PROTEINS: Structure, Function, and Genetics 23:387–397 (1995)

Successful Protein Fold Recognition by Optimal Sequence Threading Validated by Rigorous Blind Testing

David T. Jones,[1,2] Robert T. Miller,[1] and Janet M. Thornton[1]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, U.K. and [2]Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.

ABSTRACT **Analysis of the results of the recent protein structure prediction experiment for our method shows that we achieved a high level of success. Of the 18 available prediction targets of known structure, the assessors have identified 11 chains which either entirely match a previously known fold, or which partially match a substantial region of a known fold. Of these 11 chains, we made predictions for 9, and correctly assigned the folds in 5 cases. We have also identified a further 2 chains which also partially match known folds, and both of these were correctly predicted. The success rate for our method under blind testing is therefore 7 out of 11 chains. A further 2 folds could have easily been recognized but failed due to either overzealous filtering of potential matches, or to simple human error on our part. One of the two targets for which we did not submit a prediction, prosubtilisin, would not have been recognized by our usual criteria, but even in this case, it is possible that a correct prediction could have been made by considering a combination of pairwise energy and solvation energy *Z*-scores. Inspection of the threading alignments for the $(\alpha\beta)_8$ barrels provides clues as to how fold recognition by threading works, in that these folds are recognized by parts rather than as a whole. The prospects for developing sequence threading technology further is discussed.**

Key words: protein structure prediction, threading, protein sequence analysis, protein folding, computational methods, dynamic programming algorithms

INTRODUCTION

In this paper we will describe our attempts to predict a number of protein structures submitted to the protein structure prediction experiment which recently culminated in a meeting at Asilomar this past December. The general method we used is as previously described by Jones et al.,[1] though with some refinements, which we will briefly discuss later. However, due to space constraints here, a fuller description of the methods will be presented elsewhere (D.T. Jones, in preparation). The main emphasis in this paper will be a quantitative assessment of our results and how this relates to our understanding of how protein fold recognition works. As requested, we concentrate on what went wrong and what went right, and as far as possible we will try to stick to this plan. In a couple of cases, the answer to the question "What went wrong?" relates simply to either an error in using the software or an undiagnosed error in the computer program itself. Under normal circumstances, of course, such errors would not end up in print, but such is the nature of a blind trial that these errors are present in the "published" results. We debated whether to include so-called "postdiction" results here, i.e., results obtained after the structures were revealed, and in the end decided to include them, though clearly indicating which results are pure prediction and which are postdictions. Our reasons for including this information were as follows. First, as long as these postdictions are clearly marked, the reader is free to make his or her mind up as to the degree of weighting to assign each result. Second, we felt that there was more than sufficient supporting evidence as to the success of our method in the pure prediction results alone not to worry about sensible discussion of postdiction results. Third, showing what was required to achieve the expected result is by far the simplest way to answer the question "what went wrong?" Finally, and most importantly, our software is easily obtainable from our FTP server (see later) and so readers can reexamine the results at leisure and even try the software on their own test cases.

METHODS

As stated, the listed predictions were carried out using an updated version of the optimal sequence

Received April 17, 1995; revision accepted July 17, 1995.

Address reprint requests to Dr. D.T. Jones at his present address: Department of Biological Sciences, University of Warwick, Coventry, CV4 7AL, U.K.

threading method of Jones et al.,[1] which described a method for protein tertiary structure prediction based on the fitting of sequences onto structural templates. The term "threading" is now globally (and somewhat incorrectly) used to label any method which attempts to tackle the problem of aligning two protein sequences where the structure of one of the sequences is known. However, a more accurate definition of "threading," at least as we originally intended, is the alignment of a sequence with a protein structure represented in 3-D, without regard to the sequence associated with the structure. Maintaining a 3-D representation of the template structure is the key difference between profile methods[2] and threading methods.[1,3–6]

Threading methods require 3 basic components: a library of fold templates, a method for fitting the sequence onto a fold template, and a means for assessing the goodness-of-fit of the sequence and the structure. A dynamic programming algorithm ("double" dynamic programming) was used to align the given sequence with the "real" coordinates of each structure in a library of folds, taking into account residue pairwise interactions. Matching pairwise interactions is related to the requirements of structure comparison methods. The *interaction environment* of a residue i is defined as being the sum of all pairwise interactions involving atoms in residue i and atoms in all other residues $j \neq i$. This is a similar definition to that of a residue's *structural environment*.[7] In the simplest case, the structural environment of a residue i can be defined as the set of all inter-Cα distances between residue i and all other residues $j \neq i$. Taylor and Orengo[6] developed a novel dynamic programming algorithm for the comparison of residue structural environments, resulting in very high quality structural alignments, and it is a derivation of this method that we therefore use for the effective comparison of residue interaction environments. We believe this approach offers the best compromise between the speed of single level dynamic programming and so-called "frozen approximations,"[2] which do not generally return good solutions without iterative refinement, and exhaustive searches which are computationally prohibitive.

At the heart of the evaluation 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.[1,8,9] In addition to the pairwise potentials, a solvation potential is also used.[1] This potential is determined in a similar fashion to the pairwise potentials, except that the variable in this case is relative solvent accessibility rather than interatomic distance.

A modified set of potentials was used for all predictions apart from the first 3. The main difference between the old and new pairwise terms is a correction for the size of the proteins used to generate the potentials and the protein being predicted (D.T. Jones, in preparation). This correction allows more of the pairwise interaction information to be extracted that by simply truncating the potentials at 10 Å.

