J. Mol. Biol. (1993) **233**, 123–138

Protein Structure Comparison by Alignment of Distance Matrices

Liisa Holm and Chris Sander

Protein Design Group, European Molecular Biology Laboratory
D-69012 Heidelberg, Federal Republic of Germany

(Received 27 May 1992; accepted 30 April 1993)

With a rapidly growing pool of known tertiary structures, the importance of protein structure comparison parallels that of sequence alignment. We have developed a novel algorithm (DALI) for optimal pairwise alignment of protein structures. The three-dimensional co-ordinates of each protein are used to calculate residue–residue (C^{α}–C^{α}) distance matrices. The distance matrices are first decomposed into elementary contact patterns, e.g. hexapeptide–hexapeptide submatrices. Then, similar contact patterns in the two matrices are paired and combined into larger consistent sets of pairs. A Monte Carlo procedure is used to optimize a similarity score defined in terms of equivalent intramolecular distances. Several alignments are optimized in parallel, leading to simultaneous detection of the best, second-best and so on solutions. The method allows sequence gaps of any length, reversal of chain direction and free topological connectivity of aligned segments. Sequential connectivity can be imposed as an option. The method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, even in the presence of geometrical distortions. An all-against-all alignment of over 200 representative protein structures results in an objective classification of known three-dimensional folds in agreement with visual classifications. Unexpected topological similarities of biological interest have been detected, e.g. between the bacterial toxin colicin A and globins, and between the eukaryotic POU-specific DNA-binding domain and the bacterial λ repressor.

Keywords: classification of protein folds; database searching; distance geometry; pattern recognition; protein structure alignment

1. Introduction

Proteins fold into beautiful and complicated three-dimensional (3D†) structures. To date, the tertiary structures of several hundred different proteins have been solved by X-ray crystallography or 2D (two-dimensional) nuclear magnetic resonance (NMR) spectroscopy, and the number is growing rapidly. Even when protein sequences are very different, 3D structures may be surprisingly similar. Their detailed comparison will lead to increased understanding of the principles of protein architecture.

In recent years, a number of automated methods for protein structure comparison have been developed, using different representations of structure, definitions of similarity measure and optimization algorithms (Mitchell *et al.*, 1989; Subbarao & Haneef, 1991; Vriend & Sander, 1991; Fischer *et al.*, 1992; Alexandrov *et al.*, 1992; Barakat & Dean, 1991; Taylor & Orengo, 1989; Sali & Blundell, 1990). Here, we present a general approach for aligning a pair of proteins represented by two-dimensional matrices. The result is a set of structurally equivalent residue pairs, similar to the classical notion of an alignment between two sequences but more general: equivalenced segments can be freely permuted. Below, the term alignment will be used to include also the more general case.

The utility of distance matrices, also called distance plots or distance maps, in describing and comparing protein conformations has been recognized for a long time (Phillips, 1970; Nishikawa & Ooi, 1974; Liebman, 1980; Sippl, 1982). The most commonly used distance matrix is that containing all pairwise distances between residue centers, i.e. C^{α} atoms. A distance matrix is a 2D representation of a 3D structure. The matrix is independent of the co-ordinate frame and contains more than enough

† Abbreviations used: 3D, three dimensions, three-dimensional; 2D, two dimensions, two-dimensional; 1D, one-dimensional; r.m.s.d., root-mean-square deviation of C^{α} positions; PDB, Protein Data Bank; NMR, nuclear magnetic resonance; TIM, triose phosphate isomerase.

0022-2836/93/170123-16 $08.00/0

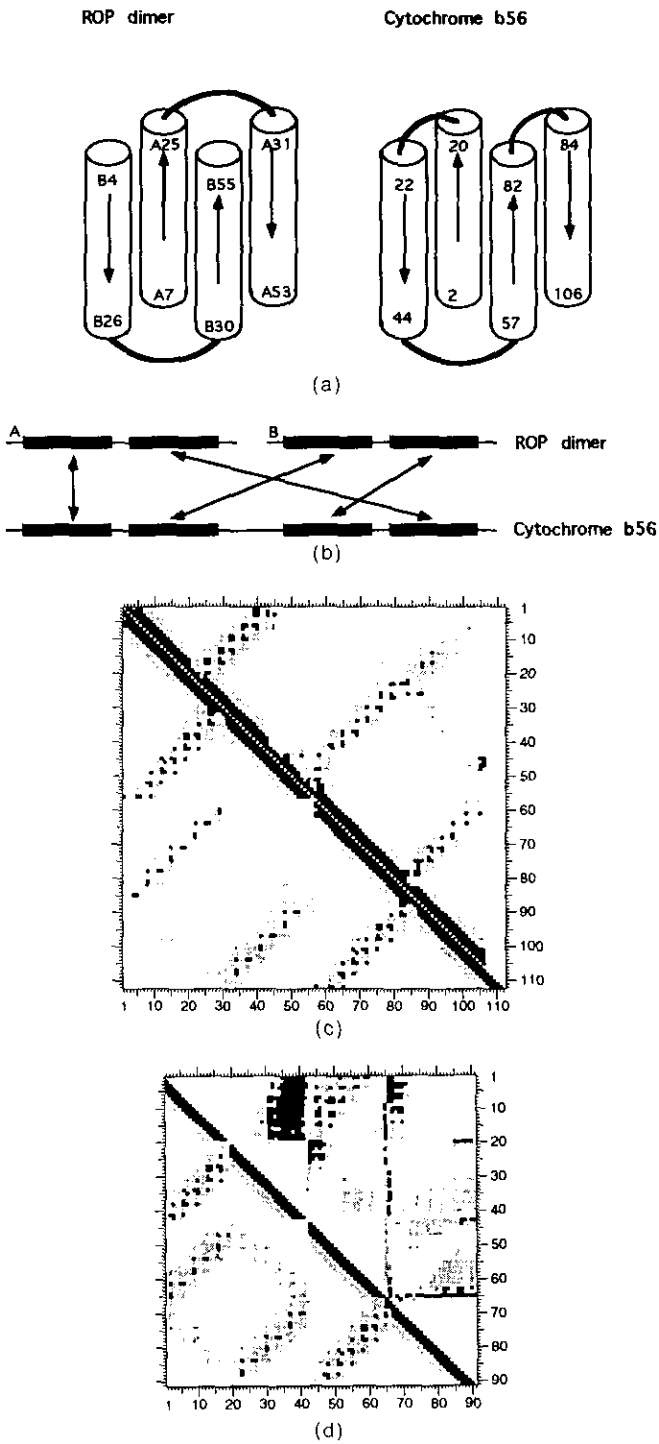


Figure 1. Comparing 3D protein structures by 2D distance matrices. Comparison of 2 4-helix bundles that differ by topological rearrangement, ROP (1ROP, Banner *et al.*, 1988) and cytochrome *b*56 (256B, Mathews *et al.*, 1979). (a) Topological cartoons of 1ROP and 256B. Helices are drawn as cylinders and loops as lines. The 4-helix bundle is formed by 2 identical helical hairpins in ROP (chains A and B) and by a single chain in 256B. Residue numbers of structurally equivalent segments are indicated on the cylinders. (b) The alignment is non-sequential. Engineering ROP into a single-chain bundle requires topological reconnections of loops (Sander, 1990). (c) Full C^{α}–C^{α} distance matrices before alignment, ROP in lower and 256B in upper triangle. All nearest-neighbor helix pairs in the bundles are antiparallel and show up as bands perpendicular to the diagonal. Distances less than 8 Å are black, 8 Å to 12 Å dark gray, 12 Å to 16 Å light gray. (d) After alignment: difference distance matrix for the structurally equivalent segments (upper triangle), using the collapsed distance matrix of ROP (lower triangle) as reference. Shading in the distance difference matrix is from white (less than 1 Å deviation) to black (more than 4 Å deviation). The black area near coordinates (35,10) indicates that the largest difference is in the relative positions of 2 of the helices in ROP compared to the structurally equivalent N and C-terminal helices in 256B. The root-mean-square positional deviation of the 91 equivalenced C^{α} atoms in optimal superimposition is 2·3 Å.

information to reconstruct the 3D structure, except for overall chirality, by distance geometry methods (Havel *et al.*, 1983).

Similar 3D structures have similar inter-residue distances. Imagine a (transparent) distance map of one protein placed on top of that of another protein and then moved vertically and horizontally. Depending on the relative displacement of the matrices, matching substructures appear as patches (submatrices) in which the difference of distances is small. Matching patches centered on the main diagonals correspond to locally similar backbone conformations, i.e. secondary structures. Matches of short distances found off the main diagonals reveal similar tertiary structure contacts. The presence of a common structural motif made up of several disjoint regions of the backbone becomes visible at one glance in a pair of "collapsed" submatrices that are obtained by deleting residues with no structural equivalent in the other structure and permuting rows and columns when topological connectivities differ (Fig. 1).

The assignment of equivalent residue pairs is a non-trivial combinatorial problem. Figure 2 illustrates the basic principle of our approach. The overall match of an alignment is evaluated by summing over the pairwise similarities of all equivalent elements in the collapsed submatrices. The first step of the algorithm divides the distance matrices into overlapping submatrices of fixed size, e.g. hexapeptide–hexapeptide contact patterns, and screens for pairs of similar contact patterns. Each contact pattern implies a subalignment involving two fragments on each protein chain. Starting from a given pair of equivalenced fragments, one can construct a chain of connected contact patterns by identifying another pair of matching contact patterns that share the previously equivalenced fragment, and so on, e.g. (a,b)-(b,c)-(c,d). We build up alignments and maximize their similarity score by iterative improvement using a random walk along the chains of paired contact patterns. Optimization of several alignments in parallel leads to automatic detection of, for example, internal repeats.

