Free Modeling with Rosetta in CASP6

Philip Bradley,[1]† Lars Malmström,[1]† Bin Qian,[1]† Jack Schonbrun,[1]† Dylan Chivian,[1] David E. Kim,[1] Jens Meiler,[2] Kira M.S. Misura,[1] and David Baker[1]*

[1]*University of Washington, Seattle, Washington*
[2]*Vanderbilt University, Nashville, Tennesee*

ABSTRACT **We describe Rosetta predictions in the Sixth Community-Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP), focusing on the free modeling category. Methods developed since CASP5 are described, and their application to selected targets is discussed. Highlights include improved performance on larger proteins (100–200 residues) and the prediction of a 70-residue alpha–beta protein to near-atomic resolution. Proteins 2005;Suppl 7:128–134.**

Key words: protein structure prediction;fragment insertion; Rosetta; CASP; full-atom refinement

INTRODUCTION

We submitted coordinate predictions for all targets in CASP6. In this article, we focus on models built without use of template coordinates—the *free modeling* prediction category. Our general approach was to build maximally diverse populations of models and rely on energy functions and clustering, together with human intervention where appropriate, to select native-like models. For several targets we incorporated information from fold recognition (FR) servers in the form of consensus beta-strand pairings. Highlights from the experiment include: successful prediction of larger all-beta (T0212) and all-alpha (T0198) proteins; refinement to near-atomic resolution of a small protein (T0281); and the successful application of the Rosetta modeling tools by other prediction groups. Here we discuss the methods that we used for free modeling in CASP6—focusing on methods developed since CASP5—and we describe several of the most interesting predictions in detail.

MATERIALS AND METHODS

The 3D-Jury Metaserver[1,2] was used for initial target classification. Targets for which neither the target nor any sequence homologs had 3D-Jury A1 scores above ~50 were modeled de novo. Domains were initially assigned by manual inspection of results from Ginzu[3] and RosettaDom (described elsewhere in this issue); in difficult cases several alternative parses were used for modeling. Sequence homologs were identified in Pfam[4] and by PSI-BLAST[5] searches against the NCBI's nonredundant protein sequence database. The target sequence and 5 to 25 sequence homologs were modeled with the Rosetta de novo protocol[6] (2000–10,000 models each). The Rosetta models were clustered and the topologies for the top 10 cluster centers were inspected visually. Secondary structure predictions (PSIPRED,[7] Sam-T99,[8] and JUFO[9]) for the target and each homolog, as well as the secondary structure content of each decoy population, were compared to define a consensus where possible and identify overconvergence in the models (e.g., loss of weakly predicted secondary structure elements). In some cases we built a second round of models with fragment sets biased toward underrepresented secondary structures to ensure a diverse population of models. For beta-sheet proteins we analyzed beta-sheet topologies graphically (see T0201 discussion) after filtering for models in which all consensus strands were paired.

We have recently developed a protocol for efficiently generating models that satisfy one or more residue-pair orientational constraints (manuscript in preparation). In this approach, the paired residues are kept in the desired relative orientation throughout folding, with breaks inserted in the peptide chain to allow fragment insertions as in the standard de novo protocol. The protein is represented as a tree (acyclic graph) composed of peptide segments together with long-range connections (a very similar tree representation is implemented in the program Undertaker developed by the Karplus group at UCSC[10]). A pseudo-energy term favoring closure of the chain breaks is included in the potential function, with a weight that increases throughout the simulation. We used this new protocol in CASP6 to efficiently produce models with long-range beta-sheet pairings. Rigid-body transformations between the coordinate systems defined by the N-C_α-C atoms of paired residues in beta-sheets were extracted from proteins of known structure. These transformations fall into four classes defined by the strand orientation (parallel or antiparallel) and an additional *pleating*

†These authors contributed equally to this article.