For specified atoms (Cβ → Cβ for example) in a pair of residues ab, sequence separation k and distance interval s, the potential is given by the following expression:

$$\Delta E_k^{ab} = RT \ln[1 + m_{ab}\sigma] - RT \ln\left[1 + m_{ab}\sigma \frac{f_k^{ab}(s)}{f_k(s)}\right] \quad (1)$$

where m_{ab} is the number of pairs ab observed with sequence separation k, σ is the weight given to each observation, $f_k(s)$ is the frequency of occurrence of all residue pairs at topological level k and separation distance s, $f_k^{ab}(s)$ is the equivalent frequency of occurrence of residue pair ab, and RT is taken to be 0.582 kcal/mol. In this work, short (sequence separation, $k \leq 11$), medium ($11 \leq k \leq 22$), and long ($k > 22$) range potentials have been calculated between the following atom pairs: Cβ → Cβ, Cβ → N, Cβ → O, N → Cβ, N → O, O → Cβ, and O → N.

The new potentials differ from the original set by rescaling the distances > 10 Å for medium and long range separations so that the mean of the rescaled pairwise distances for each structure is equal to 28.1 Å, which is the raw average for all the protein chains used to compile the potentials. This rescaling is applied both during the compilation of the potentials, and during the threading procedure itself. The aim in doing this is to correct for the size of each protein. The statistical influence of overall protein size is not apparent for interactions which are closer than 10 Å, which appear to be uniform no matter what size of protein is considered. The original solution to this size bias of longer interactions was simply to truncate the potentials at 10 Å. We find, however, that useful pairwise information does extend beyond this cut off. Unfortunately the "quality" of these distant interactions is not as high as those within a radius of 10 Å, and so medium and long range interactions are weighted by the following term:

$$W = 1, \quad d \leq 10$$
$$W = \frac{1}{1 + (d - 10)}, \quad \text{otherwise} \quad (2)$$

where d is the separation between interacting atoms.

In addition to the pairwise potentials, a solvation potential for an amino acid residue a is defined as follows:

TABLE I. Solved Prediction Targets Turned Out to Have a Previously Observed Fold*

Target	Observed class	Observed fold	Partial or complete match	Predicted class	PDB code	Z-score	Correct prediction?	Correct postdiction?
synapto	ββ	Ig	Complete	ββ	1CD8	3.89	Yes	
PBDG	αβ	TIM barrel	Complete	αβ	1ADD	3.75	Yes	
PPDK_4	αβ	TIM barrel	Complete	αβ	1GOX	3.38	Yes	
RTP	αα	Histone H5	Complete	αα	1ABK	3.37	No	Yes
kauA	αβ	TIM Barrel	Complete	αβ	1PII	3.23	Yes	
kauB	ββ	Ig	Partial	ββ	2RHE	2.61	Yes	
L14	ββ	Ribosomal S5	Partial	α+β	1PKP	2.10	Yes	
staufen3	α+β	?	Partial	α+β	1HST-A			
prosub	α+β	Ferredoxin II	Complete	—	—	—	—	No
pcna	α+β	DNA clamp	Complete	—	—	—	—	Yes
PPDK_3	α+β	Aconitase	Partial	ββ	2MFA	2.32	No	
kauG	αα	Helix bundle	Partial	αα	256B-A	2.31	Yes	
xylanase	αβ	TIM barrel	Complete	α+β	3BLM	—	No	Yes

*Key to targets: chymotrypsin/elastase inhibtor-1 by K. Huang and M. James (CE-1), urease from *Klebsiella aerogenes* chains A, B, and G by E. Jabri and A. Karplus (kauA, kauB, and kauG), catalytic core, xylanase (from *Pseudomonas fluoresences*) by J. Jenkins, extracellular endonuclease from *Serratia marcescens* by M. Miller and K. Krause (smanucecs), propeptide from subtilisin BPN' by T. Gallagher, P. Bryan, and G. Gilliland (prosub), first C2 domain of synaptotagmin by S. Sprang (synapto), domain 3 of Staufen from *Drosophila* by M. Bycroft (staufen3), biphenyl-2,3-diol 1,2-dioyxgenase by J. Bolin (bphc), replication terminator protein from *B. subtilis* by D. Bussiere and S. White (RTP), pyruvate phosphate dikinase from *Bacteroides symbiosus* by O. Herzberg (PPDK), L14 (prokaryotic ribosomal protein) from *Bacillus stearothermophilus* by C. Davies and S. White (L14), 6-phospho-β-D-galactosidase from *Lactococcus lactis* by C. Wiesman.

$$\Delta E^{a}_{\text{solv.}}(r) = -RT \ln\left[\frac{f^{a}(r)}{f(r)}\right] \quad (3)$$

where r is the percent residue accessibility (relative to residue accessibility in GGXGG fully extended pentapeptide), $f^{a}(r)$ is the frequency of occurrence of residue a with accessibility r, and $f(r)$ is the frequency of occurrence of all residues with accessibility r. Residue accessibilities were calculated using the program DSSP,[10] applied to Brookhaven coordinate files.[11] Only monomeric proteins were included in this analysis. The revised solvation potentials are based on just 5 unequal relative accessibility divisions ($r < 12\%$, $12\% \le r < 36\%$, $36\% \le r < 44\%$, $44\% \le r < 87\%$, and $r \ge 87\%$) as opposed to the 10 or 20 equal divisions as previously described. These unequal accessibility divisions have been selected through a statistical analysis of structurally similar protein chains.

The program which implements these methods, called THREADER, employs a number of different assessment criteria based on the above potentials, and these are output in columns in the output file. As a rule the entry in any column which offers the most extreme Z-score is taken to be the correct prediction. In all the cases presented here, however, this proved to be the pairwise energy column, and so the predictions given are those which offered the lowest pairwise energy match. Where time permitted, we varied the alignment parameters (gap penalties in particular), and again looked for more significant (i.e., more negative) Z-scores.

