PROTEINS: Structure, Function, and Genetics Suppl 5:119–126 (2001)
DOI 10.1002/prot.1170

Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction

Richard Bonneau,[1] Jerry Tsai,[1] Ingo Ruczinski,[1] Dylan Chivian,[1] Carol Rohl,[1] Charlie E. M. Strauss,[2] and David Baker[1]*

[1]*Department of Biochemistry, University of Washington, Seattle, Washington*
[2]*Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico*

ABSTRACT **Rosetta ab initio protein structure predictions in CASP4 were considerably more consistent and more accurate than previous ab initio structure predictions. Large segments were correctly predicted (>50 residues superimposed within an RMSD of 6.5 Å) for 16 of the 21 domains under 300 residues for which models were submitted. Models with the global fold largely correct were produced for several targets with new folds, and for several difficult fold recognition targets, the Rosetta models were more accurate than those produced with traditional fold recognition models. These promising results suggest that Rosetta may soon be able to contribute to the interpretation of genome sequence information. Proteins 2001;Suppl 5:119–126.**

Key words: contact order; Rosetta; CASP; structure prediction; clustering; fold complexity; topology

INTRODUCTION

The Rosetta method of ab initio structure prediction is based on the assumption that the distribution of conformations sampled by a local segment of the polypeptide chain is reasonably well approximated by the distribution of structures adopted by that sequence and closely related sequences in known protein structures.[1] Fragment libraries for all possible three- and nine-residue segments of the chain are extracted from the protein structure database by using a sequence profile comparison method. The conformational space defined by these fragments is searched by using a Monte Carlo procedure with an energy function that favors compact structures with paired β-strands and buried hydrophobic residues.[2] The output structures are by construction consistent with the local conformational biases inherent in the sequence and have low free energy nonlocal interactions by virtue of the Monte Carlo optimization procedure.[3] For each query sequence a large number of independent simulations are carried out by starting from different random number seeds. The resulting structures are clustered, and the centers of the largest clusters are selected as the highest confidence models.[4]

The primary improvements to the method since CASP3 fall into three classes.[5] The first class consists of improvements in the basic simulation method. Several alternative secondary structure prediction methods (PSIPRED, SAMT99, and PHD) were used to bias the fragment picking method, allowing the method to recover from mistakes in any one prediction method.[6–8] Considerable speedups in the structure generation procedure were achieved by recomputing only the subset of distances changed after each fragment substitution by using a simple neighbor list; this allowed the generation of many more structures in a given amount of CPU time. To keep the simulations fluid, the perturbations caused by fragment substitutions after the collapse of the chain were reduced either by using the local move strategy of Gunn[9] or by explicitly minimizing the perturbation by varying backbone torsion angles adjacent to the site of the substitution.

The second class of improvements consists of filters that eliminate non-protein-like conformations from large sets of simulated structures and increase the frequency of native-like conformations. The first filter eliminates conformations with poorly formed β-sheets (Ruczinski, in preparation). The second filter removes structures with contact orders lower than those of native proteins of comparable lengths (Bonneau, submitted). For the larger targets, higher contact order structures were preferentially selected from large numbers of additional simulations. The third filter (Tsai, in preparation) involves the removal of structures with poorly packed interiors as assessed by a potential function consisting of Lennard Jones, hydrogen bonding, and solvent-accessible surface area-based solvation terms. To accommodate side-chain packing, structures were relaxed by using the standard Rosetta search strategy supplemented with small random changes of the torsion angles.

The third class of improvements involves the simultaneous clustering of conformations generated independently for several sequences related to a given target sequence.[10] This procedure identifies free energy minima common to multiple members of a protein family and compensates to some extent for inaccuracies in the potential function. For most targets, the submitted structures were the centers of the five largest clusters obtained by simultaneously clustering the conformations generated for the target and a representative subset of its homologues.

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

Received 10 April 2001; Accepted 2 July 2001

TABLE I. Summary of Ab Initio Predictions for CASP4[†]

