PROTEINS: Structure, Function, and Genetics Suppl 5:127–132 (2001)
DOI 10.1002/prot.1171

Predicting Novel Protein Folds by Using FRAGFOLD

David T. Jones
Department of Biological Sciences, Brunel University, Uxbridge, Middlesex, United Kingdom

ABSTRACT **The results of applying a fragment-based protein tertiary structure prediction method to the prediction of 8 CASP4 targets are described. The method is based on the assembly of supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. Despite the significant degree of success in this case, there is clearly much more developmental work required before predictions with the accuracy of a good homology model, or even a good fold recognition model, can be made with use of this kind of approach. Proteins 2001;Suppl 5:127–132.**

INTRODUCTION

Although there are now a number of useful methods available for predicting the structure of a given protein when the native fold is already represented in the current structural data banks, the problem of predicting the structure of proteins with novel folds still persists. In many cases, however, even when a new fold is discovered, it is observed that many new folds are still composed of common structural motifs at the supersecondary structural level. In this article a method, called FRAGFOLD, is described for building up novel protein structures from recurrent supersecondary structural motifs. The strategy described attempts to greatly narrow the search of conformational space by preselecting supersecondary structural fragments from a library of highly resolved protein structures. Since we first published the results of using FRAGFOLD in the 2nd CASP experiment held in 1996,[1] we and others have been striving to make improvements to fragment-assembly methods for protein structure prediction, and it has been clear at both CASP3 and now at CASP4 that improvements have been made.

MATERIALS AND METHODS

The method used in CASP4, implemented in a program called FRAGFOLD, is based on an earlier method[1] that was relatively successful in the CASP2 experiment held in 1996. Rather than describing the method in detail again, here the main differences with the original method are described.

As with the original method, the core of the objective function used here is a set of pairwise potentials of mean force, determined by a statistical analysis of highly resolved protein X-ray crystal structures and the application of the inverse Boltzmann equation, along with a solvation potential. These potentials are identical to those currently in use by the latest versions of our threading programs: THREADER[2] and GenTHREADER.[3]

For threading applications, pairwise and solvation potentials are sufficient on their own to discriminate correct from incorrect protein folds. However, for ab initio prediction, it is also necessary to include extra terms to ensure that low-energy folds are compact, have optimal hydrogen bond networks and have no steric clashes. In threading, these additional terms are unnecessary because real protein folds are almost always compact, have no steric clashes, and have well-defined hydrogen bonding networks.

A chain fold is described as compact if the maximum distance between any pair of α-carbons ($d_{\max}$) is less than $R_{\max}$, which is defined as follows:

$$R_{\max} = \frac{5}{2}\left(\frac{3N}{0.026\pi}\right)^{1/3}$$

where N is number of residues in the protein. Noncompact folds are penalized by the term:

$$E_{\text{compact}} = d_{\max} - R_{\max}$$

Steric clash penalties are also defined in a simple fashion. For a pair of atoms in residue types a and b, separated by a distance d^{ab}

$$E_{\text{clash}} = (D^{ab}_{\min})^2 - (d^{ab})^2 \quad \text{when } D^{ab}_{\min} > d^{ab}, \text{ or}$$
$$E_{\text{clash}} = 0 \quad \text{otherwise,}$$

where $D^{ab}_{\min}$ is the minimum observed distance between residues ab in a set of highly resolved structures. Both Cβ–Cβ and Cα–Cα distances are considered.

Long-range main chain hydrogen bonding is represented by considering ideal Cα geometries of chain segments involved in either parallel or antiparallel sheets. A hydrogen bond is assumed between residue i and residue j where j > i + 5 and distances $D^{i,i\pm1}_{j,j\pm1}$ where D^i_j is the distance between Cα atoms in residues i and j) fall within the bounds derived from sheets in highly resolved structures. A maximum of two hydrogen bonding interactions are allowed per residue. The term E_{hbond} is taken as the number of residues involved in at least one and no more than two hydrogen bonded bridges.

The above energy terms are applied to a simplified representation of the polypeptide chain in which only

D.T. Jones' present address is Bioinformatics Unit, Department of Computer Science, University College London, Gower Street, London WC1E 6BT United Kingdom.

