6. Fields, S. & Jang, S. K. *Science* **249,** 1046–1049 (1990).
7. Raycroft, L., Wu, H. Y. & Lozano, G. *Science* **249,** 1049–1051 (1990).
8. O'Rourke, R. W. *et al. Oncogene* **5,** 1829–1832 (1990).
9. Zhu, H., Roy, A. L., Roeder, R. G. & Prywes, R. *New Biologist* **3,** 455–464 (1991).
10. Prywes, R., Fisch, T. M. & Roeder, R. G. *Cold Spring Harbor Symp. quant. Biol.* **53,** 739–748 (1988).
11. Kern, S. E. *et al. Oncogene* **6,** 131–136 (1991).
12. Martinez, J., Georgoff, I., Martinez, J. & Levine, A. J. *Genes Dev.* **5,** 151–159 (1991).
13. Milner, J., Medcalf, E. A. & Cook, A. C. *Molec. cell. Biol.* **11,** 12–19 (1991).
14. Friedman, P. N., Kern, S. E., Vogelstein, B. & Prives, C. *Proc. natn. Acad. Sci. U.S.A.* **87,** 9275–9279 (1990).
15. Levine, A. J. *Virology* **177,** 419–426 (1990).
16. Fried, M. & Prives, C. *Cancer Cells* **4,** 1–16 (1986).
17. Wang, E., Friedman, P. N. & Prives, C. *Cell* **57,** 379–392 (1989).
18. Weintraub, H., Hauschka, S. & Tapscott, S. J. *Proc. natn. Acad. Sci. U.S.A.* **88,** 4750–4751 (1991).
19. Soussi, T., Caron de Fromentel, C. & May, P. *Oncogene* **5,** 945–952 (1990).
20. Gannon, J. V., Greaves, R., Iggo, R. & Lane, D. P. *EMBO J.* **9,** 1595–1602 (1990).
21. Milner, J. & Medcalf, E. A. *Cell* **65,** 765–774 (1991).
22. Stenger, J. E. *et al. Molec. Carcinogenesis* **5,** 102–106 (1992).
23. Mercer, W. E., Shields, M. T., Lin, D., Appella, E. & Ullrich, S. J. *Proc. natn. Acad. Sci. U.S.A.* **88,** 1958–1962 (1991).
24. Santhanum, U., Ray, A. & Sehgal, P. B. *Proc. natn. Acad. Sci. U.S.A.* **88,** 7605–7609 (1991).
25. Ginsberg, D., Mechta, F., Yaniv, M. & Oren, M. *Proc. natn. Acad. Sci. U.S.A.* **88,** 9979–9983 (1991).
26. Ptashne, M. *Nature* **335,** 683–689 (1988).
27. Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J. & Howley, P. M. *Cell* **63,** 1129–1136 (1990).
28. Crook, T., Tidy, J. A. & Vousden, K. H. *Cell* **67,** 547–556 (1991).
29. Murakami, Y., Asano, M., Satake, M. & Ito, Y. *Oncogene* **5,** 5–14 (1990).
30. Kern, S. E. *et al.* Science **256,** 827–830 (1992).
31. Zambetti, G. P. *et al. Genes Dev.* (in the press).

ACKNOWLEDGEMENTS. We thank B. Vogelstein and S. Kern for sharing information and reagents, and E. Freulich for technical assistance. This work was supported by grants from the NIH (C.P.), March of Dimes Foundation and Searle Scholars Program (R.P.), and the Damon Runyon Fund (J.B.).