Target	Len	2′ class	Contact order of native	No. of runs	Multiple homologs folded	Cluster threshold (Å)	Best decoy generated RMSD (len)	Best submitted models	
								Dali Z-score	Good fragments RMSD (len)
T086	164	α/β	20.2	115	n	10.4	5.3 (77)	0	—
T087-A	192	α/β	15.6	35	y	8.3	5.3 (124)	3.1	6.5 (128)
T087-B	118	α/β	11.1	47	n	6.3	4.8 (124)	3.5	6.1 (85)
T091	109	α/β	6.9	50	y	3.9	3.1 (90)	7.0	4.9 (90)
T094	181	α/β	22.9	170	n	12.0	5.7 (88)	0	—
T095	244	α	10.1	25	y	2.4	3.8 (178)	2.4	5.0 (139), 2.9 (120)
T096-B	160	α	12.3	10	n	6.5	4.9 (123)	3.5	5.7 (63)
T097	105	α	11	10	y	4.1	3.8 (100)	4.6	4.6 (81)
T098	121	α	9.2	10	y	2.9	4.2 (114)	2.6	3.9 (63)
T102	70	α	9.23	12	#	3.9	3.2 (70)	5.3	3.56 (70)
T105	94	α/β	7.2	65	n	7.4	4.6 (86)	1.8	5.9 (53)
T106	128	α	10.4	45	y	7.9	4.3 (103)	2.6	6.4 (106)
T107	188	β	29.1	6	n	14.0	5.1 (82)	0	—
T110	128	α/β	10.1	18	y	4.6	2.9 (95)	3.8	4.6 (79)
T115	300	α/β	27.7	2	n	11.0	5.6 (92)	0.6	5.3 (90)
T116-A	116	α/β	13.8	33	y	6.1	4.3 (96)	1.8	6.0 (85)
T116-B	155	α/β	15.8	25	y	5.8	4.0 (91)	6.0	215 (69)
T116-C,1	150	α	22.2	8	n	10.0	4.5 (95)	2.4	4.7 (50)
T116-C,2	150	α	22.2	40	n	6.4	2.5 (69)	2.4	4.5 (57)
T118	149	α/β	—	47	n	10.0	4.9 (70)	0	—
120	336	α/β	12.3	56	n	7.0	—	0	—
T121-B	372	α/β	20.8	65	n	7.0	4.3 (89)	0	—
T124	242	α	15.28	8	n	8.4	3.0 (132)	5.0	4.52 (117)
T126	163	α/β	24.7	127	n	9.1	5.2 (85)	0	—

[†]Results are shown for each target attempted with Rosetta. The length, secondary structure class, and contact order of each target are shown in columns two, three, and four, respectively. Each run in column 5 generated ~ 1,000 decoys. If multiple homologs were independently folded and simultaneously clustered, a "y" is indicated in column 6. The "#" in column 6 for target 102 indicates the folding of multiple cyclic permutants. The clustering threshold obtained upon clustering the filtered decoys is shown in Å in column 7 (a smaller threshold indicated tighter clustering). Column 8 provides the RMSD and length of the best generated decoy in the prefiltered decoy populations. The last two columns provide the Dali Z-score and RMSD (over the length of the correctly predicted region) of the best submitted model for each target (ref.). Dali Z-scores of zero indicate that no significant correctly predicted regions were found in any of the five submitted models.

CASP4 METHOD OVERVIEW

As a first step, each target was parsed into domains, when possible, based on the PSI-BLAST[11] generated multiple sequence alignment. Parsing was done by identifying groups of homologs that aligned to subregions of a given targets alignment. Domains that could be linked with some confidence to a protein of known structure either by using PSI-BLAST, by considering the function and the secondary structure of the protein, or for the later targets, by consulting the CAFASP consensus,[12] were modeled by using a version of Rosetta adapted for comparative modeling (Rohl et al., in preparation) and will not be considered further in this article. This procedure yielded 26 domains to be modeled using the standard ab initio version of Rosetta. The β and α-β domains over 200 amino acids were subdivided to avoid the drastic slowdown in the search for large proteins (resulting only once in an unrecoverable error, parsing T0120 incorrectly into three domains). For domains with sufficient diversity in their multiple sequence alignments, several homologs were selected and independently folded, to be simultaneously clustered prior to model submission (Table I).[10]

Fragment libraries were generated for each target and the selected homologs. The Monte Carlo fragment substitution protocol[13] was used to build structures from these fragment libraries. Approximately 40,000 fragments substitutions were attempted for each structure generated. Depending on the size of the protein and the extent of convergence of the simulations, 6,000–150,000 independent conformations were produced for each target.

A critical bias in raw Rosetta populations is the overabundance of low contact order (CO),[14] overly local, structures (Bonneau, submitted). To correct for this bias, we removed all decoys with absolute CO less than that seen for 95% of native structures of comparable length and secondary structure class. Unfortunately, we did not possess the computer resources necessary to fully normalize the CO distributions for larger targets—enforcing only this lower CO cutoff still left higher CO conformational spaces relatively undersampled. Many simulations result in structures with unpaired strands or strand arrangements not seen in the database, despite the fact that strand pairing is part of Rosetta's scoring function. After the minimal CO filter was applied, structures with non-protein-like strand arrangements (unpaired strands, strands aligned with more than two other strands within 6 Å, poorly aligned strand pairs, left-handed turns between sequentially adjacent parallel strands, etc.) were removed from the decoy