Grant sponsor: Howard Hughes Medical Institute (to P.B., B.Q., J.S., D.B.); Grant sponsor: NIH NCRR; Grant number: P41 RR11823 (to L.M.); Grant sponsor: Helen Hay Whitney Foundation (to K.M.S.M.); Grant sponsor: NIH SGPP initiative (to D.C. and D.K.); Grant sponsor: the Human Frontier Science Program (to J.M.).

*Correspondence to: David A. Baker, University of Washington, Department of Biochemistry and HHMI, Box 357350, Seattle, WA 98195. E-mail: dabaker@u.washington.edu

Received 22 April 2005; Accepted 22 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20729

The article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

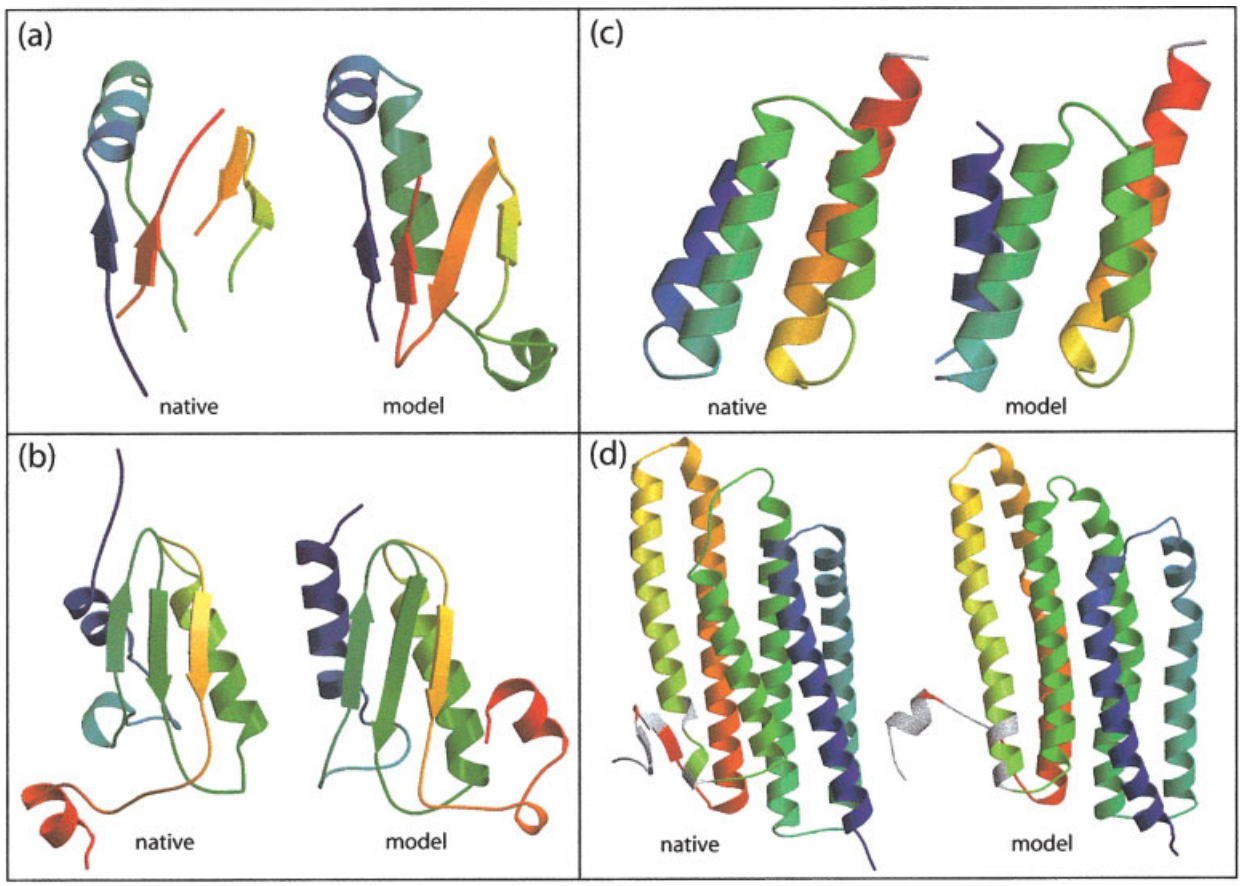

Fig. 1. Predictions using standard Rosetta. Native structures are shown on the left, models on the right. (**a**) T0209_2 model 1; 3.6 Å C_α-RMSD over 51 residues. (**b**) T0230 model 3; 3.7 Å over 86 residues. (**c**) T0248_1 model 4; 3.3 Å over 77 residues. (**d**) T0198 model 2; 4.0 Å over 210 residues (model 1: 3.94 Å over 198).