The result is a powerful and flexible method for the detection of spatial similarities in protein structures, with or without the sequential constraint. We

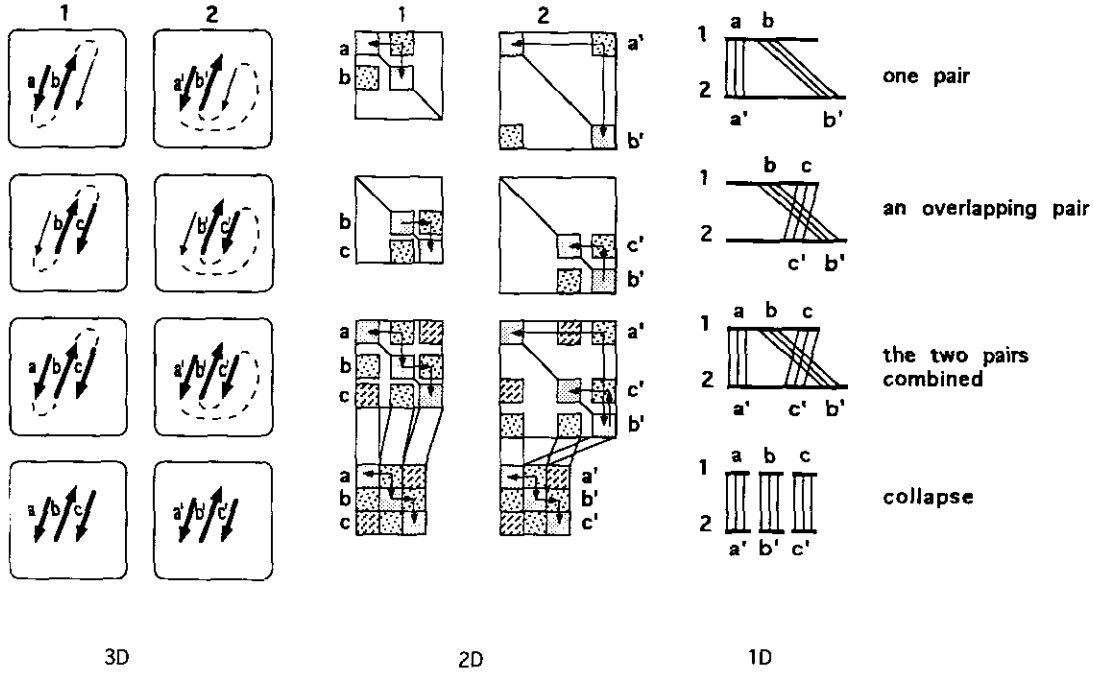


Figure 2. How to maximize the structural overlap of 2 proteins? The algorithm can be followed from top to bottom in 3 schematic representations: left, 3D chain trace; middle, 2D distance matrices; right, 1D sequence alignments. Two topologically different 3-stranded β-sheet proteins (idealized) are being compared. The structurally equivalent fragments are labeled a, b, c in one protein (protein 1) and a′, b′, c′, respectively, in the other (protein 2). In the distance matrices, similar contact patterns are filled with the same pattern, boxes on the main diagonal correspond to intra-fragment distances and off-diagonal boxes to inter-fragment distances. The similarity score of an alignment is calculated from the pairwise differences of all equivalent elements of the 2 distance matrices. Top row: an alignment is initiated from matching contact patterns that equivalence the hexapeptide–hexapeptide pair a-b with a′-b′. Second row: fragments b and b′, which are part of the previous alignment, are used to look for additional fragments by which to extend the alignment. The fragments c and c′ are identified because the contact patterns (b,c) and (b′,c′) are similar. Third row: (a,b)-(a′,b′) and (b,c)-(b′,c′) are merged into the alignment (a,b,c)-(a′,b′,c′). Although the search builds on substructures, the similarity score of the alignment depends on the fitness of each of the equivalenced fragments in the context of all others in an alignment. In this sense, there is a co-operative effect built into the optimization. Bottom row: the final agreement of hexapeptide–hexapeptide contact patterns after the removal of insertions/deletions and reordering of the aligned segments b′ and c′ in the 2nd protein. The resulting 1D alignment is at the lower right. The comparison of contact patterns is independent of sequence gaps or shuffling of segment order and can also identify matches with reversed chain direction.

report tests on a wide range of examples, including an all-against-all alignment of more than 200 representative protein structures, and discuss potential applications to algorithmically related problems in protein folding.

2. Methods

(a) *Definitions*

(i) *Formulation of the problem*

Consider 2 proteins labeled A and B. The match of 2 substructures is evaluated using an additive similarity score S of the form:

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i,j), \qquad (1)$$

where i and j label pairs of equivalent (matched) residues, e.g. $i = (i_A, i_B)$, L is the number of such pairs (the size of each substructure), and ϕ is a similarity measure based on some pairwise relationship, here the C^α–C^α distances d_{ij}^A and d_{ij}^B. Unmatched residues do not contribute to the overall score. For a given functional form of $\phi(i,j)$, the largest value of S corresponds to the optimal set of residue equivalences.

(ii) *Rigid similarity score*

Structural similarity searches can be divided into 2 categories: (1) the search for occurrences of a predefined structural pattern in a structure database, and (2) the search for the largest common substructure between 2 proteins. In the former case, it is natural to define the object function such that it minimizes dissimilarity. In the more general 2nd case, addressed here, we need to define a similarity measure that balances 2 contradictory requirements, that of maximizing the number of equivalenced residues and that of minimizing structural deviations. A simple form of such a residue-pair similarity score ϕ is:

$$\phi^R(i,j) = \theta^R - |d_{ij}^A - d_{ij}^B|, \qquad (2)$$

where the superscript R stands for rigid, d_{ij}^A and d_{ij}^B are equivalenced elements in the distance matrices of proteins A and B and $\theta^R = 1{\cdot}5$ Å is the zero level of similarity. Some of the actual distance deviations within the optimal

set of equivalences can be larger than the similarity threshold θ if compensated for by good fits elsewhere in the alignment, analogous to embedded residue type mismatches in 1D sequence alignment.

(iii) *Elastic similarity score*

The use of relative rather than absolute deviations of equivalent distances makes the elastic (superscript E) variant of the residue-pair score, ϕ^E, more tolerant to the cumulative effect of gradual geometrical distortions:

$$\phi^E(i,j) = \begin{cases} \left(\theta^E - \dfrac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*}\right) w(d_{ij}^*), & i \neq j \\ \theta^E, & i = j \end{cases} \qquad (3)$$

where d_{ij}^* is the average of d_{ij}^A and d_{ij}^B, θ^E is the similarity threshold, and w is an envelope function. We chose $\theta^E = 0{\cdot}20$, i.e. 20% deviation. This means that, e.g. adjacent strands in a β-sheet (typical distance 4 to 5 Å) should match to within 1 Å, while 2 to 3 Å displacements are well tolerated for strand–helix or helix–helix contacts (typical distances 8 to 15 Å). Since pairs in the long distance range are abundant but less discriminative, their contribution is weighted down by the envelope function: $w(r) = \exp(-r^2/\alpha^2)$, where $\alpha = 20$ Å, calibrated on the size of a typical domain. Unless stated otherwise, we report alignments generated using the elastic similarity measure of eqn (3), without imposing the constraint of strictly sequential alignment.

(b) *A greedy algorithm*

The alignment algorithm has 2 steps. The 1st step is a systematic pairwise comparison of all elementary contact patterns in the 2 distance matrices. In this work, we use hexapeptide–hexapeptide contact patterns ($i_A \ldots i_A+5$, $j_A \ldots j_A+5$) in protein A paired with ($i_B \ldots i_B+5$, $j_B \ldots j_B+5$) in protein B, where the hexapeptide $i_A \ldots i_A+5$ is equivalenced with $i_B \ldots i_B+5$ and the hexapeptide $j_A \ldots j_A+5$ is equivalenced with $j_B \ldots j_B+5$. Similar contact patterns are stored in a non-exclusive list of pairs (the "pair list"), the raw material for structural alignment. The goal of the 2nd step is to assemble pairs of contact patterns into larger consistent sets of pairs (alignments), maximizing the similarity score of eqn (1). A Monte Carlo algorithm is used to deal with the combinatorial complexity of building up alignments from contact patterns. The effort spent on generating the pair list pays back in rapid initial build-up of alignments, which is followed by refinement in 2 stages. A detailed description of the strategy to gain speed yet maintain accuracy follows. The basic idea of the algorithm (Fig. 2) is valid independent of implementation details.

(c) *Step 1: decomposition of distance matrices*

The area of a distance matrix grows as the square of the length N of the sequence, and the number of possible comparisons between contact patterns in 2 matrices grows as the product of the areas ($N_A^2 N_B^2$). As this is highly redundant compared to the number of possible residue pair equivalences ($N_A N_B$), we consider only a subset of contact patterns. Search space is reduced (1) by restricting the number of hexapeptide–hexapeptide contact patterns in each protein, and (2) by restricting the number of pairs of such patterns.

