PROTEINS: Structure, Function, and Genetics Suppl 3:171–176 (1999)

AB INITIO: PREDICTION REPORTS

Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA

Kim T. Simons,[1] Rich Bonneau,[1] Ingo Ruczinski,[2] and David Baker[1]*
[1]*Department of Biochemistry, University of Washington, Seattle, Washington*
[2]*Department of Statistics, University of Washington, Seattle, Washington*

ABSTRACT **To generate structures consistent with both the local and nonlocal interactions responsible for protein stability, 3 and 9 residue fragments of known structures with local sequences similar to the target sequence were assembled into complete tertiary structures using a Monte Carlo simulated annealing procedure (Simons et al., J Mol Biol 1997; 268:209–225). The scoring function used in the simulated annealing procedure consists of sequence-dependent terms representing hydrophobic burial and specific pair interactions such as electrostatics and disulfide bonding and sequence-independent terms representing hard sphere packing, α-helix and β-strand packing, and the collection of β-strands in β-sheets (Simons et al., Proteins 1999;34:82–95). For each of 21 small, ab initio targets, 1,200 final structures were constructed, each the result of 100,000 attempted fragment substitutions. The five structures submitted for the CASP III experiment were chosen from the approximately 25 structures with the lowest scores in the broadest minima (assessed through the number of structural neighbors; Shortle et al., Proc Natl Acad Sci USA 1998;95:1158–1162). The results were encouraging: highlights of the predictions include a 99-residue segment for MarA with an rmsd of 6.4 Å to the native structure, a 95-residue (full length) prediction for the EH2 domain of EPS15 with an rmsd of 6.0 Å, a 75-residue segment of DNAB helicase with an rmsd of 4.7 Å, and a 67-residue segment of ribosomal protein L30 with an rmsd of 3.8 Å. These results suggest that ab initio methods may soon become useful for low-resolution structure prediction for proteins that lack a close homologue of known structure. Proteins Suppl 1999;3:171–176.** © 1999 Wiley-Liss, Inc.

Key words: protein structure prediction; knowledge-based scoring functions; fragment assembly; critical assessment of structure prediction experiment (CASP)

INTRODUCTION

The picture of protein folding that motivates our approach to protein tertiary structure prediction is that sequence-dependent local interactions bias segments of the chain to sample distinct sets of local structures, and that nonlocal interactions select the lowest free energy tertiary structures from the many conformations compatible with these local biases. In implementing the strategy suggested by this picture, we use quite different models to account for the local and nonlocal interactions. Rather than attempting a physical model for local sequence structure relationships, we turn to the protein database and take the distribution of local structures adopted by short sequence segments (less than 10 residues in length) in known three-dimensional structures as an approximation to the distribution of structures sampled by isolated peptides with the corresponding sequences. Nonlocal interactions in proteins we assume can be reasonably well modeled at the level required for folding without a detailed representation of the side chains. The primary nonlocal interactions we consider are hydrophobic burial, electrostatics, disulfide bonding, main chain hydrogen bonding, and excluded volume; because of uncertainties over parameter values we derive these from known protein structures using a self-consistent Bayesian approach.[1] The lack of detail in the model potentially facilitates identification of the determinants of protein folding, and importantly, makes ab initio structure prediction of small proteins computationally feasible.

METHODS

Prediction of the structures of CASP 3 targets with fewer than 160 residues began with a multiple sequence alignment generated by PSI-BLAST[2] using default parameters. Sequences with less than 25% or more than 90% sequence identity to the target were removed. Structure prediction for those sequences obviously related to proteins of known structure was not attempted (more than 25% sequence identity). The edited multiple sequence alignments were sent to the PHD Web server,[3] and 25 fragments from known structures with sequences and

*Correspondence to: David Baker, Department of Biochemistry, Box 357350, University of Washington, Seattle, WA 98195. E-mail: baker@ben.bchem.washington.edu
Received 5 February 1999; Accepted 21 May 1999

