PROTEINS: Structure, Function, and Genetics 53:457–468 (2003)

Rosetta Predictions in CASP5: Successes, Failures, and Prospects for Complete Automation

Philip Bradley[†], Dylan Chivian[†], Jens Meiler[†], Kira M.S. Misura[†], Carol A. Rohl[†], William R. Schief[†], William J. Wedemeyer[†], Ora Schueler-Furman, Paul Murphy, Jack Schonbrun, Charles E.M. Strauss, and David Baker[*]

Department of Biochemistry, University of Washington, Seattle, Washington

ABSTRACT **We describe predictions of the structures of CASP5 targets using Rosetta. The Rosetta fragment insertion protocol was used to generate models for entire target domains without detectable sequence similarity to a protein of known structure and to build long loop insertions (and N-and C-terminal extensions) in cases where a structural template was available. Encouraging results were obtained both for the de novo predictions and for the long loop insertions; we describe here the successes as well as the failures in the context of current efforts to improve the Rosetta method. In particular, de novo predictions failed for large proteins that were incorrectly parsed into domains and for topologically complex (high contact order) proteins with swapping of segments between domains. However, for the remaining targets, at least one of the five submitted models had a long fragment with significant similarity to the native structure. A fully automated version of the CASP5 protocol produced results that were comparable to the human-assisted predictions for most of the targets, suggesting that automated genomic-scale, de novo protein structure prediction may soon be worthwhile. For the three targets where the human-assisted predictions were significantly closer to the native structure, we identify the steps that remain to be automated. Proteins 2003;53:457–468.** © 2003 Wiley-Liss, Inc.

Key words: protein structure prediction, fragment insertion, ROSETTA, CASP, full-atom refinement

INTRODUCTION

Rosetta was developed originally as an approach to the problem of de novo protein structure prediction, which sought to incorporate insights from experimental studies of protein folding.[1,2] After promising results in CASP3[3] and in light of the rapid rate of experimental structure determination, Rosetta was extended to model evolutionarily variable regions (such as long loops, domain insertions, and N- and C-terminal extensions) in the context of a template built by classical comparative modeling methods. In CASP4, Rosetta-built models (both with and without templates) were good in many cases.[4]

For CASP5, we followed the CASP4 approach of attempting to build complete models using Rosetta for every target sequence. To generate template-based models, we used homologous structure information; insertions, loops, and extensions with low-sequence similarity to the homologue were modeled by using the fragment insertion method in the context of the template. When convincing homology information was not detected, we modeled the entire sequence with our de novo fragment insertion method.

Here we describe the methods used to generate the de novo domain and long insertion predictions, with an emphasis on improvements made since CASP4 and the factors most likely to have contributed to both our successful and unsuccessful predictions. With the long-term goal of developing an accurate, completely automated procedure, we identify the contributions of human expertise to our predictions by comparing with results from a completely automated version of our protocol.

MATERIALS AND METHODS

Improvements in Rosetta Since CASP4

The Rosetta method of de novo protein structure prediction is based on the assumption that the distribution of conformations available to each three-and nine-residue segment of the chain is reasonably well approximated by the distribution of structures adopted by the sequence of the segment (and closely related sequences) in known protein structures. Fragment libraries for each three-and nine-residue segment of the chain are extracted from the protein structure database using a sequence and secondary structure profile–profile comparison method. The conformational space spanned by these fragments is then

Grant sponsor: Howard Hughes Medical Institute; Grant sponsor: Helen Hay Whitney Foundation; Grant sponsor: Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation; Grant sponsor: Human Frontier Science Program; Grant sponsor: National Institutes of Health; Grant numbers: NRSA AR08558 and T32 HG 00035.

†P. Bradley, D. Chivian, J. Meiler, K.M.S. Misura, C.A. Rohl, W.R. Schief, and W.J. Wedemeyer contributed equally to this work.

C.A. Rohl's present address is Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064.

W.J. Wedemeyer's present address is Departments of Biochemistry and Physics, Michigan State University, East Lansing, MI 48824.

C.E.M. Strauss present address is Biosciences Division, Los Alamos National Laboratory, Los Alamos NM 87544.

*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Box 357350, J-567 Health Sciences, Seattle, WA 98195-7350. E-mail: dabaker@u.washington.edu

Received 1 March 2003; Accepted 23 June 2003

searched using a Monte Carlo procedure with an energy function that favors hydrophobic burial and strand pairing and disfavors steric clashes. For each target sequence, large numbers of decoy structures are generated with this protocol and then clustered; the five largest clusters are generally chosen as our predictions. Further details may be found in the Supplemental Materials.

Between CASP4 and CASP5, advances were made in several aspects of the Rosetta method. The first improvement was the incorporation of two filters that remove conformations with non-protein-like properties. The first filter removes overly local, low contact order conformations,[5] whereas the second removes conformations with β-strands not properly assembled into β-sheets.[6] These filters were applied to large populations of decoy conformations and were very useful during CASP4 and have since been incorporated into the standard Rosetta procedure. The numbers of conformations generated for each CASP5 target given below refer to conformations that pass both filters.

The second area of improvement is in the energy function used during the search of conformational space. The atomic radii for backbone atoms and distances of closest approach between centroids used in the original Rosetta force field[1,2] were obtained from the distances of closest approach of atom pairs in a large set of protein structures. During the subsequent development of full-atom refinement methods, we noted that many of the decoys produced by the initial low resolution search contained significant backbone clashes. This resulted from artificially small atomic radii that derived from unrealistically short distances of closest approach in low-resolution crystal structures and NMR solution structures in the protein data set used to obtain the radii. Recomputation of these parameters using a set of high-resolution crystal structures resulted in more physically realistic larger values, and incorporation of this information into Rosetta reduced the number of backbone-atom clashes significantly and (most likely) the frequency of overly compact (but low scoring) conformations. The environment-dependent pair term in the original centroid mode-scoring function[2] was replaced by an environment-independent term to eliminate binning artifacts. Defects in the r-sigma term[2] that gave rise to incorrect β-strand register were fixed, as well as mistakes in the logic associated with restricting the backbone hydrogen bonding of a given segment of a β-strand to backbone atoms of not more than two other strands. Numerous other small bugs were corrected and speedups were implemented, along with the incorporation of modules for loop modeling, backbone refinement, domain assembly, and protein design, which were useful in some of the special cases described below.

The methodology for picking fragments from the protein structure database (the program NNMAKE) was also improved by ensuring that an appropriate diversity of secondary structures is present in the fragment library for regions with weak propensity to adopt a single secondary structure. In the Rosetta picture of folding, the secondary structure ultimately adopted by such regions will reflect the nonlocal interactions in the low-energy tertiary structures; hence, it is important that a diversity of conformations be present in the fragment libraries for these regions. Diversity is ensured by using three secondary structure predictions independently to supplement the sequence profile score used by NNMAKE. Between CASP4 and CASP5, a quota system was introduced to ensure that the percentages of sheet, helix, and coil in the fragments matched those of the input secondary structure predictions, and a new prediction method, JUFO,[7,8] was added.