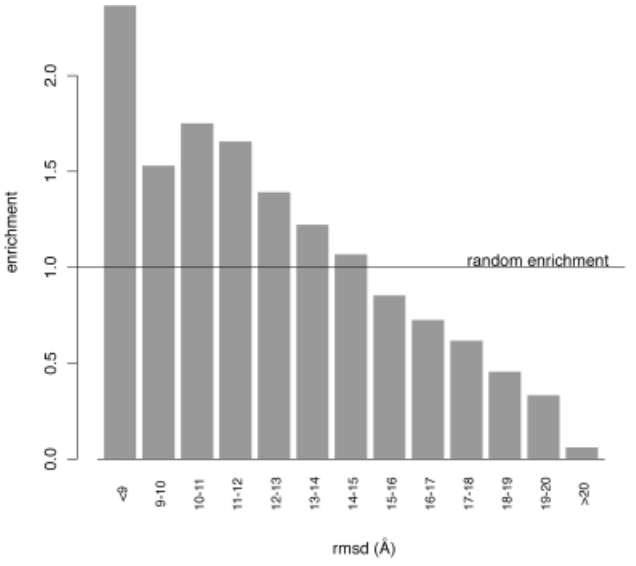

Fig. 1. The effect of the strand arrangement filter for target 87 domain A. The enrichment value is the percentage of the final population in a given RMSD bin divided by the percentage of the original decoy population in the bin. The enrichment expected from a random filter is 1.0 and is indicated with a solid line. For this target, the filter more than doubles the proportion of decoys $<$ 9 Å while rejecting almost all structures $>$ 18 Å.

populations (Ruczinski, in preparation). The occurrence of non-protein-like strand arrangements becomes more frequent as proteins become larger, for the majority of the large targets containing multiple strands, this filter removed 30–90% of the decoys. Figure 1 shows the performance of this filter for target 87 domain A.

After the successive application of these two filters (or just the CO for all-α proteins), the reduced model (backbone plus one centroid per residue) was expanded by the addition of side chains by using Dunbrack's backbone-dependent rotamer library.[15–17] The structures were then relaxed to accommodate and properly pack side chains using rotamer substitutions, single torsion angle moves, and small perturbation fragment insertions. An all atom, physically based, potential was used to identify well-packed, low-scoring models (Tsai, in preparation). In many cases, there was not enough computer time to add side chains to all structures prior to clustering, and side chains were added only after clustering the decoys.

The final step in the procedure is the simultaneous clustering of the filtered populations of conformations for a given target and its homologs. For clustering, two structures are considered structural neighbors if within a given root-mean-square deviation (RMSD) cutoff. This cutoff is reduced iteratively until the largest cluster (the group of conformations closer than the cutoff to the conformation with the most neighbors) contains $\sim$ 100 conformations. Once the cutoff (clustering threshold) has been determined, the members of the largest cluster are removed from the population, and the next largest cluster is determined by using the same procedure. Clustering stops when the next cluster under consideration contains fewer than 10 members. Clustering thresholds for each target are shown in Table I; smaller thresholds denote tighter convergence. Cluster centers were ranked by size, and in most cases, the five largest unique clusters were submitted. For larger proteins, the ranking of cluster centers was modified after manual inspection to remove persistent non-protein-like models (i.e., models containing recognizable features common to incorrect Rosetta models), and the resulting top five unique models were submitted. The degree to which cluster centers were manually reordered is roughly correlated with the clustering threshold.

RESULTS FOR TARGETS FOLDED WITH THE AB INITIO PROTOCOL

Overall summaries of the quality of the predictions for each target are provided in Table I and by the Hubbard plots in Figure 2.[18] For 17 of the 21 domains attempted with fewer than 300 residues, fragments of $>$50 residues were predicted with RMSDs of $<$6.5 Å. This is a considerable increase in consistency over our CASP3 results.[19] For all but two of the domains, there were segments of $>$60 residues with $<$6 Å RMSD to the native structure in conformations that were generated but not submitted. Inspection of the Hubbard plots in Figure 1 reveals that several of the predictions, notably T091, T106, and T116, were considerably better than any other predictions made for the target at CASP4. In most cases, at least one of the Rosetta predictions was among the best made for that target. Despite this success on a relative scale, the Hubbard plots make very clear that there is still a long way to go: for most targets the RMSD over the full length of the target is quite high.

Because of the availability of additional information and differences in length and secondary structure class, slightly different procedures were followed for each target. These are described below along with the factors that are likely to have contributed to the success or failure of the predictions.

T097, T098 (Small α)

These proteins were well within the size and complexity limits of Rosetta. Decoy populations were generated, side chains were added, the structures were scored in the presence of the full atom model, and the lowest 30% of decoys according to the all-atom score were clustered. The centers of the top five clusters were submitted. Our first model for target 97 (Fig. 3) was quite accurate for most of the structure (4.6 Å RMS over 81 residues) and was one of the best predictions overall for this protein (Fig. 2). For both targets 97 and 98, significantly better predictions were generated and not submitted (Table I), illustrating the need for better discrimination for small domains via continued development of the all-atom potential.