TABLE I. Evaluation of the CASP3 Structure Predictions†

Target	Fold	Length	Success level	Best submitted			Best generated	
				rmsd	Residues	Model	rmsd	Residues
46 γ adaptin	**β**	119	*	4.5	1-37	5	4.9	1-58
56 DNAB	**α**	114	***	4.7	26-103	5	4.7	26-103
59 SMD3	**β**	71		5.3	5-30	4	4.5	5-35
61 HDEA	**α**	76	**				4.0	25-85
63 IF5a	**αβ/β**	135	*/**				5.4	83-138
64 SinR	**α**	102	****	3.0	1-63	2	2.0	1-63
65 SinI	**α**	31	****	3.8	9-39	1	2.0	9-39
71 **α** adaptin	**β/αβα**	125/113	*/				4.4	71-125
74 EPS15	**α**	95	****	6.0	6-100	2	3.8	9-100
75 ETS-1	**α**	110	*	5.3	32-54	4	3.9	57-108
77 L30	**αβα**	104	***	3.8	14-80	4	3.8	14-80
79 MarA	**α**	129	****	6.4	9-109	2	4.3	9-124
81 MGSA	**αβα**	151	***	5.5	23-83	2	3.0	65-114
83 Cyanase	**α/βα**	156	*/	4.8	7-48	5	4.1	1-50

†The classification used in the "Success level" column is nothing correctly generated or submitted (no stars), large fragment correct in a generated structure (*), globally correct generated structure (**), large fragment correct in a submitted structure (***) and globally correct submitted structure (****). For each target, the longest low-rmsd segments (all rmsds reported in this article are for standard sequence-dependent structural superpositions) in the five submitted and the 1,200 generated are described in the "Best submitted" and "Best generated" columns (for example, residues 1-37 in model 5 for target 46 had an rmsd to native of 4.5 Å). For target 66, only the portion of target 64 that interacts with target 65 was considered.

secondary structures similar to the sequences and predicted secondary structure of each three- and nine-residue segment of the target sequence were identified using a scoring function similar to that of Fischer and Eisenberg.[4] The secondary structure of the fragments is biased toward but not constrained to that suggested by the secondary structure prediction. Protein tertiary structures were generated from these sets of three- and nine-residue fragments using the fragment insertion-simulated annealing strategy described previously[5] with 100,000 attempted fragment insertions per structure. The scoring function is composed of sequence-dependent terms representing hydrophobic burial, electrostatics, and disulfide bonds and sequence-independent terms describing α-helix and β-strand packing and the assembly of β-strands into β-sheets.[2] A term linear in the radius of gyration was added during the annealing process to generate compact structures, but it was not used in the final evaluation of the structures. For each target, 1,200 structures were generated, for a protein of 120 residues. This took approximately 2 days on five workstations (Alpha 533 MHz). The top approximately 25 structures with the lowest scores and in the broadest minima (as assessed by the number of the other 1,199 structures within 7 Å rmsd[6]) were visually inspected for single hydrophobic cores, few hydrophobic surface residues, unpaired buried polar amino acids, compactness, and regular supersecondary structures, and the top five structures were submitted. When the experimentally determined structures became available, the DALI Web server[7] was used to identify structural similarity between these structures and the five submitted structures. We call our approach to ab initio structure prediction ROSETTA, after the stone that allowed the deciphering of Egyptian hieroglyphics.[8]

RESULTS AND DISCUSSION

We predicted the structures of 21 of the 43 targets available in CASP3 that did not have obvious homologues with known structures. Of these targets, eighteen experimentally determined structures, which cover the gamut of secondary structure composition, are available for comparison to the predicted structures. When the experimentally determined structures were made available before the CASP3 meeting, we evaluated our predictions by searching for native-like substructures using the DALI server[8] in both the large set of 1,200 structures and the five submitted structures. As indicated in Table I, significant structural matches were found for a number of the targets. After the CASP3 meeting, the predictions made by other groups provided a convenient reference point for evaluating the relative strengths and weaknesses of the method; the comparison is perhaps the best way to take into account the differences in the difficulties of the different targets. We found the graphical summaries prepared by Tim Hubbard[9] particularly useful for this purpose and have reproduced his summaries in Figure 1. Each line in the plots represents one prediction; our predictions are shown in color (red indicates the first prediction).

What Went Wrong

For a number of targets (81, 83a, 54, 46, and 61) our predictions were poor compared with those of other groups. Most of these were larger proteins that turned out to be threading targets; in these cases our de novo conformational searching method could not compete with the vast reduction in the size of the search space implicit in threading methods. Improvement in the search strategy is clearly necessary for proteins of more than 100 residues.

