PROTEINS: Structure, Function, and Genetics Suppl 3:104–111 (1999)

FOLD RECOGNITION: PREDICTION REPORTS

Successful Recognition of Protein Folds Using Threading Methods Biased by Sequence Similarity and Predicted Secondary Structure

David T. Jones,* Michael Tress, Kevin Bryson, and Caroline Hadley
Department of Biological Sciences, University of Warwick, Coventry, United Kingdom

ABSTRACT **Analysis of our fold recognition results in the 3rd Critical Assessment in Structure Prediction (CASP3) experiment, using the programs THREADER 2 and GenTHREADER, shows an encouraging level of overall success. Of the 23 submitted predictions, 20 targets showed no clear sequence similarity to proteins of known 3D structure. These 20 targets can be divided into 22 domains, of which, 20 domains either entirely match a previously known fold, or partially match a substantial region of a known fold. Of these 20 domains, we correctly assigned the folds in 10 cases. Proteins Suppl 1999:3:104–111.** © 1999 Wiley-Liss, Inc.

Key words: protein structure prediction; molecular modeling; protein folding; sequence analysis; dynamic programming

INTRODUCTION

In the wake of various mass sequencing projects, the development of reliable methods for predicting protein structure from sequence is becoming of paramount importance in the field of molecular biology. Experimental procedures are still relatively time-consuming, and this increase in sequence information from genome sequencing projects is magnifying the already growing gap between sequence and structural information.

Several different approaches have been used to predict protein structure from sequence, with varying levels of success. Ab initio methods encompass any means of calculating coordinates for a protein sequence from first principles—that is, without reference to existing protein structures. Relatively little success has been seen in this area, however. Comparative, or homology, modeling, attempts to predict protein structure on the strength of a protein's sequence similarity to another protein of known structure (based on the premise that similar sequence implies similar structure). This is a generally reliable approach to protein structure prediction, but there are several limitations to this method, not least of which is its dependence on the existence of a close homologue of known structure and the subsequent accuracy of alignment. Indeed, at the time of writing, only 3% of the 30,000 or so currently known sequence families includes a member of known 3D structure. Sensitive sequence profile methods can extend the range of comparative modeling to the point where around 12% of the families can be assigned to a known 3D structural superfamily, but in contrast to this, it is estimated that the probability of a novel protein having a similar fold to a known structure is currently as high as 60–70%. There is, therefore, a great deal to be gained from attempting to solve the problem of building models for very remotely related proteins, and for proteins which have merely analogous folds, and so are beyond the scope of normal comparative modeling strategies. To date the most successful approach to solving this problem has been that of threading, where a sequence is fitted onto a structural framework and the goodness of fit evaluated by an empirical energy function or some other probabilistic scoring scheme.

In this paper we describe the results of applying a set protocol to protein structure prediction based on fold recognition methods developed in our laboratory.

METHODS

For our fold recognition predictions in CASP3, two methods were routinely applied to each target sequence: THREADER (version 2.5) and GenTHREADER (version 3.0).

THREADER 2.5

THREADER 2.5 is the latest version of our threading program,[16] and although it now incorporates a number of new features (including options for locating domains in target sequences), and a more refined set of potentials, the overall concept of the method remains more or less unchanged since CASP2. First, a library of unique, continuous protein domain folds is derived from the database of protein structures. The test sequence is then optimally fitted to each library fold (allowing for relative insertions and deletions in loop regions), using a double dynamic

*Correspondence to: David T. Jones, Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail: jones@globin.bio.warwick.ac.uk
Received 19 February 1999; Accepted 8 June 1999

© 1999 WILEY-LISS, INC.

programming algorithm,[1] with the "energy" of each possible fit (or threading) being calculated by summing the proposed pairwise interactions and solvation parameters. The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match.

Recent features added to the method allow sequence information and predicted secondary structure information to be considered in the fold recognition process. Sequence information is weighted into the fold recognition potentials using a transformation of a mutation data matrix.[1] By carefully selecting the weighting of the sequence components in the scoring function it is possible to balance the influence of sequence matching with the influence of the pairwise and solvation energy terms. In contrast to this, secondary structure information is not incorporated into the sequence-structure scoring function. In this case, secondary structure information is used to mask regions of the alignment path matrix so that the threading alignments do not align (for example) predicted β strands with observed α helices. A confidence threshold is applied to the secondary structure prediction data so that only the most confidently predicted regions of the prediction are used to mask the alignment matrix. All secondary structure predictions were carried out using our own neural network-based method: PSIPRED[13] (URL: http://globin.bio.warwick.ac.uk/psipred), which has proven to be more accurate than previous secondary structure prediction methods both in CASP3 itself and in benchmarking trials. In CASP3, PSIPRED achieved a Q3 score of 77%, which was 4 percentage points higher than the next best method.