Numerous other methods currently in development in our group were tested on subsets of the targets. Increased production of complex topologies was achieved in part through development of methods for detecting diverging turns and penalizing the formation of hairpins in such regions as well as for promoting nonlocal sheet contacts (J.M., in preparation). A method was used for recognizing evolutionarily conserved functional patches.[9] We also used cluster centers from our decoy population to search the PDB for structurally similar regions using the structure comparison method MAMMOTH.[10] Finally, significantly larger numbers of decoys were made for the targets in CASP5 compared to those in CASP4, resulting in a greater likelihood of producing native-like (and possibly topologically complex) decoys.

In the following subsections, we describe our standard prediction protocol; deviations from this standard protocol will be noted in the description of the individual targets.

Target Classification

A sequence was classified as a de novo or template-based target with use of PSI-BLAST[11] and Pcons2[12] (also described in this volume). If the E-value of the top PSI-BLAST hit was worse than 0.001 and the score of the top Pcons2 hit was worse than 1.5, the sequence was predicted to be a new fold/difficult fold recognition target and was modeled with Rosetta's de novo method. Otherwise, the target was classified as comparative modeling/easy fold recognition target and was modeled using Rosetta's template-based by approach. Sequences that received borderline Pcons2 scores were modeled separately using both methods and the most plausible models were submitted.

Domain Parsing

Target sequences were parsed into domains using multiple-sequence information and matches to known structures as described in the accompanying article on the Robetta server in this issue. For large targets, we attempted to use regions of low-sequence conservation to determine segment boundaries; however, in cases in which multiple-sequence information was uninformative, we split the sequence into roughly equal lengths. Models were generated for each predicted domain independently.

De novo (Fragment-Insertion) Modeling: Fragment Selection and Model Generation

See supplemental materials.

Clustering and Model Selection

For each target, fragment libraries and sets of decoy structures were generated both for the target sequence and for up to three homologous sequences identified with PSI-BLAST. Twice as many models were generated for the target sequence as for the homologues; the resulting models from the target and homologous sequences were pooled and then clustered as described previously.[13] Models of the target sequence were selected from the largest resulting clusters.

For clustering to succeed, a sufficient number of native-like decoys must be present among the models generated. Unfortunately, formation of native-like structures can be a low-probability event for larger, more complicated proteins; in such cases, the population is generally dominated by non-native conformations. To improve model selection for proteins with at least three predicted β-strands, we used a test set of mixed α/β proteins of >130 residues to develop a filter that enriched for native-like structures in our model populations. With the requirement that the three most native-like models (assessed by C^{α} RMSD) remain in the final population, we experimented with iterative filtering methods using individual terms of the total energy function as selection criteria. The most successful protocol was to select the top third of the population based on the β-strand pairing score and the third of those models with the smallest radius of gyration.

All-Atom Refinement of Models

For targets under 100 residues, the submitted predictions were chosen without clustering, as follows. The top 15% lowest-energy models were refined by using an improved version of the full-atom refinement protocol described previously,[14] which couples Monte Carlo minimization of the backbone and side-chain conformations. The full-atom energy function is dominated by Lennard-Jones interactions, an orientation-dependent hydrogen-bonding potential, and an implicit solvation model. Typically, 5,000–20,000 decoys were refined, and the five decoys with the lowest energies that belonged to different clusters were submitted.

Template-Based Modeling: Sequence Alignment

Our alignment method "K*Sync" (D.C., in preparation) produces large sets of candidate alignments (via a modified Smith-Waterman alignment algorithm[15]) by systematically varying the weights on score terms representing multiple-sequence information for both the query and the parent, the predicted and observed secondary structure, and the obligateness of a region to the fold (see the accompanying article on the Robetta server for more details).

The ensemble of sequence alignments was converted to an ensemble of three-dimensional template structures, and short-to-medium unaligned regions (<17 residues) were modeled in the context of these templates using an abbreviated insertion-modeling procedure (described in the next subsection). Alignments containing insertions that failed to produce conformations in agreement with the geometry of the template stems were discarded from the ensemble. The remaining alignments were ranked by evaluation of the structural models by several energy criteria. Human intervention was used to either select one of the high ranking alignments or to produce a new alignment by recombining the preferred features of multiple high ranking alignments.

Template-Based Modeling: Insertion Modeling

Unaligned regions corresponding to gaps in the sequence alignment as well as regions judged likely to show significant structural divergence from the parent structure were modeled by the Rosetta fragment insertion protocol in the context of the fixed template.[20] For regions of <17 residues, roughly 300 initial conformations were selected from a database of known structures using similarity of sequence, secondary structure, and stem geometry. Initial conformations for longer regions were built from 3-mer and 9-mer fragments. The conformations of all variable regions were then optimized by using fragment insertion and random dihedral angle perturbations. A gap closure term in the potential in combination with conjugate gradient minimization was used to ensure continuity of the peptide backbone. Optimization of variable regions was accomplished by using the standard Rosetta potential with centroid representation of side-chains, followed by optimization with explicit side-chains. All variable regions were optimized simultaneously, starting from a random selection of initial conformations. Generally, ~1000 independent optimizations were conducted. Variable regions were ranked independently by energy, and low-energy conformations for each variable region were combined into a final model.

Domain Assembly and Side-Chain Packing

For targets containing more than one domain, the separate domain models were combined into one full-length model. This was accomplished by splicing each domain together into a single chain, followed by fragment insertion into the linker region surrounding the splice site. The last step consisted of packing the side-chains using a backbone-dependent rotamer library[16] with a Monte Carlo conformational search procedure similar to that used in the all-atom refinement procedure described above.[17]

RESULTS AND DISCUSSION

Table I summarizes the results for the Rosetta CASP5 predictions, which used the fragment insertion de novo modeling procedure to build either the entire model or long insertions in the context of a fixed template. Targets for which Rosetta predictions were successful are addressed individually below. We compare our predictions to the native structure and discuss the specific methods used for each target in relation to the standard protocol described in Materials and Methods (Fig. 1). We address the usefulness of these alterations by comparing our submitted models to those generated with a fully automated version of the standard protocol. Targets for which the predictions were unsuccessful are then discussed together in an effort

TABLE I. Summary of Results for CASP5 Targets Predicted With Fragment Insertion by the Rosetta Algorithm