(i) *Reduced distance matrix*

Neighboring contact patterns may overlap by as much as 11 of 12 residues. To suppress repetitive overlaps, the chain of a protein is partitioned into segments, conceptually similar to secondary structure elements. Successive hexapeptide fragments (starting at residue $i, i+1, \ldots$) that repeat a strongly similar contact pattern along the main diagonal are merged into longer segments, e.g. along an α-helix. For each hexapeptide along the chain, tertiary contacts in the "reduced" distance matrix are represented by just 1 contact pattern per contacting segment. The contact pattern that is retained between a given hexapeptide and the hexapeptides that belong to a given segment, is the one with the smallest mean intra-pattern distance.

(ii) *Pair list of matching contact patterns*

The contact patterns in each distance matrix are sorted according to mean intra-pattern distance so that the pair list can be built up starting from short-range tertiary contacts and moving towards longer-range interactions. The reduced distance matrix of A is first compared against the full distance matrix of B, then the reduced matrix of B is compared against the full matrix of A, and redundant pairs are removed. Pairs corresponding to reversing the chain direction of 1 or 2 hexapeptides can be included by permuting the residue indices as appropriate. If a protein is being compared to itself, the trivial diagonal pairs are forbidden. To gain speed, filters on the total, row and column sums of distances are used to exclude grossly incompatible pairs of contact patterns from calculation of the similarity score. The filters require that the deviation of the sum of distances in one contact pattern is in the range -18% to $+22\%$ compared to that of the other contact pattern. The pair list is closed when either (1) the mean intra-pattern distance reaches 25 Å (longer intra-protein distances are ignored), or (2) 80,000 pairs with a positive similarity score have been recorded. After sorting, 40,000 pairs with the highest scores are passed on to the alignment step (independent of protein size). As an example of the effectiveness of the filters, the number of possibilities screened at each step of the algorithm in the comparison of bacteriophage T4 and hen egg-white lysozymes are summarized in Table 1.

(d) *Step 2: assembly of alignments*

(i) *Monte Carlo optimization*

The key idea of Monte Carlo optimization is iterative improvement by a random walk exploration of the search space, with occasional excursions into non-optimal territory. A move is a randomly chosen change in the configuration of the system. The probability p of accepting a move is $p = \exp(\beta^*(S' - S))$, where S' is the new score and S is the old score and β is a parameter (Metropolis *et al.*, 1953). In physical simulations, β is inversely proportional to the temperature of the system. Moves that improve the score are always accepted, but the higher the temperature, the more probable are excursions downhill.

The basic moves in the present implementation are the addition and deletion of residue equivalence assignments, with the corresponding increment or decrement of the total similarity score. The chain of configurations generated during a Monte Carlo search, i.e. sets of residue pair equivalences (alignments) is here called a trajectory. The alignment with the highest score along each trajectory is remembered. To preclude spurious matches of single residues at the intersection of 2 chain segments that just happen to cross at some point in space, we constrain the moves to units of tetrapeptides. Tetrapeptides may

overlap, so that one move can result in a net increase or decrease of between 1 and 4 residues in the alignment. Each hexapeptide fragment in the pair list generates 3 overlapping tetrapeptide fragments (residues 1 to 4, 2 to 5, 3 to 6).

The optimization starts from a seed alignment (one or more pairs of equivalenced residues). The Monte Carlo algorithm has 2 basic modes of operation. In the expansion mode, an alignment is incremented using contact patterns that overlap with it. More precisely, if the current alignment contains the residue pair (i_A, i_B), then all pairs of matching contact patterns that include this residue pair provide possible extensions of the alignment. One expansion cycle corresponds to testing all prospective candidates in the pair list in random order. As the alignment must be a one-to-one mapping between A and B, addition of a new fragment pair may require the tentative removal of inconsistent previous equivalence assignments. If the new fragment is accepted, the removals become permanent. Because similarity is defined as a sum over all equivalent intramolecular pairs (eqn (1)), the scores of all residue pairs included in the alignment change as the alignment changes. The trimming mode removes from the alignment any tetrapeptide fragments (but not necessarily their overlapping neighbors) that give a net negative contribution to the total similarity score.

(ii) *Selection protocol*

To cover a broad range of potential optima, several trajectories are optimized in parallel, with selective removal of redundant and lower scoring alignments. The range of alignments is narrowed onto the highest scoring one(s) in 3 stages. Each stage consists of 1 or more expansion/trimming cycles described above (Table 1). A trimming cycle is performed after the 1st and every 5 subsequent expansion cycles. Initially, and after each cycle that results in a new high of the similarity score, a low temperature ($\beta = 50$) is used for "steepest ascent" optimization (analogous to energy minimization). Otherwise, expansion cycles are performed with β equal to 0·1 divided by the square root of the current best score of the trajectory. This accounts for the smaller relative increment as an elementary pattern is merged into a growing alignment and gives a roughly constant acceptance ratio for alignments of different length.

In stage 1, a large number of seed alignments are generated. The pair list is screened for all triplets of non-overlapping hexapeptides. For example, in Fig. 2 the pairs (a,b)-(a′,b′), (a,c)-(a′,c′) and (b,c)-b′,c′) could form the triplet (a,b,c)-(a′,b′,c′). Seeds for the alignment are generated from all singlets, e.g. a-a′, contained in the triplets. Singlets that overlap and have the same relative sequence shift are merged into 1 seed. The maximal number of seeds is of the order of 100. Pairs of strongly similar structures yield fewer and longer seed alignments. Each seed is used to initialize a trajectory that goes through exactly 1 expansion/trimming cycle. If the equivalence assignments in 2 trajectories converge to closer than 50% identity, the one with the lower score is eliminated. The alignments are sorted and the top scoring ones are retained. Keeping the ten highest-scoring trajectories gives good results in practice.

In stage 2 optimization is continued in parallel until all alignments have settled in an optimum, i.e. until the score has not improved for 20 expansion/trimming cycles. To gain speed and variety, trajectories are eliminated if equivalence assignments are more than 80% identical with a higher-scoring trajectory, or if the score lags too far behind that of the leading alignment. Trajectories with a score below a certain fraction of the best overall score (checked every 20 expansion/trimming cycles) are eliminated. The fraction asymptotically approaches 1·0 according to the series $n/(n+1)$, where n is the number of the check. (If suboptimal alignments are desired, comparison is relative to the Nth best score.)

The 3rd stage consists of refining the best alignment. The alignments from the previous stage are basically complete, but may have suboptimally aligned segments frozen in. This is because correlated shifts of several segments are difficult to achieve with moves of tetrapeptide fragments, a limited number of steps and a strongly co-operative similarity score. To explore the local surroundings of a near-optimal alignment, the best alignment is used to initialize 10 parallel trajectories with 30% of aligned blocks randomly removed. These are optimized as in stage 2. The trajectories are reinitialized after every 20 expansion/trimming cycles, until the optimal score (best of the 10) no longer improves.

(e) *Computer implementation*

The algorithm was implemented in a Fortran-77 program called DALI. The program has topology options to constrain the alignments to be sequential or disallow matches in reversed chain direction. Typically, pairwise alignments can be generated in 5 to 10 minutes of computer time on a Sparc-1 CPU.

3. Results

(a) *Robustness of the algorithm*

(i) *Global or local optimum?*

To test the reproducibility of the Monte Carlo search, pairwise comparisons were repeated 100 times using different random number seeds. The global optimum was defined as the highest score found. Comparisons of T4 lysozyme to hen egg white lysozyme, a distantly related pair of the $\alpha+\beta$ class, converged to within 2% of the global optimum score (score = 256, 86 equivalent residues) with 96% fidelity. In comparing colicin A against arcid clam hemoglobin, two 3-on-3 helical sandwich folds (Holm & Sander, 1993*a*,*b*), there are two nearly equal possibilities to align one of the two helical layers, differing by a shift of one helical turn. A total of 2% of the test runs found the global optimum (score = 472, 118 equivalent residues) and 94% of the runs settled into the second-best optimum (score = 447, 114 equivalent residues).

(ii) *Sensitivity to starting point*

To test the radius of convergence of the algorithm, fully optimized alignments were generated from all seeds (Table 2). (Normally, trajectories compete with one another and suboptimal trajectories are gradually discarded until only 1 alignment remains.) The screening of triplets identifies similar local neighborhoods, conceptually similar to "supersecondary structure". The vast majority of seeds correspond to incorrect optima. Most alignments remained stuck in local optima, but in a number of cases, a path was found from initially incorrect equivalence assignments to the correct final align-

Table 1
Simplifying combinatorial complexity in the comparison of hen egg-white lysozyme (1lyz) with T4 lysozyme (2lzm)

A. Distance matrices	
1lyz	
No. of overlapping hexapeptides	124
Total no. of contact patterns	7626
No. of contact patterns in reduced distance matrix	5332
2lzm	
No. of overlapping hexapeptides	159
Total no. of contact patterns	12,561
No. of contact patterns in reduced distance matrix	4709
B. *Pair list*	
Total no. of pairs of contact patterns	96×10^6
Total no. of pairs of contact patterns after reduction	71×10^6
No. of checks by filters on row/column sums†	9×10^6
No. of residue-by-residue similarity score calculations	2×10^5
No. of kept pairs of contact patterns after ranking by score	4×10^4
C. *Monte Carlo optimization*	
Screening	
No. of parallel trajectories	80
No. of expansion/trimming cycles‡	1
No. of kept alignments after ranking by score	10
Optimization of divergent alignments	
No. of parallel trajectories	10
No. of expansion/trimming cycles‡	80
No. of kept alignments after ranking by score	1
Refinement of best alignment	
No. of parallel trajectories	10
No. of expansion/trimming cycles‡	40
No. of kept alignments after ranking by score	1

At each step of the algorithm, the search tree is heavily pruned to overcome the combinatorial explosion. See Methods for details.

