporting Figure S4). Upon noticing this expansion during the prediction season, the lowest energy conformations for these targets were filtered by a surface-area-based solvation score[25] for a subset of the predictions. These conformations were more compact than the best scoring predictions based on the Rosetta all-atom energy alone and, correspondingly, turned out to be the best submissions from our group. These submissions were typically more accurate than other groups by high resolution criteria but worse by low resolution criteria (see Supporting Figure S4 for an example from T0382), illustrating both the benefits and pitfalls of all-atom refinement. Increasing the flexibility of bond lengths and angles as well as directly minimizing the volume of large voids inside the protein core (see Supporting Information Figure S2) are being explored as potential ways to counteract model expansion during all-atom refinement.

### The limits of conformational sampling

Even with a distributed computing network of many thousand processors, there are numerous cases in which reasonable sampling of the all-atom energy function cannot be achieved. Conformational space is larger for proteins of longer lengths as well as for proteins that have less certain secondary structure; both issues affected our prediction efforts.

While little can be done to restrain the conformational space for large free modeling targets (e.g., T0287, T0296, T0356), template-based modeling for large targets (Class II) can be aided by taking advantage of information from multiple templates. In this class of targets, the excellent models of the Zhang group (top line in Figure 3B) particularly stand out. Currently, evolutionary information is clearly more powerful than detailed physical chemistry for larger proteins as the all-atom sampling problem becomes exceptionally difficult with the vastly increased size of the search space. Further supporting this conclusion, there is a rough boundary in the success of our refinement approach at sequence length 200 [Fig. 2(D)]: below this length, we frequently improved over the template, while above this length, we had more failures than successes.

For free modeling targets in CASP7, uncertain secondary structure prediction was a further confounding factor that greatly increased the size of the conformational space that needed to be sampled. A general paucity of sequence homologs may be correlated with this uncertainty in secondary structure. For example, T0285 gave no sequence hits upon a BLAST or PSI-BLAST search; the target sequence itself was not in the database. The resulting inadequate sequence profile contributed to an incorrect secondary structure prediction, which made confident identification of an existing template and convergence of the free modeling protocol difficult. Other free modeling cases, drawn from the structural genomics initiatives (targets T0304, T0314, T0320, T0349, T0353, and T0386), contained regions where different secondary structure methods were either uncertain or conflicting. For these targets we carried out different batches of simulations forcing different secondary structures (see Supporting Figure S5 for models from T0353). However, this strategy appeared to “over-diversify” the search. Native topologies were not sampled closely enough to appear in the set of low energy conformations (Supporting Figure S5C). At the present level of sampling, highly accurate free modeling appears to require that for each region of the sequence at least two of the four input secondary structure predictions are reasonably correct. With the accuracy of secondary structure predictions approaching 80%,[20] this condition is met for most proteins but not this subset of CASP7 targets. In the future, more computational power coupled with extensive sampling of alternative secondary structures may allow all-atom refinement to use signals from favorable tertiary interactions[32] to surmount a prior lack of information on secondary structure.

## CONCLUSIONS

All-atom refinement using extensive sampling with Rosetta@home contributed to producing excellent models for many, but certainly not all, targets of less than 200 amino acids to which it was applied (Classes III–V described above). In template-based modeling efforts, aggressive refinement with the Rosetta all-atom energy function led to improvement of at least one of the submitted models over the best available template for the majority of proteins under 200 residues. In free modeling efforts, large amounts of sampling led to several high resolution predictions, in some cases reaching accuracies better than 2 Å. An additional benefit of all-atom refinement is that it automatically generates physically realistic features such as hydrogen bonds with native-like geometries.

Several observations strongly suggest that more efficient sampling will lead to more consistent success in structure prediction. In the cases where all-atom refinement could not reach the native structure, our methodology did not typically converge. Further, in these cases, native crystal structures generally appear lower in energy than our models in the absence of extensive contacts with crystal neighbors. In principle, each of the confounding factors described above—loss of compaction, secondary structure uncertainty, even oligomeric systems—could be overcome if a more effective strategy could be found to optimize the rugged landscape produced by the all-atom energy function. More efficient sampling strategies would also reduce the computational effort required for each prediction. Such strategies are being pursued and will see their most rigorous test in CASP8.

## ACKNOWLEDGMENTS

We thank participants in Rosetta@home for their contributions of computing power, which made it possible to use all-atom refinement to predict the CASP7 structures in the limited time available. Predictors who identified low-energy models are identified in the Supplementary Material. We thank the CASP organizers and contributing structural biologists for providing the targets. We are grateful to Jim Havranek for advice on coiled coils. We thank Keith Laidig for developing the computational infrastructure underlying Rosetta@home; the Department of Energy INCITE program for access to Blue Gene/L at Argonne National Lab and the IBM Blue Gene Watson supercomputers; and the San Diego Supercomputer Center for supercomputing time.

## REFERENCES

1. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. Proteins 2005;61(Suppl 7):128–134.
2. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
3. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34(Database issue):D247–D251.
4. Anderson DP. BOINC: A system for public-resource computing and storage, In 5th IEEE/ACM International Workshop on Grid Computing, Pittsburgh, USA, November 8, 2004.
5. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
6. Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated Robetta protocols. Proteins 2005;61(Suppl 7):157–166.
7. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 2005;61(Suppl 7):193–200.
8. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
9. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.
10. Brooks BR, Bruccoleri RE, Olafson BD, States JD, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minmimization, and dynamics calculations. J Comp Chem 1983;4:187–217.
11. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
12. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
13. Plewczynski D, Pas J, Von Grotthuss M, Rychlewski L. Comparison of proteins based on segments structural similarity. Acta Biochim Pol 2004;51:161–172.
14. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004; 55:656–677.
15. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 2003;12:963–972.
16. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. Protein Sci 2005;14:1328–1339.
17. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 2006;34:e112.
18. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
19. 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.
20. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
21. Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;7:360–369.
22. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001;17:713–720.
23. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162–1176.
24. Bradley P, Baker D. Improved β-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 2006;65:922–929.
25. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
26. Wollacott AM, Zanghellini A, Murphy P, Baker D. Prediction of structures of multidomain proteins from structures of the individual domains. Protein Sci 2007;16:165–175.
27. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci USA 1995;92:8259–8263.
28. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
29. O'Shea EK, Klemm JD, Kim PS, Alber T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 1991;254:539–544.
30. Purushothaman SK, Bujnicki JM, Grosjean H, Lapeyre B. Trm11p and Trm112p are both required for the formation of 2-methylguanosine at position 10 in yeast tRNA. Mol Cell Biol 2005;25:4359–4370.
31. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61(Suppl 7):91–98.
32. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA 2003;100:12105–12110.

DOI 10.1002/prot