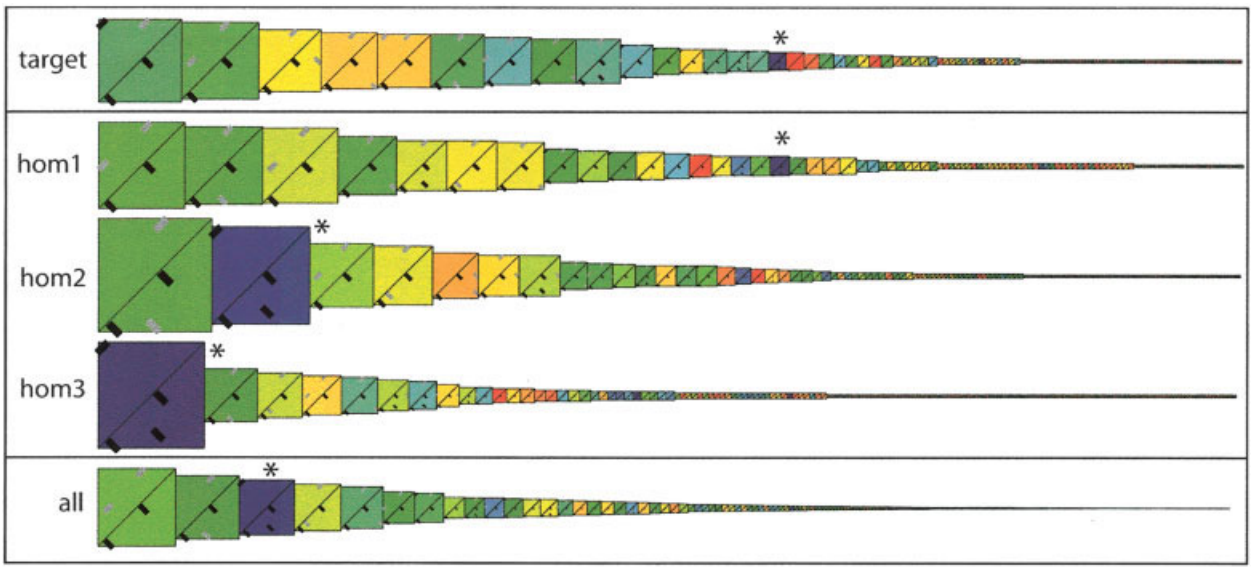

Fig. 2. Beta-sheet topology distributions for target T0201 and three sequence homologs. Each box depicts a beta-sheet topology, with strand pairings represented as they would appear in a contact map; parallel pairings are drawn above the diagonal, antiparallel below. Native pairings are black; nonnative pairings are gray. The dimensions of each box are proportional to the number of times that topology was seen in the decoy population. The boxes are colored by average C_α-RMSD to native for the decoys with that topology; blue indicates lower RMSD values and red indicates higher values. The top row of boxes represents the topologies sampled in de novo simulations with the target sequence. The middle three rows represent three sequence homologs. The bottom row describes the total set of models generated. The native topology is marked with an asterisk.