† The pair list is built up starting from the smallest intra-pattern mean distance, and closed when the maximal number of pairs has been stored.

‡ One cycle means 1 pass through (a subset of) the pair list.

ment. Thus, the optimization procedure is not overly sensitive to the choice of initial alignment.

(iii) *Detection of structurally meaningful multiple optima*

Since several alignments are optimized in parallel, suboptimal solutions can be printed out as an option. $(\beta\alpha)_8$ barrels are a tough test of robustness of the optimization algorithm because of their large size and multiple optima due to the 8-fold symmetry of the structures. In the comparison of tryptophan synthase to itself, the algorithm finds the expected seven cyclically permuted alignments (Table 3). The highest score is for a shift of four $\beta\alpha$ units, consistent with the idea that $(\beta\alpha)_8$ barrels may have evolved by repeated duplication of simpler units.

In the light of these three severe tests, the overall robustness of the algorithm is more than satisfactory.

Table 2
Seed test

	Correct alignment (no. of runs)	Incorrect alignment (no. of runs)
T4 lysozyme (2lzm) – hen egg-white lysozyme (1lyz)		
Correct seed	4	0
Incorrect seed	2	74
Colicin A (1colA) – ark hemoglobin (1sdhA)		
Correct seed	6	0
Incorrect seed	16	86

As a test of the radius of convergence of the algorithm, the build-up was followed from the initial seed alignment to the final optimized and refined alignment. The alignment with the highest score (and close variants) was classified as correct, and seeds were classified as correct if they overlapped with the correct full alignment. Optimization of the similarity score can lead to the correct alignment even though the alignment is initialized from an incorrect seed.

Table 3
Internal symmetry of a $(\beta\alpha)_8$ barrel

$(\beta\alpha)$ units 1-2-3-4-5-6-7-8 aligned with	Similarity score	No. of equivalenced residues	r.m.s.d. (Å)
5-6-7-8-1-2-3-4	1194	171	2·9
7-8-1-2-3-4-5-6	1101	177	3·4
8-1-2-3-4-5-6-7	989	177	3·2
4-5-6-7-8-1-2-3	970	167	3·0
6-7-8-1-2-3-4-5	944	170	3·2
3-4-5-6-7-8-1-2	818	169	3·7
2-3-4-5-6-7-8-1	757	159	3·4

Alignment of a protein against itself reveals internal repeats of structural elements. For the $(\beta\alpha)_8$ barrel of tryptophan synthase (1WYSA; Hyde *et al.*, 1988), the method finds 7 non-trivial cyclically permuted alignments. The program was run disallowing matches with reversed chain direction, and asking for the 20 best alignments without overlaps in their sets of equivalenced residue pairs.

(b) *Quality of the alignments*

(i) *Verification of accuracy*

Conserved functional residues provide anchor points by which one can verify the accuracy of the purely structural alignment between members of divergent protein families. For example, DALI aligns correctly the GxGxxG (G = Gly, x = any amino acid) fingerprints of several dehydrogenases, the conserved disulfide bridges and central tryptophan in immunoglobulins, the DTG (Asp,Thr,Gly) triplet in aspartic proteases, and the metal ligands in blue copper proteins. DALI's alignment of plastocyanin with azurin is essentially the same as that by Taylor & Orengo (1989), whereas Fischer *et al.* (1992) have several strands shifted by one or two residues. Between hen egg-white lysozyme and T4 phage lysozyme, DALI aligns glutamic acids 11 and 25, which are the principal catalytic residues (Fig. 3), whereas the alignment by Taylor & Orengo (1989)

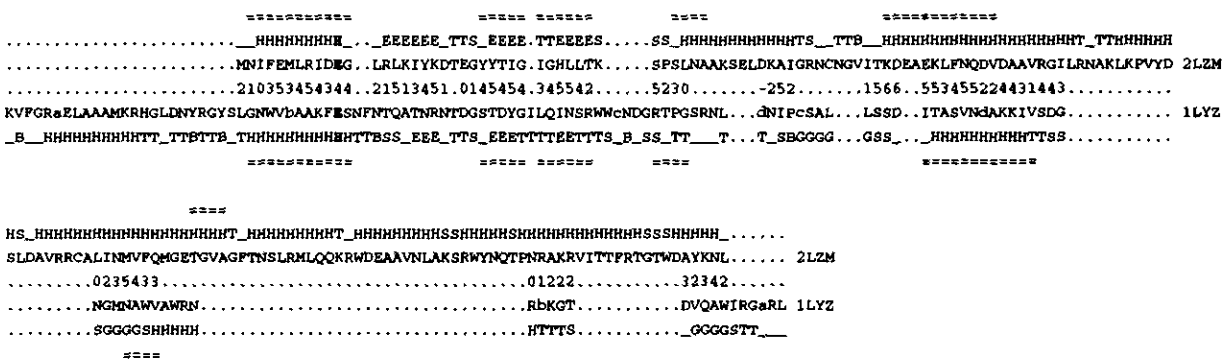


Figure 3. Structural alignment of hen egg-white and T4 lysozymes. Hen egg-white lysozyme (upper, PDB dataset 1LYZ, Blake *et al.*, 1965) and T4 phage lysozyme (lower, PDB dataset 2LZM, Weaver & Matthews, 1987) are a classical example of remotely related proteins. Alignment using the elastic score (eqn (3)) recognizes a common core of 77 residues (4·2 Å r.m.s.d.), indicated by numbers between the sequence lines. The glutamic acid residues required for catalytic activity are bold. Use of the rigid similarity score (optimization using eqn (2)) leads to considerably fewer equivalences: 42 residues (2·2 Å r.m.s.d.) mainly in 2 helices and a β-hairpin around the active site (shown by = characters in the top and bottom lines). Sequential order of aligned segments was imposed as a constraint in both cases. Secondary structure assignments (Kabsch & Sander, 1983) are shown next to the sequences: H, α-helix; G, 3_{10}-helix; E and B, β-sheet, T, turn; S, bend; _, non-hydrogen-bonded structure. Lower-case sequence symbols are cysteine residues participating in disulfide bridges (a with a, b with b, etc.). Dots mark gaps, non-aligned segments and trailing ends. The average similarity score per residue (middle row) indicates the quality of the fit in different regions along the chain. The average score is reported as the percentage of the maximal elastic score (applying eqn (3) with all distance deviations set to zero) at 10% intervals: "0" for 0 to 10%, "1" for 10 to 20% and so on; − for negative values.

has a shift of one residue in this region. Structural homology is less clear-cut in the peripheral regions and residue equivalences here vary between methods that optimize different scores (e.g. Rossmann & Argos, 1976; Subbarao & Haneef, 1991). A valuable feature of the elastic similarity score in the comparison of homologous proteins is that it captures the relative movements of structural elements, leaving only loops in radically different conformations unaligned, e.g. for the globin family, we find very good agreement with the manually derived alignments described by Bashford *et al.* (1987).

(ii) *Detection of inter-domain motion*

Actin, heat shock protein hsp70 and hexokinase are three functionally diverse proteins with a common ATPase domain. The ATP-binding site is in a cleft between two subdomains connected by a hinge formed by a helix–helix contact point (Holmes *et al.*, 1993). The crystal structures of actin and heat shock protein are in a "closed" conformation and that of hexokinase in an "open" conformation. Comparison of the closed and open form is normally done separately for each domain. Here, the full length of the chains is aligned in a single comparison between actin and heat shock protein (288 equivalent residues), as well as between actin and hexokinase (232 equivalent residues), in spite of the hinge motion. The resulting structural alignment brings into register functionally important residues; a non-trivial result in a very difficult case (Fig. 4).

(iii) *Extent of common core*

The elastic similarity score detects 3D similarities at the domain level (Figs 3 to 5). Decreasing the value of the similarity threshold shifts emphasis from longer alignments to a more stringent definition of the structural core. For example, if $\theta^E = 0{\cdot}20$ and sequentiality is imposed, the pair actinoxanthin–superoxide dismutase (1ACX-2SODO) has 80 equivalent residues yielding an r.m.s.d. of 2·7 Å. If the similarity threshold is decreased to $\theta^E = 0{\cdot}15$, eight equivalent residue pairs are excluded and the r.m.s.d. is reduced to 2·3 Å. The rigid similarity score (eqn (2)) is qualitatively similar to measures based on rigid-body 3D superimposition. For 1ACX-2SODO, the rigid score with $\theta^R = 1{\cdot}5$ Å yields a common core of 58 residues with 1·8 Å r.m.s.d. This result compares favorably with the manual superimposition by Hazes & Hol (1992) who obtained a core of 49 residues with 1·9 Å r.m.s.d. using a 3·5 Å cutoff on positional deviations after rigid-body superimposition.

(c) *All-against-all alignment of protein structures*

An all-against-all structure comparison was carried out for 225 representative protein structures, a total of 25,200 pair comparisons. The representative set was selected from the Protein Data Bank (Bernstein *et al.*, 1977) so that all pairs have less than 30% sequence identity (Hobohm *et al.*, 1992). We searched for general 3D similarities allowing shuffled and reversed segments, i.e. no topological constraints were imposed. Here, we report four types of results extracted from the large number of pair comparisons: different topologies in similar 3D folds, structural families, novel structural resemblances, and observations on sequence-structure patterns.

