PROTEINS: Structure, Function, and Genetics 53:480–485 (2003)

Assembling Novel Protein Folds From Super-secondary Structural Fragments

David T. Jones* **and Liam J. McGuffin**
Department of Computer Science, Bioinformatics Unit, University College London, London, United Kingdom

ABSTRACT **The results of applying a fragment-based protein tertiary structure prediction method to the prediction of 14 CASP5 target domains are described. The method is based on the assembly of supersecondary structural fragments taken from highly resolved protein structures using a simulated annealing algorithm. A number of good predictions for proteins with novel folds were produced, although not always as the first model. For two fold recognition targets, FRAGFOLD produced the most accurate model in both cases, despite the fact that the predictions were not based on a template structure. Although clear progress has been made in improving FRAGFOLD since CASP4, the ranking of final models still seems to be the main problem that needs to be addressed before the next CASP experiment. Proteins 2003;53:480–485.**
© 2003 Wiley-Liss, Inc.

Key words: protein structure prediction; folding; simulation; novel folds; fold recognition

INTRODUCTION

Although threading methods are now widely used tools for building 3-D models for newly characterized proteins,[1] no matter how much these methods are improved, they suffer from the fundamental limitation of only being able to predict folds which have already been observed. Methods for predicting novel folds are rendered somewhat easier than otherwise might be the case because even when a new fold is discovered, it is generally observed that the fold will still be composed of common structural motifs at the supersecondary structural level. In this paper a new version of a method, called FRAGFOLD, is described for modeling protein structures from recurrent supersecondary structural motifs. The strategy described attempts to greatly narrow the search of conformational space by pre-selecting supersecondary structural fragments from a library of highly resolved protein structures. Since we published the results of using FRAGFOLD in the 2nd CASP experiment held in 1996,[2] we and others[3] have been striving to make improvements to fragment-assembly methods for protein structure prediction, and it has been clear at both CASP4 and now at CASP5 that significant improvements have been made.

Prediction Method

The method used in CASP5 (FRAGFOLD), is based on an earlier method[2] which was used in the CASP2 experiment held in 1996 and the CASP4 experiment held in 2000.[4] Here only the most recent alterations to the method will be described.

As with the original method, at the heart of the objective function 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 to convert observed frequencies of residue pair interactions to free energy changes. In addition to the pair potentials, a solvation potential is also employed. These potentials are identical to those currently in use by the latest versions of our threading programs: THREADER[5] and GenTHREADER.[6]

For threading applications, pairwise and solvation potentials are often 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.

In previous versions of FRAGFOLD an explicit term was used for estimating the degree of compactness of a generated fold. However, when we re-evaluated some of our predictions in CASP4 we found that including this term in the potential function was producing folds which were too globular, particularly in the case of small proteins. For our CASP5 predictions, therefore, we omitted any explicit compactness measure and instead relied upon the solvation potential to produce reasonably compact conformations. This proved to be successful.

Steric clashes were again penalized in a simple fashion. For a pair of atoms in residue types a and b, separated by a distance $d^{ab,}$

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

$$E_{clash} = 0 \qquad \text{otherwise,}$$

*Correspondence to: David T. Jones, Department of Computer Science, Bioinformatics Unit, University College London, Gower St., London, UK WC1E 6BT. E-mail: dtj@cs.ucl.ac.uk

Received 17 February 2003; Accepted 3 April 2003

© 2003 WILEY-LISS, INC.

where D_{min}^{ab} is the minimum observed distance between residues i and j 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 C-α geometries of chain segments involved in either parallel or anti-parallel sheets. A hydrogen bond is assumed between residue i and residue j where $j > i+5$ and distances $D_{j\pm1}^{i\pm1}$ fall within the bounds derived from sheets in highly resolved structures. A maximum of 2 hydrogen bonding interactions are allowed per residue. $-E_{hbond}$ is taken as the number of residues involved in at least 1 and no more than 2 hydrogen bonded bridges.

The above energy terms are applied to a simplified representation of the polypeptide chain in which only main chain heavy atoms and beta-carbon atoms are considered, and summed thus:

$$E_{\text{total}} = W_1 E_{\text{short-range}} + W_2 E_{\text{long-range}} + W_3 E_{\text{solv}} + W_4 E_{\text{steric}} + W_5 E_{\text{hbond}}$$

where $W_{1..5}$ are adjustable weights.