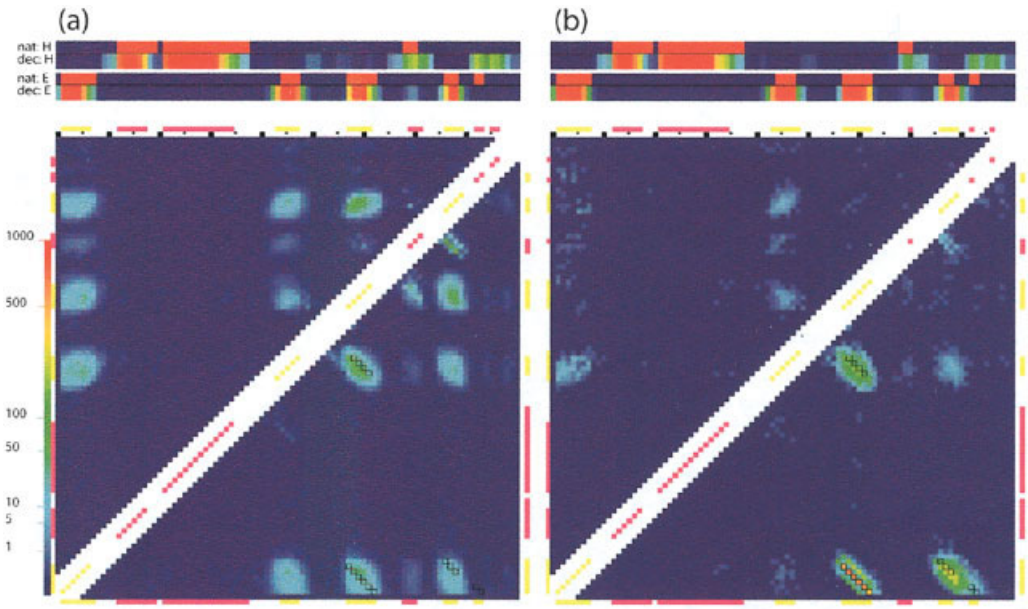

Figure 3.