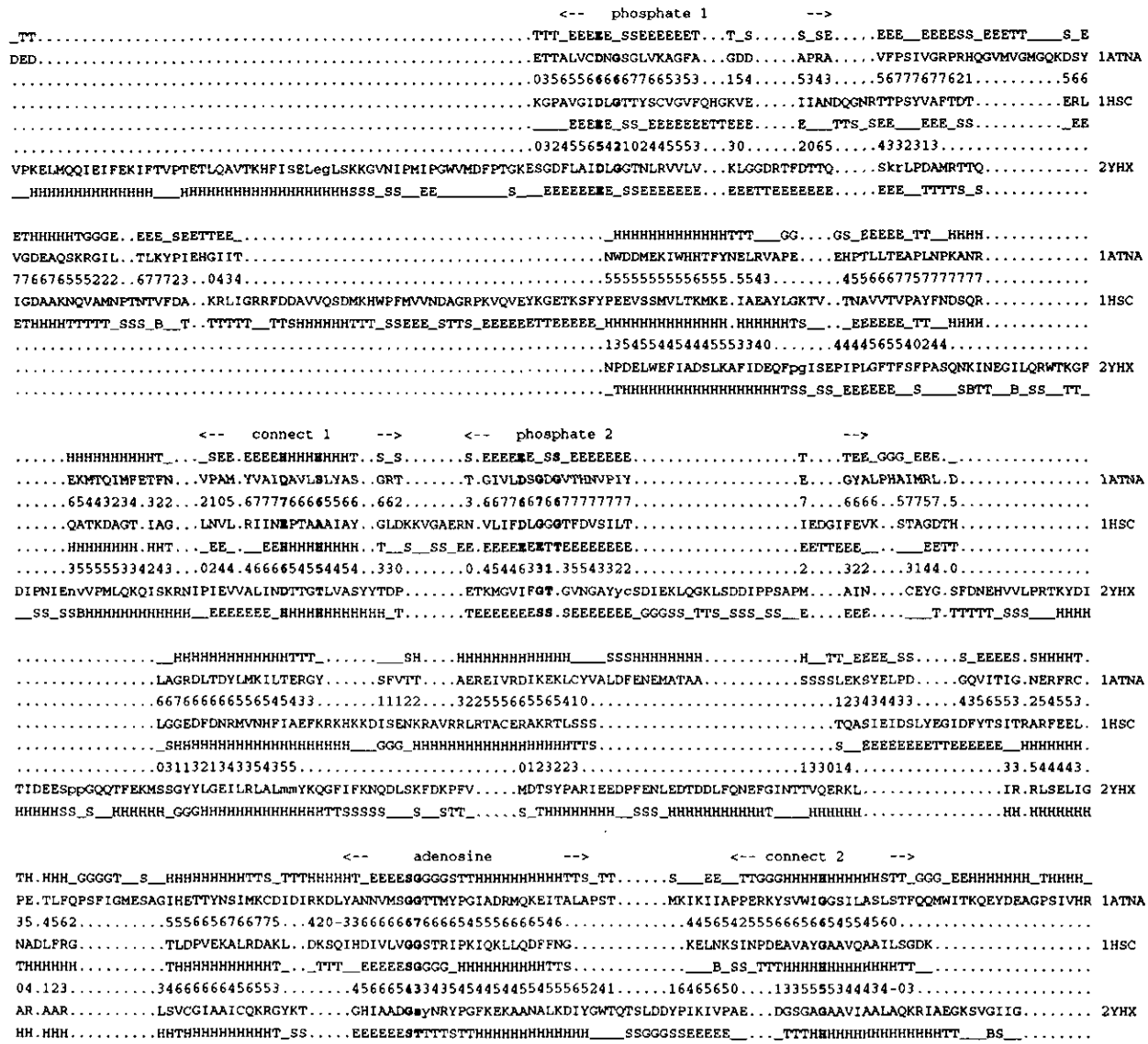


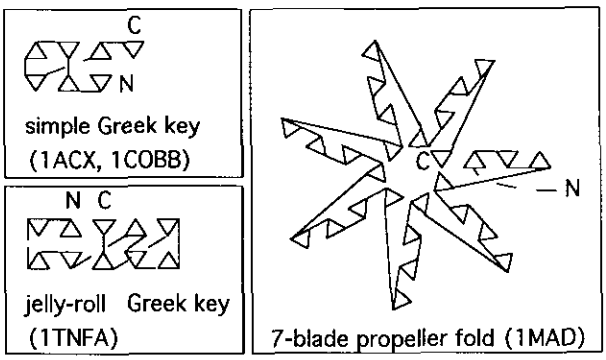
Figure 4. Structural alignment of heat shock protein 70 and hexokinase with actin. Actin (1ATNA, Kabsch *et al.*, 1990), heat shock protein 70 (1HSC, Flaherty *et al.*, 1990) and hexokinase (2YHX, Anderson *et al.*, 1978) use a common ATPase domain to perform diverse functions in the cell. Sequence patterns conserved across the 3 protein families are confined to 5 regions indicated above the sequences (phosphate 1, etc.), with the most strongly conserved residues in bold type (see Bork *et al.*, 1992, for details). The connect 1 and connect 2 helices form a hinge between 2 subdomains. Superimposition in 3D gives an r.m.s.d. of 3·1 Å for the actin–heat shock protein pair. The 2 subdomains of the actin–hexokinase pair give comparable r.m.s.d. values when superimposed separately, but because of hinge motion the r.m.s.d. is as high as 5·1 Å for the pair as a whole. The PDB dataset 2YHX contains a tentative amino acid sequence deduced from electron density. Shown here is the cloned sequence of yeast hexokinase aligned to the crystallographic structure as described by Bork *et al.* (1992). Insertions are bounded by lower-case characters. Other notation is as in Fig. 4.

(i) *Different topologies, yet similar structures*

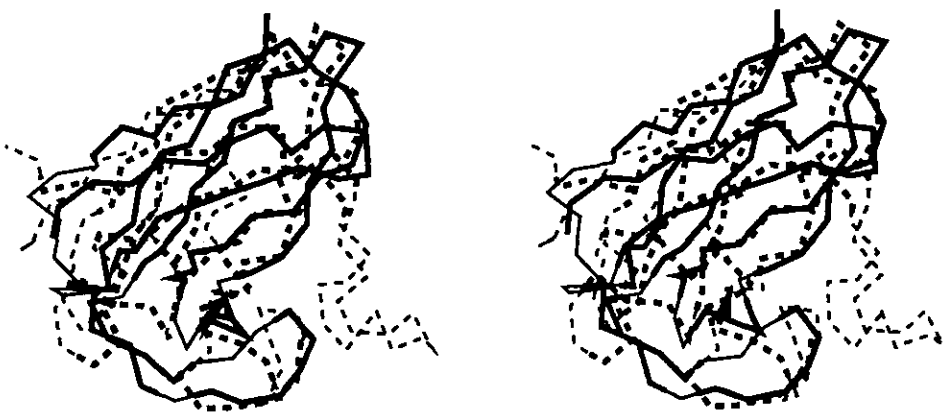
The packing of secondary structure elements can be strikingly similar in spite of different topological arrangements of the polypeptide chain. Figure 5 shows three examples among antiparallel β-barrels, as a result of a database scan with actinoxanthin. When topological constraints are relaxed, the core of actinoxanthin is found to match not only other "simple Greek keys" (e.g. superoxide dismutase), but also parts of "jelly-roll Greek keys" (e.g. tumor necrosis factor) and "7-blade propeller" folds (e.g. methylamine dehydrogenase).

(ii) *Structural families: trees, clusters*

The raw result of the all-against-all search is the set of all pair similarities. Intuitively speaking, the similarities can be used to position each structure type in a high-dimensional space. Within this space, folds with some architectural similarity cluster together, e.g. proteins with compact all-helical


C
N
simple Greek key
(1ACX, 1COBB)
N C
jelly-roll Greek key
(1TNFA)
C
N
7-blade propeller fold (1MAD)

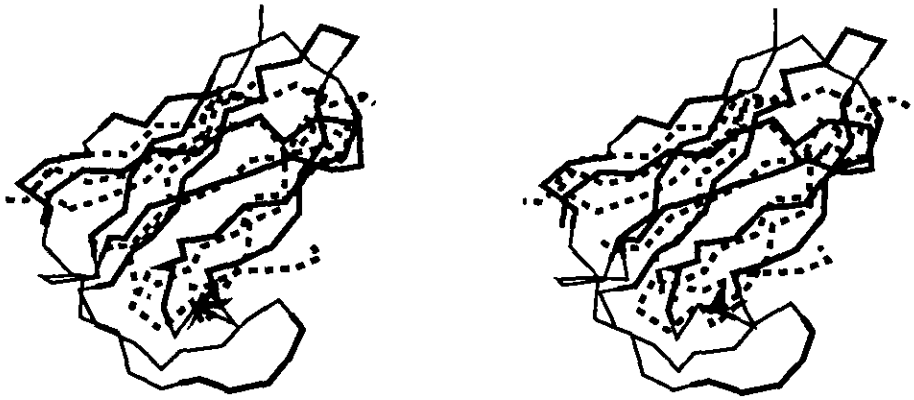
(a)



(b)

(c)

Fig. 5.