Received 5 April 2001; Accepted 2 July 2001

main-chain heavy atoms and β-carbon atoms are considered and summed thus:

$$E_{\text{total}} = W_1E_{\text{short-range}} + W_2E_{\text{long-range}} + W_3E_{\text{solv}} + W_4E_{\text{steric}} + W_5E_{\text{hbond}} + W_6E_{\text{compact}}$$

where W_{1-6} are adjustable weights, $E_{\text{short-range}}$ represents the sum of mean force pair potential terms for residue separations of 6 or less, $E_{\text{long-range}}$ represents the sum of mean force pair potential terms for residue separations of 7 or more, and E_{solv} represents the sum of single residue solvation energies.[1]

The above potential terms are calculated across multiply aligned sequences, rather than just a single sequence. In this case, the energy terms are calculated for each sequence in the aligned family of sequences fitted onto the given main-chain trace and then averaged. For the steric terms, however, the maximum of the steric energies is taken so that a favorable main-chain conformation is compatible with all the aligned sequences in the family.

Preselection of Fragments

The first stage of the folding simulation involves the selection of favorable supersecondary structural fragments at each residue position along the target sequence. Supersecondary structures are defined by taking two or three sequential secondary structures from a library of protein structures. Currently, the following supersecondary structures are defined:

α-*hairpin*	consecutive α-helices in a compact arrangement
α-*corner*	consecutive α-helices in a noncompact arrangement
β-*hairpin*	hydrogen-bonded consecutive β-strands
β-*corner*	non-hydrogen-bonded consecutive β-strands
β-α-β *unit*	parallel hydrogen-bonded β-strands with intervening α-helix
split β-α-β *unit*	parallel non-hydrogen-bonded β-strands with intervening α-helix

The fragment selection stage of the folding procedure involves the summation of pair potential terms and solvation terms for the target sequence (and aligned homologues) threaded onto each supersecondary motif, at each position in the sequence. So for a target sequence of length L, and a motif of length M, $L - M + 1$ ungapped threadings are considered. A new feature of fragment selection used for the CASP4 predictions was to skip any threadings that contradicted the reliable regions of predicted secondary structure. Secondary structure was predicted with an updated version of PSIPRED,[4] which is a simple, yet accurate secondary structure prediction method, incorporating two feed-forward neural networks that perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated-BLAST).[5] Version 2.0 of PSIPRED includes a new algorithm that averages the output from up to four separate neural networks that have been trained on different training set subsets to increase prediction accuracy. The PSIPRED predictions for all of the CASP4 targets were also submitted in the secondary structure prediction category and achieved an average Q_3 score of 80.6% across the 40 submitted target domains with no obvious sequence similarity to structures present in PDB.[6] These predictions, along with the associated PSI-BLAST multiple sequence alignments, were used as inputs to FRAGFOLD. Note that apart from biasing the selection of fragments, secondary structure prediction information was not used elsewhere in FRAGFOLD. The objective function does not currently include terms that relate to these secondary structure constraints.

In addition to the sequence-specific fragment list, a general fragment list is also constructed from all tripeptide, tetrapeptide and pentapeptide fragments from the library of highly resolved structures. These smaller fragments are not preselected.

Having selected the starting fragment lists, a single folding simulation progresses in the following way. First, a random conformation for the target sequence is generated by selecting fragments entirely randomly. Fragments are spliced by superposing the α-carbon and the main-chain nitrogen and carbonyl-carbon atoms of the C-terminus of one fragment on the equivalent atoms of the N-terminus of the other fragment. Each randomly selected fragment is spliced onto the end of the growing chain until all N residues have been covered. Having generated a random conformation for all N residues, a simple steric check is performed, and the conformation is rejected if any pair of atoms is closer than a predetermined minimum distance. A residue-specific table of minimum distances was used, which was compiled from a set of highly resolved protein structures (resolution was better than 1.5 Å). If parts of the randomly generated chain overlap according to the table of minimum distances, then the conformation is rejected and another randomly generated conformation is selected by using the same procedure. This continues until the starting conformation has no steric clashes.

Before the simulation starts, it is necessary to calculate weights for the components of the energy terms. To select an optimum set of energy component weights, a number of random-chain conformations are generated by using the above procedure. However, after generating a conformation with no clashes, the conformation is modified by randomly selecting a random fragment conformation from the fragment lists. For each of these conformations, the component energy terms are calculated, and standard deviations of each component are calculated. The ratios of these standard deviations to that of the short-range pairwise potential component are used as weights.