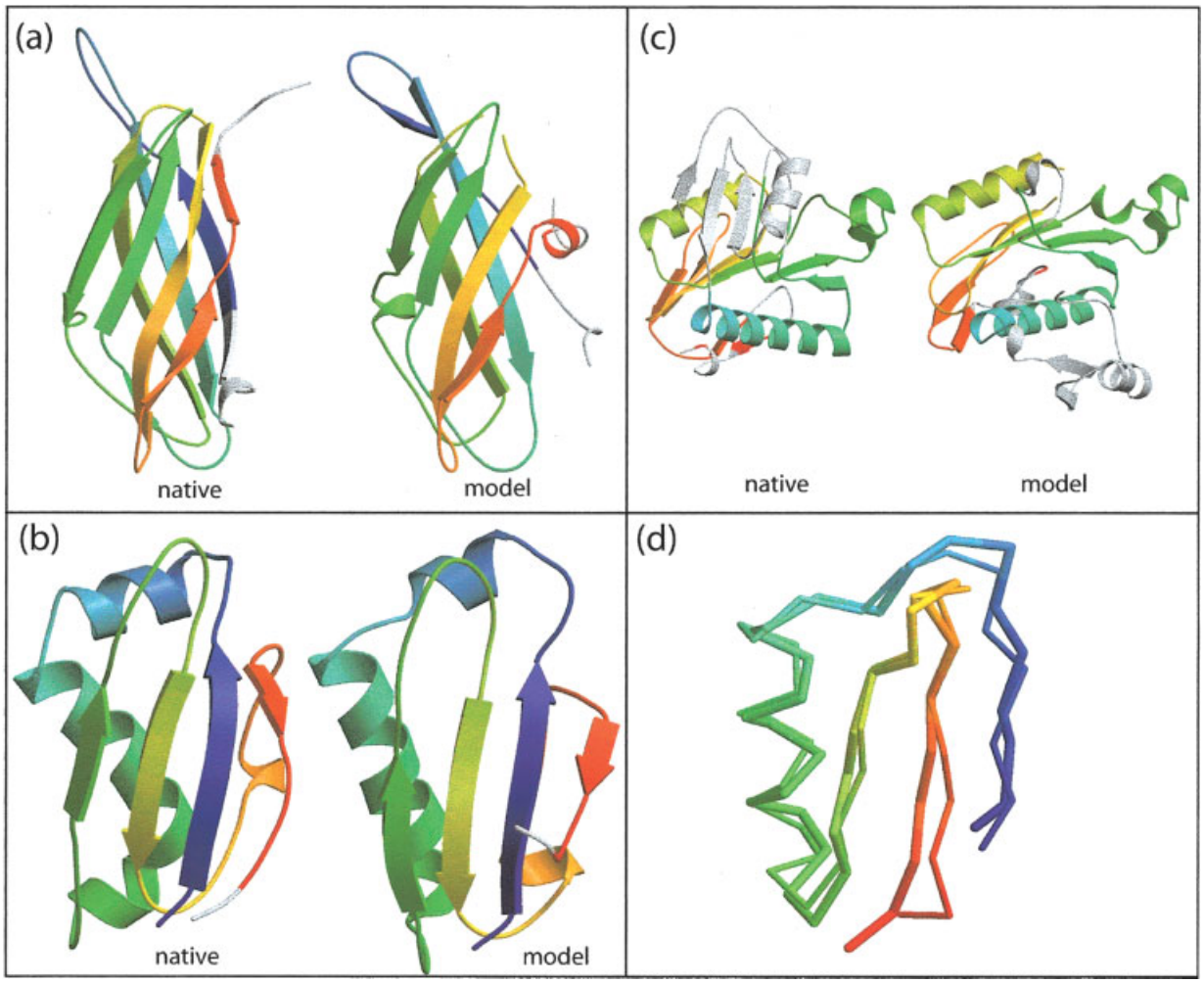

Figure 4.

term (−1 or 1) that specifies the orientation of the beta-carbons (which can point either above or below the beta-sheet). Four representative rigid-body transformations, one from each class, were used to define the relative orientation of beta-sheet pairings.

Given this machinery for constructing models with specified pairings, how do we select a set of target pairings for a folding simulation? For CASP6 we selected beta-sheet pairings by two approaches: resampling long-range (high contact order) pairings from an initial set of de novo models; and resampling pairings that were seen frequently in models from automated FR servers. In either case, beta-sheet pairings were identified in a set of input models and binned according to the residues paired and the orientation (parallel or antiparallel). Frequencies for each pairing were calculated and the most frequent nonlocal (separated by more than a single loop or alpha-helix) pairings were chosen for resampling. We focused on nonlocal pairings because Rosetta is already quite successful at generating sequence-local beta-sheet pairings, typically hairpins or beta–alpha–beta structures. Nonlocal pairings, on the other hand, are more difficult to sample in the standard protocol.

For proteins under 100 residues we used the Rosetta high-resolution refinement protocol[11] to generate low-energy all-atom structures starting from the Rosetta de novo models, as we did in CASP5.[12] Experience in our lab suggests that native and very near-native (under ~1.5 Å C_α-RMSD) structures tend to have lower all-atom energies than misfolded models.[11] Because structural change during refinement is typically small, however, starting models must already be fairly close (under ~2.5 Å C_α-RMSD) to the native to sample this native energy basin and be selected by the all-atom energy function. Thus, for all-atom refinement to be useful in model selection, the starting population of de novo models should sample conformational space as widely as possible. A new approach in CASP6 was the refinement of models from multiple homolog as a means of diversifying the starting population for refinement. We threaded the target sequence onto de novo models for sequence homologues, rebuilding insertions and deletions with the Rosetta loop-modeling protocol.[13] These models were added to the population of target-sequence models used as starting points for high-resolution refinement.

Fig. 3. Beta-strand pairing frequencies in de novo models (left panel) and decoys built by resampling nonlocal beta-strand pairs (right panel) for target T0272, domain 1. Sequence position increases from left to right and from bottom to top. The square at position (*i,j*) is colored according to the frequency with which residues *i* and *j* are paired in a beta-sheet in the models; parallel pairings are shown above the diagonal, antiparallel pairings below. Native pairings are boxed. Secondary structure (yellow boxes for strand, pink boxes for helix) is shown along the left and top (decoy consensus) and right and bottom (native) margins and the diagonal (decoy consensus), and also by fraction strand (E) and helix (H) across the top of the plots.