(d)

Figure 5. Structural similarities among antiparallel β-barrels with different topologies. Three examples of similar packing of secondary structure elements in spite of different loop connections were selected from a database scan with actinoxanthin (1ACX; Pletnev *et al.*, 1982). The stereo plots show 1ACX depicted by continuous lines, the matched protein by broken lines and the structurally equivalent segments in bold. (a) Topology diagrams. β-Strands are drawn as triangles whose apex points up or down according to the direction of the strand. (b) Superimposition of actinoxanthin with superoxide dismutase (1COBB; Djinovic *et al.*, 1991). Seven β-strands match sequentially, as both proteins are simple Greek keys. In addition, 3 loop segments match in reversed chain direction. A total of 90 residues are aligned, yielding an r.m.s.d. of 2·9 Å. PDB residue numbers of aligned segments follow (1ACX/1COBB): 1 to 8/13 to 20, 15 to 25/26 to 36, 39 to 42/59 to 62, 43 to 46/80 to 77 (reversed loop), 48 to 64/82 to 99, 66 to 69/104 to 101 (reversed loop), 73 to 78/69 to 64 (reversed loop), 80 to 83/107 to 110, 87 to 95/112 to 120, 96 to 106/139 to 149. (c) Actinoxanthin with tumor necrosis factor (1TNFA; Eck & Sprang, 1989). The alignment is non-sequential and has 2 reversed strands. A total of 80 residues are aligned, yielding an r.m.s.d. of 3·6 Å. PDB residue numbers of aligned segments follow (1ACX/1TNFA): 1 to 8/98 to 91 (reversed strand), 13 to 16/127 to 130, 17 to 24/83 to 76 (reversed strand), 28 to 36/149 to 157, 41 to 44/11 to 14, 48 to 51/15 to 18, 55 to 63A/140 to 131 (reversed loop), 64 to 67/48 to 51, 68 to 71/41 to 38 (reversed strand/loop), 75 to 79/6 to 10, 85 to 94/53 to 62, 97 to 106/117 to 126. (d) Actinoxanthin with methylamine dehydrogenase (1MAD; Vellieux & Hol, 1989). The match includes the 1st and 2nd blades of the propeller fold of 1MAD. The entire structure of 1ACX is shown but, for clarity, only the aligned segments are shown for 1MAD. A total of 74 residues are aligned (non-sequentially), yielding an r.m.s.d. of 3·7 Å. PDB residue numbers of aligned segments follow (1ACX/1MAD): 4 to 7/69 to 72, 9 to 12/73 to 76, 16 to 21/77 to 82, 27 to 37/44 to 54, 40 to 44/55 to 59, 47 to 53/60 to 66, 55 to 58/86 to 83 (reversed loop/strand), 59 to 65/97 to 104, 74 to 78/356 to 352 (reversed loop), 87 to 96/32 to 41, 97 to 106/360 to 369.

domains. The emerging universal distribution of protein folds is presented here in two classical representations. A planar projection of points in protein structure space, using correspondence analysis, reveals an overall grouping in terms of the all-α, all-β and α/β structural classes (Fig. 6). A dendrogram, which approximates the pair similarities by the length of branches, reproduces essentially all known structural classifications, such as the different types of β-sandwiches in the all-β class (Fig. 7, presented in 6 parts). The fact that an automatic procedure was used opens the way toward future automatic classification of protein substructures and folding domains. When more than 1000 different protein structures will be known, automatic classification will be essential.

(iii) *Unexpected similarities: four examples*

The POU protein is a bipartite eukaryotic transcription factor consisting of a C-terminal homeodomain and an N-terminal POU-specific domain. Surprisingly, the POU-specific domain (Dekker *et al.*, 1993) has the same topology of fold as the λ repressor, 434 repressor and cro proteins. The common core comprises four helices, sequentially aligned, with an r.m.s.d. of 2·3 to 2·5 Å and 22 to 26% sequence identity in 54 to 59 residues. Apparently, the similarity was not detected by sequence comparisons because of a six-residue insert in the POU-specific domain compared to the canonical helix-turn-helix motif in DNA-binding proteins. The similarity has both evolutionary and functional implications.

The membrane-insertion domain of the bacterial toxin colicin A has the same topology of fold as globins and phycocyanins, with six helices sequentially aligned. The three protein families are functionally diverse and lack detectable sequence similarity, suggesting physical convergence to a stable folding motif, the 3-on-3 helical sandwich (Holm & Sander, 1993*a*).

The fold of the monomer of neutrophil defensin, a lytic peptide, is fully contained in that of sea anemone neurotoxin, a sodium channel inhibitor. Superimposition of residues B2 to B17, B18 to B21, B25 to B31 in defensin (1DFN, Hill *et al.*, 1991) with residues 18 to 33, 35 to 38, 39 to 45 in neurotoxin (1SH1, Fogh *et al.*, 1990) yields an r.m.s.d. of 1·6 Å for 27 residues. Both proteins are rich in disulfide bridges, but only one of these is structurally equivalent.

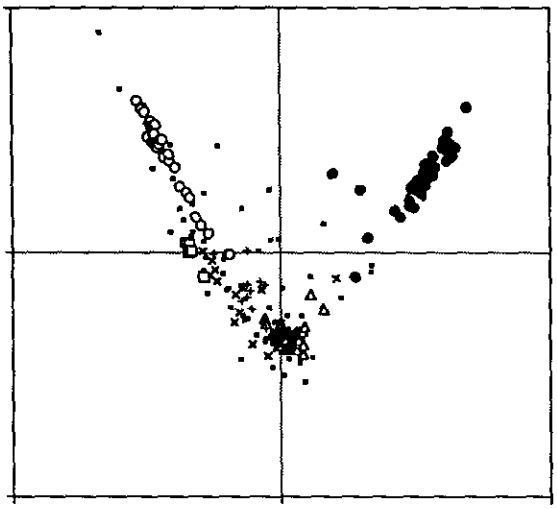


Figure 6. Clusters in protein structure space: a 2D projection of protein similarities from the all-against-all comparison in the representative set of 225 proteins by correspondence analysis (Hill, 1973). Structures are labeled according to the groups in Fig. 7. For any 2 proteins, the closeness in 2D approximately represents pair similarity. The miscellaneous group consists mostly of short chains. Structures with helical cores are found on the right, β-proteins on the left and α/β proteins form an extremely tight cluster in the lower middle (overlapping open triangles appear black). The region between the helical and α/β cores is almost empty (different helix packing), whereas the transition from α/β proteins to all-β proteins is more gradual (similar sheet geometry). The plot was generated from the 225×225 matrix of normalized similarity scores, solving an eigenvalue problem using a program described by Holm (1986). The eigenvalues are 0·55 for the horizontal axis and 0·34 for the vertical axis. Similarity scores for a given search structure were reported in units of standard deviations (σ) above database background to make the results of database scans with proteins of different size comparable. Outliers above 2σ were iteratively excluded from the calculation of mean and σ. Values above 10σ were truncated to 10 and values below 1σ were set to zero. Structural classes: ■, miscellaneous; ●, alpha; ○, β-barrel; △, α/β; ×, twisted β; +, β-sandwich; □, β-trefoil.

Figure 8 shows the remarkable local coincidence of a structural module in several proteins, consisting of about 20 residues in four segments and containing two disulfide bridges and a small β-sheet. The module is present in five diverse families: (1) plant inhibitors of carboxypeptidase and trypsin; (2) the cellulose-binding domain of fungal cellobiohydrolase I (Kraulis *et al.*, 1989); (3) wheat germ agglutinin; (4) the erabutoxin family (grouped together by Richardson, 1981); (5) human neurophysin, which differs in loop connections.

(iv) *Sequence-structure patterns*

Multiple structure alignments of proteins with low overall sequence similarity and no functional overlap provide a source of information about sequence patterns of importance to structure formation. Most residue identities occur at positions with a particular backbone conformation, e.g. with glycine, serine, or proline acting as capping residues and turn promoters. In the hydrophobic cores of particular structural classes, one sees occasional identities, e.g. of leucines and alanines in helices, or of β-branched side chains in β-sheets, in accordance with secondary structure propensities. In spite of these observations, the sequence-structure code remains elusive in practice. For example, scanning sequence databases with 1D sequence profiles (Gribskov *et al.*, 1987), we find that profiles derived from subsets of $(\beta\alpha)_8$ barrels (also called TIM barrels) pick up sequences of other $(\beta\alpha)_8$ barrels with very low sequence similarity, but the signal is too noisy for reliable detection of remote structural homologues (data not shown).

4. Discussion

We have developed a novel algorithm for the alignment of proteins represented by two-dimensional matrices, and have applied the algorithm to comparing protein structures given their C^α co-ordinates. The method is fully automatic, general, reasonably robust and conceptually simple. Defining similarity in terms of agreement of intramolecular distances accounts for conformational flexibility of protein structure in a way that is out of reach for methods based on rigid-body superimposition. Although Monte Carlo optimization cannot be guaranteed to yield the global optimum, our method in practice accurately aligns divergently related protein pairs and reliably detects common 3D folding motifs in database searches, as assessed by visual inspection (Figs 3 to 8). There is no case-by-case parameter fiddling: one set of parameters works well for all structural classes. Because of the sensitivity of the method, a database scan with one structure currently takes an overnight run on a workstation. Speed could be increased by a prefiltering step at a higher level than hexapeptides, e.g. secondary structure elements.