Given a random starting conformation and an appropriate set of weights on the component energy terms, the simulation proper is started. A simulated annealing approach is used to minimize the energy function. A random move is made by either selecting a locally optimum (supersecondary structure) fragment from the 10N lists of preselected fragments at each position in the target sequence, or a completely free choice is made from the additional list of small fragments. Half of the moves made involve a locally

TABLE I. A. Summary of the First Models Submitted to the CASP4 Server

Target	Model	Fold class	Length of target/domain	Overall RMSD	Length of best fragment	RMSD of best fragment	Z-score
T0094	1	α + β	177	16.97	34	4.78	1.8
T0096	1	α	155	17.43	64	4.92	7.6
T0097	1	α	105	9.76	59	4.99	6.6
T0098	1	α	119	14.16	60	4.83	6.8
T0102	1	α	70	4.90	70	4.90	8.8
T0110	1	α + β	95	6.95	68	4.97	8.4
T0126	1	α + β	162	19.67	33	4.78	1.6
T0127	1	α + β	130	15.75	46	4.99	4.1

TABLE I. B. Other Good Submitted Models

Target	Model	Fold class	Length of target/domain	Overall RMSD	Length of best fragment	RMSD of best fragment	Z-score
T0096	2	α	155	17.03	66	4.97	8.0
T0097	4	α	105	11.25	61	4.99	7.0
T0098	3	α	119	10.68	61	4.83	7.0
T0126	2	α+β	162	18.62	46	4.98	4.1
T0127	2	α+β	130	13.23	49	4.99	4.7

RMSDs are calculated by using the standard expression: $\sqrt{\sum_{i=1}^{n} d_i^2/n}$ where d_i is the distance between equivalenced atoms at position i in two structures of length n.
The "best fragments" are defined as the longest contiguous segments in both predicted and experimental structures which can be superposed to < 5.0 Å RMSD.
Z-scores (number of standard deviations above the mean) were calculated by using statistics obtained from superposing fragments of protein pairs that have different overall folds. The longest fragments from these dissimilar protein pairs, which could be superposed to <5.0 Å, were found to have a mean length of 24.6 residues, with a standard deviation of 5.2 residues.

optimum fragment, and half involve a free selection from the small fragment list. In this way, 50% of the time, the random move will generate a locally optimum structure at the given position of the target sequence, and 50% of the time a completely random local conformation will be generated. The rationale here is that there is a very high chance that a piece of the target protein sequence will be folded into a locally optimum supersecondary structure in any single move. This single move represents a large number of possible small moves that a protein might, hypothetically at least, make very early on the folding process. This allows the very crude simulation presented here to at least represent the current thinking on rapid local nucleation of structure during the folding process. Computationally, this allows a great deal of conformational space searching to be bypassed, because it is not necessary to construct supersecondary structures from a large number of small single torsion angle changes to the chain fold.

A standard simulated annealing algorithm is used in running the simulation. FRAGFOLD has the option of using a genetic algorithm as its optimization algorithm rather than simulated annealing, but this option was not used here. Random moves are made as detailed above but are only accepted if the Metropolis criterion is met. The starting temperature for the simulation is selected by making 500 random moves to the starting conformation and calculating the largest absolute energy change between any two moves. The simulation is started at a temperature corresponding to 10 times this ΔE, that is, from $E = kT$, $T_{start} = 10\ \Delta E/k$. The temperature is halved after either 5,000 random moves have been accepted by the Metropolis criterion, or a total of 50,000 moves have been tested. When every random move is rejected by the Metropolis criterion, then it is assumed that the current structure is "frozen." At this point, the temperature is set to zero, and another 50,000 random moves made to allow the system to be "quenched."

Selection of Final Structures

For each prediction target, 20 separate simulations (using different random number seed values) were carried out simultaneously across a small Linux-based compute farm. The final 20 conformations were clustered automatically by using the NMRCLUST program,[7] running with default parameters. The representatives of the five most populated clusters were submitted to the CASP4 assessment. In some cases, however, fewer structures were submitted because lesss than five clusters were produced.

RESULTS

The methods described above were applied to eight of the CASP4 targets. These targets were selected as those that seemed most likely to be novel folds on the basis of fold recognition results from THREADER and GenTHREADER.[8] As described, up to five predictions were submitted for each target based on the results of the structure based clustering. Table I summarizes the submitted predictions for the eight targets, and Figure 1 shows the best submitted fragments for each.