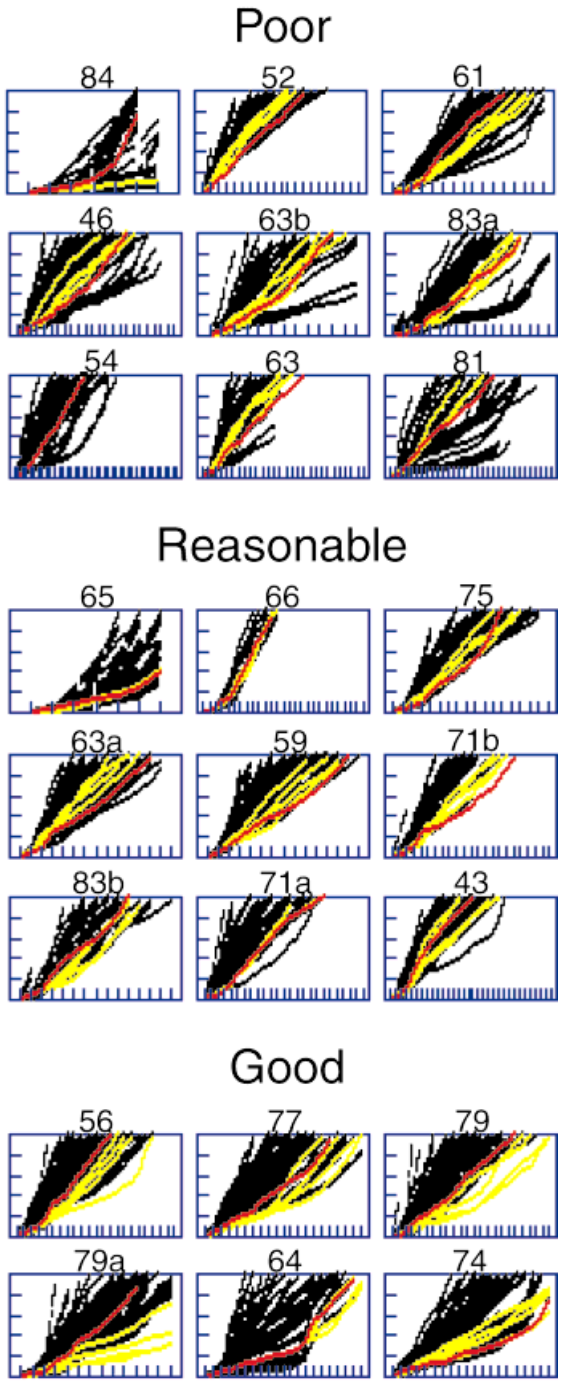


Fig. 1. Graphical analysis of the predictions of CASP3. Each prediction for each target is indicated by a line; the x axis is the maximum number of residues that can be superimposed on the native structure for a given rmsd threshold (y axis). Good models have large numbers of residues superimposable on the native structure for relatively low rmsd cutoffs and thus are represented by the lines closest to the x axis. Black, all models submitted by other groups; red, our model 1; yellow, our models 2–5. Axis labels were omitted to avoid cluttering the figure; fully labeled versions of these figures are available at the CASP3 Web site: http://predictioncenter.llnl.gov/casp3/results/.

The method may also work less well for proteins with atypical local sequence features; for target 52, for example, the local structure prediction was quite poor.

For four targets—cyanovirin (target 52), HPPK (target 43), the second prediction of RLZ (target 84), and VanX (target 54)—variations on the standard method were used to try to incorporate additional information. These were largely unsuccessful; a method for using such additional information is clearly an area for future work.

In cases in which successful predictions were made, our subjective ranking system failed to identify the best predictions in the five submitted structures; the number one predictions were usually relatively poor. This was particularly dramatic for MarA, target 79, in which models 2–5 were all much better than model 1 (we choose model 1 because it was more compact). Better results would probably have been obtained if we had resisted the temptation to intervene at the last step and completely automated the selection procedure.