RESULTS

In general, we attempted to predict as many of the threading targets as possible, and the complete list of targets tackled is shown in Tables I and II. The first three announced targets (xylanase, bhted, and smanucecs) were predicted using a method essentially identical to that previously described,[1] though with a more up to date fold library containing 244 protein chains for xylanase and 253 for the other two. For all subsequent predictions, however, the modified procedure described earlier was used. The fold library was also updated for the last 6 predictions (L14, PBDG, PPDK, RTP, staufen3, and synapto), extending the library to 266 representative chain folds.

As might be expected, many of the 17 chains folds we attempted to recognize turned out to be previously unobserved folds. There are of course many criteria one can apply in the classification of proteins, and in some indistinct cases classifications can be subjective. By our criteria, using the program SSAPc[12] to structurally align and classify the structures, we determined that 6 of the 17 chain folds had previously observed folds, with another 6 offering significant local matches.

C2 Domain of Synaptotagmin (synapto)

The C2 domain of synaptotagmin produced a very clear prediction of an immunoglobulin-like domain fold, with the top 3 chains all sharing this fold, and the top scoring fold (PDB entry 1CD8) having a Z-score of 3.89. The Molscript[13] ribbon diagrams for 1CD8, synaptotagmin are shown in Figure 1.

TABLE II. Predicted Targets Which Were Found to Have a Novel Fold

Target	Observed class	Predicted fold	Predicted class	PDB code	*Z*-score
PPDK_2	α+β	Ig	ββ	1CD8	3.32
BPHC	α+β	Beta propellor	ββ	1NSB-A	3.05
PPDK_1	αβ	TIM barrel	αβ	1PII	2.88
smanucecs	α+β	Doubly-wound αβ	αβ	5CPA	2.37
ce-1	ββ	Null prediction	ββ	1AAJ	1.85
bhted	α+β	Enterotoxin A	α+β	1LTS-A	(Low)

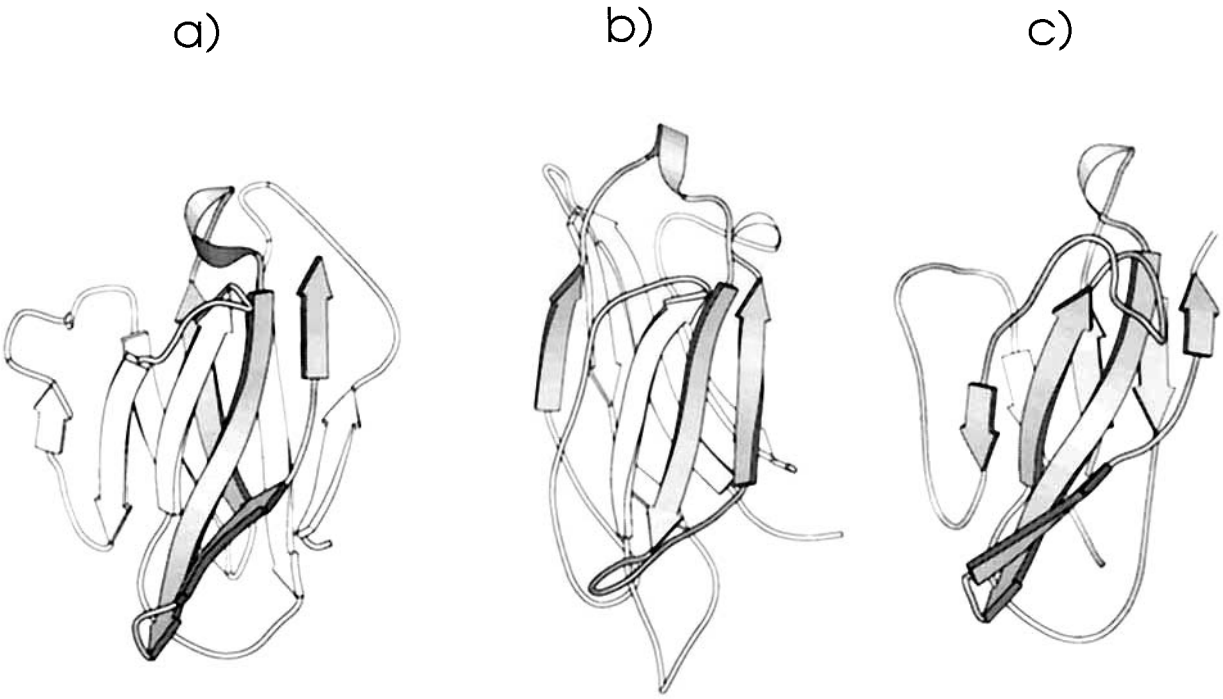

Fig. 1. (**a**) A ribbon diagram of PDB entry 1CD8. The shaded secondary structures show the region over which the sequence–structure alignment of the synaptotagmin sequence is in exact agreement with the structural alignment. (**b**) The X-ray structure of synaptotagmin C2 domain. (**c**) The implied 3-D model for synaptotagmin based on the tenascin structure (PDB entry 1TEN).

Prediction of 6-Phospho-β-D-galactosidase (PBDG)

PBDG clearly predicted as being a TIM barrel, with a *Z*-score of 3.75, and the top 3 scoring folds all being $(\alpha\beta)_8$ barrels.

Urease

The A chain of urease is clearly dominated by a classic $(\alpha\beta)_8$ barrel fold, which was correctly predicted with a reasonable top *Z*-score (3.23). The fold suggested as being most similar by the crystallographers was that of adenosine deaminase (PDB entry 1ADD, which was not in the database at the time of the prediction deadline). Threading the sequence onto this fold did indeed produce a lower energy model than the original best match (1PII).

The overall topology of the B chain did not seem to resemble anything in the current databank, and certainly nothing in the library we used. However at least 3 strands of the observed distorted β sandwich could be aligned with strands in a typical immunoglobulin domain, though whether this similarity is significant is somewhat doubtful. A similarity with immunoglobulin domains would concur with our prediction, however.