The latest feature we have been experimenting with is the option to combine threading results for a family of related proteins in order to enhance sensitivity. In this case a BLAST[2] search was performed with the target sequence against a non-redundant data bank of protein sequences. Matched sequences with an E-value below 0.001 were considered as significant, and were extracted from the data bank. These sequences were then multiply-aligned with the target sequence to delineate the start and end positions for each sequence (this was to take into account the fact that in some cases the target sequence was itself a domain). Each trimmed sequence was then independently threaded using THREADER, and the individual results files were then combined into a single consensus file by averaging along each column.

GenTHREADER

GenTHREADER[3] is our latest fold recognition method which has been designed to be both fast and reliable, and is particularly aimed at automated genome annotation. The method uses a sequence profile-based alignment algorithm to generate alignments which are evaluated by threading techniques. As a final step, each threaded model is evaluated by a neural network in order to produce a single measure of confidence in the proposed prediction. The speed of the method, along with its sensitivity and very low false-positive rate makes it ideal for automatically predicting the structure of all the proteins in a translated bacterial genome. The method has been applied to the genome of *Mycoplasma genitalium*, and analysis of the results shows that as many as 46% (now 51%) of the proteins derived from the predicted protein coding regions have a significant relationship to a protein of known structure.[1]

In making CASP3 predictions, GenTHREADER was used as a pre-filter. Where GenTHREADER was able to make a confident prediction (generally in cases where a clear evolutionary link is apparent between the target protein and an entry in the fold library), this fold was assumed correct and THREADER was used to generate the final alignment (though with appropriate sequence weighting options). In cases where GenTHREADER did not produce an unambiguous result, full runs were performed with THREADER itself in order to deduce a set of likely folds.

In making our predictions, both GenTHREADER and THREADER were used in the following set protocol:

1. Initial prefilter using GenTHREADER to detect family and close superfamily relationships.
2. Initial threading run carried out using predicted secondary structure as a constraint.
3. Homologous sequences collected and separate threading runs carried out for each sequence.

RESULTS AND DISCUSSION

Table I summarizes all of the threading predictions we submitted to CASP3. Where possible, we attempted to submit predictions for every target that we considered suitable for threading, including three targets for which there was an obvious homologue of known 3D structure but for which the sequence-structure alignment was non-trivial. Due to the early deadline we were unable to submit a prediction for target T0052, but all other targets were analyzed, giving a total of 23 predictions submitted. The CASP3 targets which turned out to have previously observed folds can be divided into three categories: easy targets with weak sequence similarity to a template structure (family targets), medium difficulty targets with a probable superfamily relationship to a template structure, and hard targets involving purely analogous fold similarities.

For the easiest category of prediction targets, with clear sequence similarity to a protein of known 3D structure, the problem of fold assignment is not present. Models were submitted for targets T0082, T0064, T0068, and in these cases, the problem was to generate an accurate sequence-structure alignment. This aspect of our CASP3 results will be discussed later.

Superfamily Targets

The next easiest targets were those with insignificant sequence similarity to a protein of known 3D structure, but which were clearly members of an existing structurally determined superfamily. A total of nine targets were considered to be in this category by the assessors, and of these nine targets, we assigned correct folds to seven.