What Went Wrong on an Absolute Scale But Not so Bad on a Relative Scale

Our internal evaluation before the meeting and the postmeeting evaluation in light of the predictions of other groups were generally in accordance with regard to the most successful predictions but differed with regard to the next tier of predictions. The latter set of targets includes 83b, 71a, 43, 59, 71b, 75, 65, 63a, and 66. We considered these predictions poor, but they turned out to be acceptable when the difficulty of the target was considered and measures other than contiguous superimposable residue matching were used (Fig. 1). Thus, the method appears to capture some nontrivial features of these structures even when long, contiguous, low-rmsd fragments were not generated.

What Went Right

Particularly good predictions were made for four targets. The submitted structures were selected on the basis of the number of other structures within 7 Å rmsd and the score. As shown in Figure 2, both measures turned out to be moderately correlated with the rmsd to the native structure. The correlation between rmsd and score (Fig. 2, left panels) is reasonably good given the performance of the scoring function in recognizing native-like structures in large decoy sets[2] (the scoring function was extensively optimized in the structure-generation procedure, and thus, many structures in low-scoring false minima could have been generated). The use of the "number of structural neighbors" statistic to assess the breadth of the free energy minimum populated by a structure[7] proved to have been useful in this blind ab initio structure prediction context: the lowest rmsd structures almost always had a large number of structural neighbors (Fig. 2, right panels).

For two of the four targets, large fragments were predicted reasonably well. For target 77 (L30), the best predicted structure (model 4), which was the third most common topology among the low-energy structures, had a 67-residue fragment with an rmsd to the experimentally determined structure of 3.8 Å (Fig. 3). This and an