					Number of amino acids with an RMSD below 4 Å/6Å[e]		
Name[a]	class[b]	co[c]	α/β[d] [%]	Length	Human[f]	Standard[g]	Best[h]
129	nf	30.1	64/0	170	108/153	87/116	111/159
149_2	nf	34.6	23/35	116	52/71	44/62	76/92
161	nf	33.7	53/11	154	45/83	57/79	55/95
162_3	nf	24.6	36/38	168	58/79	—	68/95
181	nf	25.1	30/18	111	35/59	52/65	65/103
146_1	fr/nf	31.4	28/25	107	28/51	—	42/54
146_2	fr/nf	29.2	23/26	89	45/60	—	70/76
146_3	fr/nf	21.9	0/10	56	27/31	—	26/39
146_4	fr/nf	9.2	19/0	47	23/30	—	33/40
170	fr/nf	16.3	60/0	69	64/67	60/64	66/68
172_2	fr/nf	24.7	54/0	101	52/62	—	90/101
173	fr/nf	55.1	35/15	287	127/149	60/84	127/149
186_3	fr/nf	5.2	0/33	36	28/32	—	—
187_1	fr/nf	42.7	42/19	187	57/85[i]	—	76/114
135	fr/a	31.7	34/30	106	83/98	54/64	94/105
148_1	fr/a	23	38/32	71	62/64	57/62	65/66
148_2	fr/a	23.1	41/27	91	73/74	75/77	80/90
162_1	fr/a	13.1	76/0	56	56/56	—	56/56
162_2	fr/a	16.3	0/25	51	33/43	—	38/40
187_2	fr/a	38	38/14	227	51/85	—	85/120
191_1	fr/a	28.1	45/21	139	80/100	85/98	102/105
174_1	fr/h	47.2	38/26	197	54/64	—	52/67
174_2	fr/h	34.6	36/25	155	44/47	—	47/62
156	fr/h	46.4	18/32	156	59/88[j]	71/88	81/107
130[k]	fr/cm	—	39/20	100	79/90	—	—
172_1[k]	cm	—	43/23	192	129/159	—	—
186_2[k]	cm	—	40/18	250	142/186	—	—

[a]CASP identification number.
[b]Assessor classification (nf, new fold; fr/nf, fold recognition/new fold; fr/a, fold recognition analog; fr/h, fold recognition homologue).
[c]Contact order.
[d]Fraction of amino acids in α-helix or β-strand conformation.
[e]The number of residues (Cα atoms) of the model superimposable (using a variant of MaxSub[19] which uses RMSD as the threshold) on the native structure within a 4 Å RMSD cutoff (left) and within a 6 Å cutoff (right).
[f]Best Rosetta model submitted for CASP5.
[g]Best fully automated prediction using standard CASP5 protocol.
[h]Best ROSETTA model in decoy population before filtering.
[i]The best submission for T187 was a comparative model based on template 1 vpe with 57 and 116 residues aligned within 4 Å and 6 Å, respectively.
[j]The best submission for T156 was a comparative model based on template 1 dik with 78 and 107 residues aligned within 4 Å and 6 Å, respectively.
[k]Modeled with a template (130, 1f5aA; 172_1, 1ej0A; 186_2, 1gkpA) and fragment insertion (see text).

to highlight the main sources of problems for the method at its present stage of development.

The similarity of the true native structure to the best of the five CASP5 Rosetta predictions, the best of the five models selected by the fully automated protocol, and the best prediction in a large set of Rosetta-generated structures is indicated in Table I. The table includes proteins in the fold recognition category for which reliable parents were not identified by PSI-BLAST or Pcons2, as well as all targets modeled exclusively using the fragment insertion protocol. T172 and T186 are template-based predictions and are included in Table I because they provided a context for a de novo modeled domain insertion. T130 was also modeled using a template and is included in Table I because the submissions contained significant regions modeled by fragment insertion. Figure 2 shows Global Distance Test (GDT[18]) plots for select de novo targets of the five submitted models and the five models generated by the completely automated standard protocol. Figures 3 and 4 show ribbon diagrams of the best submitted model and the native structure for selected targets.

T129–170 Residues, All-α-Protein With Two Subdomains

Straightforward application of the standard Rosetta protocol yielded excellent results for this all-α-helical protein. Although most decoys were generally non-native by C^{α} RMSD (median: 16.1 Å), the density of decoy clustering correlated well with RMSD, as shown in Figure 5; the near-native decoys are more densely clustered than

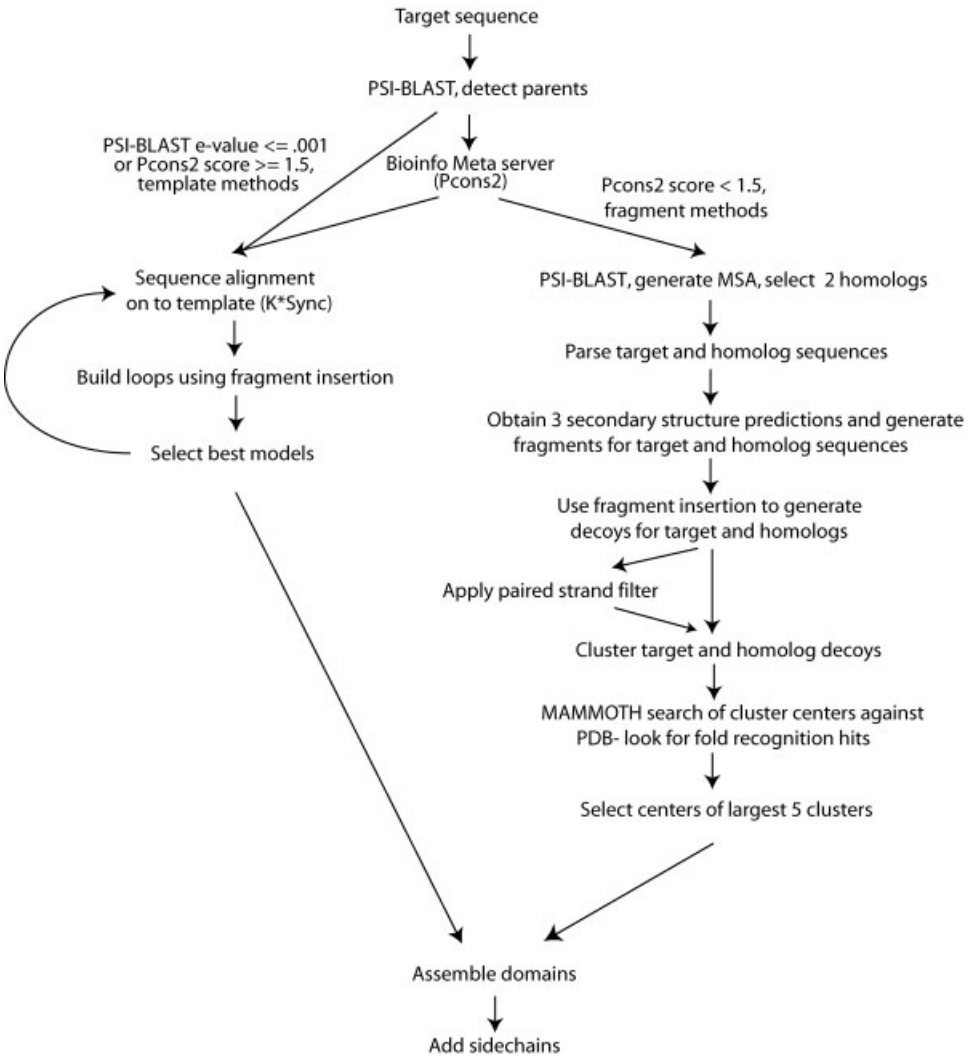

Fig. 1. Flowchart of general Rosetta protocol. Starting with obtaining the target sequence, steps for target identification, decoy generation, and selection are outlined for both the template-based approach (used for targets with homologous structures available in the PDB) and for the fragment insertion approach (used for new fold and difficult fold recognition targets).

the general population. As a result, the cluster centers represented the best decoys generated; for example, model 4 ranked second out of 11,075 decoys by C^{α} RMSD to native. Long-range contacts between the second and fifth helices were predicted correctly in these models, corresponding to approximately correct assembly of the two subdomains. Automated predictions for this target are of comparable quality to the manual submissions (Fig. 2, Table I), which is not surprising because the manual submissions were chosen with minimal deviations from the standard protocol.