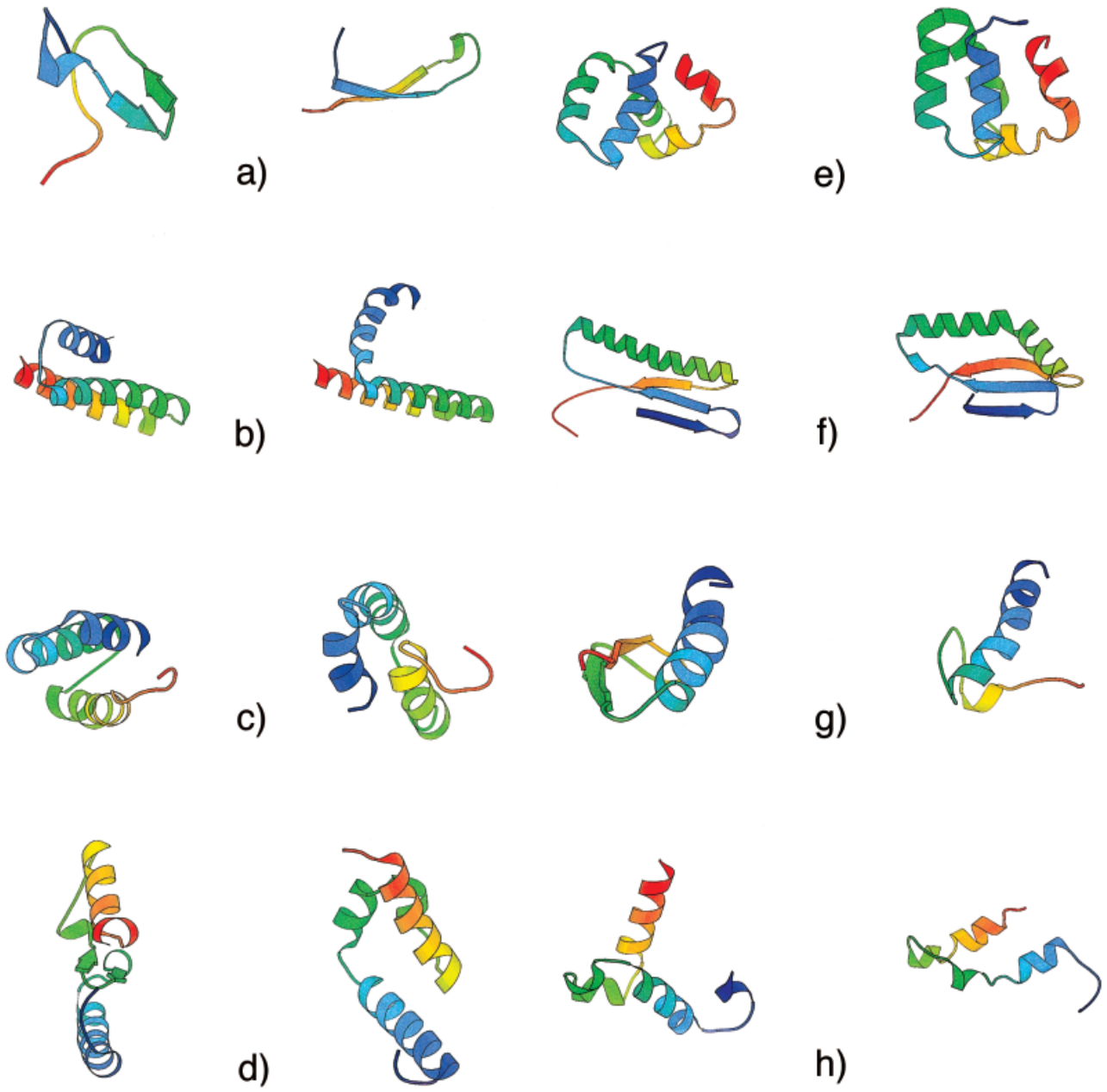

Fig. 1. The best predicted fragments taken from the first ranked models are shown alongside the equivalent fragments in the experimental structures. Structures were plotted by using Molscript[12] and colored according to residue position from the N-terminal residue in the fragment (in blue) to the C-terminal residue (in red). **a:** Target T0094—residues 145–178. **b:** Target T0096—residues 165–228. **c:** Target T0097—residues 170–228. **d:** Target T0098—residues 195–258. **e:** Target T0102—residues 1–70. **f:** Target T0110—residues 29–96. **g:** Target T0126—residues 118–163. **h:** Target T0127 residues 303–351.