A new approach to protein fold recognition

D. T. Jones*†, W. R. Taylor† & J. M. Thornton*

* Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular Biology,
University College, Gower Street,
London WC1E 6BT, UK
† Laboratory of Mathematical Biology, National Institute for Medical Research,
The Ridgeway, Mill Hill, London, NW7 1AA, UK

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs[1] or folds[2–4] in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

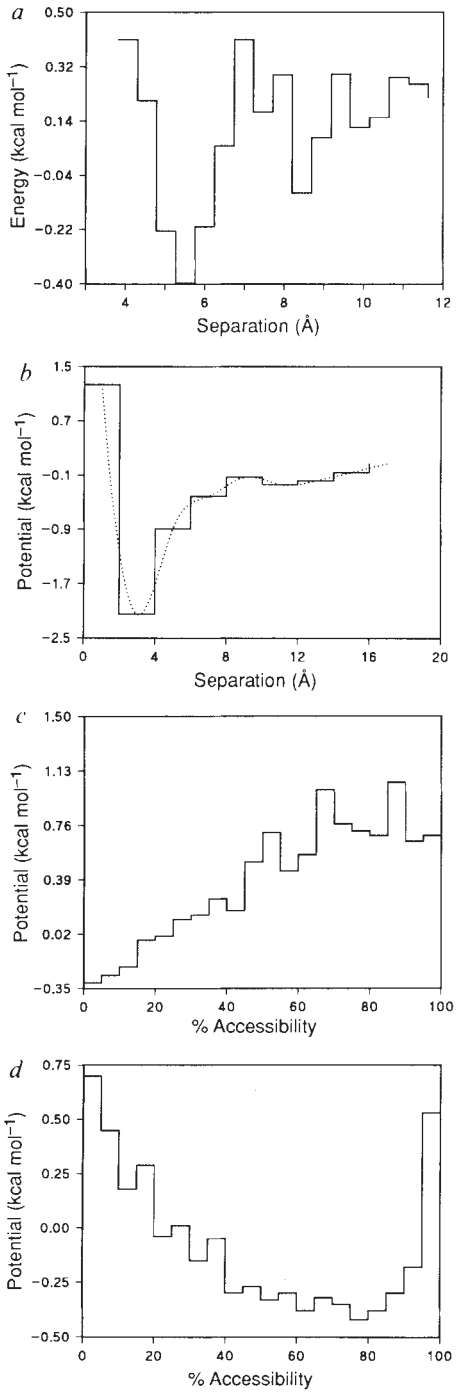
In outline our method is simple. A library of different protein folds is derived from the database of protein structures. In our case, the library contained all the unique, moderately well resolved chains (sequence identity $< 30\%$, resolution ≤ 2.8 Å) in the July 1991 release of the Brookhaven database[5], totalling 102 chains. Each fold is considered as a chain tracing through space; the original sequence being ignored completely. The test sequence is then optimally fitted to each library fold (allowing for relative insertions and deletions in loop regions), with the 'energy' of each possible fit (or threading) being calculated by summing the proposed pairwise interactions. 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.

In previous work, the difficult problem of optimizing the threading of the test sequence onto a structure with respect to the detailed pairwise interactions has been avoided by matching at the sequence level. At its most basic, this involves matching the sequence of the given fold with the test sequence, scoring each residue–residue match by means of a score matrix such as the Dayhoff matrix[6]. But there are now many examples of proteins exhibiting high structural similarity yet little or no similarity in their sequences (sequence identity $< 15\%$). In view of this, several groups have attempted to match sequences to folds by describing the fold not in terms of its amino-acid sequence, but in terms of the environment of each residue location in the structure[2–4]. The environment (for example, the local secondary structure and solvent accessibility) of a particular residue tends to be more highly conserved than the identity of the residue itself, and so methods that match each residue in the test sequence to the environments of each residue in a protein fold are able to detect more distant sequence–structure relationships than purely sequence-based methods. Finkelstein and Reva attempted to take pairwise interactions into account by addressing the problem of fitting a sequence onto idealized lattice models of 8-stranded β-sandwich folds using an iterative procedure[3,7]. We have used a dynamic programming-based algorithm[8,9] capable of optimizing pairwise interactions, which was originally applied to the problem of structural comparison. This algorithm uses a standard sequence alignment method to optimize the threading of the sequence onto the structure around each residue in turn, finally computing the best threading through the whole structure by means of a shortest-path algorithm.