T130-H10073 From *H. influenzae*, Four-Stranded Sheet Flanked by Three Helices

The parent structure 1fa0B (Yeast Poly-A Polymerase) used by Robetta was detected by Pcons2. For our human group predictions, we chose instead to use 1fa0B's close structural relative 1f5aA (Bovine Poly-A Polymerase) because of what appeared to be slightly closer sequence homology to the target. The default K*Sync alignment indicated the loss of a hairpin but retention of a helix (labeled "liberated helix" in Fig. 4) that packed against the hairpin in 1f5aA. We elected not to model this helix as part of the template and, instead, allowed the fragment-based loop-modeling protocol to build the helix and connecting turns (residues I52–R77). This permitted the adjustment of the helix that we supposed must occur in the absence of the hairpin. In addition, the parent 1f5aA possessed a much longer C-terminal helix than T130 appeared to have; therefore, we allowed the loop-modeling protocol to build the entire C-terminal portion of the model (residues D82–L114), unfortunately failing to capture the C-terminal strand. However, both of the de novo modeled helices were quite accurate (Fig. 4).

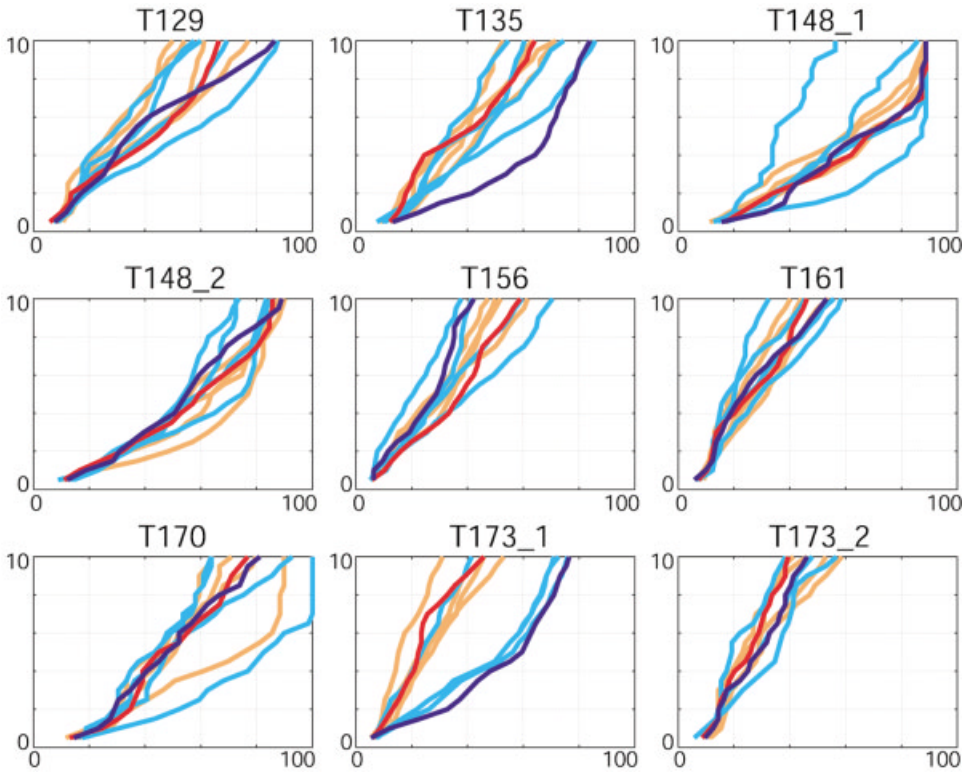

Fig. 2. Global distance test (GDT[18]) plots for selected targets comparing the CASP5 Rosetta submissions with predictions made with a fully automated version of the same protocol. Cyan (models 2–5) and dark blue (model 1) represent the CASP5 submissions, orange (models 2–5) and red (model 1) represent models made with a fully automated version of the CASP5 protocol (see Materials and Methods). The *y* axis represents a C^{α} RMSD cutoff under which to fit the model to the native structure, and the *x* axis represents the percentage of the model that will fit below that cutoff value.

T135-α/β Ferredoxin Fold

Because of substantial variations in the secondary structure predictions for this target and the failure of conformations generated with the standard protocol to cluster well (which can indicate that the true structure has a high contact order[5]), as well as weak Pcons2 matches to ferredoxin folds, we deviated from the standard protocol. As suggested by secondary structure predictions for most homologues, but not the prediction for T135 itself, β-strand fragments were favored in the region corresponding to the second strand. During the simulations, nonlocal β-strand contacts were favored to try to produce higher contact order structures (J.M., in preparation). The first submitted model has the correct topology and agrees with the native structure below 4 Å C^{α} RMSD over 80 amino acids. However, the fourth β-strand is shifted relative to the native structure, which prevents the sequence-dependent superimposition of this part of the model.

T148–162-Residue, Domain-Swapped, Double Ferredoxin Fold

T148 is a long sequence for Rosetta, but it has a deep multiple-sequence alignment and confident secondary structure predictions. Because the predicted βαββαβ signature indicates the ferredoxin fold, we hypothesized that T148 was a tandem ferredoxin fold. We parsed the sequence and folded the halves separately, generating tight ferredoxin-fold clusters for the C-terminal domain (retrospective analysis shows that some decoys in the ensemble aligned to the native structure within 2 Å C^{α} RMSD over 70 residues) but dispersed clusters with predominately local topologies for the the N-terminal domain. We abandoned this approach because of the incorporation of significant human bias. The domain swap of strand 1 made this a difficult target and caused difficulties in determining domain structure.

None of the models submitted for T148 were parsed into domains, and the standard protocol was followed with the addition of the strand score/gyration radius filters described in Clustering and Model Selection. When assessed by a variant of MaxSub,[19] which uses RMSD as the threshold, the five submitted models were similar to those picked by the standard protocol alone (Table I). However, the filters led to the selection of a model with the correct fold in the C-terminal domain (but not the correct topology, due to the strand swap) as our first model. It is of interest that Rosetta predictions of full-length T148 produced many models with secondary structure elements segregated into two domains; some models (including the best decoy generated, Fig. 3) even had the correct segregation, including the domain swap.