Fig. 4. Predictions using the beta-sheet resampling protocol. Native structures are shown on the left, models on the right. (**a**) T0212 model 2; 3.97 Å over 109 residues (model 1: 4.0 Å over 104). (**b**) T0272_1 model 1; 3.4 Å over 85 residues. (**c**) T0273 model 2; 3.97 Å over 126 residues (model 1: 3.98 Å over 111). (**d**) N-terminal subdomain of T0273, model 4; 1.55 Å over 40 residues (model 1: 3.71 Å over 36).

RESULTS AND DISCUSSION

The prediction targets can be grouped into categories according to the modeling techniques we applied. In the first category are the targets to which we applied the standard Rosetta de novo protocol. This resulted in accurate predictions for several targets, described further below; it is encouraging that the automated server Robetta did as well as the human group on these targets. In a second category are targets (such as T0198) to which we applied the standard de novo protocol together with human intervention at the model selection stage on the basis of additional information. For predictions in a third category we used the broken-chain protocol described above to construct models with specified beta-strand pairings. These pairings were taken from de novo models or FR servers, as described below. In the final category are small proteins for which we used the Rosetta all-atom refinement protocol and energy function in model selection.

Standard Protocol

For several targets we followed the standard Rosetta de novo prediction protocol and achieved results comparable to the fully automated server Robetta. Highlights among these cases include T0209 domain 2 (57 residues alpha/beta; model 1: 3.6 Å C_α-RMSD over 51 residues), T0230 (102 residues alpha/beta; model 3: 3.7 Å over 86 residues, model 1: 3.9 Å over 71), T0248 domain 1 (79 residues, all-alpha; model 4: 3.3 Å over 77 residues, model 1: 2.4 Å over 62) [Fig. 1(a–c)]—all relatively small proteins with simple topologies. For the 294 residue target T0248, the accurate domain parse from RosettaDom allowed us to fold more tractable subsequences. Target T0209_2 was recognized as a domain based on a Pfam match to the first domain. For target T0230, we incorrectly concluded from previously published NMR data[14] that there was a beta strand at the N-terminus. As a result, we deviated from the model ordering suggested by clustering, and moved our best model from first to third.

Standard Protocol Plus Extra Information

Based on the domain architecture assigned by Pfam—two copies of the PhoU domain—we recognized that target T0198 (235 residues, all-alpha) likely consisted of a structural repeat. This conclusion was supported by the fact that isolated PhoU domains are found in a number of proteins. Secondary structure predictions suggested that each domain was composed of three long alpha-helices. We folded the domains as independent units and generated primarily right and left-handled helical bundles. As with target T0129 in CASP5, we anticipated that the challenge would be in correctly packing the two domains, and that the constraints of generating a stable fold for the full protein might help distinguish between alternative topologies for the domains. For this reason we focused our attention on folding simulations of the entire protein.

Assuming that the two domains would have similar structures, we filtered the full-sequence models based on RMSD between the first and second domains, using a sequence mapping derived from the individual alignments to Pfam. We clustered the models after this filtering step, and selected topologies from among the largest clusters. The native fold was very well sampled after filtering, and as a result models 1 and 2 both had the correct topology [Fig. 1(d)]. Both models superimpose well to the native structure over almost the entire length—model 2 matches to 4.0 Å C_α-RMSD over 210 residues, and model 1 to 3.94 Å over 198 residues.

Target T0201 (94 residues, alpha/beta) was a member of a Pfam sequence family (DUF 464); we folded all 14 members of the family to generate a diverse set of models. The consensus secondary structure prediction for the family consisted of five beta-strands and two alpha-helices. Figure 2 shows the distribution of five-stranded beta-sheet topologies for the target sequence and three of the sequence homologs. Topologies were ranked by their frequency of occurrence across all homologues (Fig. 2, "all"), with manual reordering to account for differences in contact order that bias the sampling frequency distribution.[15] After reranking, the native topology moved from third to first, as the two more frequently sampled topologies were significantly lower in contact order. We submitted two models with this beta-sheet topology (model 1, 3.97 Å C_α-RMSD over 75 residues; model 3, 3.99 Å over 74 residues).