T102 (Small α, with Constraints)

A great deal of experimental information was available prior to CASP4 for this small helical protein. The secondary structure was known from partially assigned NMR data,[20] and it was known that the protein's N- and C-termini are covalently linked after synthesis.[20] To average the effects the termini had on the outcome of the

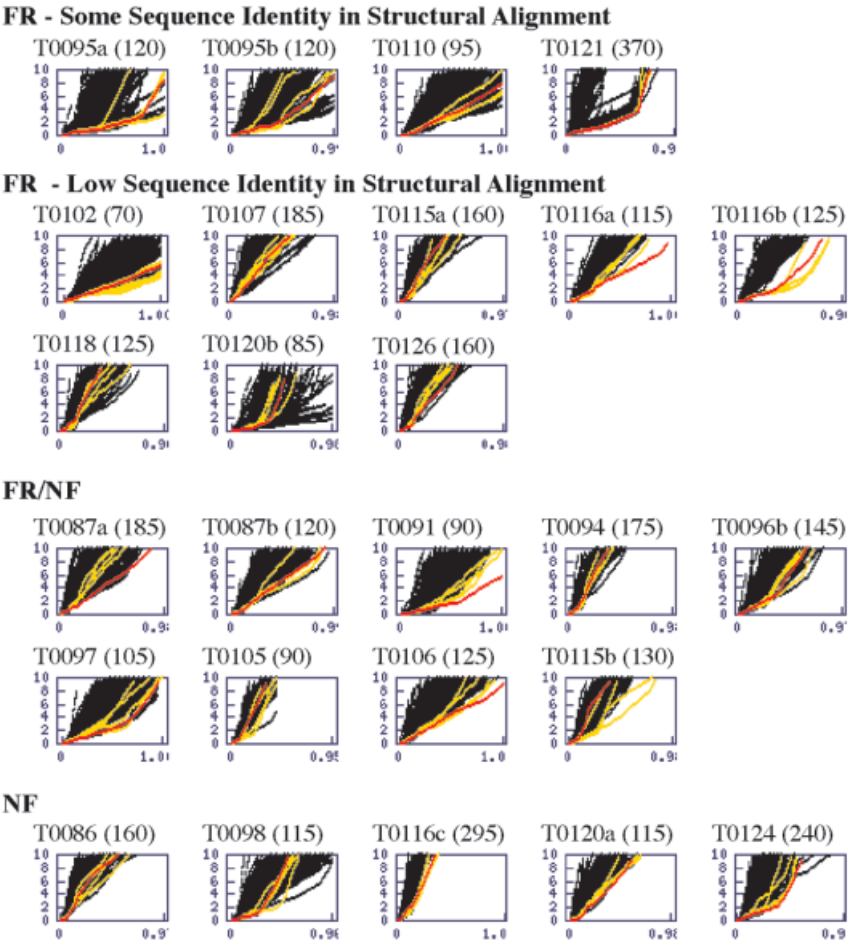

Fig. 2. Hubbard plots. The y-axis is the RMSD in Å to the native structure of the best superimposable region for contiguous segments with length (expressed as a fraction of the length of the native protein) indicated on the x-axis. Our first model is indicated by the red line in each plot, as models 2 to 5 are indicated in yellow and the predictors of all other groups are in black. For multidomain targets, plots corresponding to the full length and to each domain are shown. Targets are grouped into four categories: (a) fold recognition (FR) with moderate sequence identity within the structural alignment, (b) FR with little sequence similarity between target and the correct template within the structural alignment, (c) FR with no significant sequence similarity, and (d) new folds (NF) structures with previously unknown folds. For none of these targets was PSI-BLAST able to detect the correct fold in the PDB.

simulations, two cyclic permutants of the sequence were folded (one spanning residues 1–70 and one composed of residues 7–70 followed by residues 1–6). Conformations with large distances between the N- and C-termini were removed from the decoy populations, side chains were added, and the conformations were relaxed by using the all-atom energy function and the N-to-C-termini constraint. The relaxed filtered populations converged strongly, upon simultaneous clustering of the two cyclic permutants, and the top five unique clusters were submitted after manual inspection. Our fourth model for this target was closer to the correct native structure than was the closest structural homolog (1NKL) and was also sufficiently close to 1NKL to select it from the protein database prior to the native structures release. This result demonstrates the power of Rosetta combined with even limited experimental information, and the identification of the functionally related protein 1NKL, which has a completely unrelated sequence, illustrates how Rosetta predictions may be useful for functional annotation.