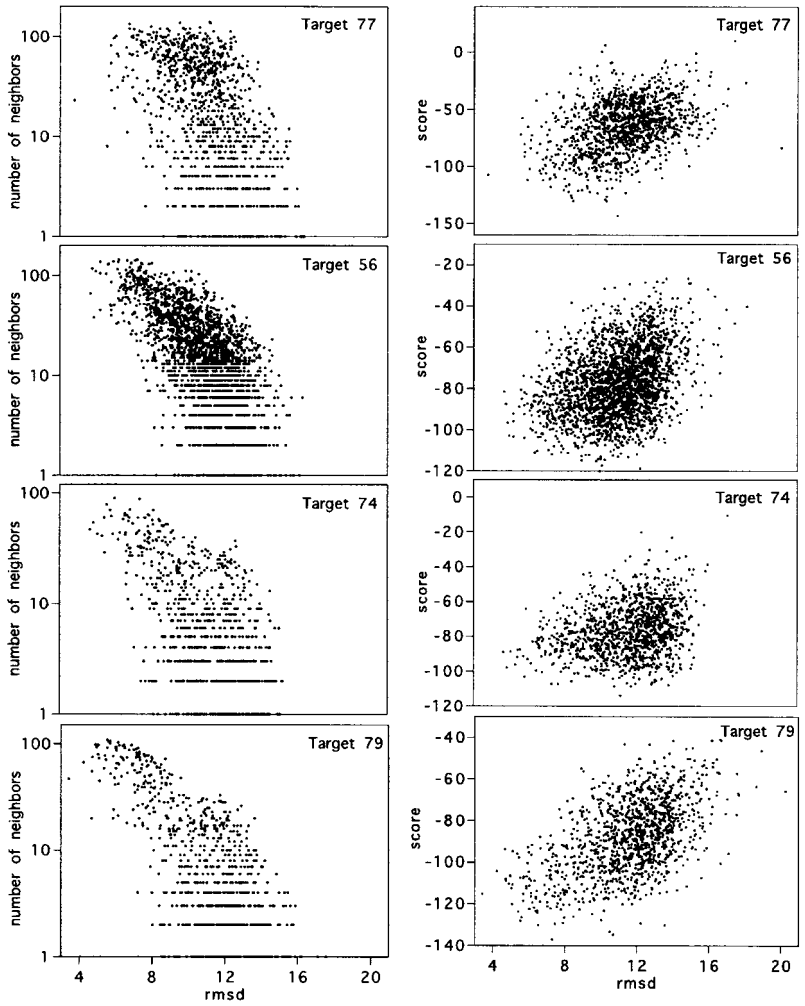


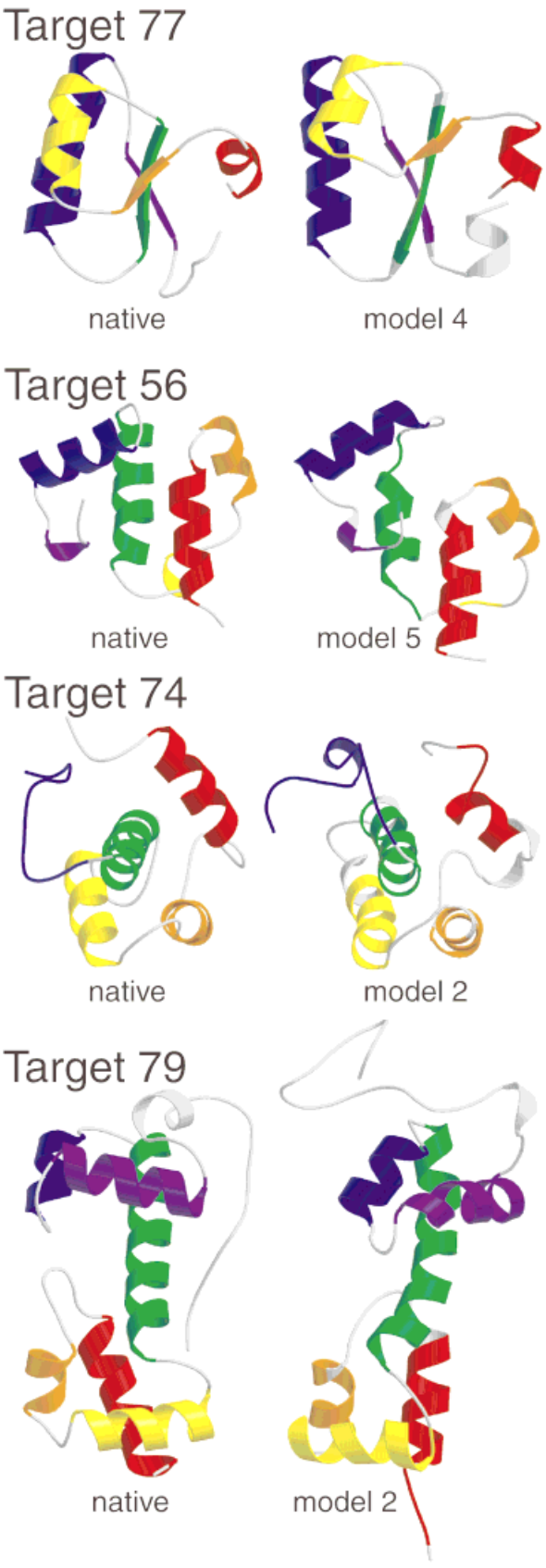
Fig. 2. Score and neighbor density versus rmsd to native. The score using the function minimized in the simulated annealing protocol (left panels) and the number of neighbors within 7 Å rmsd (right panels) for the generated structures are plotted versus the rmsd to native. The rmsd is measured over 67 residues for target 77, 75 residues for target 56, 95 residues for target 74, and 99 residues for target 79.

excellent prediction by the Skolnick group were competitive with the predictions made for target 77 using threading methods (Fig. 1, target 77).

One of our predictions for target 56 (DNAB[8]), a true ab initio target (no similar structures are found in the PDB using DALI[8] or VAST[11]), was quite good over a fairly large segment (Fig. 1 and Fig. 3). This substructure occurred frequently in the set of 1,200 simulated structures (Fig. 2), but surprisingly no globally correct structures were generated. Either the fragment set consistently turns the chain in the wrong direction after the blue helix in Figure 3, or native-like structures are not as low scoring as the alternatives. The secondary structure was known for this target in advance of the prediction, but this information was not used in the generation of the structure shown in Figure 3. As is clear from Figure 1, this prediction is better than any of the other predictions made by us or by others for this target.