The disulfide connectivity and domain structure for target T0237 had been previously published,[16] and we used this information in our simulations of this challenging target. The first two domains were folded de novo, with terms incorporated into the scoring function to reward formation of the desired disulfide bonds. The resulting decoys were filtered for satisfaction of the disulfide constraints (at low resolution) and clustered to select models for submission. Although these models have low GDT-TS scores (22.5 for domain 1 and 27.2 for domain 2), they were among the best submitted for this target and capture low-resolution features of the secondary-structure packing in this complex fold.

Beta-Sheet Folding

Target T0272 was large enough (211 residues) to necessitate parsing, although no single cutpoint was clearly preferred. To maximize diversity, we built models with two alternative parses (one of which turned out to be correct) for the target and six homologs. The secondary structure predictions were not highly confident, particularly for the second domain, and varied widely across the homologs. By comparing topologies across the simulations we defined a set of core beta-strands (four per domain) and rebuilt models using only beta-strand fragments at these positions. Analysis of the topologies in these decoys suggested the possibility of a ferredoxin domain repeat, which was supported by the weak predictions for the second strand in each domain (edge strands are generally harder to predict), but the sampling of this (nonlocal) topology was limited [Fig. 3(a)]. To enhance sampling of high contact order beta-sheets, we built a third round of decoys using the broken-chain protocol with long-range beta-strand pairings extracted from the second round. For the first domain, the native ferredoxin fold was well-sampled at this stage [Fig. 3(b)], with the correct register and pleating in the nonlocal pairings. We selected models for submission by clustering the low-energy decoys [Fig. 4(b)]. For the second domain, the target sequence did not converge as well as several of the shorter homologs. Models for the second domain were built by threading the target sequence onto homolog cluster centers and using the Rosetta loop modeling protocol. Finally, full-chain models were assembled using fragment insertions in the linker while keeping the domains rigid.

For several targets we applied the broken-chain resampling protocol to build models with consensus beta-sheet pairings taken from FR servers. Target T0212 was a 126 residue all-beta protein. Although the top 3D-Jury score was rather low (22.0), there was strong consensus among the servers as to the existence, register, and pleating of a nonlocal beta-strand contact between the second and fifth strands. A second contact between the third and sixth strands was predicted with less confidence and some uncertainty as to the register. We built models by forcing one or both of these pairings, clustered the low-scoring decoys, and selected the centers of the largest clusters for submission. Models 1 and 2 had the native beta-sandwich topology, with model 2 having the correct register in five of the six beta-strand pairings [Fig. 4(a)]. Although the consensus beta-pairings from FR servers were correct, the global folds of the majority of models from which these pairings were taken were not; this highlights the potential strength of this contact-based approach to combining fold recognition and de novo structure prediction.

Based on the results of the 3D-Jury metaserver and analysis of conserved residue patterns, we determined that target T0273 (187 residues, alpha/beta) likely had a restriction endonuclease-like fold. Because the 3D-Jury scores for the FR matches were rather low (20–30) and the alignments had long deletions, we decided to model this target with the broken-chain Rosetta fragment assembly method rather than starting with one of the template-based structures. We extracted three beta-strand pairs (representative pairings: 70–88a, 91–113a, and 116–143p) from those server models that matched to the endonuclease fold, and constructed models using the beta-sheet resampling protocol. Despite the limited template information, these models were the best submitted for this challenging target [Fig. 4(c)]. In addition, although the orientation of the 40-residue N-terminal subdomain relative to the rest of the fold was not correctly predicted, the internal structure of this subdomain was modeled quite well: four of the five submitted models had C_α-RMSDs under 4A, with model 4 matching to 1.55 Å [Fig. 4 (d)]. These results together with those for T0212 illustrate that de novo methods supplemented with information from remote homologs can do consider-

Fig. 5. T0281 model 1 superimposed onto the native structure (1.59 CA-RMSD) showing core side chains and colored by sequence.

ably better than template-based methods in the distant fold-recognition regime.