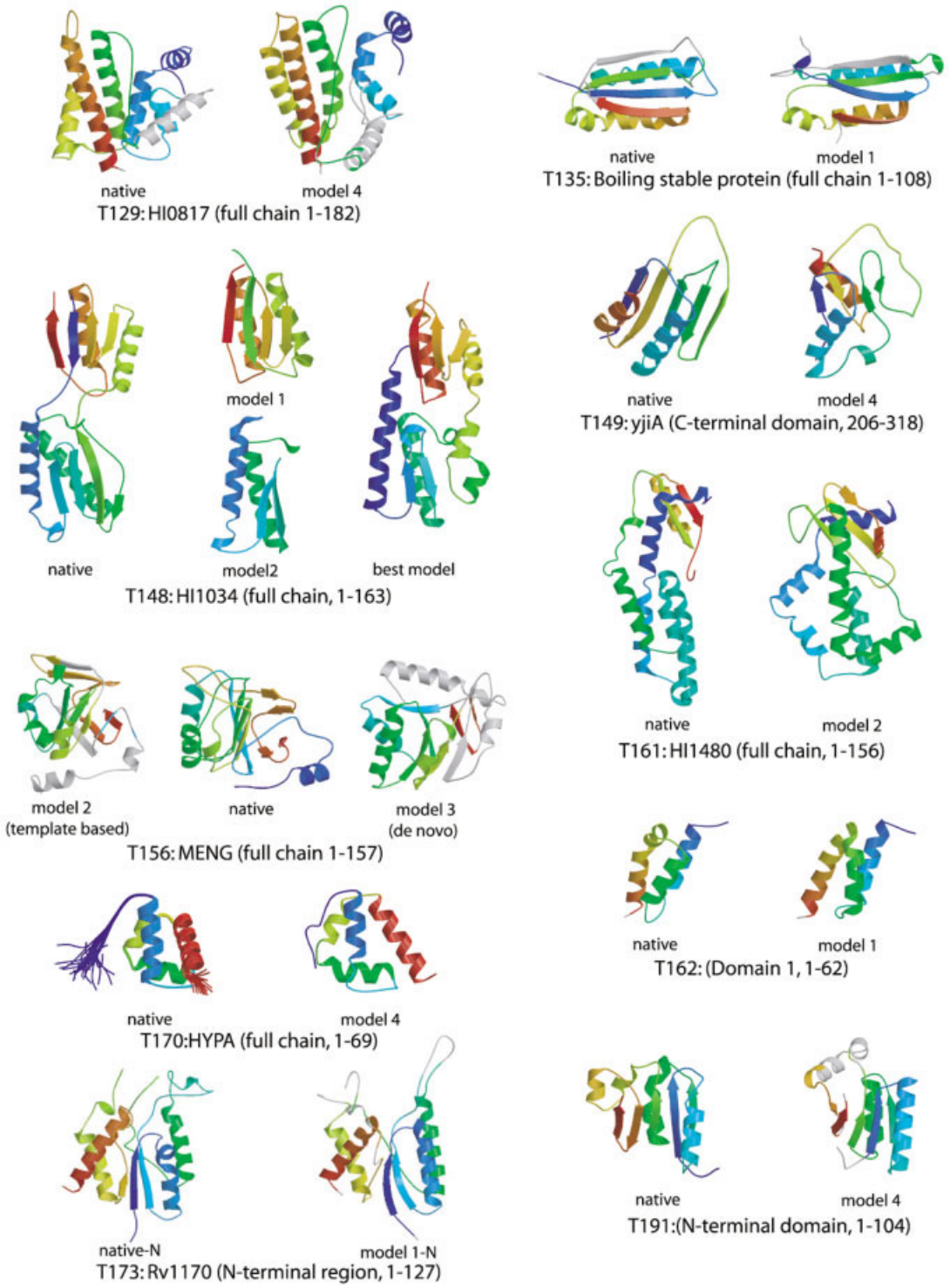

Fig. 3. Ribbon diagrams of predictions made by using the fragment insertion approach. The native structure and best submitted model are shown colored from the N-terminus (blue) to C-terminus (red). For T148, the best generated model is also shown, and for T156, both template-based and fragment insertion-based models are shown. For targets T173, T135, T156, and T191, colored regions deviate from the native structure by $\leq$4 Å, and gray regions deviate by $>$4 Å. For targets T129 and T156, colored regions deviate from the native structure by $<$ 6 Å C^{α} RMSD, whereas the gray regions deviate by $>$6 Å.