The G chain appears to be very close to being a 4-helix bundle (as can be seen in the topology cartoon[16] shown in Fig. 3b) but with a β-hairpin in place of most of the final helix. Our prediction of a 4-helix bundle seems to be pretty reasonable under the circumstances—though obviously the fold is unique as a whole. Unfortunately the wrong alignment was submitted for urease G, the last considered alignment was submitted rather than the one which generated the lowest energy match. Looking at the correct alignment it is clear that recognition

was achieved by aligning the initial helix hairpin in the urease G chain with the initial hairpin in cytochrome b_{562}, though with a shift of one helix turn relative to the correct structural alignment for both helices.

Xylanase

The misprediction of xylanase turned out to be due to a mistake in running the prediction program. In retrospect it appears that xylanase is a fairly easy TIM barrel to detect, probably due to the fact that xylanase comprises a single "pure" $(\alpha\beta)_8$ domain, and indeed in subsequent tests we have found our method quite able to detect this match (matching adenosine deaminase, PDB code 1ADD, with a Z-score of 3.78). Due to the rather uninteresting nature of the cause for failure in this case we will not go into any further detail.

Replication Termination Protein (RTP)

Structure comparison between RTP and structures in the Brookhaven database using the program SSAPc indicated a very strong similarity with histone H5, although with an additional long C-terminal helix in the case of RTP. The submitted prediction for RTP was based on the structure of endonuclease III (PDB code 1ABK), which while being an all-α fold with some superficial similarity to histone H5, clearly has the wrong fold. When we looked back at the search results we were surprised to see that the raw score for histone H5 was better than that of endonuclease III, and that the alignment of RTP and the histone was substantially correct. The failure to predict RTP correctly was due to an incorrect requirement for predicted structures to align over at least 70% of the sequence and 50% of the structure. Due to the absence of a match for the long C-terminal helix of RTP in the histone structure, only 68% of the RTP sequence was aligned with the histone structure, and hence this match was rejected. As a result of this we have amended the cutoff, and current versions of the program require only 55% of the sequence to be matched.

Prediction of Pyruvate Phosphate Dikinase

The prediction of PPDK was greatly complicated by the multiple domain boundaries. Fortunately tentative domain boundaries were given, otherwise it would not have been possible to make a prediction. Overall it appeared that PPDK included a TIM barrel domain, one or two antiparallel β barrels (possibly resembling Ig domains), and another parallel α/β domain possibly doubly-wound but again most probably another TIM barrel. Domain 1 matched one of the two TIM barrels in 1PII, though with a lower than expected Z-score (and significant overlap with doubly-wound chains). Domains 2 and 3 both matched a range of β sandwich/barrel structures, in particular Ig-like folds. Finally, domain 4 principally matches 1GOX, a TIM barrel enzyme, but again with some overlap with doubly-wound domains. One possible explanation for the results obtained would have been a discontinuous domain arrangement—in particular a discontinuous TIM barrel (where it is possible that the barrel spans both domain 1 and domain 4). There is a precedence for this—pyruvate kinase which includes a split TIM barrel, a β domain, and a doubly-wound α/β domain, and we did consider the possibility that PPDK structurally resembled pyruvate kinase.

The crystal structure revealed that 2 of the 4 domains had a previously observed fold. The first domain seems to be a unique domain architecture with some superficial resemblance to Rossmann-like domains, though we were unable to find any significant structural similarities in our fold library. Domain 2 seems to share a common motif with protein G, though overall the fold seems to be unique. Domain 3 is again rather unusual. Its architecture is an α–β–β sandwich, incorporating several $\beta\alpha$ units in a singly-wound arrangement. Despite the unusual nature of this domain, however, it does exhibit weak similarity with the C-terminal domain of aconitase (the SSAPc score in this case was 68, which is only just below the usual threshold score of 70 for significant structural similarity). Domain 4, while somewhat elaborated, contains a recognizable $(\alpha\beta)_8$ barrel.

Domain 4 was correctly predicted as being an $(\alpha\beta)_8$ barrel, with a reasonable Z-score of 3.38. The similarity of domain 3 to the C-terminal domain of aconitase was not, however, recognized. Again, as with the match between RTP and histone H5, the match between aconitase and PPDK domain 3 involved only 17% of the total aconitase structure and was thus filtered out from the final list of scores. However, unlike RTP, the score of this match was not at the top of the list of raw unfiltered scores, and furthermore the alignment between aconitase and PPDK domain 3 was completely wrong, with sections of PPDK domain 3 aligned over the entire length of aconitase. This is, of course, to be expected, as threading is a global alignment procedure, and cannot therefore be expected to identify local similarities. As we are careful to point out when discussing threading methods, the problem of how to handle domain boundaries in both target sequence and the fold library is as yet unsolved.

Other Recognizable Structures

We were not provided with coordinates for staufen3, and so were unable to come to any conclusions as to its similarities with existing structures. The assessors have, however, assigned staufen3 to a known fold family, and we have counted it as such in Table I.

For some of the targets we did not submit a prediction, and yet some of these turned out to have a

previously observed fold. In the case of pcna (the DNA clamp), we took note of the crystallographers comments regarding their suspicion that this protein had a similar fold to the previously determined structure for a DNA clamp.[14] When we checked the release of PDB available to us at the time, however, we found there were no coordinates for the original clamp structure. A threading prediction run produced no significant matches, and because of this we declined to submit a prediction in the belief that the crystallographers were probably right. Subsequently we have been able to add the structure of the original DNA clamp to our fold library (PDB entry 2POL, chain A) and have found that the structure would have been easily recognized with a *Z*-score of 3.29.