The above potentials terms are summed across multiply aligned sequences, rather than just a single sequence. For the steric terms, however, the maximum of the steric energies is taken so that the generated main chain conformation is compatible with all of the aligned sequences in the family.

Preselection of Fragments

The first stage of the folding simulation involves the selection of favourable supersecondary structural fragments at each residue position along the target sequence. Supersecondary structures are defined by taking 2 or 3 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 non-compact 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, $L-M+1$ threadings are considered. A new feature of fragment selection used for the CASP5 predictions was to eliminate any threadings which contradicted the reliable regions of predicted secondary structure. Secondary structure was predicted with an updated version of PSIPRED.[7] The PSIPRED predictions for all of the CASP5 targets were also submitted in the secondary structure prediction category, and achieved an average Q_3 score of 79.5% across the submitted target domains with no obvious sequence similarity to structures already present in PDB.[8] These predictions, along with the associated PSI-BLAST multiple sequence alignments were used as inputs to FRAGFOLD. Note that apart from biasing the selecting of fragments, secondary structure prediction information was not used elsewhere in FRAGFOLD. The objective function does not currently include terms which 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 pre-selected.

Having selected the starting fragment lists, a single folding simulation progresses in the following way. Firstly a random conformation for the target sequence is generated by selecting fragments entirely randomly. Fragments are spliced by superposing the α-carbon, 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 clash check is carried out, and the conformation is rejected if any pair of atoms (main chain and β-carbon) is found to be 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 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 selected using the same procedure. This continues until the starting conformation has no steric clashes.

Before the simulation starts, it is necessary to calculate the relative weighting of the components of the potential function ($W_1..W_5$). To find these weights, a number of random chain conformations (typically 1000) are generated using the above procedure. For each of these conformations, the component energy terms are calculated, and the 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, a simulated annealing algorithm is used to minimise the energy function. A random move is made by either selecting a locally optimum fragment from the $10N$ lists of pre-selected fragments at each position in the target sequence, or a completely free choice is made from the

TABLE I. Examples of the Best FRAGFOLD Models Submitted to the CASP 5 Server

Target	Model	Fold class	Length of target	Total RMSD	Best predicted subset size	RMSD of best predicted subset	MaxSub score	GDT_TS
T0129_1	2	α	89	7.26	24	1.89	2.16	38.48
T0129_2	4	α	94	6.67	44	2.00	3.63	49.47
T0139	1	α	62	8.75	26	1.61	3.57	44.36
T0149_2	2	α + β	116	14.93	25	1.37	1.91	26.29
T0161	3	α + β	156	11.46	33	2.70		23.87
T0170	2	α	69	7.08	27	2.02	3.05	47.83
T0181	4	α + β	111	13.05	33	2.63		26.57
T0187_1	3	α + β	187	18.06	24	1.15	1.18	16.97

Examples of the best FRAGFOLD models submitted to the CASP5 server in the new fold (NF) and new fold/fold recognition (NF/FR) categories. RMSDs are calculated by using the standard expression:

$$\sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

where d_i is the distance between equivalenced α-carbon atoms at position i in two structures of length n. Best subset RMSD and MaxSub scores were calculated by using the MaxSub program developed by Siew *et al.*[9] GDT_TS (Global Distance Test Total Score) was calculated by using the method of Zemla *et al.*[10] and taken from the CASP 5 summary tables found at URL http://predictioncenter.llnl.gov/casp5.

additional list of small fragments. Half of the moves made involve a locally optimum fragment, and the other half of the moves involve a free selection from the small fragment list. In this way, approximately half of the random moves will result in forming a super-secondary structural motif at the selected position in the polypeptide chain. Computationally, these large fragment moves allow a great deal of conformational space searching to be bypassed.