To evaluate the energy of a sequence in a particular conformation we need a set of potentials for residue interactions that do not require explicit modelling of all side-chain atoms. Previous work[10] has shown that classical potentials (for example, CHARMm[11]) cannot identify proteins that have been folded into non-native conformations. For these reasons we use a set of knowledge-based potentials and explicitly consider the degree of residue solvation, both of which do in fact identify such misfolded proteins[12,13]. In particular, we use a set of pairwise potentials similar to those described by Sippl[14] which are derived from a statistical analysis of known protein structures (see Fig. 1 legend for details). For a given pair of atoms, a given residue sequence separation and a given interaction distance, these potentials provide a measure of pseudo-energy, which relates to the probability of observing the proposed interaction in native protein structures. By dividing the empirical potentials into sequence separation ranges, specific structural significance may be tentatively conferred on each range. For instance, the short-range terms predominate in the matching of secondary structural elements. By threading a sequence segment onto the template of an α-helical conformation and evaluating the short-range potential terms, the probability of the sequence folding into an α-helix may be evaluated. In a similar way, medium-range terms mediate the matching of super-secondary structural motifs, and the long-range terms, the tertiary packing. Some sample potentials are shown in Fig. 1*a–d*.

Our medium and long-range pairwise potentials differ from those proposed in ref. 14 in that interactions beyond 10 Å are ignored. These interactions are not residue-specific and are determined simply by solvation effects. In place of these long-distance terms, we substitute a 'solvation potential'. This potential measures the propensity of each amino-acid type for a certain degree of solvation, approximated by the residue solvent-accessible surface area.

Many studies have shown that only the cores of distantly related structures are conserved, therefore in calculating the energy of a given threading we ignore all pairwise terms involving loop residues. Loop positions are evaluated by the solvation potential alone, which takes into account the tendency for loop regions to be exposed to solvent.



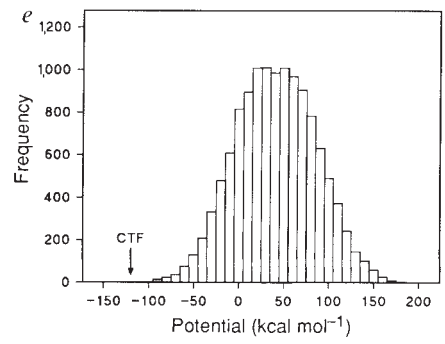


FIG. 1 Samples of the statistically derived potentials are shown. *a*, Short-range ($k = 3$) Ala-Ala $C\beta \rightarrow C\beta$ interaction. Low-energy states are observed for distances around 6 Å, corresponding mainly to α-structure, and 9 Å, corresponding mainly to β-structure. *b*, Long-range ($k > 30$) Cys-Cys $C\beta \rightarrow C\beta$ interaction. The most significant energy minimum around 4 Å corresponds to disulphide bridge formation. *c*, Solvation potential for leucine, and *d*, solvation potential for glutamic acid. *e*, Threading histogram for the C-terminal ribosomal protein fragment, 1CTF. All possible threadings of the CTF sequence on the CTF structure were computed (secondary structure gaps disallowed) and the energies of each threading calculated. The native threading is indicated, and was found to be the lowest energy threading.
METHODS. The calculation of pairwise pseudo-energy terms has been described[14]. For specified atoms ($C\beta \rightarrow C\beta$ for example) in a pair of residues *ab*, topological level (sequence separation) k and distance interval s, the potential is given by the following expression