The algorithm has a number of possible further applications. (1) Comparing residue–residue interaction energies in place of distances between residue centers would provide a more physical view on structure comparison. (2) Realistic structural models could be generated by comparing designed, predicted or experimentally determined contact maps against a database of known structures. (3) We look forward to applying the method to the problem of sequence-structure alignment (Ouzounis *et al.*, 1993), evaluated using effective amino acid pair potentials and taking full account of the pair dependencies that prevent the application of 1D dynamic programming algorithms.

The classification of 3D folds as a result of the all-against-all comparison in the representative set may be useful in the analysis, design and prediction of protein structures. The highest scoring alignments

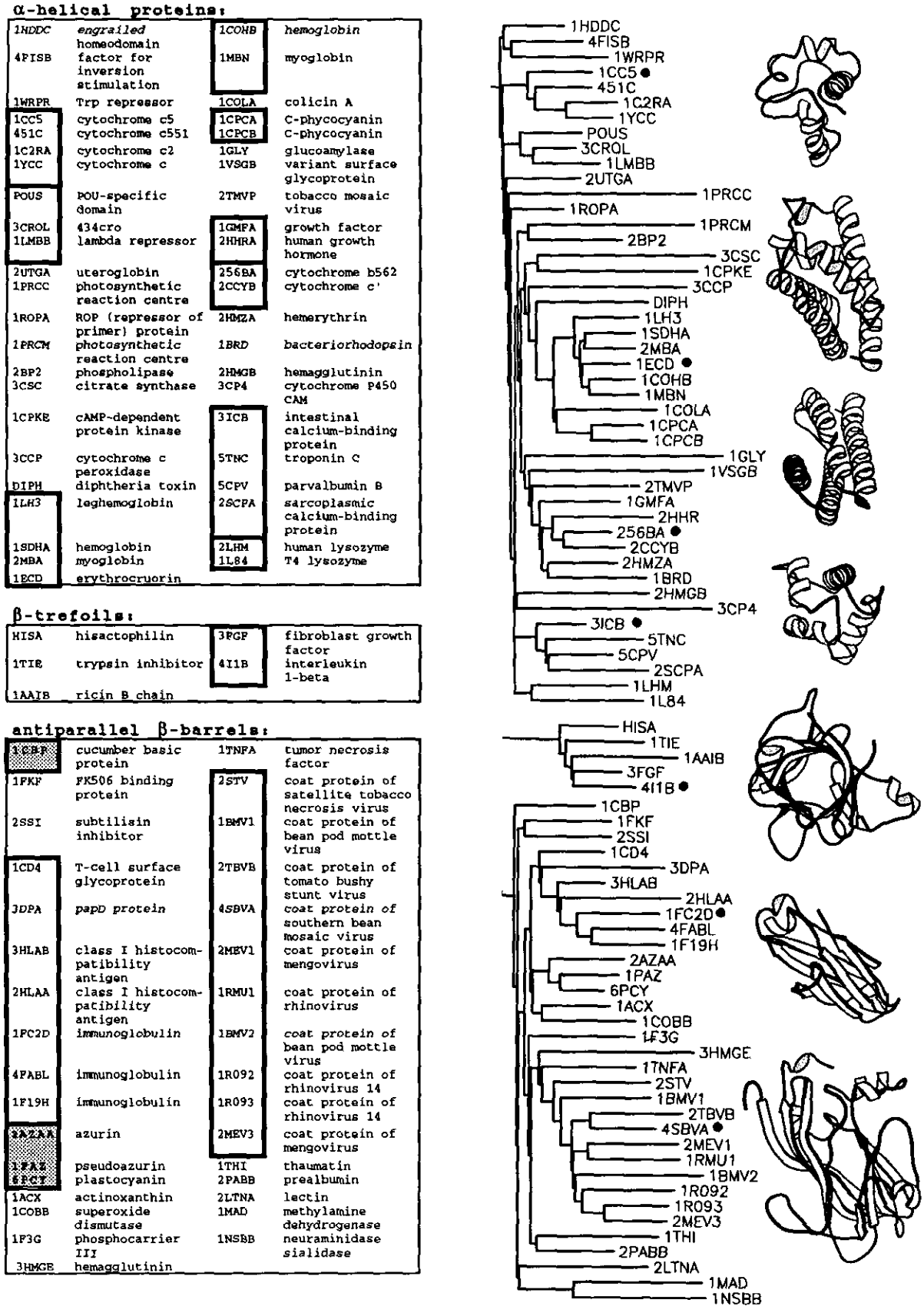


Figure 7. Protein structure family trees by average linkage clustering. Pairwise comparisons were generated allowing shuffled and/or reversed segments. Branch lengths represent similarity scores and are plotted on a logarithmic scale. A total of 6 branches were cut off from the complete tree. Divergently related families (entire proteins or domains) with similar function are boxed. The most frequently recurring all-helical folds are the 3-on-3 helical sandwich (DIPH . . . 1CPCB; Holm & Sander, 1993*b*) and the 4-helical bundle motif (1GLY . . . 1BRD), which is found in many topological variations and often as part of larger helical proteins, e.g. the 7-helical bundle in 1BRD. All-β proteins are divided in a number of classes. The largest is the antiparallel β-barrel branch, in which most structures have the simple Greek key or jellyroll topology (Richardson, 1981). Proteins with the simple Greek key fold are so similar to each other that a number of other branches get merged with the larger blue copper proteins (2AZAA, 1PAZ, 6PCY) just before the smallest blue copper protein 1CBP (86 residues) would join its cousins. Practically all structures in the α/β class have in

α/β-proteins:

1RNH	ribonuclease H	1ACE (**)	acetylcholine esterase
1HRHA	ribonuclease H	4DPRB	dihydrofolate reductase
1GAL	glucose oxidase	3DFR	dihydrofolate reductase
1COX	cholesterol oxidase	2REB	RecA protein
1PHH	parahydroxybenzoate hydroxylase	3AAT	aspartate aminotransferase
1LPFA	lipoamide dehydrogenase	3ADK	adenylate kinase
3GRS	glutathione reductase	1GKY	guanylate kinase
2CLA	chloramphenicol acetyltransferase	3TS1	tyrosyl-tRNA synthetase
2GLSA	glutamine synthetase	3CPA	carboxypeptidase A
5ACN	aconitase	1LAP (***)	leucine aminopeptidase
2YHX	hexokinase	1TPT	thymidine phosphorylase
1ATNA	actin	1R1E	Eco RI endonuclease
1HSC	heat shock protein 70	6ICD	isocitrate dehydrogenase
3FBPB	fructose-1,6-bisphosphatase	2ATCA	aspartate transcarbamoyltransferase
2RVEB	Eco RV endonuclease	1PFKA	phosphofructokinase
5XIAB	xylose isomerase	8ADH	alcohol dehydrogenase
2TAAA	taka-amylase A	1GD1R	glyceraldehyde-3-phosphate dehydrogenase
1MLE	muconate lactonizing enzyme	1SIC	subtilisin
1YPIA	triosephosphate isomerase	2GBP	galactose binding protein
1ALD	aldolase	2LIV	Leu/Ile/Val binding protein
2RUSA	RUBISCO	3PGM	phoshoglycerate mutase
1WSYA (*)	trypthophan synthase A chain	1PGD	6-phosphogluconate dehydrogenase
3TRX	thioredoxin	5LDH	lactate dehydrogenase
1GP1A	glutathione peroxidase	4MDHB	malate dehydrogenase
1WSYB	tryptophan synthase B chain	2FCR	flavodoxin
1FNR	ferredoxin reductase	1FX1	flavodoxin
3PGK	phosphoglycerate kinase	1NIPA	nitrogenase iron protein
1TGL	triacylglycerol acylhydrolase	1ETU	elongation factor Tu
2SC2	serine carboxypeptidase	5P21	ras p21

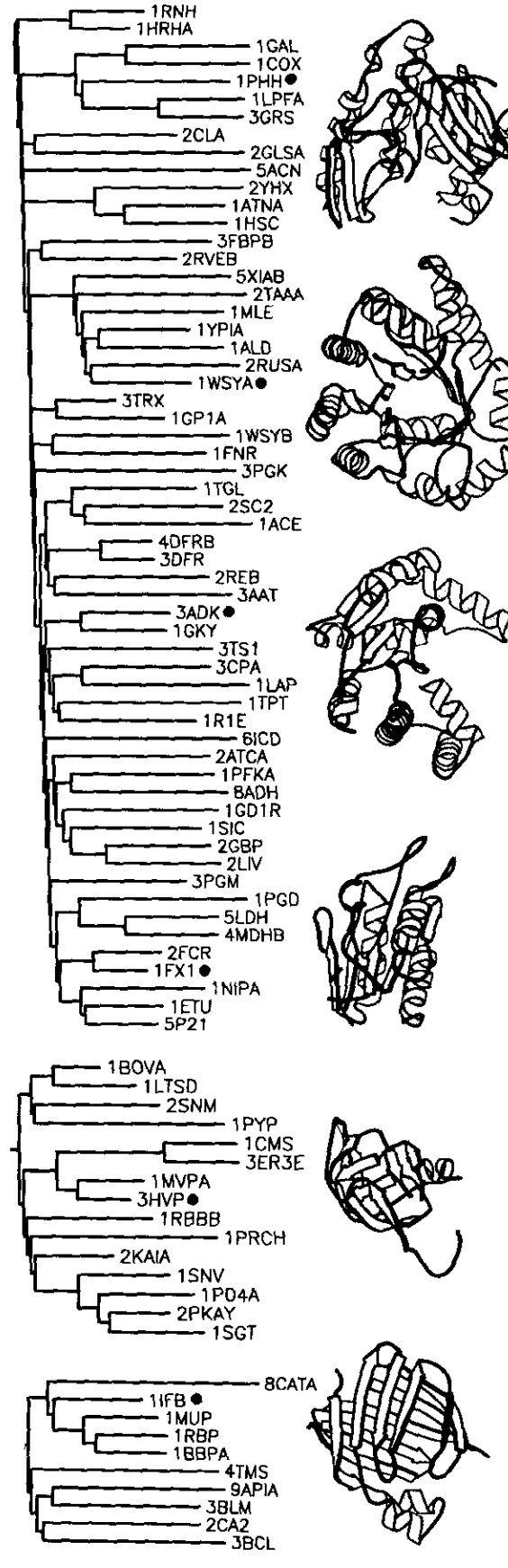
(*) Wilmanns & al.(1991) propose divergence because of a common phosphate-binding loop.
(**) α/β hydrolase family described by Ollis & al. (1992).
(***) Resemblance noted and divergence proposed by Artymiuk & al. (1992).