TABLE I. Summary of Submitted Fold Recognition Predictions†

Target	Predicted fold	Actual fold
T0043	3.40.50 1bmtA — Rossmann fold	3.30.70 α-β Plaits
T0044	3.40.360 1scuB — Succinyl-COA Synthetase, chain B, dom 1	3.65.10 Udp-n-acetylglucosamine 1-carboxyvinyl-transferase; Chain: A, domain 1
	3.40.50 1scuB — Rossman fold	3.40.50 Rossmann fold
	3.40.50 1scuB — Rossmann fold	3.65.10 Udp-n-acetylglucosamine1-carboxyvinyl-transferase; Chain: A, domain 1
T0045	3.30.70 1hal – α-β Plait	Not solved
	3.30.70 1hal – α-β Plait	
T0046	2.60.40 1ten — Immunoglobulin-like fold	2.60.40 Immunoglobulin-like fold
T0051	3.20.25 1onr — Transaldolase B, chain A	Not solved
T0053	3.40.50 1akl — Rossmann fold	3.40.50 — Rossmann fold
	3.40.50 1akl — Rossmann fold	3.40.50 — Rossmann fold
T0054	3.40.50 1iow — Rossmann fold	Co-ordinates not available
	3.30.470 1iow – D-AA Aminotransferase, chain A, dom 1	Superfamily similarity to d-Ala-d-Ala peptidase
	3.30.150 1iow — Biotin Carboxylase, subunit A, domain 2 (1vhh second ranked fold)	1vhh (sonic hedgehog protein)
T0056	1.10.340 Endonuclease III, domain 1	1.10.? — Unique
	1.10.380 Endonuclease III, domain 2	
T0059	2.40.50 1mjc – OB-fold	2.30.30 Pleckstrin domain fold
T0061	NEW FOLD	Unique
T0062	2.40.10 2cnd — Thrombin, subunit H	Not solved
	3.40.50 2cnd — Rossmann fold	
T0063	2.40.10 1eft — Thrombin, subunit H	2.40.?
	2.40.50 1ah9 – OB-fold	2.40.50 — OB-fold
T0064	1.10.260 434 Repressor	1.10.260 434 Repressor
T0067	3.90.80 2prd — Inorganic Pyrophosphatase	3.?.? — Partial similarity to immunoglobulin fold
T0068	2.160.20 1rmg — Pectate Lyase C-like	2.160.20 Pectate Lyase C-like
T0071	2.60.160 Diphtheria Toxin, domain 3	2.20.20 Anthopleurin-A
	NEW FOLD	
T0072	2.60.40 Immunoglobulin-like	Not solved
T0074	1.10.238 Recoverin, domain 1	1.10.238 Recoverin, domain 1
T0075	1.10.472 Cyclin A, domain 1	1.10.?
T0077	3.40.50 Rossmann fold	3.30.? — Unique/Not in library/β-α-β unit
T0078	NEW FOLD	Not solved
T0079	1.10.260 434 Repressor	1.10.? — Unique (contains HTH motifs)
T0080	NEW FOLD	3.10.25 Met-tRNA Fmet Formyltransferase; chain A, dom 2
T0081	3.40.50 Rossmann fold	3.40.50 Rossmann fold
T0083	1.10.260 434 Repressor	1.10.260 434 Repressor
		3.30.? Irregular α/β domain
T0084	4.10.220 GCN4 Leucine Zipper, subunit C	4.10.220 GCN4 Leucine Zipper
T0085	1.10.? 1fgjA	1.10.? — similar to 1fgjA

†For each prediction target, the predicted and observed folds are given. Where possible, CATH[5] codes are given: (C)lass, (A)rchitecture, (T)opology and (H)omology.

Targets T0074, T0083, and T0085 were initially identified at the GenTHREADER pre-filtering stage. Target T0074 (the EH2 domain of human EPS15) was clearly identified as a member of the EF-hand calcium binding protein superfamily with a "certain" confidence rating. Target T0083 (cyanase) was identified as a member of the helical repressor superfamily (quite a surprising result given its function), with a "high" confidence rating. Of the targets which were identified by GenTHREADER, target T0085 was least confidently assigned, with only a "medium" confidence rating (which indicates a 30% chance of the match being a false positive).

For the remaining superfamily targets it was necessary to make use of the full threading approach as described. For targets T0081, T0053, and T0079 the threading results very clearly pointed at a single (correct) fold and so no human input was required in the decision-making process. For the remaining targets, however, the initial results did not point to a single candidate fold, and so some human intervention was required.

Target T0063

Target T0063 turned out to be surprisingly challenging. Initial threading runs produced no clear result, apart from a clear bias towards all-β structures (which was to be expected given that the protein was predicted to be an all-β structure). Before abandoning T0063, however, a threading search was carried out using a restricted fold library comprising all-β folds which are also known to be involved in nucleic acid binding. With this limited selection of folds,

a single candidate structure did become apparent: the OB-fold. Of the OB-folds which were occurring most frequently at the top of the sorted results tables, the fold which appeared most plausible on the basis of functional role was that of PDB entry 1AH9 (initiation factor 1). In almost all of the alignments involving OB-folds, the alignments involved the C-terminal half of the target protein. The principal model for T0063 was therefore taken to be an OB-fold in the C-terminal half of the target. Further searches were then carried out with the remainder of the target sequence to see if another fold could be assigned to the putative N-terminal domain. The results for the N-terminal domain were not conclusive, although the fold which was prominent when the threading results were summed across all members of the target sequence family was a domain from another functionally plausible protein (1EFT – elongation factor Tu).