Prosubtilisin (prosub) also produced no significant hits in the searches we performed. Reading the prediction notes we were concerned by the fact that prosub only folds in the presence of the subtilisin domain, and on the assumption that our potentials would not be suitable for such cases, again declined to enter a prediction. When the structure of prosub was revealed, it turned out to be structurally similar to the ferredoxin II-like $\alpha+\beta$ fold family, which also includes the fold of procarboxypeptidase (PDB entry 1PBA). When we looked back at the program output for prosub we found that 1PBA (though not 5FD1—ferredoxin II) had the lowest combined (solvation + pairwise) *Z*-score, but was not favored by the pairwise *Z*-scores which we normally use to assess significance (by this criterion alone, ferredoxin II was at position 3, beaten by PDB entries 1ABA and 2NCK, chain L, though the latter chain also has a fold similar to ferredoxin II).

We declined to enter predictions for the mystery protein, which we recognized "by eye" as being a designed $(\alpha\beta)_8$ barrel. We discussed our concerns about predicting a designed protein with the organizers and subsequently decided not to submit a prediction (hits to a wide range of $\alpha\beta$ proteins were in fact obtained, including both singly- and doubly-wound structures, but with no clear separation between them).

Other Structures

The rest of the target structures proved to have unique folds, and despite some tenuous similarities between some of the predictions and the native structures, nothing much can be extracted from these results. In some cases supersecondary structures were matched reasonably well, and in others only secondary structure content appeared to be recognized. One case is worthy of mention, however. The fold for ribosomal protein L14 appears to be unique. Despite the poor *Z*-score, we were originally quite hopeful for our prediction in this case, in that we predicted a similarity to PDB entry 1PKP, which is the structure for ribosomal protein S5, and for which there would have been a good functional explanation for the similarity. When we looked closely at the structural alignment between L14 and S5, however, we were very surprised to note a quite striking similarity between the N-termini of these two folds. Using the SSAPc-derived alignment we were able to superpose 28 equivalent Cα atoms to just 2.2 Å rmsd, corresponding to a region spanning the first 3 strands of 1PKP. Whether this has any functional or evolutionary relevance is unclear, however, it should be noted that the function of the target protein would have been correctly assigned had the origins of the sequence been unknown (a hypothetical gene product from a large scale genome project for example).

Discussion of Alignments

Looking at the alignments generated for the top scoring hits it is clear that fold recognition can be accomplished without requiring the alignment to be correct over the entire length. Judged solely on the results submitted for the top scoring folds in this evaluation exercise, it can be concluded that threading achieves recognition by finding high scoring *local alignments*. For example, in the case of the TIM barrels, none of the barrels is aligned entirely correctly—typically only 3 or 4 of the 8 $\beta\alpha$ units are matched—and not always to the equivalent units in the other structure. The threading alignment for PBDG with 1ADD is shown in Figure 2a. In this case $\beta\alpha$ units 1, 2, 3, and 5 in 1ADD are equivalenced to units 2, 4, 5, and 6 in PBDG. However, this is not surprising given the symmetry of the barrels, and indeed it has been shown[15] that permuted alignments often equivalence more residues than the correct alignments. It should however be pointed out that due the very large number of secondary structure insertions in adenosine deaminase relative to PBDG, aligning this pair of structures is a difficult task even for structural alignment methods.

The threading alignment for synaptotagmin (Fig. 2b) shows a similar pattern, with one-half of the structure being correctly aligned, but with the other half of the structure offset by one strand position across the other sheet in the sandwich. In contrast to this, however, it is interesting to see the threading alignment generated for the tenascin, which is another Ig domain in the fold library, but which was placed at position 10 in the sorted energy list. In this case the entire domain is correctly aligned. Tenascin is in fact more similar to synaptotagmin than CD8 (which was the lowest energy match) in terms of Cα rmsd after superposition, and in terms of the overall topology (as shown in the TOPS[16] topology cartoons in Figure 3a, along with the ribbon diagram of the implied model based on the threading alignment with tenascin shown in Fig. 1c) and this highlights a problem with the potentials used. Ideal threading potentials should produce energies which correlate