T105 (Small α-β)

Although target 105 is only 95 residues and the strand arrangement filter and CO filter passed >50% of the decoys, the conformations produced by our initial simulations clustered poorly. Our best prediction was only correct for 53 residues corresponding to the local four stranded sheet present in this protein, consistent with the general observation that the extent of clustering is a reasonable predictor of model quality.

T091 (Small α-β Homodimer)

The strand arrangement and CO filters passed a large percentage (~80%) of the decoys generated for target 91 due to the limited number of possible strand arrangements. Multiple homologs were folded and simultaneously

clustered, producing several tight clusters. The top five unique clusters were submitted, with no manual intervention. The superposition in Figure 3 shows that our best submitted model (4.9 Å over the full length of the native structure) has a global fold almost indistinguishable from the native structure. The fold of target 91 is novel, and thus our predictions were considerably better than predictions made using fold recognition methods.

T110 (Small α-β)

Although several other groups correctly aligned this target to a known fold, our predictions for this protein were generated by using Rosetta and no template. Our ab initio predictions, the top five unique cluster centers, were of comparable quality to those generated via fold recognition-alignment methods.

T106 (Nonspecific Disulfide Constraints, Medium α)

Target 106 was known to be secreted, and we thus assumed that the 10 absolutely conserved cystine residues participated in disulfide bonds. We independently folded six homologs along with the query sequence and selected conformations with at least four pairs of cysteines close together. These conformations were then relaxed using the all-atom potential supplemented with a cysteine pairing term. Our first model properly paired all five cysteines and was correct to within 6.4 Å over 106 residues, including large loop regions with surprising similarity to the native structure (see superposition in Fig. 3), despite the absence of any structurally related protein with such loops in the protein structure database. The success of the procedure used for this target has led us to investigate the general applicability of this protocol for secreted proteins containing multiple cysteines.

T095, T124 (Large α)

For target 95, several sequence homologs as well as the exact target sequence were folded independently and clustered simultaneously. Target 95 clustered very tightly over the first 130 residues and over the last 110 residues but did not cluster when the global RMS between decoys was used as a distance metric. The top three unique cluster centers from the clustering based on the first 130 residues and the top two clusters from the clustering based on the last 110 residues were submitted. For both regions of target 95, good fragments with lengths >100 residues were submitted (Fig. 3 and Table I). A model that was generated but not submitted had an RMSD of 3.8 Å over 178 residues (Table I). Our best prediction for target 124, one of the best made for this target (Fig. 2), had an RMSD of 5.0 over 117 residues, comprising two very long helices.

T096-B (Large α Plus Homology Domain)

For this protein, we folded the A domain via our comparative modeling procedure and the B domain by using Rosetta. Little effort was devoted to the proper assembly of the domains, because we deemed this beyond the state of the art, given the errors inherent in both our ab initio and comparative modeling methods. A 63-residue segment of domain B was predicted with sufficient accuracy to recognize the correct fold with a Dali Z-score of 3.5.[21–23] Significantly larger correct predictions existed in our decoy populations, demonstrating again that for small proteins and low contact order proteins the most pressing current problem is discriminating between correct and incorrect decoys.

Little manual intervention was used for the targets described above. For these small, low complexity targets, our clustering procedure largely produced protein-like cluster centers that were submitted with no modification or reordering of cluster ranks. For the targets described below (larger, more complex folds), Rosetta tended to converge less strongly, and the cluster centers were reordered by manual inspection prior to submission.

T087 (Two-Domain α-β)

Target 87 was parsed into two domains based on the PSI-BLAST-generated multiple sequence alignment, and predictions were generated for the first 164 residues (domain A) and the last 194 residues (domain B) separately. For these domains, generating and filtering very large numbers of structures using the strand arrangement filter and CO filter were absolutely crucial. Our good predictions (submitted as model 1 for domain A and model 3 for domain B), although in the top 20 cluster centers, were not in the top five cluster centers, and only through manual intervention did we produce good predictions for these domains. β-sheet containing proteins of this size (164 and 192 residues in a two-domain context) are clearly on the cusp of what Rosetta can predict and beyond what Rosetta can predict automatically without further modification to our protocol. These predictions represent the largest correct blind ab initio predictions of portions of β-sheet-containing structures to date that we are aware of. Our predictions for this target are competitive with those generated by using fold recognition methods (Fig. 2).