Missed Superfamily Targets

In three superfamily cases, we failed to submit the correct fold as our rank 1 prediction. The case of target T0054 in our case was a particularly interesting one, as we submitted a correct prediction as our second ranked prediction. For this target, complete threading searches were carried out as described, but no clear indication of a single fold was apparent. Interestingly, the top scoring fold from the initial threading results was 1ZNB-A, which turned out to be a zinc β-lactamase. Given that it was known that T0054 was a zinc metalloprotease and an antibiotic resistance protein, this result was initially taken as the most promising working hypothesis. However, when the threading results were averaged over the whole target sequence family, the overall top fold was 1IOW. This was again a plausible match, being a D-Ala:D-Ala ligase. Although the threading results were not clear-cut at this stage, 1IOW was considered to be the most likely match. At this point, we consulted the literature for clues to the function of T0054 and found that it had been proposed as a member of the same superfamily as the N-terminal domain of the sonic hedgehog signalling protein (1VHH), and a zinc D-Ala:D-Ala peptidase (1LBU).[4] In fact, a sequence motif had been described for this superfamily, though the apparent sequence similarity was very weak. Using a recent feature of THREADER, the sequence motif SxHxxGxAxD was used to bias the threading alignments generated for target T0054, and with this constraint, 1VHH became the clear top scoring fold. At this point we had to decide between the unbiased threading prediction, and the new prediction based on "prior knowledge" of a conserved sequence motif. Under normal circumstances we would propose both models as possibilities, but for CASP3 we had to make an arbitrary decision and finally submitted the "unbiased" model as our principal model.

The remaining two targets (T0044 and T0080) were not correctly predicted, mainly due to omissions (both avoidable and unavoidable) in our fold libraries. In the case of T0044, the fold family had not been entered into the CATH[5] classification database at the time the predictions were carried out, and so the most similar folds for this target were not present. In the case of target T0080, the only correct match for this target in the current release of PDB was a small 74-residue domain from 1FMT (Met-tRNA formyltransferase) which accounted for only 34% of the target sequence. Threading alignments matching less than 50% of the target sequence are rejected automatically by THREADER 2, so we would not have been able to identify this match.

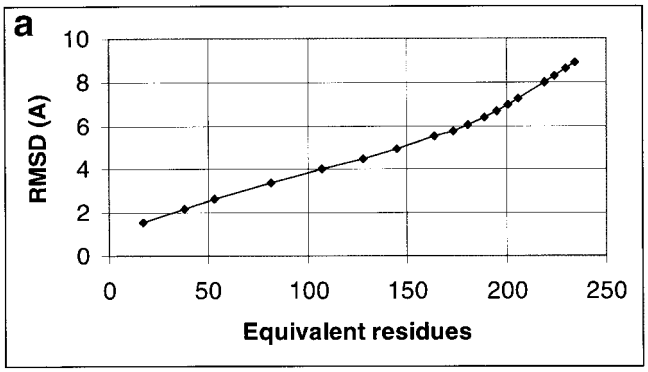
Fold-Level Similarities

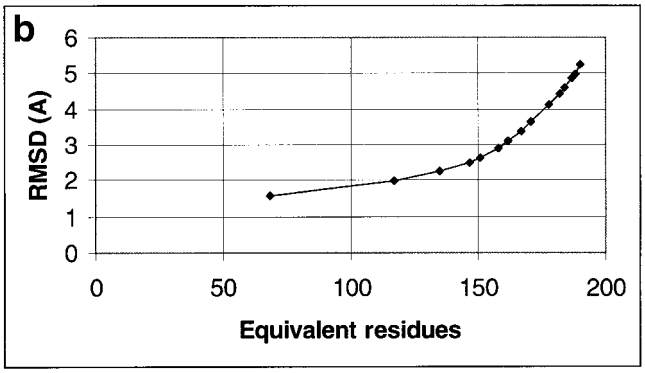
The hardest class of protein structure similarity for protein fold recognition is that corresponding to analogous protein folds, i.e., proteins with similar folds, but no apparent common ancestry. Of the seven targets (including both domains of target T0071) which fell into this category, correct folds were assigned to two (T0046 and T0071nt), and a partially similar fold assigned to one other target (T0077). In all three cases there was little room for misinterpretation in the compiled threading results. Both T0046 and T0071nt clearly had folds similar to the classic immunoglobulin fold, and target T0077 matched a number of doubly-wound α/β folds quite strongly. For both target T0043 and T0077, although the overall correct fold was not identified, templates were selected which allowed at least part of the target protein to be modeled accurately. In the case of T0043, which has a ferredoxin-like fold (α-β plait) comprising split β-α-β motifs, we assigned a Rossmann fold, which is comprised of classic β-α-β motifs. Despite the different chain topology, the submitted model for T0043 had an RMSD of 5.88 Å over 44 residues, which accounts for 28% of the target protein chain. A Rossmann fold was also assigned to target T0077, and again a partial similarity is evident. In this case a β-α-β unit is common between target T0077 and the chosen template structure, and the model fitted the experimental structure with an RMSD of 5.84 Å over 50 residues. In this case, however, the partial similarity accounted for 48% of the target protein chain.