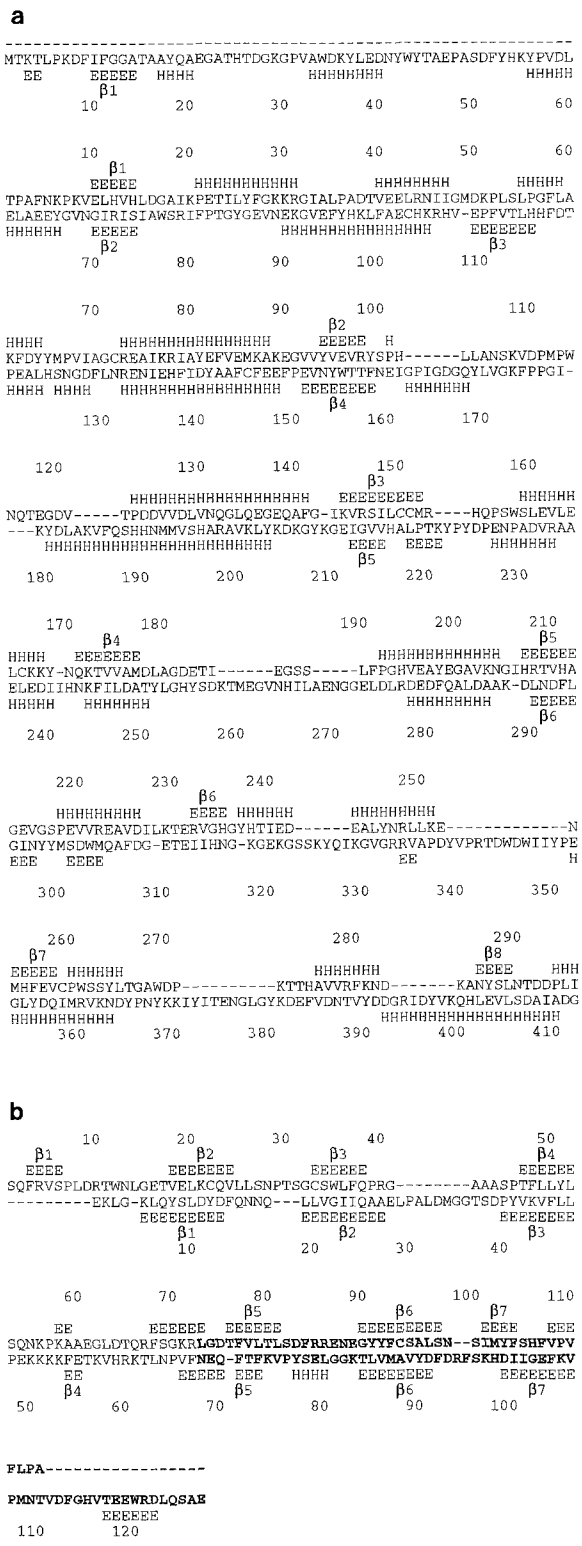

Fig. 2. (**a**) Excerpt from the threading alignment between 1ADD (adenosine deaminase, top) and PBDG (bottom). The barrel strands are labeled (β1..β8) in both cases. (**b**) Threading alignment between 1CD8 (top) and synaptotagmin C2 domain (bottom). The core 7 strands are labeled (β1..β7) in both cases. The section of the alignment in bold shows the region over which the alignment is in agreement with the full structural alignment, and shown shaded in Figure 1a and b.

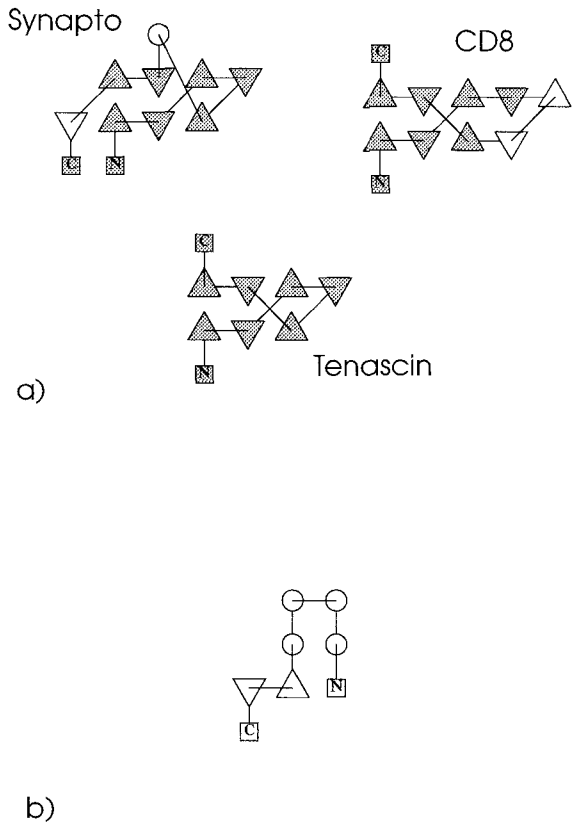

Fig. 3. (**a**) The topology cartoons for the C2 domain of synaptotagmin, CD8, and tenascin are compared. The strands comprising the core 7-stranded immunoglobulin motif is shown shaded in all three cases. These 7 strands are labeled in Figure 1b. (**b**) The cartoon for the G chain of urease (kauG) showing the almost complete helix bundle, along with the terminal β-hairpin.

strongly with the degree of similarity between the native and template structures, but it is clear that the potentials in use at present do not fully achieve this goal. We have observed that the threadings which offer the best structural match (and the best alignments) are those which produce the greatest number of very low energy pairwise interactions. Normally, these very low energy interactions are swamped by the many other interactions that contribute to the overall pairwise energy sum. It might prove useful to consider these outlying interactions separately when deciding which structure to use for modelling.

Discussion of *Z*-Scores

A high success rate is not necessarily the only factor to consider when assessing a prediction method. Another important factor to consider is whether the method offers any means to assign a confidence to the generated predictions. The only real claim made for our original threading method was that it was able to rank the folds in the library in order of their likelihood of being similar to the

native structure of the test protein. Analyzing the results from the current method we observe a reasonable correlation between the pairwise energy Z-scores and whether or not the predicted protein structure is correct. For the normal fold library of around 300 chains, we can assign confidences to the Z-scores as follows:

$Z < -3.5$	Very significant—probably a correct prediction
$Z < -30$	Significant—good chance of being correct
$-2.7 < Z < -3.0$	Borderline significance—possibly correct
$-2.0 < Z < -2.7$	Poor score—could be right, but needs other confirmation
$Z > -2.0$	Very poor score—probably no suitable folds in the library

It should be noted, however, that these Z-score ranges depend on the number of folds in the fold library. If a much smaller library were to be used, then the resulting Z-scores would not be useful measures of significance.

Given that we observe a relationship between the estimated confidence of a prediction and the actual degree of success achieved, it is of interest to see why some hits were confidently predicted and yet found to be wrong. Clearly in these cases, the method employed has identified some statistically significant property, and yet has been led astray. By far the most interesting of the mispredictions is that of the replication termination protein. This is a small protein, and would therefore be expected to have a relatively small threading search space. For reasons discussed earlier, the match to histone H5 was rejected for the simple reason that an insufficient amount of the structure had been aligned. However, looking at the threading alignment in this case, it is clear that the program has done an excellent job of finding an optimal threading, and this agrees well with the structural alignment. The rmsd for the implied threaded model is only 4.5 Å when compared with the crystallographic model (secondary structure Cα atoms only), showing again that the threaded model is in good agreement with the native structure. Everything appears to be right at this stage in that the alignment is right and the model has a low energy. However, is the energy gap separating this model from the next best in the fold library significant? The answer to this is unfortunately no: if a sufficiently extensive threading search is performed and the gap penalties adjusted to provide the lowest possible energy structure, another chain fold does indeed still provide a model almost indistinguishable in terms of energy, and again this chain fold is found to be endonuclease III. The lowest absolute pairwise energy for the optimal threading of replication termination protein onto endonuclease III is found to be −247 kcal/mol and yet the best model based on histone H5 is found to have an energy of −239 kcal/mol. Both these models are reasonable in terms of the energy evaluation, but clearly only the histone H5 match can provide a correct model.