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

where m_{ab} is the number of pairs *ab* observed at topological level 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, and $f_k^{ab}(s)$ is the equivalent frequency of occurrence of residue pair *ab*. RT is taken as 0.582 kcal mol^{-1}. Short- (sequence separation, $k \leq 10$), medium- ($11 \leq k \leq 30$) and long- ($k > 30$) range potentials have been calculated between the following atom pairs: $C\beta \rightarrow C\beta$, $C\beta \rightarrow$ N, $C\beta \rightarrow$ O, N $\rightarrow C\beta$, N $\rightarrow$ O, O $\rightarrow C\beta$ and O $\rightarrow$ N. Similarly, the solvation potential for an amino-acid residue *a* is defined as

$$\Delta E_{solv}^{a}(r) = -RT \ln \left[\frac{f^a(r)}{f(r)}\right]$$

where r is the % 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[19] applied to Brookhaven coordinate files. For multimeric proteins, only the chains explicitly included in the coordinate files were taken into account.

An obvious first test of these potentials was to attempt to thread a sequence onto its own native structure. Taking a number of small structures, for which it was practical to evaluate every possible threading (disallowing gaps in regions of regular secondary structure), we have found that the native threading of a sequence onto its own structure is usually found to be the lowest energy threading. As an example, the native threading histogram for the C-terminal ribosomal protein fragment (CTF) is shown in Fig. 1*e*. One small protein for which the native threading does not have the lowest evaluated energy is crambin, but this is attributable to the fact that this protein is not soluble in water, and consequently the solvation effects are not correctly modelled by our solvation potential.

To demonstrate the capability of our method for recognizing protein folds and generating an accurate sequence-structure alignment, we consider here the example of C-phycocyanin. The striking feature of the chain fold of C-phycocyanin is that the globular portion (helices A-H) closely resembles the globin fold[15]. Despite the similarity in fold, the sequence homology between the globins and C-phycocyanin is very low, with only 14 identities between the 174 β-chain residues of C-phycocyanin and sperm whale myoglobin. So far, sequence analysis methods have proved unable to detect the globin fold in C-phycocyanin. For example, despite success in constructing templates to select almost every available globin sequence, it has not been possible to match these templates against the phycocyanins[16]. Using the optimal threading algorithm, the C-phycocyanin sequence was threaded on each member of the library of protein folds to find its most compatible fold. The two lowest-energy folds were found to be sea hare myoglobin (−451 kcal mol^{-1}) and midge