Alignment Accuracy

So far we have only discussed the fold assignment aspects of our CASP3 predictions. However, fold assignment is not sufficient in its own right. Given a correct fold assignment the next step is to generate an accurate sequence-structure alignment and to use this alignment to generate an accurate 3D model for the target protein. Rather than submitting just a sequence-structure alignment in CASP3, we elected to go as far as we could towards an all-atom model for the target sequence. For cases where we felt confident that a correct fold had been assigned and we were at least somewhat optimistic about the quality of the alignment, the alignment was passed to an automatic comparative modeling program (MODELLER3[6]) so that loops and side chains could be built in. For other cases, a crude backbone-only model was constructed with no attempt being made to model loops.

Table II shows a summary of our submitted models for CASP3. Reassuringly, the models submitted for the three targets (T0064, T0068, and T0082) for which there was an available template structure with detectable sequence





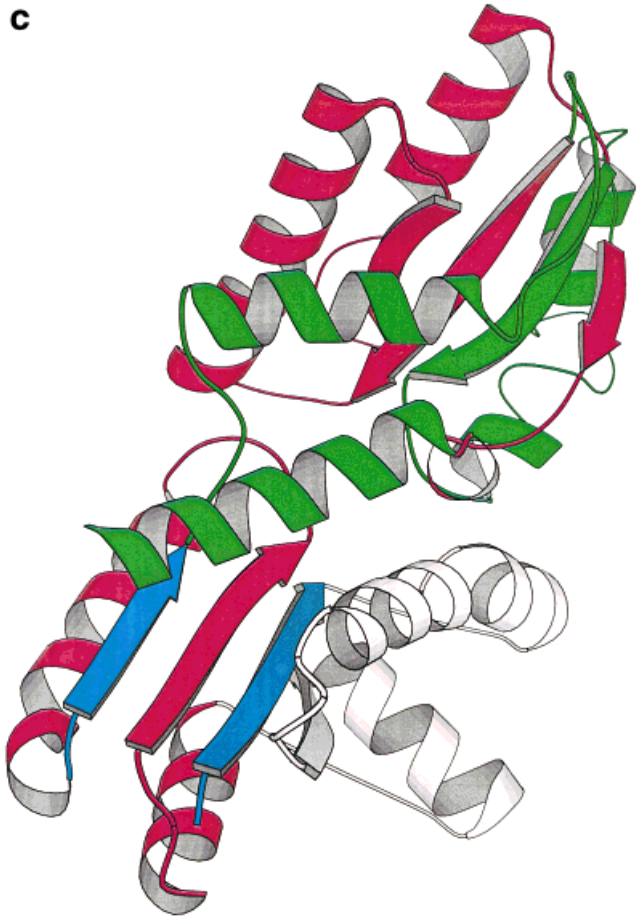


Fig. 1. **a)** Coverage/accuracy plot (RMSD against number of equivalent residues) for target T0053. **b)** Coverage/accuracy plot for T0082 (difficult homology modeling target). **c)** Molscript[11] diagram, showing which regions of target T0053 were most accurately aligned. Red indicates a region with an alignment shift of zero and blue indicates a shift of < 3. Regions with an alignment shift ≥ 3 but where the alignment was not shifted by more than a whole element of secondary structure are shown in green.