To find out why the energy for the endonuclease model is as low as that for the histone model, we took the two models and compared them with the X-ray structure. In terms of Cα rmsd, the model based on endonuclease is 11.3 Å, compared to 4.5 Å in the case of the histone model. However, the calculation of threading energy is based solely on *distances,* rather than 3-D coordinates. In the case of the endonuclease model, are all the distances wrong to the same degree (resulting in a total rmsd of 11.3 Å) or are some distances accurately matched and some matched very badly. The answer to this is clear when the distributions of distance errors are examined. The largest distance error in the model based on histone H5 is only 14.5 Å, but for the endonuclease model, some of the distance errors are as great as 32 Å. However, in stark contrast to this observation, is the observation that the endonuclease model has a much larger number of very accurately matched distances. Overall, the histone model has only 118 distances correct to less than 0.5 Å, and yet the endonuclease model has as many as 152 distances correct to this degree. These data highlight a major deficiency in the threading potentials used, and it is probable that this same fault permeates every other threading potential developed to date. This deficiency can perhaps be expressed most succinctly by the idiom "a miss is as good as a mile"—in other words, the threading potential correctly assigns low energies to interactions that are very accurately modeled, but does not distinguish interactions that are only slightly inaccurately modeled from those which are grossly wrong. The fact that the endonuclease model has the most correctly matched distances is all that is effectively recognized, but this alone is not enough to guarantee that the native structure resembles that of the model. To ensure that the folds match there should be no gross distance errors; these gross errors indicate a radical difference in the overall topologies of the two structures.

A similar pattern is observed for the prediction of the synaptotagmin C2 domain. In this case, the fold was of course predicted correctly, but the fold expected to be the best match for synaptotagmin (tenascin) was not found at the top of the ranked list. In comparison with tenascin, CD8 has a long β hairpin near the C-terminus, highly similar to the equivalent structure in synaptotagmin. This is responsible for providing a large number of identical distances, and in this part of the threading alignment is in precise agreement with the structural alignment.

DISCUSSION AND CONCLUSIONS

The "take home" message that is clear from the results presented here and at the meeting is that threading techniques are very effective. The degree of predictive success achieved is even more remarkable when it is realized that practical methods for fold recognition were not in existence prior to 1991, marked by the publication of the profile method of Bowie et al.,[2] and then the subsequent work on threading methods.[1]

Considering the pure prediction results alone, our method has achieved a very acceptable success rate, recognizing 7 out of 11 structures for which a substantial degree of similarity was observed with a previously solved structure. Adding the postdiction results for xylanase, rtp, and pcna (which were mispredicted for trivial reasons) to this total gives us 11 out of 13. The only targets which we probably would not have predicted correctly under ideal conditions would be staufen3 (though we cannot say this for certain as we do not have coordinates for this target) and prosub, and even in this latter case the prediction could have been rescued by considering the combined *Z*-score rather than the pairwise *Z*-score alone.

From these results one might reasonable conclude that the fold recognition problem is more or less solved. However, this would not be correct. The lack of variety in the prediction targets makes a thorough assessment of the relative merits of different methods fairly difficult. We have certainly collected some examples of sequence-structure pairs in our laboratory that would easily defeat any existing fold recognition method, including our own. Despite these reservations, however, we are clearly achieving a very high recognition rate, and our method can quite reliably recognize TIM barrels, and immunoglobulin domains. It has been pointed out that had someone predicted all the folds to be TIM barrels then they would have made more correct predictions than any other group. This is not quite true, but it is in good agreement with the notion of "superfolds,"[17] although it should be emphasized that we made only a single false-positive TIM barrel assignment. In some respects, recognizing TIM barrels and Ig domains might be considered "easy," and more weight be assigned to predictions of more unusual folds (such as the match between the prosubtilisin target and the ferredoxin II fold). However, recognizing common folds is a necessary first step, and it is important that these folds be recognized reliably.[17] It must be realized that the problem of recognizing common folds must be solved *before* the problem of recognizing more esoteric folds be considered. If a method fails to systematically recognize common folds then this must be seen as a very negative trait. There is really insufficient data available to evaluate the performance of the prediction methods on esoteric folds, but our method is certainly well on the way to solving the problem of recognizing common folds.

As discussed earlier, the quality of the alignments considered globally was not as good as might be expected. However, the correctness of alignments in specific regions tells us much about the way in which threading works. Clearly matches are not required across the whole fold. In some cases, however, smaller folds are correctly aligned across the entire fold (especially for cases which include helices as an integral part of the architecture). For larger structures, such as TIM barrels, it is clear that only a few supersecondary structures are correctly aligned, and this seems sufficient to allow the fold to be recognized. These results suggest that key structural features in protein folds act as "fingerprints" for the given motif, and these features mostly emphasize clusters of short distance interactions (close contacts). Structures which can provide the largest number of ideal close contacts for the given sequence will be the structures that are selected from the library, and for this reason it is not necessarily the structures which offer the lowest global RMSD score which are selected by the threading procedure. If this is the basis of fold recognition by threading, then prospects are not good that accurate models will be able to be built for larger structures. It would appear that results for threading small domains are good enough to allow low resolution models, but for structures as complex as TIM barrels, for example, there is still some way to go. However, these early results are very encouraging, and it is certain that radical improvements will be made to these methods over the coming years. One possibility we are investigating is the construction of idealized average structural models for the superfolds, which incorporate knowledge on the kinds of structural variation allowed within the given fold family.