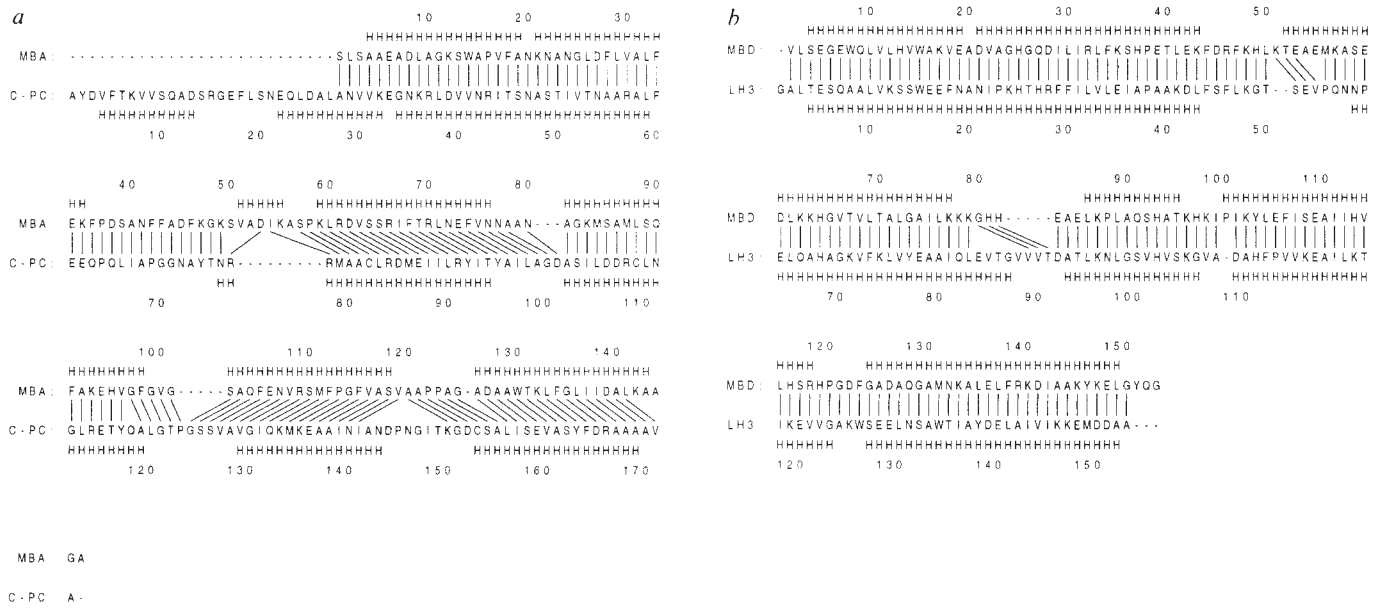


FIG. 2 *a*, Alignment of sea hare myoglobin (1MBA) with C-phycocyanin (C-PC) β-chain from *Mastigocladus laminosus* (SWISSPROT code PHCB$MASLA), found by optimal threading. Author-assigned secondary-structure codes are shown. The alignment is compared to the structurally determined alignment by Pastore and Lesk, where lines are drawn between structurally equivalent residue pairs as determined in the reference alignment[17]. *b*, Optimal threading of yellow lupin leghaemoglobin (LH3) on the structure of sperm whale myoglobin (Brookhaven code 1MBD). Alignment of protein structures is compared with the structural alignment obtained using the program SSAP[9].

erythrocruorin (-356 kcal mol^{-1}) followed by several other all-α protein folds. Figure 2*a* shows the alignment corresponding to the optimal threading of the C-phycocyanin β-chain sequence onto the best matching fold (sea hare myoglobin). For comparison, the optimal threading alignment of myoglobin and leghaemoglobin is shown in Fig. 2*b*. In terms of sequence, myoglobin and leghaemoglobin are only distantly related (17% sequence identity), but their structural similarity is much higher than in the case of phycocyanin and myoglobin, leading to a relatively unambiguous alignment. It should be borne in mind that even the structural alignment of phycocyanin and myoglobin is uncertain[17]. The fact that the optimal threading algorithm finds sea hare myoglobin to be the best model for C-phycocyanin is in accordance with a report[17] in which the

TABLE 1 Summary of trial fold-recognition searches

Test protein	Source	Fold	Best match	ΔE	% Sequence identity	Matches
C-phycocyanin β (C-PC)	Red algae	Globin	1MBA	101	7	1, 2, 9, 18, 25
Glycolate oxidase (GOX)	Spinach	TIM barrel	1WSY(A)	52	10	1, 3, 49
Muscle aldolase (ALD)	Human	TIM barrel	4XIA(A)	80	6	1, 2, 3
Lactate dehydrogenase (LDH)	Dogfish	Rossmann	4MDH(A)	87	15	1*
Elastase (EST)	Pig	Trypsin	4PTP	110	35	1, 14
CD4	Human	Ig	2FB4(H)	87	10	1, 2, 31
Stellacyanin (STEL)	Varnish tree	Cu binding	2AZA(A)	18	14	1, 6, 20
Cytochrome B562 (B562)	*E. coli*	4-helix bundle	2MHR	78	6	1
Trypsin inhibitor DE-3 (ETI)	Kaffir tree	Interleukin 1β	1I1B	14	5	1
PapD-chaperonin	*E. coli*	Ig	2FB4(L)	64	15	1, 5, 9, 35
70K, Heat-shock cognate (HSC)	Cow	Actin	1ATN(A)	94	9	1
Hexokinase B (HEX)	Yeast	Actin	1ATN(A)	0	12	1