Although FRAGFOLD offers the option of using a genetic algorithm to search conformational space, for all of the CASP5 predictions, simulated annealing was used as follows. Random moves are made as detailed above, but are accepted with a probability $e^{-\triangle E/kT}$, where $\triangle E$ is the energy change caused by the move. The starting temperature (T_0) 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 i.e. from $E = kT$, $T_o = 10\ \Delta E/k$. The temperature is reduced by 5% after M random moves have been allowed, or a total of N moves have been tried. The values of M and N are chosen according to the length of the target protein as follows:

$$\begin{array}{ll} l < 50 & M = 1000, N = 10000 \\ 50 \leq l < 120 & M = 5000, N = 50000 \\ l \geq 120 & M = 10000, N = 100000 \end{array}$$

When every random move is rejected at a given temperature, then it is assumed that the current structure is "frozen". At this point, the temperature is set to zero, and a further 50000 random moves made to allow the system to "quench".

Selection of Final Structures

For each prediction target, between 200 and 2000 separate simulations (using different random number seed values) were run on a Linux-based compute farm. The final conformations were clustered by rigid body superposition. A single–linkage clustering algorithm was used with a 6 Å cut-off. The representatives of the 5 largest clusters were submitted to the CASP5 assessment. In some cases, however, less than 5 clusters were produced. In one case (target T0161) no clusters were found at all. In these cases, the remaining models were selected using a simple sequence shuffling test as follows. The thinking behind this shuffling test was to try to evaluate the specificity of the sequence-structure match i.e. to identify the structures for which the target sequence is in some sense optimal. In this test, the target sequence was randomly shuffled 1,000,000 times and the pairwise energy calculated for the shuffled sequences threaded (without gaps) onto the fold being evaluated. The mean (μ) and standard deviation (σ) was calculated for the energies (E) of the shuffled sequence models in order to calculated a Z-score for each fold ($Z = (\sigma - E) / \mu$). The folds with the highest Z-score were then selected to give a total of 5 submitted models for each target.

RESULTS

The methods described above were applied to a total of 16 of the CASP5 targets. These targets were selected as those which seemed most likely to be novel folds on the basis of fold recognition results from THREADER[5] and GenTHREADER.[6] As described, up to 5 predictions were submitted for each target based on the results of the structure based clustering. Table I summarises the submitted predictions for the 8 target domains which turned out to have novel folds or folds with very limited similarity to existing folds, and Figure 1 shows the best submitted fragment for each.

Target T0129

Target T0129 comprises 2 all-helical domains. The first domain is an orthogonally packed bundle of four helices,