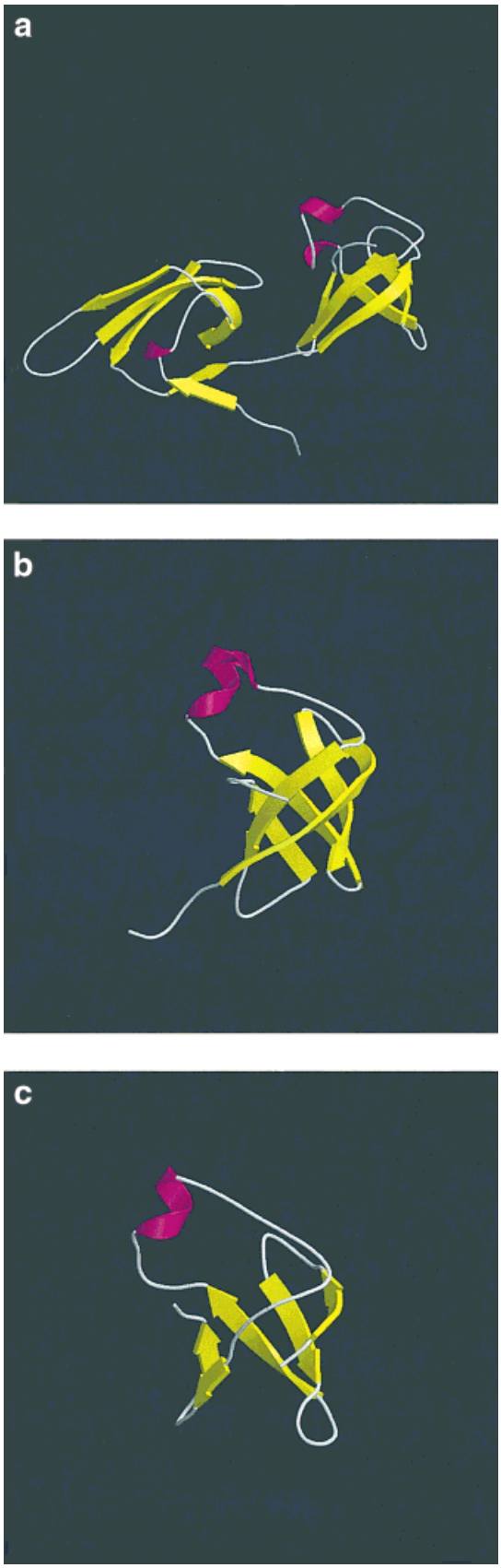
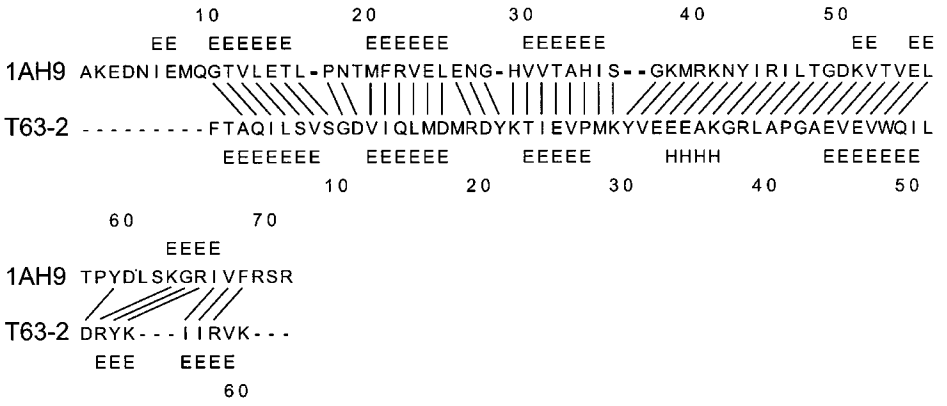


Fig. 2. **a)** X-ray structure for target T0063.[12] **b)** Template structure for model (PDB code 1AH9). **c)** Submitted model for T0063 domain 2. **d)** Initial threading alignment for domain 2 of target T0063 when threaded onto the template structure 1AH9. **e)** Alignment based on residue equivalences inferred from a rigid body superposition of the final model and the template structure onto the experimental structure, which was held fixed.

d)



e)