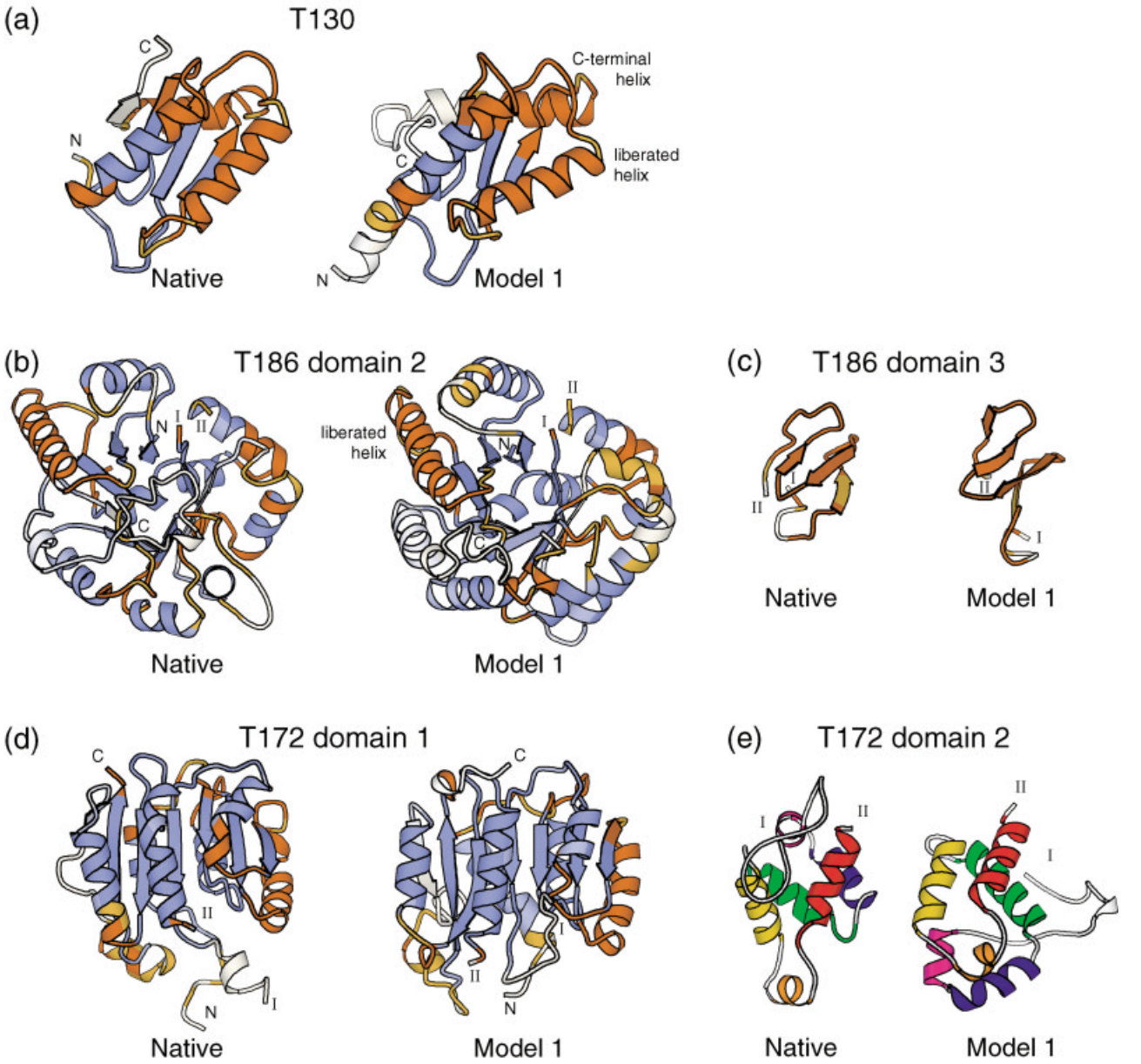

Fig. 4. Ribbon diagrams of targets predicted by using a combination of template-based and fragment insertion approaches. **a:** The native structure and the best model for T130, built by following our template-based protocol. The different shades of blue indicate regions that were modeled as template using coordinates from the homologue parent structure 1f5aA, whereas red, yellow, and white indicate regions that were modeled as loops with our modified de novo protocol that takes into account the context of the template. Dark blue and red show those residues that are within 4 Å, light blue and yellow deviate <8 Å, and ice-blue and white are >8 Å away from one another in the fit. **b:** T186 domain 2 native-model pair illustrates the good quality of the alignment for this TIM barrel domain, following the color scheme in (a). The success in the alignment for T186_2, particularly in the stem regions indicated by (I) and (II), provided the opportunity to build a good model for the minidomain insertion (**c**), accomplished with our long loop-modeling protocol. **d:** T172 possessed a domain insertion between strand 4 and the helix that precedes strand 5 [the stems are indicated by (I) and (II)], which was long enough to justify modeling following our full de novo domain-modeling protocol. **e:** The best model for the inserted domain T172_2 captures the helical elements well and is in quite good agreement with the native over the second half (the green, yellow, orange, and red helices).

T149_2–116-Residue Quasi-Ferredoxin Fold

The C-terminal domain of T149 was a challenging target due to the number of nonlocal contacts and the weak secondary structure prediction for strand 4. In our submitted model 4, strand 4 is not well formed and the C-terminal helix is on the wrong side of the sheet; nevertheless, the overall topology is similar to the native protein.

T156–158-Residue Methyltransferase

This target has a contact order of ~46 and Rosetta rarely generates decoys with such high contact orders. The five models generated by the fully automated protocol exhibited contact orders between 22 and 30. Because of the suspected complexity of the fold (a weak Pcons hit was to the methyltransferase 1dik; one submission was modeled

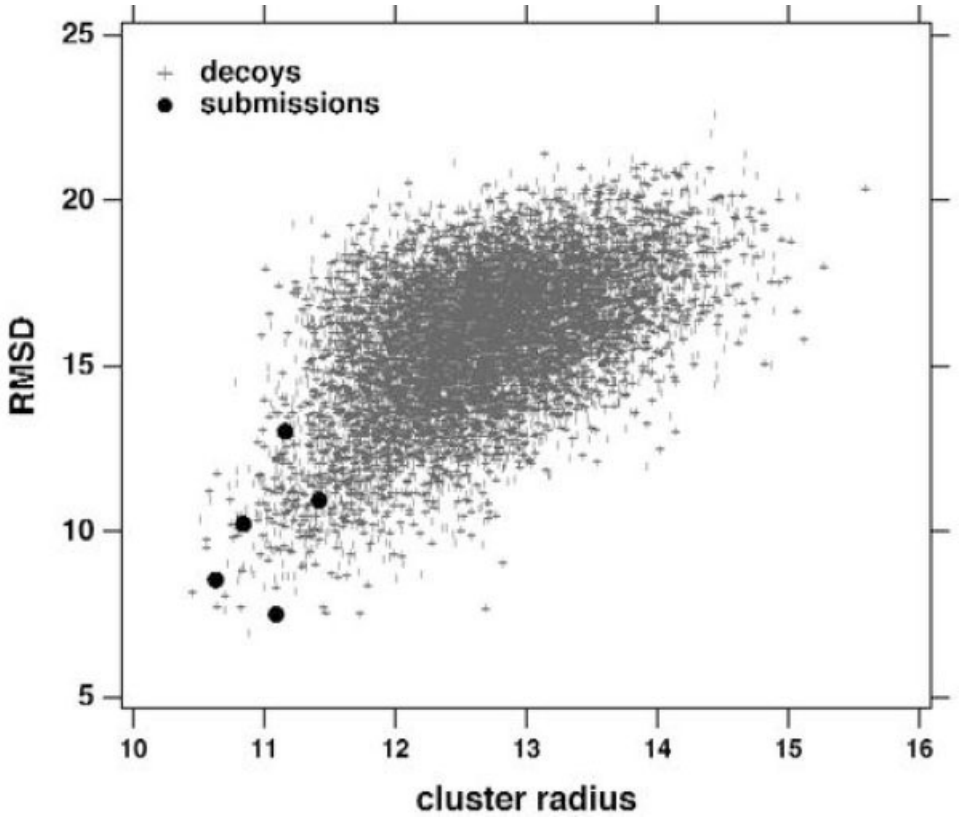

Fig. 5. Correlation between clustering density and model accuracy for T129. The 11,075 decoys produced for target T129 (a 182-residue, all-α-protein) are plotted on the basis of their global C^{α} RMSD to the native structure (*y* axis) and the density of nearby structures in the population (*x* axis). The density is calculated by comparing each decoy to all the others and recording the C^{α} RMSD to the 100th nearest neighbor. This distance is termed the cluster radius; smaller values indicate a higher density of neighbors. The five submitted models are shown as filled circles.

by using this template), we deviated from the standard protocol during decoy generation by adding a term that promotes nonlocal strand pairing (as in T135); during clustering, we also used a contact order filter to enrich the population for high contact order models. These alterations shifted the range of contact orders to 25–38. Our first submitted model has 59 amino acids below 4 Å C^{α} RMSD and has an overall topology similar to the native structure. However, the best submitted model (model 3) was built by comparative modeling using 1 dik as template (Table I).

T161–154 Residues, Novel Fold, Helical Motif Capped by β-Sheet

T161 is an elongated protein consisting of a helical motif capped by a three-stranded β-sheet, through which the protein dimerizes. We were unable to find sequences with convincing homology to T161, and the secondary structure predictions were weak over several regions. All of the submitted models were folded as a single chain, and the standard protocol was used with the strand score/gyration radius filters. Models 1–4 correctly predict a helical motif and a capping sheet, whereas none of the models would have contained the β-sheet had we not used these filters. Although our submitted models are generally of low quality, as assessed by GDT[18] and a MaxSub[19] variant, the overall fold of model 2 is in reasonable agreement with the experimental structure and was perhaps the best model produced for this protein in CASP5. In general, the models were more globular and the β-strands less exposed than in the native structure. These strands form the dimerization interface in the native structure and are shielded from solvent in the dimer. Dimerization is clearly difficult to model during the de novo folding protocol.