We can extend this interpretation of our own results to the wider question of how other methods compare with our own. Ideally, fold recognition would result from a consideration of all the pairwise interactions in a protein fold, not just the close interactions, but the very long distance interactions as well. However, this would require a "magical" potential function which gives a meaningful energy for the interaction of say a serine and a leucine when separated by as much as 40 Å. It is clear that physically, the presence of a leucine 40 Å away is going to have little (if any) influence on the presence of a serine at a given site, and so we cannot expect this information to come from a statistical analysis of the database. If we were able to construct such a function then the protein folding problem per se would be solved. Unfortunately we are left with the situation that sequence–structure-specific information is abundant only at close range (where hydrophobic contacts and ion-pair contacts are dominant). The main difference between the methods proposed for

fold recognition to date has been the degree of accuracy used to model this interactions. In the original work of Bowie et al.[2] a 1-D profile of the structure was constructed, based on averaged properties (accessibility and secondary structure). This effectively averaged across the details of specific pair interactions, and while these profiles are very computationally efficient, they do not offer the greatest degree of discrimination between folds. The next level up from this is to consider contacts only. Considering errors in matched distances again, this would relate to concentrating purely on distance variations of between 0 and 1.5 Å. Unfortunately these contacts alone are not diagnostic of a given fold—as can be seen, for example, for replication termination protein and endonuclease III, more suitable contacts can be found in a structure entirely different from the native structure. The next step appears to be to consider interactions beyond contacts and to take the distances into account. This approach appears to be the most effective, but there are problems. It is important not to swamp the short distance contact information with the much less reliable interactions at distances beyond 10 Å. There is clearly going to be more sequence-specific information at distances closer than 10 Å than beyond this distance where only general electrostatic and solvation preferences will dominate, but it is not clear how the relative contributions should be weighted. Contacts are the most important, but the less distinct interactions beyond contacts must not be ignored. It should be noted in passing that the problem of being able to recognize only short distance interactions has a parallel in the processing on NMR spectroscopic data, where NOEs can be detected only for protons separated by at most 5 Å or so. In NMR structural determination, as with the threading results discussed above, the overall topology of a molecule can be uncertain even after a large number of close contacts have been assigned. It is possible that improvements in NMR data analysis could be applied to threading methods, or perhaps vice versa.

Perhaps the overriding factor that came across in the results from the meeting is that there is no substitute for biological knowledge. Is the predicted fold a superfold,[17] or is the target protein functionally related to the predicted fold? Is the generated model plausible, or are there unreasonably large cavities present? Can information be gleaned from multiple sequences? Is there any experimental evidence that can be brought into play? These are just a few of the factors that should be considered when making a prediction in the "real world." The conditions of the contest were fairly close to real world conditions, but there were gaps which would normally be usefully filled by discussing the predictions with the people who have the most knowledge of the protein in question (the experimentalists). Results from, for example, threading analysis should be used only to suggest further study or to make experimentally *testable* hypotheses.

The current version of the threading software (THREADER), developed by D. Jones, is available via anonymous FTP from server ftp.biochem.ucl.ac.uk in directory pub/THREADER. The software is free to academic users. A graphical user interface to THREADER which allows easy manual assessment of threading results, developed by R. Miller, is also available from the same site in directory pub/px. Our fold classification resource, CATH (Class Architecture Topology Homology) which has proved very useful for interpreting threading results, can also be accessed via World Wide Web (http://www.biochem.ucl.ac.uk/bsm/index.html).

ACKNOWLEDGMENTS

We thank Willie Taylor and Christine Orengo for useful discussions. We are also grateful to Christine Orengo and Alex Michie for help with structural alignments and their CATH structure classification database. We are, of course, as always grateful to the crystallographers who have contributed structures to the databank, and in particular those who contributed unpublished coordinate sets to the prediction target library.

REFERENCES

1. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature (London) 358:86, 1992.
2. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
3. Godzik, A., Skolnick, J. Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. Proc. Natl. Acad. Sci. U.S.A. 89:12098–12102, 1992.
4. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein-sequence through the folding motif. Proteins Struct. Funct. Genet. 16:92–112, 1993.
5. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. 232:805–825, 1993.
6. Nishikawa, K., Matsuo, Y. Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. Protein Eng. 6:811–820, 1994.
7. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
8. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol. 213:859–883, 1990.
9. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J. Identification of native protein folds amongst a large number of incorrect models: The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216: 167–180, 1990.
10. Kabsch, W., Sander, C. Dictionary of protein secondary structure—pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2377–2637, 1983.
11. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-base archival file for macromolecular structures. J. Mol. Biol. 112:535–543, 1977.
12. Orengo, C.A., Brown, N.P., Taylor, W.R. Fast structure

alignment for protein databank searching. Proteins. 14: 139–167, 1992.
13. Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24:946–950, 1991.
14. Kuriyan, J., ODonnell, M. Sliding clamps of DNA polymerases. J. Mol. Biol. 234:915–925, 1993.
15. Fischer, D., Wolfson, H., Nussinov, R. Spatial, sequence-order-independent structural comparison of alpha/beta proteins: Evolutionary implications. J. Biomol. Struc. Dyn. 11:367–380, 1993.
16. Flores, T.P., Moss, D.S., Thornton, J.M. An algorithm for automatically generating protein topology cartoons. Protein Eng. 7:31–37, 1994.
17. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature (London) 372:631–634, 1994.