T116 (A Four-Domain Homodimer)

Target 116 is an 811-residue DNA repair protein that forms homodimers (with four domains per monomer) that completely surround double-stranded DNA. Despite its large size, we were able to separate this protein into five sequence-contiguous domains (domain C was incorrectly parsed into two domains, referred to as C1 and C2 in Table I) and make correct predictions for three of the four sequence-contiguous domains present in the native protein. Domain D predictions were generated by using our comparative modeling technique, whereas predictions for domains A, B, and C were generated by using Rosetta. Domains A and B represent the largest, most complex, topologies we can realistically expect to generate with the Rosetta protocol for CASP4. The strand arrangement filter combined with the CO filter (which together removed all but 10% of the initial population) and increased sampling were crucial to our success for this domain, whereas the folding of multiple homologs and all-atom relaxation probably played a lesser role (due to the rarity of generating correct topologies). Our top models for T116A and T116B (Fig. 3) were considerably better than other predictions for these domains (Fig. 1). The predictions for domain B were

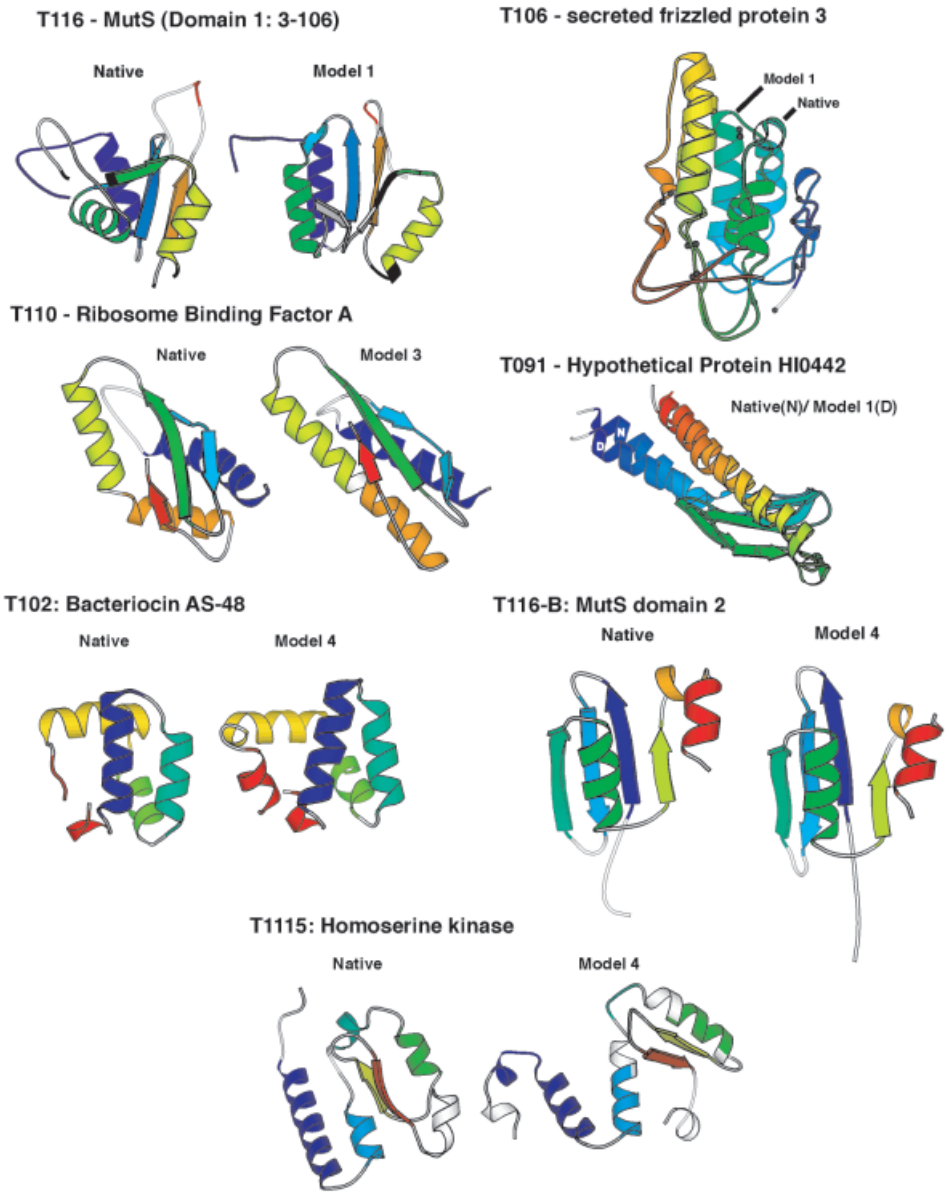

Fig. 3. Comparison of predicted and native structures. Corresponding sequence regions are colored identically in the model and in the native. Uncolored regions correspond to regions over which our prediction was incorrect. For target 106 and 91, the native and best model are shown as a superposition. Properly paired cysteines are shown as black spheres for target 116. For target 116-A, regions of our models corresponding to regions of missing density in the native structure are shown in black.

closer to the native structure than were any of the possible template structures in the Ribonuclease-H fold family.