T162_1–Domain 1(56 Residues) of F-Actin Capping Protein a-1 Subunit From Chicken

Owing to uncertainties in the domain parsing for T162, three variations of the de novo protocol were used to produce decoys for T162_1. T162_1 was parsed and folded as residues 1–60, as part of a larger segment of the protein (residues 1–109), and as an N-terminal extension of de novo decoys previously produced for a central segment of the protein (residues 61–219). The first protocol most frequently produced decoys with a broken third helix, in disagreement with consensus secondary structure prediction and probably caused by our filter on the radius of gyration. Because decoys from the second and third protocols were in better agreement with the secondary structure predictions, only those decoys were submitted. Both the second and third protocols produced good models (submitted models 2 and 1, respectively). Model 1 had a C^{α} RMSD of 2.8 Å over the entire 56-residue

domain, and model 2 had a C^{α} RMSD of 2.7 Å over 53 residues (4.0 Å over 56 residues).

T170–69-Residue α-Helical FF Domain

Given the small size of this target, we applied the full-atom refinement protocol to the initial models generated by the standard fragment insertion protocol. The five submitted models were selected by lowest full-atom energy, after clustering to determine that the models were not too similar. Model 4 was the best prediction with a C^{α} RMSD of 4.2 Å. Only 40 models of the 15,000 refined were significantly better than model 4 (the best generated is 2.9 Å C^{α} RMSD to native), indicating that the energy function was fairly effective at selecting the best models produced.

T172-Conserved Hypothetical Protein From *T. maritima*, Two Domains

Predictions for this target were made by using a hybrid template-based and de novo modeling approach. The parent, 1ej0 chain A (Ftsj methyltransferase from *E. coli*), did not contain coordinates for residues R116–N216, and this was assumed to be a domain insertion. Because of the length of the insertion, it was modeled as an independent domain using the de novo protocol. The models for the second domain were filtered for the ability of the N-and C-termini to span the gap in the template-based models for domain 1 (labeled I and II in Fig. 4).

T173–303 Residue α/β Protein

T173 is difficult because of its length, and we made several attempts at domain parsing. The MSA over the first half (roughly 1–165) of the sequence was deep and showed regions of strong conservation, allowing us to identify several homologues that had shorter loops and were more tractable than the target sequence. We generated models for the target and a nonredundant set of the 10 shortest proteins in this region of the MSA. One of the homologue sequences folded more successfully than the others (i.e., produced a greater fraction of decoys that passed the contact order, gyration radius, and strand score filters) and gave rise to decoys with a reasonably well-formed four-stranded sheet surrounded by helices. The centers of the top clusters were selected, and models of the target protein were built by using the loop-modeling protocol (described in Materials and Methods) to map on the original sequence and fill gaps. These models were manually inspected, at which point it became clear that the fourth cluster center brought together two highly conserved segments at the ends of the first and fourth strands, via a 3214 strand topology. Based in part on the extent of clustering of conserved residues,[9] this model was selected as our top submission for the N-terminal half of the protein. The C-terminal segment, with longer regions of weak secondary structure prediction, proved more difficult to fold; however, a domain parse beginning at residue 223 folded consistently to a subdomain in the native structure consisting of an α-helix and β-meander (42 residues under 4 Å C^{α} RMSD).

T186–N-Acetylglucosamine-6-Phosphate Deacetylase From *T. maritima*

Target 186 was modeled by using our template-based protocol using the parent structure 1gkp chain A (D-Hydantoinase from *Thermus* sp.), a parent detected by PSI-BLAST. T186 possesses a minidomain insertion (residues I257–T292), classified as domain 3, within the TIM barrel domain 2. Our alignment for T186 to the TIM barrel portion of 1gkp was quite good overall and possessed the correct alignment at the stem portions of the template adjoining domain 3 (in our model residues S256 and F294, labeled I and II in Fig. 4) to allow for long loop modeling using our fragment insertion protocol in the context of the template. As can be seen in Figure 4, the insertion was modeled quite well. We believe this is the first example of a successful long loop modeling in the CASP experiments and, hence, particularly exciting. In addition, in the TIM barrel domain, flexible modeling of a helix not fixed to the starting template resulted in the correct packing register of the helix against the β-sheet template.

T191_1–Domain 1 of Shikimate 5-Dehydrogenase From *M. Jannaschii*

T191 was parsed into two domains (residues 1–105 and 106–282) based on homology to the protein 1gpjA. Because the sequence alignment between target and parent was rather poor in the first domain, three of the models submitted for the first domain were produced by using the standard de novo protocol (two comparative modeling models were also submitted). A large number of de novo decoys were initially produced (76,000). Before clustering, the decoy population was reduced to 4600 decoys by using the strand score and radius of gyration filters discussed in Clustering and Model Selection in Materials and Methods. Many of the largest cluster centers had one of two defects: either the third and fourth helices were merged into one, in disagreement with the consensus secondary structure prediction or a hairpin was disconnected and packed poorly with the rest of the protein, a relatively common Rosetta pathology. The decoys chosen for submission were the centers of the largest clusters that did not possess these defects. The best of our five submitted models was de novo; it has 100 amino acids below 4 Å and an overall C^{α} RMSD of 5.9 Å. Of all the initial decoys generated, only 1.2% were better by C^{α} RMSD, and only 0.5% had more aligned residues under 4 Å.

What Went Wrong?

In the new fold and fold recognition categories, our least accurate predictions were for targets T146, T162, T174, T181, and T187. Domain parsing was a problem for T146, T162, T174, and T187. All of these proteins are large (325, 286, 417, and 417 residues, respectively) and contain two or more domains that were (for the most part) not identified correctly during CASP5.

T162, T174, and T187 also had complex topologies. Rosetta produced a single four-stranded sheet for the second domain of T162, rather than the more complex native topology of a sandwich of two hairpins, but rela-

tively long fragments were correctly predicted for domains 1 and 3 (Table I). A feature of T174 and T187 that presents a serious challenge for Rosetta is the swapping of secondary structural elements between large domains. In T174, the first domain includes the 176 C-terminal residues along with a single N-terminal strand, whereas the second domain is composed of the intervening I60 residues. In T187, the first domain includes a single N-terminal helix with 168 residues from the C-terminal, whereas the 227 intervening residues comprise the second domain.