Target T0094—Cyclic Phosphodiesterase

Target T0094, the largest target predicted using the described method, is a protein with a fairly irregular $\alpha + \beta$ fold and has no strong similarities to any protein in the current PDB. The FRAGFOLD prediction for this target was poor, with only a few small fragments being correct.

Target T0096—FadR

Target T0096 is a two-domain protein comprising a winged DNA binding domain and a mainly helical domain with no apparent overall similarity to any known folds. The FRAGFOLD prediction shown in Figure 1 is only for the larger all-helical domain. In this case, the prediction is reasonably good, with a fragment comprising two helical hairpins being particularly well modeled. Although the overall fold of the protein was not accurately reproduced, the architecture of the predicted fold (large regular bundle of helices) was in reasonable agreement with that of the experimental structure.

Target T0097—ERp29 C-Terminal Domain

Target T0097 is a helical protein comprising an irregular bundle of five helices. Again, a fragment comprising two helix hairpin motifs is well predicted, and this fragment represents >50% of the fold.

Target T0098—Spo0A C-Terminal Domain

Target T0098 is an interesting case. It was originally listed as a protein with a unique all-α fold, but in fact, it exhibits similarity to members of the DNA-binding repressor family. Despite the similarity surrounding the DNA binding motif, the overall chain fold is not well represented by any structure currently in PDB. A fragment comprising around 60 residues is well modeled in the first submitted model for this target.

Target T0102—Bacteriocin AS-48

Target T0102 is a globular protein with an all-α fold, but with the interesting feature that it is a cyclic peptide—the N- and C-termini are joined together post-translationally. To constrain this feature in FRAGFOLD, a simple harmonic distance constraint was added to constrain the Cα-Cα distance between the N- and C-termini to 3.8 Å. This was added as a seventh weighted term in the objective function.

The models submitted for T0102 were the best results we submitted from FRAGFOLD. The overall chain fold was correctly predicted with an overall RMSD of 4.9 Å. However, it turned out that this fold was not unique. It is surprising that T0102 turned out to be a likely distant relative of NK-lysin,[9,10] even though NK-lysin is not a cyclic peptide. Coincidentally, NK-lysin was the same structure that FRAGFOLD predicted accurately in CASP2.

Target T0110—Ribosome-Binding Factor A

Target T0110 is a protein with an α + β class fold. The fold comprises a three-stranded antiparallel sheet against which is packed two helices. The FRAGFOLD prediction for this target was good, with the overall fold being reproduced. Again, unfortunately, this fold turned out not to be unique. Nevertheless, this is a good example of a case where FRAGFOLD has produced a reasonable model for a protein with an α + β class fold.

Target T0126–Olfactory Marker Protein

Target T0126 has a complex α + β fold, which comprises an incomplete nine-stranded barrel and two helices. As might be expected from the complex topology and the size of this target, the FRAGFOLD predictions were very poor. Essentially, the submitted structures had no real resemblance to the correct structure.

Target T0127–Magnesium Chelatase

Target T0127 is a two-domain protein with the first domain being a member of the P-loop hydrolase superfamily. The FRAGFOLD predictions were made for the C-terminal domain, which is mainly helical. Although some small fragments were correctly predicted, the overall chain fold for this domain was not reproduced. The fact that this domain is intimately associated with the hydrolase domain probably accounts for the difficulty in this case.

DISCUSSION

Although most of the discussion is addressed to the central CASP theme of "what went right and what went wrong," it is worth making a few general comments first. FRAGFOLD is a completely automatic method for generating plausible protein folds from libraries of fragments. At no point was any human intervention used to rank the submitted models or to manually adjust any of the program's parameters. This is an important point to make if this kind of approach is to be useful in predicting structures on a genomic scale, and an important point to allow a reasonable evaluation of the success of the method.

What Went Right?

For two of the targets (T0110 and T0102), the overall chain fold was correctly predicted with reasonable accuracy. Although these two proteins turned out not to have unique folds, no similar folds were present in the set of proteins used to compile the fragment library or in the compilation of the knowledge-based potentials. In the case of target T0110, the prediction was particularly interesting because the protein contains a β-sheet with a nontrivial topology and so shows that good predictions can be made for αβ proteins as well as all-α proteins.

Of the remaining targets, T0097 and T0098, both α-helical proteins, stand out as >50% of the chain folds were reproduced with an RMSD < 5 Å. Target T0096 is also reasonably well predicted, with a fragment of 64 residues being modeled to similar accuracy, but this only represents 41% of the target protein.