T086, T094, T115, and T126 (Large Complex α-β Domains)

The protocol used for generating models for these targets was identical to the protocols used on targets 87-A and B and targets 116-A and B. Table I shows that for all of these targets structures were generated with large low RMSD segments, but these segments were generated too infrequently to be detected by the clustering procedure even after the enrichment step. For Target 115, we were able to produce a 71-residue segment within 4.6 Å RMSD to native (Fig. 3) despite the fact that relatively few simulations were carried out because the CASP documentation stated that the protein contained known folds.

T107 (Large β)

This protein had the highest CO native and, with 188 residues, was simply too large and too complex for Rosetta to fold. Because of the difficulty we experienced getting simulations to even pass the lower CO and upper Rg filters, during preliminary runs, efforts to predict this protein were stopped after just 6,000 decoys were made. Nothing good was submitted or generated, largely because of the breakdown of our search strategy.

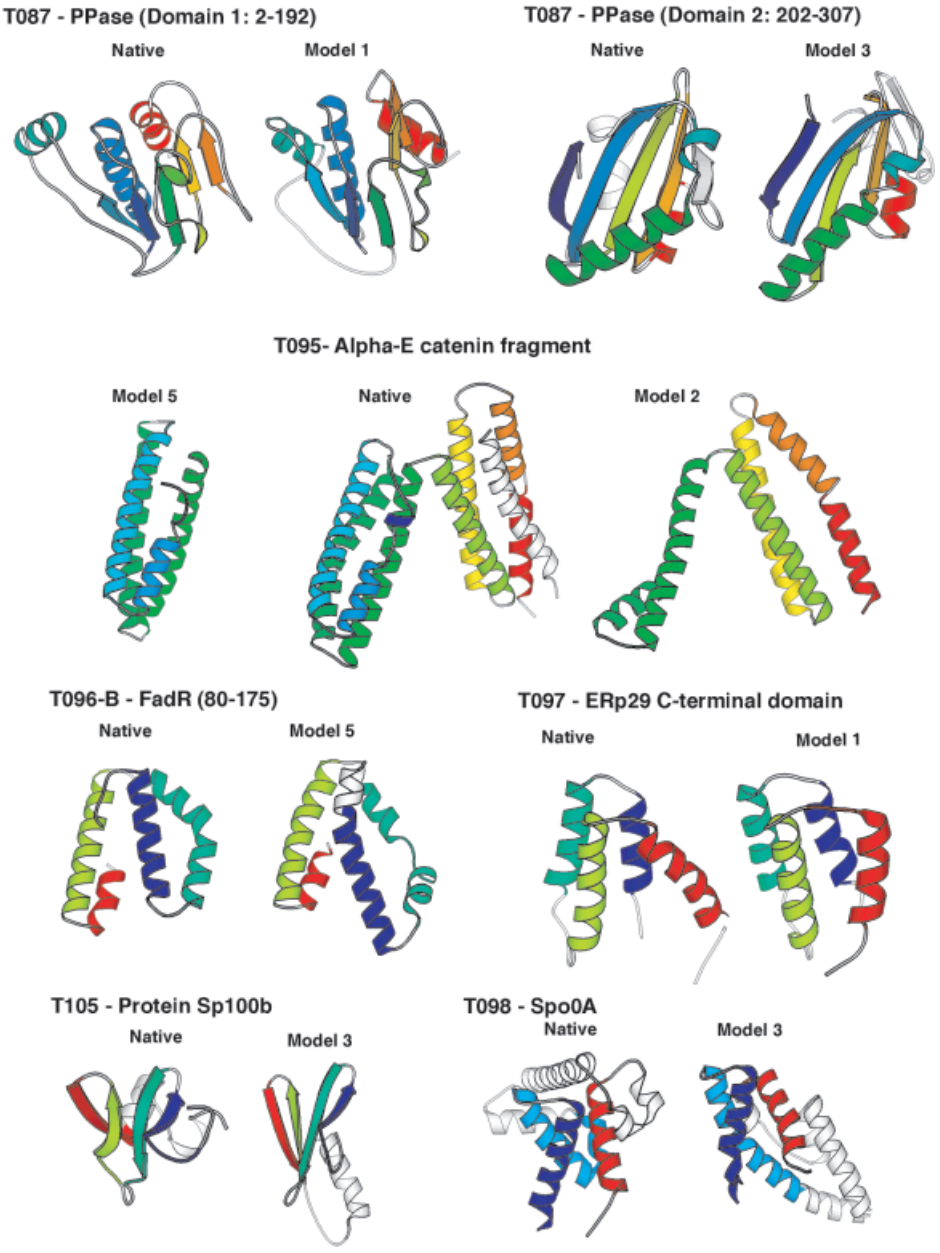

Figure 3. (Continued.)

T120, T116-C (Incorrect Domain Parsing)

For these targets, the quality of our predictions was severely compromised by our incorrect parsing of the query sequences into subdomains. The automatic parsing of large ORFs into domains, a crucial first step to any structural genomics project, is something we and other groups are currently working on. Small (~50 residue) correct fragments were submitted, despite this mistake, for the C domain of target 116.