Of the targets for which complex topology and domain parsing were not issues, the most obvious failure is T181, which contains a strand that was almost always modeled as a helix in the Rosetta predictions. This was due to a bias toward helix in this region in the secondary structure predictions contributing to fragment selection. It is of interest that a new 3D structure-based secondary structure prediction method (JUFO-3D) predicts this region as a strand because it is spatially close to a β-hairpin and in the correct position to form hydrogen bonds with an adjacent strand. Potentially, a second round of Rosetta models made by using this improved secondary structure prediction could have been much more accurate. The JUFO-3D neural network uses the three-dimensional structure of Rosetta decoys in addition to the sequence information. It leads to a 4% increase in the Q_3 measure of secondary structure prediction accuracy with respect to the sequence-only analog for the CASP targets we modeled de novo.[21]

What We Learned

First, the CASP5 results show that Rosetta can produce models of increasingly complex topologies (i.e., of higher contact order) that are often roughly correct. Because of the relatively small number of new fold targets, progress from CASP4 to CASP5 is difficult to evaluate quantitatively; however, several successfully predicted proteins in CASP5 had higher contact orders than any successful CASP4 de novo predictions.

Second, the plausible model of the long insertion in T186 using de novo methods suggests that the coupled de novo/template-based method could be useful for modeling evolutionary novelties in protein families with a representative of known structure.

Third, the fully automated standard protocol produced models for many targets comparable in quality to the human-assisted Rosetta predictions (Table 1). (As noted elsewhere in this issue, the implementation errors in the Robetta server make the Robetta predictions a worse standard for comparison). This finding suggests that human intervention did not significantly improve model quality, at least at the level of the numerical assessment. However, the human-assisted predictions were clearly better in three cases: T135, T170, and T173.

What was the critical departure from the automated protocol for these three targets, and could it be incorporated into future automated protocols? For T135 and T173, the key was a more extensive use of the sequences of homologous proteins. The automated protocol does make use of homologous sequence information by generating models for two homologous sequences as well as the native sequence and subsequently clustering the models for the sequences together simultaneously. This automatically imposes distance constraints in regions of large deletions in one or both homologues (the residues flanking the deletion must be close in space) and introduces variation in secondary structure prediction in homologues. However, for T135 and T173, we made additional use of homologue information. For T135, it was recognized that the secondary structure prediction for the query sequence was likely to be incorrect because it differed from those of most homologue. For T173, modeling efforts were focused on a homologous sequence lacking several large insertions, and the model for the query was then built from these models. Both are potentially automatable–for large domains, an automated procedure could focus on building a good model of the smallest member of the family, whereas alternative secondary structure predictions found for most members of a family could be given more precedence in modeling a query sequence. The recognition that a model for T135 was plausible because it resembled a ferrodoxin fold could be readily automated by using MAMMOTH.[10] For T170, the human-assisted protocol used the full-atom refinement procedure, which has not yet been incorporated into the automated protocol. As the refinement protocol matures, it should be straightforward to incorporate it into a future automated protocol.

Finally, CASP5 highlights the primary challenges facing de novo structure prediction. For large proteins, domain parsing is a formidable problem. Promising results for α and α/β proteins suggest that Rosetta itself may be useful as a domain-parsing tool (David E. Kim, unpublished, and results from T148); however, there is clearly much still to be done in this area. For single-domain proteins, two key areas need work: assembling complete structures for complex domains and full-atom refinement to improve the accuracy and ranking of models for proteins below 100 amino acids. A long-term goal of de novo structure prediction is clearly to produce models of atomic-level accuracy for small proteins.

ACKNOWLEDGMENTS

The authors thanks the CASP5 organizers and assessors, the experimentalists who participated in CASP5, and Nick Grishin and Ming Tang for analysis of the full Rosetta decoy sets. We would also thank Keith Laidig for flawless administration of our computing resources, and the members of the Baker laboratory for helpful discussions. We acknowledge funding from the following sources: Howard Hughes Medical Institute (W.J.W., J.S., D.B), Helen Hay Whitney Foundation (K.M.S.M.), Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation (O.S-F.), NIH NRSA AR08558 (W.R.S.), the Human Frontier Science Program (J.M.) and NIH Training Grant T32 HG00035 (P.B.). D.C. is a PMMB fellow, administered by the Florida State University with funding from the Burroughs-Wellcome Fund.

SUPPLEMENTAL INFORMATION

Fragment Selection

For each sequence, two sets of fragments are generated. The first has 25 fragments of length 9 for every residue (except for the last 8 residues), and the second contains 200 fragments of length 3. Fragments are selected on the basis of the agreement of their sequence with the MSA profile of the target, as well as the agreement between the predicted secondary structure with the DSSP secondary structure assigned to the fragment in its PDB file. Chemical shifts were available for the fold recognition target T0138 and were used to produce the fragment files for loop modeling, as has been described previously.

Decoy Generation

The fragment files were used to build models by the Rosetta protocol,[1–4] which has not changed significantly since CASP4. Briefly, Rosetta is a five-stage, fragment insertion Metropolis Monte Carlo method. Backbone atoms are represented explicitly and their connectivity is maintained, whereas side-chains are approximated by centroids. 1) The first stage begins with a fully extended chain and inserts 9-mer fragments at random positions for at least 2000 steps, until every backbone dihedral angle has been altered at least once. The only component of the potential function considered at this stage is a steric-clash term that prevents close approaches of backbone atoms and centroids.[1,2] 2) The second stage also consists of 2000 9-mer fragment insertions, but the scoring function includes residue-environment and residue–residue scores favoring hydrophobic burial and specific pair interactions, as well as secondary structure-packing scores.[1,2] 3) The third stage consists of 10 iterations of 2000 9-mer fragment insertions during which the local strand-pairing score is cycled on and off to promote formation of nonlocal β-strand pairing over local strand kinetic traps, whereas the local atom density is pushed toward that of native protein structures. 4) In the final stage, three iterations of 4000 3-mer fragment insertions are conducted out; a term linear in the radius of gyration is added to help condense the model and a higher resolution model of strand pairing is used. 5) The final decoy is stored only if it passes several filters designed to eliminate common Rosetta pathologies, such as decoys with an overly high radius of gyration or unpaired β-strands. Between 10,000 and 400,000 independent simulations were conducted for each target sequence, starting from different random number seeds.

REFERENCES

1. 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.
2. 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 Function, and Genetics 1999;34:82–95.
3. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using Rosetta. Proteins 1999;Suppl 3:171–176.
4. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;45 Suppl 5:119–126.
5. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
6. Ruczinski I, Kooperberg C, Bonneau R, Baker D. Distributions of beta sheets in proteins with application to structure prediction. Proteins 2002;48:85–97.
7. Meiler J. JUFO: Secondary structure prediction for proteins. www.jens-meiler.de/jufo.html 2002.
8. Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;7(9):360–369.
9. Schueler-Furman O, Baker D. Conserved residue clustering and protein structure prediction. Proteins 2003;52:225–235.
10. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2611.
11. Altschul SF, Madden TL, Schäffer 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.
12. Lundstroem J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network–based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
13. Bonneau R, Strauss CEM, Baker D. Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 2001;43:1–11.
14. Tsai J, Bonneau R, Rohl C, Baker D. An impared protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76–87.
15. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
16. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
17. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
18. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13–21
19. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16:776–785.
20. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins 2003. In Press.
21. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. PNAS 2003. In Press.