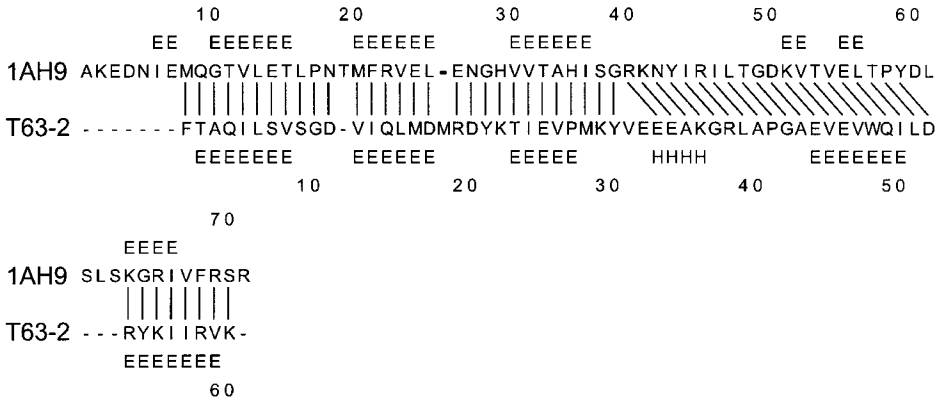


Figure 2. (Continued.)

similarity, were found to be accurately modelled. Despite the very low level of sequence similarity between T0068 and the template structure (1RMG), the submitted model was in fact somewhat better than the overall RMSD value suggests, with 250 residues fitting with an RMSD of 3.43 Å.

For the superfamily cases, the alignments, while not being as accurate as those for the family level relationships, were nonetheless surprisingly accurate. As a representative example, consider the model built for target T0053 (CBIK). Figure 1a shows a coverage/accuracy plot for our first model for target T0053, based on a threading alignment against 1AK1 (ferrochelatase from *B. subtilis*). In this case, 105 residues (40%) in the model can be superposed onto the experimental structure with an RMSD of 4 Å. As a comparison, Figure 1b shows the coverage/accuracy plot for target T0082, a hard comparative modeling target, where 170 residues (89%) can be fitted to a similar level of accuracy. Figure 1c shows the regions of target T0053 which were aligned with different degrees of alignment error relative to the structural alignment as calculated by SSAP.[7]

For the two fold-level cases (T0071 N-terminal domain, and T0046) the submitted models were far from satisfactory, with average alignment shifts of 5.9 and 6.2 respectively.

TABLE II. Summary of Alignment Accuracy for Correctly Assigned Targets

Target	Template PDB code	Model RMSD (Å)	AlnShift[a]	% Aln1[b]
T0046	1TEN	21.47	6.2	11.80
T0053	1AK1	10.71	4.1	24.49
T0054	1VHH	11.45	5.7	36.63
T0063ct	1AH9	4.80	1.2	33.33
T0064	1R69	1.81	0.0	100.00
T0068	1RMG	9.88	0.9	69.15
T0071nt	1DDT	16.91	5.9	8.90
T0074	1AVS-A	4.66	0.0	100.00
T0077	1SRR-A	11.83	0.7	35.58
T0079	1R69	11.78	23.0	0.86
T0081	2CMD	10.48	1.8	27.81
T0082	1BOL	5.24	0.2	80.00
T0083	1R69	10.05	0.0	70.23
T0084	2DGC-A	0.60	0.0	96.00
T0085	1FGJ-A	19.00	5.3	0.05

[a]AlnShift: Average shift between the structural alignment between the target and template structures and the threading alignment.
[b]%Aln1: Percentage of threading alignment with zero shift relative to the correct structural alignment.

The submitted model for target T0063 presents an interesting footnote to the discussion on alignment accuracy. As stated earlier, we chose to submit 3D models (prediction format TS) rather than just sequence-structure alignments (prediction format AL). After the experimental structure for T0063 was made available to us, we found that the alignment suggested by rigid body superposition of the submitted model and the template structure onto the fixed experimental structure was in fact better than the alignment we had used to build the model in the first place (Fig. 2 a–e). Clearly, the structural constraints imposed by MODELLER had in some way corrected for the small shifts in the initial alignment. This highlights the fact that the gap-penalty scheme employed by THREADER does allow non-physical alignments to be produced (i.e., alignments which are not consistent with maintaining a complete polypeptide chain). Although the current algorithm does include some heuristics to ensure that deletions cannot span points in the structure that are too far apart, this scheme does not take into account the fact that the gap might be bridged by a non-local change in the protein's fold. It might be hoped that a more intelligent gap penalty scheme would enhance the selectivity of THREADER, but whether a sufficiently flexible scheme could be devised which would not impact on the algorithm's execution time is open to question. In the case of target T0063 (C-terminal domain) the modeling procedure was able to make corrections in 3D, but this was made relatively easy due to the small size of the domain fold and the fact that the initial alignment was already almost correct.