For two targets, good predictions were made over almost the entire length of the protein. A large portion of target 74, the EH domain of EPS15[12] (Fig. 3), is structurally similar to a number of calcium-binding proteins; thus, it was a relatively easy threading target. Although this information was not used explicitly in the fragment selection, fragments of EF-hand motifs from known structures contributed to the fragment library and undoubtedly improved the tertiary structure prediction. Our prediction has several interesting features. First, the relative positioning of the helices is better than that of many of the templates used for threading. Second, a rare variant of an EF-hand motif found between the yellow and green helices

Fig. 3. Cartoons of native structures and successful predictions. Molscript[14] and raster3d[15] were used to create ribbon diagrams of target 77 (67 residues of 104 total), target 56 (75 residues of 114 total), target 74 (all 95 residues) and target 79 (all 116 residues). From N- to C-terminal, the order of secondary structures in the structural comparisons is red, yellow, orange, green, blue, and purple. The colored segments are specified by the native secondary structure.



is well modeled. Finally, the structurally unique C-terminal portion of the structure is reasonably well positioned (blue colored segment; the threading predictions are truncated because the C-terminal part of the molecule is not found in previously determined structures). Our success on this target suggests that a combination of our method and conventional threading/homology modeling methods may be quite useful for predicting structures in cases in which a subset of the protein is clearly related to a protein of known structure.

Our most successful prediction was for target 79, the transcriptional activator MarA.[13] The correct fold occurred very frequently in the set of simulated structures (Fig. 2); our failure to rank it number one attests to the weakness of our "visual inspection" based ranking of the top five structures. Models 2 and 4 were considerably better than any of the other predictions made for this target, both over the entire length and throughout the first domain (Fig. 1, targets 79 and 79a). Importantly, our predictions for this target are good enough to have functional implications; the two domains contain helix-loop-helix motifs that are spaced appropriately to bind DNA (as pointed out by Alexey Murzin, this should have made it obvious that models 2–5 were better than model 1). Fragments of DNA-binding proteins appear not to have contributed to the final conformations: the twenty largest contributors of fragments are not DNA-binding proteins.

What We Learned and Future Directions

1. Low-resolution structure prediction can succeed without explicit representation of side chains and side chain-side chain packing. However, modeling of side chains may help improve the selection and generation of better predicted structures.

2. Human intervention can be bad (cf target 79); all steps of the method should be completely automated.

3. At the CASP3 meeting there were complaints that the method was "nonphysical." We strayed from the physically plausible picture of folding described in the introduction by using multiple sequence information in the fragment selection. It is unclear how much multiple sequence information contributed to the predictions, but future work aimed at modeling folding will clearly have to start with single sequences and preferably reduce the reliance on the protein database as a source of parameter estimates.

4. The target 74 and 79 results suggest that the method, perhaps with additional information from threading/sequence searches implemented in the form of pseudopotentials or by fixing large substructures, potentially can be useful for threading/homology modeling in cases of very low sequence similarity and for large insertions/extensions not included in the template structure.

ACKNOWLEDGMENTS

We thank the organizers and assessors of the CASP3 experiment for their valuable contributions to the structure prediction field. We thank Alexey Murzin for numerous discussions and Tim Hubbard for the plots in Figure 1. We also thank Chris Bystroff, Ed Thayer, Vesteinn Thors-

son, and Charles Kooperberg for discussions and help with multiple sequence alignment preparation. This work was partially supported by a fellowship to K.T.S. from the Public Health Service, National Research Service Award T32 GM07270 from National Institute of General Medical Sciences; a Howard Hughes Medical Institute Predoctoral fellowship to R.B., the NSF Center for Molecular Biotechnolgy; and young investigator award to D.B. from the NSF and the Packard Foundation.

REFERENCES

1. Simons KT, Ruczinki I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
2. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
3. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
4. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.
5. 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.
6. 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.
7. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
8. Encyclopedia Brittanica, 1998.
9. Zemla A, Venclovas C, Reinhardt A, Fidelis K, Hubbard TJ. Numerical criteria for the evaluation of ab initio predictions of protein structure. Proteins Suppl 1997;1:140–150.
10. Weigelt J, Brown SE, Miles CS, Dixon NE, Otting G. NMR structure of the n-terminal domain of *E. coli* DNAB helicase: implications for structure rearrangements in the helicase hexamer and its biological function, to be published.
11. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
12. DeBeer T, Carter RE, Lobel-Rice KE, Sorkin A, Overduin M. Structure and Asn-Pro-Phe binding pocket of the Eps15 homology domain. Science 1998;281:1357–1360.
13. Rhee S, Martin RG, Rosner JL, Davies DR. A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. Proc Natl Acad Sci USA 1998;95:10413–10418.
14. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.
15. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.