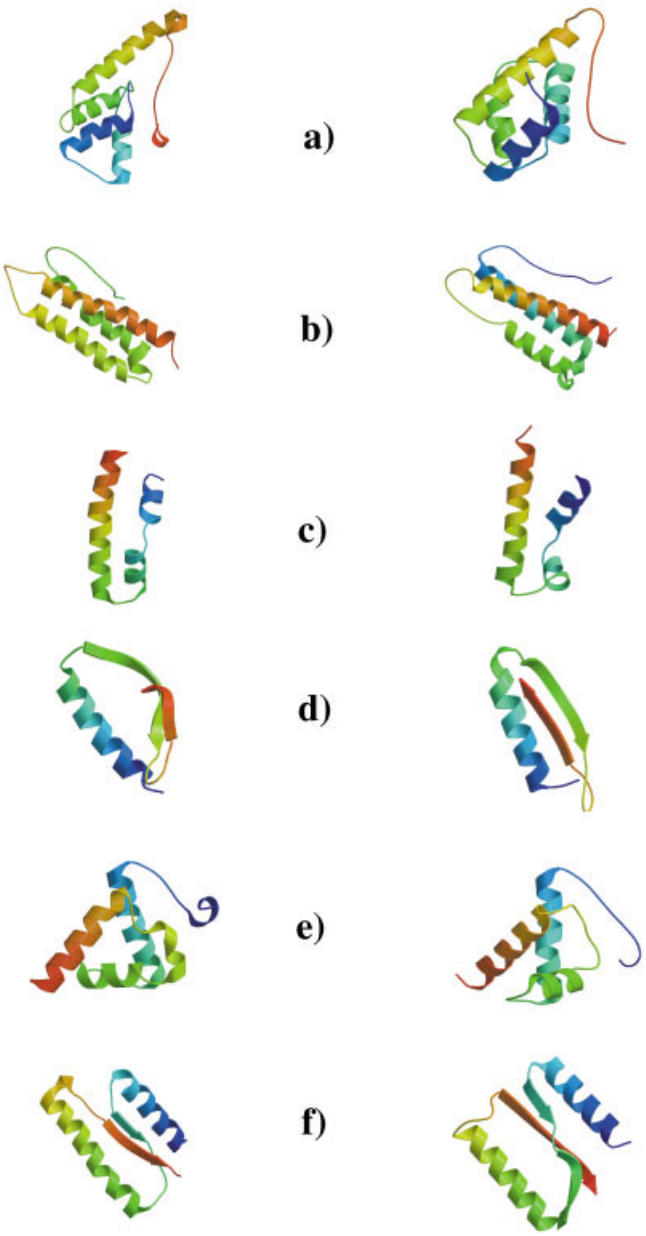

Fig. 1. Predicted models (left) and equivalent residues in experimental structure (right). Structures were plotted using Molscript[11] and Raster3D[12] and colored according to residue position from the N-terminal residue (blue) to the C-terminal residue (red) **a**, Target T0129 domain 1 - model 2. **b**, Target T0129 domain 4 - model 4. **c**, Fragment of target T0139 - model 3, residues 257-300. **d**, Fragment of target T0149 domain 2 - model 2, residues 235-276. **e**, Target T0170 - model 2. **f**, Fragment of target T0187_1, residues 250-313.

the second an up-down bundle of three helices. No domain boundaries were detected in the target before the models were submitted, and so all 5 submitted models were for the whole protein chain. Both of the domains were reasonably well predicted by FRAGFOLD, though the best domain model for each domain was found separately in two of the submitted full-chain models (model 2 and 4). None of the submitted models offered a satisfactory conformation for the entire protein chain.

Target T0139

Target T0139 is a small all-helical fold which was not part of the official CASP assessment due to the early publication of the structure. Nevertheless, the 3rd FRAGFOLD model was reasonably close to the experimental structure with 53 out of 83 residues giving an RMSD of 5.8 Å.

Target T0149 (Domain 2)

This target domain comprises an anti-parallel sheet with 5 strands and two helices packed against one side of the sheet. Despite some locally correct regions, the overall topologies of the FRAGFOLD predictions were not correct, with all of the models having each of the helices on opposite sides of the central sheet. Clearly, the accurate prediction of sheets with complex topologies is still a bottle-neck in the method.

Target T0161

Unfortunately no structure for this target has been released to predictors and so little analysis of the fold or the FRAGFOLD predictions are possible. The fold was quite complex, however, and PSIPRED secondary structure prediction accuracy for this target was relatively low (Q_3 score of ~65%). Nevertheless, in terms of GDT_TS scores, FRAGFOLD model 3 was ranked in second place. Just under half of this model could be superposed on the experimental structure with an RMSD of 6.0 Å.

Target T0170

Target T0170 has a simple 4-helical fold with weak similarity to already known folds. The FRAGFOLD predictions for this target were reasonable, with model 2 being the best giving an RMSD of 5.97 Å for 64 out of 69 residues. Most of the error in this structure was in the orientation of the last helix axis, which was off by approximately 60°.

Target T0187 (Domain 1)

This target has a discontinuous $\alpha+\beta$ domain fold, which was the longest FRAGFOLD domain prediction which was submitted to CASP5. The central sheet comprised 6 strands and so not surprisingly, the FRAGFOLD predictions for this domain were rather poor.

Other Targets

Table II summarises the submitted predictions for 6 further domains which were classified as fold recognition targets. Surprisingly, in most cases the FRAGFOLD predictions are competitive with the best fold recognition results submitted for these targets. Even more surprisingly, in two cases (T0162_1 and T0193_1) FRAGFOLD produced the top ranked models (based on GDT_TS score) even though no template structures were used.

TABLE II. Examples of FRAGFOLD Models for Fold Recognition Targets

Target	Model	Fold class	Length of target	Total RMSD	GDT_TS
T0135	1	α	108	13.11	32.78
T0148	4	α + β	163	12.15	33.33
T0156	2	α + β	157	11.75	44.36
T0162_1	4	α	56	-	70.54
T0162_2	5	α + β	156	-	40.12
T0193_1	1	α	74	4.59	65.88

The best FRAGFOLD models submitted to the CASP 5 server in the fold recognition (FR/A and FR/H) categories.