In each case the database included 102 protein chains, except where the test protein was itself in the database, in which case it was excluded. Templates for each chain were constructed as described in the text, with residues not in helices or strands (as calculated by DSSP[19]) assigned as loop residues. For the 70K heat-shock cognate protein and hexokinase searches, the coordinates for actin were also included (coordinates deposited under the code 1ATN, but not yet released). Proteins with >25% sequence identity to the test protein were also excluded from the calculation of potentials. The pairwise and solvation terms were summed and stored separately, and standard deviations (s.d._{pair} and s.d._{solv}) for the two contributing factors calculated over the set of 102 folds. To balance the contributions of the pairwise and solvation terms, the final energy was taken as $E = E_{pair} + WE_{solv}$, where $W = (\text{s.d.}_{pair}/\text{s.d.}_{solv})$. The 'confidence' of the match (ΔE) is given in terms of the absolute energy difference between the top scoring fold and the next highest scoring, different, fold. The 'best match' column gives the Brookhaven ID of the best matching chain fold (including chain identity where appropriate), along with the sequence identity between the best matching chain and the test protein. Positions in the sorted list of threading energies of similar folds are also shown. A constant set of alignment parameters (gap penalty for example) was used for all databank searches shown. Typical execution times for a single search of 102 chains are around 100 minutes on a Unix workstation (Solbourne 5/602). The 102 chains used were as follows: 351C, A256B, 2AAT, 1ABP, A5ACN, 8ADH, 3ADK, A8ATC, B8ATC, A2AZA, 3BLM, 1BP2, 2CA2, A7CAT, 1CC5, 1CCR, A2CCY, 1CD4, 2CDV, 3CLA, 2CNA, I4CPA, 5CPA, 2CPP, 4CPV, 1CRN, 2CRO, E1CSE, I1CSE, 1CTF, 1CY3, 2CYP, 3DFR, A4DFR, A1DHF, 1ECA, E2ER7, H2FB4, L2FB4, 1FD2, 1FX1, 3FXC, 4FXN, A3GAP, 2GBP, 1GCR, O1GD1, 3GRS, A3HHB, 1HIP, A2HLA, B2HLA, 1HOE, 1I1B, 3ICB, 3ICD, 1L01, 2LBP, 6LDH, 1LH1, 31LRD, A2LTN, 1LZ1, 1MBA, 1MBD, A4MDH, 2MHR, 2OVO, A2PAB, 9PAP, 1PAZ, 1PCY, A1PFK, 3PGK, 3PGM, 1PHH, 5PTI, 4PTP, 1RHD, 2RHE, 2RNT, 7RSA, 4RXN, 2SGA, I4SGB, 1SN3, 2SNS, O2SOD, 2SSI, 2STV, I1TGS, E2TMN, 4TNC, A1TNF, 1UBQ, 1UTG, A9WGA, R2WRP, A1WSY, B1WSY, A4XIA, A1YPI.

* Other topologically similar (yet structurally different) parallel $\alpha\beta$ folds were positioned at 3, 7, 11, 12, 13, 17, 19, 31, 34, 82.