High-Resolution Refinement

At 70 residues, target T0281 was within the size limits for application of the Rosetta all-atom refinement methodology. We generated de novo models for the target sequence and refined these models in the Rosetta high-resolution potential. In addition, we folded sequence homologues and built target-sequence models using Rosetta loop modeling, with the aim of generating a more diverse population of starting models for refinement. We clustered the low-energy models after refinement and selected low-scoring representatives of a variety of topologies for submission. Our model 1 submission (which came from a de novo simulation for one of the sequence homologues) had a C_α-RMSD to the native of 1.59 Å and recovered the native core packing moderately well (Fig. 5). All-atom refinement improved this model significantly: the C_α-RMSD before refine-

ment was 2.2 Å. Analysis of the target sequence (which did not fold well to the native topology) revealed several hydrophobic residues at exposed positions in the native structure that are replaced with polar residues in other sequences in the family. This observation offers a possible explanation for the folding success of other family members, and provides a rationale for this approach to high-resolution structure prediction.

What Went Wrong?

In addition to the successful predictions described above, there were several targets for which we failed to generate good predictions. Sources of difficulty included discontinuous domains (T0241), secondary-structure prediction failures (T0239, T0215), topologically complex, high contact order beta-sheets (T0242), and dimers with large interfaces (T0238). For some targets we used template-based approaches where free-modeling might have produced better results (T0216); for others the reverse was true (T0213, T0214). In some cases, small domain insertions or extensions in comparative modeling targets should probably have been folded as independent units (T0280_2).

CONCLUSIONS

As in years past, the CASP6 experiment was a fruitful experience for our group. We tested and refined newly developed methods. In addition, by working closely on individual targets we were led to several promising new avenues of research. One particular highlight was the success of high-resolution refinement for target T0281. The approach we developed for this target has since been tested on a larger benchmark with encouraging results. Another highlight is the success of other prediction groups in using and modifying the Rosetta code for their own modeling efforts. Examples include target T0215, in which the Ginalski group used Rosetta de novo simulations for sequence homologs to select models with the correct topology (our efforts were hampered by insufficient use of homologous sequence information); and comparative modeling target T0271, for which the GeneSilico group correctly modeled a missing C-terminal helix using Rosetta simulations in the context of a multimeric model. We look forward to the CASP7 experiment.

ACKNOWLEDGMENTS

The authors would like to thank the structural biologists who contributed structures to the CASP experiment, and the CASP organizers and assessors. We thank Keith Laidig for flawless administration of our computing resources.

REFERENCES

1. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
2. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. Bioinformatics 2001;17:750–751.
3. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53(Suppl 6):524–533.
4. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32(Database issue):D138–141.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
6. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
8. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R. Predicting protein structure using only sequence information. Proteins 1999;Suppl 3:121–125.
9. Meiler J, Mueller M, Zeidler A, Schmaeschke F. JUFO: secondary structure prediction for proteins. www.jens-meiler.de; 2002.
10. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 2003;53(Suppl 6):491–496.
11. Misura KM, Baker D. Progress and challenges in high-resolution refinement of protein structure models. Proteins 2005;59:15–29.
12. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 2003; 53(Suppl 6):457–468.
13. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55:656–677.
14. Almeida MS, Peti W, Wuthrich K. 1H-, 13C- and 15N-NMR assignment of the conserved hypothetical protein TM0487 from *Thermotoga maritima*. J Biomol NMR 2004;29:453–454.
15. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
16. Hodder AN, Crewther PE, Matthew ML, Reid GE, Moritz RL, Simpson RJ, Anders RF. The disulfide bond structure of Plasmodium apical membrane antigen-1. J Biol Chem 1996;271:29446–29452.