Overall, six of the eight targets predicted with FRAGFOLD have significant similarities to the experimental structures (Z-scores > 4.0), with three having particularly high Z-scores (> 7.0). Of the two targets that had poor first models, however, in one case (T0126) the second ranked model was significantly better than the first.

Unlike the predictions entered into CASP2, there is little evidence this time of overprediction of β-sheet. This is no doubt due to the use of predicted secondary structure in the preselection of fragments.

What Went Wrong and Why?

For two of the eight targets, the predictions had very little similarity to the correct chain folds with Z-scores < 2.0. This was probably due to the fact that these were the two largest targets tackled and that both included a high proportion of β-sheet structure. As a consequence, the conformational spaces of these targets could not be explored adequately by FRAGFOLD.

Even for the better-predicted structures, it is clear that even where the secondary structure was correctly predicted, the detailed packing of the secondary structural units was usually not accurately modeled. Given the simplicity of the representation of the protein chain (main-chain heavy atoms and β-carbon atoms only), errors in

detailed packing are almost inevitable. It is likely that accurate prediction of a native structure will require an all-atom model with specific side-chain interactions being accurately modeled. It might be the case, however, that a detailed representation of side-chain interactions can be brought into a simulation at a late stage, as long as the rough topology is correctly predicted by using the simplified representation.

It is hoped that these issues can be addressed in time for the next CASP experiment, where it is also to be hoped that many more suitable prediction targets will be available. It is very hard to come to any reasonable conclusions about the generality of fragment-assembly methods based on a very small number of "new folds." Nevertheless, there is much to learn from the results of the CASP4 experiment both from the results shown here and from the results shown elsewhere in this issue. It is interesting to compare this method with the method of the Baker group (University of Washington) who ostensibly are using a very similar method[11] but with somewhat more success, particularly in predicting folds with β-sheets. Apart from the obvious differences in the choice of potentials and the precise optimization algorithm, there are a few differences in overall philosophy that are worthy of attention. First, Baker et al. do not use a specialized fragment library. Instead, they apparently use a library of nine-residue fragments, selected essentially by sequence similarity. Although it is intellectually pleasing to preselect particular fragments that correspond to supersecondary motifs, it may well be the case that this biased selection process turns out not to be optimal. Second, Baker et al. run a very large number of short simulations, rather than a small number of much longer simulations. This may perhaps relate well to the "molten globule" state in which the fold of the final fold is apparent as an ensemble of partially folded states rather than a single instance of a fully folded state. The final area of major difference appears to be in the final clustering procedure. Baker et al. not only cluster a large number of folds, but they also apply a number of filters to eliminate folds that are not likely to occur.

CONCLUSIONS

It is clear from the results that are presented here that FRAGFOLD is capable of generating compact structures with significant similarity to the experimentally determined structures. This has been shown for both α-helical proteins and proteins that include β-sheets. It is, however, apparent that there still remains a large gap between the quality of structures produced by FRAGFOLD and fold recognition or comparative modeling techniques. Although it is somewhat difficult to envisage routine usage of this kind of approach for predicting protein structure on a large scale (e.g., in structural genomics), the current level of success shown here and elsewhere by other groups shows a great deal of encouraging progress toward this goal.

ACKNOWLEDGMENTS

I thank the organizers, assessors, and other participants of the CASP4 experiment for their hard work and useful discussions. In particular, I also thank the experimentalists for making their structures available.

REFERENCES

1. Jones DT. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins 1997;Suppl 1:185–191.
2. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
3. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
4. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
6. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-base archival file for macromolecular structures. J Mol Biol 1977;112:535–543.
7. Kelley LA, Gardner SP, Sutcliffe MJ. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. Protein Eng 1996;9:1063–1065.
8. Jones DT, Tress M, Bryson K, Hadley C. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins 1999;Suppl 3:104–111.
9. Andersson M, Gunne H, Agerberth B, Boman A, Bergman T, Sillard R, Jornvall H, Mutt V, Olsson B, Wigzell H, Dagerlind A, Boman HG, Gudmundsson GH. NK-lysin, a novel effector peptide of cytotoxic T and NK cells. Structure and cDNA cloning of the porcine form, induction by interleukin 2, antibacterial and antitumour activity. EMBO J 1995;14:1615–1625.
10. González C, Langdon GM, Bruix M, Gálvez A, Valdivia E, Maquaeda M, Rico M. Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. Proc Natl Acad Sci 2000;97:11221–11226.
11. 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.
12. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24: 946–950.