CONCLUSIONS

The above targets can be roughly separated into two categories: those proteins for which the search strategy was the limiting factor (large, complex proteins) and those proteins for which decoy discrimination was the limiting factor (small alpha and alpha-beta proteins). Our pre-CASP4 efforts were divided roughly equally between dealing with search and discrimination limitations.

Our increased success on larger proteins was due to the improvements in the basic simulation procedure, which allowed more complete sampling in less CPU time, and the contact order and strand arrangement filters, which eliminated a large fraction of the non-protein-like conformations. Although these developments clearly extend the reach of the method, more extensive conformational searching is needed if Rosetta's size and complexity limits are to be extended. A promising observation is that the extent to which simulations converge (the clustering threshold) is

correlated with the CO of the native state. Thus, in the future it should be possible to focus more conformational sampling on the higher contact order structures and perhaps improve performance on this class of proteins, to which most of our CASP4 failures belong. The clustering threshold is also correlated with model accuracy, allowing for the assessment of the reliability of predictions based on protein length, secondary structure class, and clustering threshold.

For smaller, less complex proteins, our increased success was due to a combination of the all-atom relaxation and the folding and clustering of multiple homologs. These improvements decreased the probability of incorrect convergences and thus increased the likelihood of submitting correct predictions. For small proteins (e.g., T098, T097, and T102), the most important areas for improvement involve further increasing model reliability and accuracy, which is likely to require improvements in all-atom sampling methods and more accurate potential functions.

For small α and α-β proteins, our procedure was essentially automated, whereas for larger, more complex targets, for which our simulations did not converge, there was some manual intervention used to select more protein-like structures. Where the manual intervention helped or hurt is not clear, but our predictions for these larger targets were not outstanding.

Rosetta outperformed traditional fold recognition algorithms on several targets that were classified by assessors as fold recognition targets (as it did during CASP3).[24] For cases like T102 and the second domain of T116, our models were closer than the closest template in the PDB, indicating that the lack of dependence on a template, while limiting the size and complexity of the structures that can currently be predicted, can allow substantial increases in model accuracy. The globally correct models for proteins with novel folds, such as targets 91 and 106, shows that the building up of protein structures from fragments does not limit the procedure to structures which have been seen before.

The results reported here show dramatic progress in ab initio structure prediction in the 4 years since CASP2, where it was concluded that such methods were unlikely to progress to the point of being useful in a reasonable time frame.[25] We believe that the CASP process itself—with the invaluable objective evaluations and challenges it has provided—has played a critical role in catalyzing much of this progress.

ACKNOWLEDGMENTS

We thank all of the structural biologists who contributed structures to CASP4 and the organizers who made CASP4 possible. We thank Brian Kuhlman, David Kim, Eric Alm, and all members of the Baker laboratory for useful discussions during CASP4. We also thank Keith Laidig and Formix for effective and innovative administration of the computer resources necessary during CASP4. R.B. acknowledges support from a Howard Hughes Predoctoral fellowship. J.T. acknowledges the support of an NSF biological informatics fellowship. This work was supported in part by young investigator awards from the NSF and the Packard foundation and by the HHMI.

REFERENCES

1. Han KF, Bystroff C, Baker D. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci 1997;6:1587–1590.
2. 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.
3. 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.
4. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
5. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA [In Process Citation]. Proteins 1999;Suppl 3:171–176.
6. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
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;37:121–125.
9. Gunn JR. Hierarchical minimization with distance and angle constraints. Proc Int Conf Intell Syst Mol Biol 1998;6:78–84.
10. Bonneau R, Strauss C, Baker D. Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 2001;43:1–11.
11. 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.
12. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M. CAFASP-1: critical assessment of fully automated structure prediction methods [In Process Citation]. Proteins 1999;Suppl 3:209–217.
13. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
14. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
15. Dunbrack RL, Jr, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1:334–340.
16. Dunbrack RL, Jr, Karplus M. Backbone-dependent rotamer library for proteins: application to side-chain prediction. J Mol Biol 1993;230:543–574.
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. Hubbard TJ. RMS/Coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. Proteins 1999;37:15–21.
19. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction [In Process Citation]. Proteins 1999;Suppl 3:149–170.
20. Gonzalez C, Langdon GM, Bruix M, Galvez A, Valdivia E, Maqueda M, Rico M. Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin [In Process Citation]. Proc Natl Acad Sci USA 2000;97:11221–11226.
21. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
22. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19:256–268.
23. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
24. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets [In Process Citation]. Proteins 1999;Suppl 3:88–103.
25. Lesk AM. CASP2: report on ab initio predictions. Proteins 1997;Suppl 1:151–166.