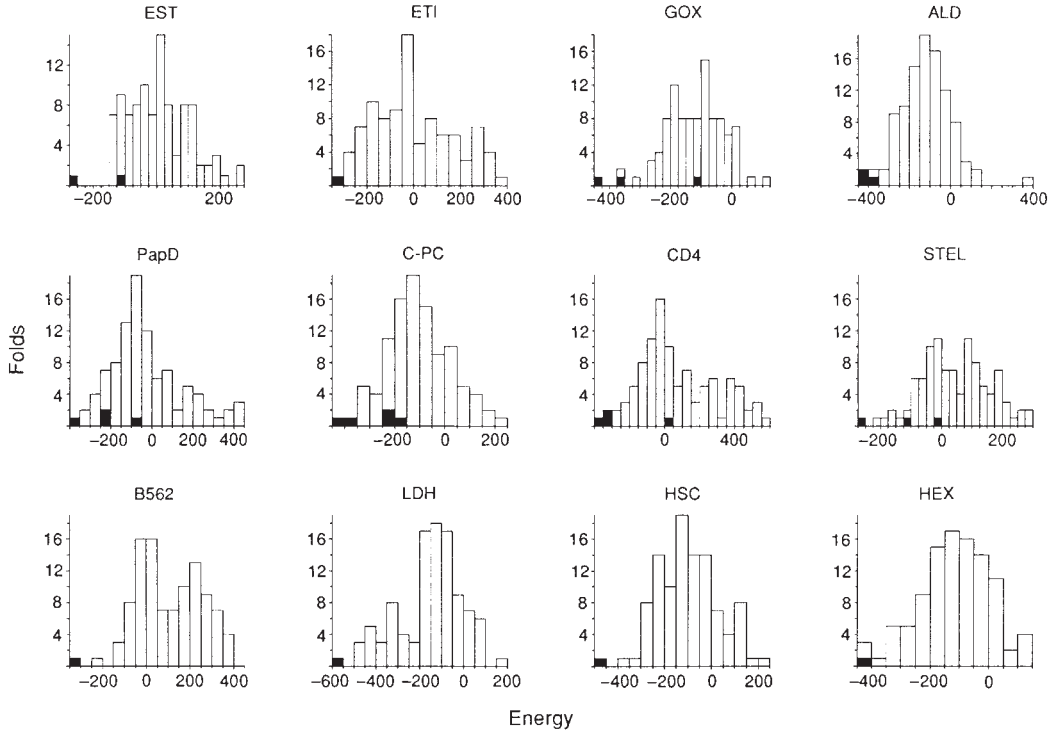


FIG. 3 For a number of test cases (see Table 1) the histogram of energies for optimally threading onto each of the 102 folds is given. In each histogram, the positions of folds expected to match the given sequence (that is, those folds similar to the known fold of the test sequence) are shown as filled bars. For example, in the case of LDH (lactate dehydrogenase), the expected match in the database of folds is MDH (malate dehydrogenase). This match is shown as a single filled bar representing an energy of -577 kcal mol^{-1}, an energy which is lower than that achieved by any other fold. As noted in the text, in some cases expected folds are apparently not detected. This occurs for two reasons: either the expected structures are not sufficiently similar to the native fold, or the optimization method does not succeed in producing a satisfactory alignment. The C-PC results demonstrate the former case. A number of unrelated highly helical proteins (carp parvalbumin and T4 lysozyme, for example) score better than the low-scoring globins. The worst-scoring globin is in fact human haemoglobin, in which case the poor score is due not only to the substantial secondary structural shifts relative to C-PC, but also to the fact that the calculated accessibilities are for the complete tetramer. The second situation arises in the results for GOX, where the algorithm fails to find an optimal threading of GOX onto XIA (xylose isomerase), resulting in an unexpectedly poor score. Of particular note in the results shown are the $(\alpha\beta)_8$ (TIM) barrel and trypsin inhibitor DE-3 examples. The degree of sequence homology between different $(\alpha\beta)_8$ barrel enzyme families and between trypsin inhibitor DE-3 and interleukin-1β is extremely low (5–10%). As a consequence of this, again, sequence template methods have been unable to detect these folds. Also of note are the results for the 70K heat-shock cognate protein (HSC70), and yeast hexokinase B. The N-terminal ATPase fragment of the heat-shock cognate protein has an almost identical structure to that of actin, but the similarity between hexokinase and actin is more topological than at the level of specific structural interactions. The two degrees of similarity are borne out by the threading results for these proteins, in that although actin is the lowest energy fold for hexokinase, the separation between the actin fold and the next-best-matching fold (aspartate transcarbamylase, ATC) is almost zero (0.1 kcal mol^{-1}); the rather weak structural similarity between hexokinase and actin would therefore appear to be just at the limits of our method. In contrast, the match between HSC70 and actin is clearly significant.