CONCLUSIONS

Perhaps one of the most significant observations that has come from CASP is that a great deal of success in fold recognition can be achieved purely from knowledge of protein structure and function relationships. More commonly, information on function and other sources of information are applied to the results of a threading method as a "post-processing" step. Many of the threading submissions made to CASP do not include the raw output from a particular method, but prediction made using human intervention. Such intervention can involve visual inspection of the proposed alignment, inspection of the proposed 3D structure on a graphics workstation, comparison of proposed secondary structure with that obtained from secondary structure prediction or even consideration of common function between the target and template proteins.

Although for several of our predictions, some human input was used, this was mostly in the fine tuning stage, after the overall fold had been assigned by either THREADER or GenTHREADER.

One comment that was made regarding the CASP1 threading results was that while threading methods were clearly very able at identifying folds, they were not able to produce accurate alignments.[8,9] That was very definitely the case for CASP1, but in CASP2, a number of good alignments were submitted for some threading targets.[10] It was also observed that the targets for which the better alignments were obtained were those involving obvious evolutionary relationships. For our predictions in CASP3, this again appears to be the case, with the poorest alignments being submitted for the targets falling in the analogous fold category.

Despite this caution, however, the alignments submitted for the homologous targets were nonetheless quite accurate. It is now becoming quite difficult to see the distinction between comparative modeling and threading for these targets.

It is customary in the CASP process to summarize what went right and what went wrong in the prediction process, and so to conclude we will briefly summarize our observations in these categories.

What Went Right?

The good news for our predictions was of course the high overall success rate, with folds being correctly assigned to 10 of the 20 submitted domain folds. In addition to the success in assigning folds, where folds were assigned, accurate alignments were produced for family and superfamily cases. Overall, we were reasonably satisfied with the prediction protocol we used.

What Went Wrong?

The bad news was mainly concentrated in the analogous fold cases. Folds were not correctly assigned to either of the two targets with ferredoxin-like folds, and neither of the two SH3 domain folds was recognized. For both these cases, the incorrect folds somewhat resembled the correct folds, as discussed earlier, but nonetheless it is clear that we must focus some attention on these false-positive cases to see how they can be filtered out.

Beyond these cases, the other failures can be attributed to human error (e.g., T0054), an incomplete fold library

(e.g., T0044) and incomplete domain definitions (e.g., T0080).

Perhaps the most disappointing result for us was the continuing failure to generate accurate alignments for proteins with a β-sandwich architecture (especially the immunoglobulin fold). Poor β-sandwich alignments have been a common feature of our predictions in CASP1, CASP2, and now CASP3. It is clearly a priority for us to investigate the sources of these errors, and hopefully find a solution in time for CASP4.

Program Availability

THREADER can be downloaded from the following URL: http://globin.bio.warwick.ac.uk/~jones/threader.html

GenTHREADER and PSIPRED may be accessed as a server from the following page: http://globin.bio.warwick.ac.uk/psipred

ACKNOWLEDGMENTS

This work was supported by The Royal Society (DTJ), Zeneca (CH), and the BBSRC (MT and KB).

REFERENCES

1. Jones DT. THREADER: Protein sequence threading by double dynamic programming. In: Salzberg S, Searls D, Kasif S, editors. Computational methods in molecular biology. New York: Elsevier; 1998.
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. Jones DT. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
4. McCafferty DG, Lessard IAD, Walsh CT. Mutational analysis of potential zinc-binding residues in the active site of the enterococcal D-Ala-D-Ala dipeptidase VanX. Biochemistry. 1995;36:10498–10505.
5. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH — a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
6. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1997;1:50–58.
7. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
8. Lemer CMR, Rooman MJ, Wodak SJ. Protein structure prediction by threading methods: evaluation of current techniques. Proteins 1995;23:337–355.
9. Jones DT, Miller RT, Thornton JM. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 1995;23:387–397.
10. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins Suppl. 1997;1:92–104.
11. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J App Cryst 1991;24:946–950.
12. Peat TS, Newman J, Waldo GS, Berendzen J, Terwilliger TC. Structure of translation initiation factor 5A from Pyrobaculum aerophilum at 1.75 Å resolution. Structure 1998;6:1207–1214.
13. Jones DT. Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol 1999; in press.