TABLE III. Comparison of Best Submitted Models to Best Generated Models

Target	No. of folds generated	RMSD of best submitted model	RMSD of best generated structure
T0129_1	668	7.26	6.1
T0129_2	668	6.67	3.7
T0139	396	8.75	3.5
T0170	1638	7.10	3.1
T0187_1	270	17.86	14.5

The best submitted model RMSDs for NF and FR/NF targets are shown alongside the total number of generated folds and the best generated model RMSD.

WHAT WENT RIGHT?

For four out of eight of the submitted new fold (and borderline new fold) domain predictions the overall fold of the domain was correctly predicted with reasonable accuracy. Although two of these proteins had weak similarity to known folds, no similar folds were present in the set of proteins used to compile the fragment library or in the compilation of the statistical potentials. This is more or less a similar level of success to that we achieved at CASP4, though the CASP5 targets are arguably harder than those tackled in the earlier experiment. At worst, this demonstrates a sustained ability to successfully apply fragment-based new fold prediction methods to new protein sequences, and at best, taking the relative target difficulty into account, this demonstrates some progress.

Although this paper is focused on the new fold predictions made by FRAGFOLD, it is very interesting to note that in two fold recognition cases (T0162_1 in the FR/A category and T0193_1 in the FR/H category) FRAGFOLD produced a better model than any other method evaluated, including fold recognition methods. This suggests that FRAGFOLD might be of use in refining models even when a template structure is available.

WHAT WENT WRONG AND WHY?

One major disappointment in our predictions was that the best of the 5 submitted models was rarely the 1st submitted model. Also, Table III shows that in several cases, very good models were generated in the initial ensemble of 200-2000 structures, but were not identified by either the single linkage clustering method or the simple sequence shuffling tests. Improved methods for identifying incorrectly folded decoy structures and for identifying the most significant clusters of conformations are clearly called for. Perhaps a consensus approach could be applied whereby instead of simply clustering structures and selecting a representative from each cluster, the cluster could be used to define a consensus fold.

The other area where there continues to be a lack of progress is in the correct folding of proteins with complex beta sheets. Formation of beta sheets is a highly cooperative process requiring many regions of the polypeptide chain to converge in just one configuration, it is thus easy to see why fragment-based methods have such difficulty in handling this type of structure. A possible means forward that we are investigating is to develop a semi-template based approach which would allow different sheet topologies to be generated more efficiently. So, rather than assembling fragments of super-secondary structure, fragments of tertiary structure could be assembled in a similar way.

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 even for proteins with entirely novel folds. This has been demonstrated for both α–helical proteins and proteins which include β–sheets, though the performance on proteins with sheets is markedly worse than on mostly helical proteins. It is, however, apparent that there still remains a large gap between the quality of structures produced by FRAGFOLD and fold recognition or comparative modelling techniques. Nevertheless, in at least two cases the predictions generated by FRAGFOLD were superior to all of the predictions made by fold recognition methods even though a suitable template structure could be found in the existing databases. In previous papers we have stated that it was difficult to envisage routine usage of fragment-based approaches for predicting protein structure on a large scale. However with the increasing levels of success of these approaches and the increased availability of substantial computing resources, this statement is becoming far from certain.

ACKNOWLEDGMENTS

We would like to thank the organizers and assessors of the CASP5 experiment for their hard work, and in particular we should like to thank the experimentalists for making their structures available.

REFERENCES

1. Jones DT. Protein structure prediction in genomics. Brief. Bioinformatics 2001;2:111–125.
2. 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.
3. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated anealing and Bayesian scoring functions. J. Mol. Biol. 1997;268:209–225.
4. Joness DT. Predicting novel protein folds by using FRAGFOLD. PROTEINS 2001;Suppl. 5:127–132
5. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature. 1992;358:86–89.
6. Jones DT. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 1999;287: 797–815.
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999;292:195–202.
8. 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.
9. Seiw N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2001;16:776–785.
10. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. PROTEINS 1999;Suppl.3: 22–29.
11. Kraulis PJ. Molscript - a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 1991;24:946–950.
12. Merrit EA, Bacon DJ. Raster3D photorealistic molecular graphics. Methods in Enzymology 1997;277:505–524.