helix geometry of this globin was found to be closest to that of C-phycocyanin. Not only has the method correctly identified its globin fold, but has accurately located it in the C-phycocyanin sequence and has generated an alignment close to that obtained by careful structural alignment. It is clear that the method has identified the related folds in the database. It should be emphasized that no specific sequence information was used in the threading process: the structure was considered only as a chain of anonymous placeholders onto which the given sequence is threaded.

The results of other trial searches using the method of optimal sequence threading are shown in Table 1 and Fig. 3. From Fig. 3 it is apparent that in some cases expected matches are far from the top of the list. On inspection it was found that in these cases the threading algorithm had clearly misaligned the proteins, and had failed to find a reasonable optimum of the objective function, although this could generally be corrected by adjusting the alignment gap penalty.

The method described here shows promise as a new means for sensitively recognizing protein folds, and it is evident from the results that new information beyond sequence similarity is being exploited here. We are now exploring the generation of model folds[3,18], to escape from the limitation of only being able to predict previously observed folds, and the incorporation of multiple sequence data (from aligned sequence families) in the recognition process. □

Received 5 February; accepted 21 May 1992.

1. Taylor, W. R. & Thornton, J. M. *J. molec. Biol.* **173,** 487–514 (1984).
2. Overington, J., Johnson, M. S., Šali, A. & Blundell, T. L. *Proc. R. Soc.* B**241,** 132–145 (1990).
3. Finkelstein, A. V. & Reva, B. A. *Nature* **351,** 497–499 (1991).
4. Bowie, J. U., Lüthy, R. & Eisenberg, D. *Science* **253,** 164–170 (1991).
5. Bernstein, F. C. *et al. J. molec. Biol.* **112,** 535–542 (1977).
6. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *Atlas of Protein Sequence and Structure* Vol. 5 suppl. 3 345–352 (Natn. Biomed. Res. Fnd, Washington DC, 1978).
7. Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. *Nature* **354,** 105–106 (1991).
8. Taylor, W. R. & Orengo, C. A. *J. molec. Biol.* **208,** 1–22 (1989).
9. Orengo, C. A. & Taylor, W. R. *J. theor. Biol.* **147,** 517–551 (1990).
10. Novotny, J., Bruccoleri, R. E. & Karplus, M. *J. molec. Biol.* **177,** 787–818 (1984).
11. Brooks, B. *et al. J. comp. Chem.* **4,** 187–217 (1983).
12. Hendlich, M. *et al. J. molec. Biol.* **216,** 167–180 (1990).
13. Eisenberg, D. & McLachlan, A. D. *Nature* **319,** 199–203 (1986).
14. Sippl, M. J. *J. molec. Biol.* **213,** 859–883 (1990).
15. Schirmer, T., Bode, W., Huber, R., Sidler, W. & Zuber, H. *J. molec. Biol.* **184,** 257–277 (1985).
16. Bashford, D., Chothia, C. & Lesk, A. M. *J. molec. Biol.* **196,** 199–216 (1987).
17. Pastore, A. & Lesk, A. M. *Proteins* **8,** 133–155 (1990).
18. Taylor, W. R. *Prot. Engng* **4,** 853–870 (1991).
19. Kabsch W. & Sander C. *Biopolymers* **22,** 2577–2637 (1983).

ACKNOWLEDGEMENTS. We thank T. P. Flores, S. J. Hubbard, C. A. Orengo and M. B. Swindells for discussion, and K. C. Holmes for permission to use the coordinates for actin. D.T.J. acknowledges receipt of an SERC CASE studentship with the MRC.