twisted β-barrels:

1BOVA	verotoxin 1	1RBBB	ribonuclease B
1LTSD	heat-labile enterotoxin	1PRCH	photosynthetic reaction centre
2SNM	staphylococcal nuclease	2KAIA	kallikrein
1PYP	inorganic pyrophosphatase	1SNV	Sindbis virus capsid protein
1CMS	chymosin	1P04A	alpha-lytic protease
3ER3E	endothiapepsin	2PKAY	kallikrein
1MVPA	viral protease	1SGT	trypsin
3HVP	HIV protease		

orthogonal β-sandwiches:

8CATA	catalase	4TMS	thymidylate synthase
1IFB	intestinal fatty-acid binding protein	9APIA	alpha-1 antitrypsin
1MUP	major urinary protein	3BLM	beta-lactamase
1RBP	retinol binding protein	2CA2	carbonic anhydrase
1BBPA	bilin binding protein	3BCL	bacteriochlorophyll A protein



common a 4 or 5-stranded β-sheet flanked by helices on both sides. The large common core of $(\beta\alpha)_8$ barrels makes them a distinct subclass (5XIAB . . . 1WSYA). References to the structure determinations may be found in the headers of the PDB files; they are omitted here for space reasons. Structures marked by filled circles are shown as ribbon diagrams (Kraulis, 1991).

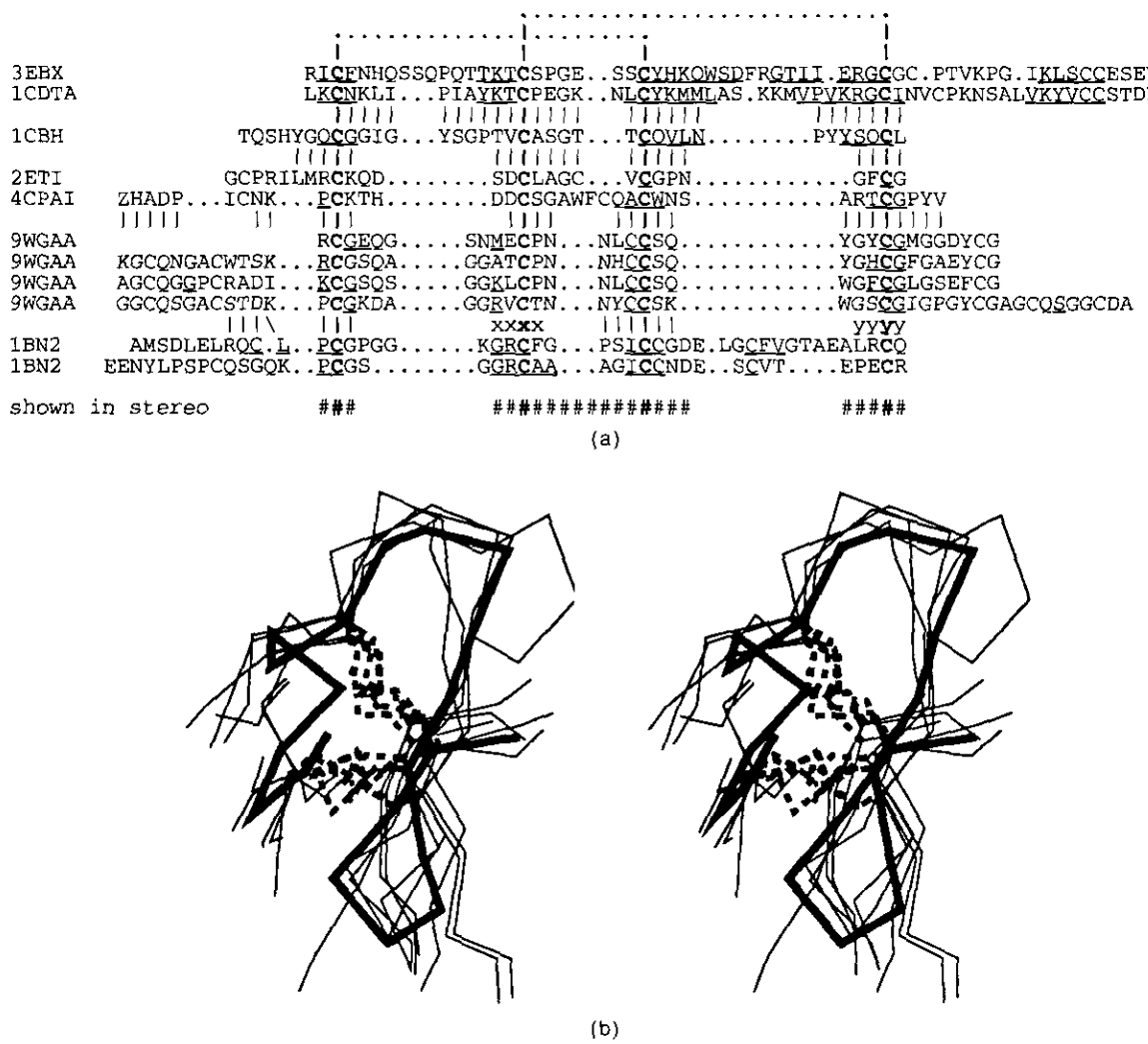


Figure 8. *Common structural module in the core of small cysteine-rich domains.* As a result of the all-against-all comparison, a folding motif described in carboxypeptidase inhibitor (4CPAI; Rees & Lipscomb, 1982) and the C-terminal domain of cellobiohydrolase I (1CBH; Kraulis *et al.*, 1989) was identified in 3 additional protein families. (a) Structural alignment of erabutoxin (3EBX; Smith *et al.*, 1988), cardiotoxin (1CDTA; Rees *et al.*, 1990), 1CBH, trypsin inhibitor (2ETI; Chiche *et al.*, 1989), 4CPAI, wheat germ agglutinin, with 4 repeats of the domain (9WGAA; Wright, 1990), and neurophysin, with 2 repeats of the domain (1BN2; Chen *et al.*, 1991). Cysteine residues forming equivalent disulfide bridges (top) are shown in bold. Underlined residues participate in β-sheets. Structurally equivalent residues are indicated by vertical bars between families. 4CPAI was aligned with the 2nd domain in 9WGAA. The alignment between 9WGAA and 1BN2 is non-sequential: box xxxx in 9WGAA is structurally equivalent with box yyyy in 1BN2, and *vice versa*. (b) 3D superimposition of residues marked by hatches in the bottom line of (a). The structures of 3EBX, 1CDTA, 1CBH, 2ETI, 4CPAI, 9WGAA 2nd domain and the 1st domain of 1BN2 were superimposed on the C^{α} atoms of the 4 common cysteine residues. C^{α} traces are thin and the common disulfides are dotted bold. The chain trace of 2ETI (residues 8 to 28, bold) may be followed from the lower to upper left, then over the top to the right-hand side, down at the back and finally up in front.

are available by anonymous ftp (file transfer protocol) from ftp.embl-heidelberg.de in the directory /pub/databases/protein_extras/fssp (Holm *et al.*, 1992). Searches of newly solved protein structures against the representative set are performed on request. Send co-ordinates by electronic mail to holm@embl-heidelberg.de. A list of 3D alignments with proteins similar to the input structure will be returned.

We wish to thank friends in the Protein Design Group for discussing structural alignment algorithms. We are very grateful to crystallographers and NMR spectroscopists who submit co-ordinates to the Protein Data Bank. We thank R. Kaptein, D. Eisenberg and T. A. Holak for access to co-ordinates not yet in the Protein Data Bank. L.H. began this work as an EMBO fellow and was supported further by the Human Frontiers Science Program.

References

Alexandrov, N. N., Takahashi, K. & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.

Anderson, C. M., Stenkamp, R. E. & Steitz, T. A. (1978).

Sequencing a protein by X-ray crystallography. II. Refinement of yeast hexokinase B. Co-ordinates and sequence at 2·1 Å resolution. *J. Mol. Biol.* **123**, 15–33.

Artymiuk, P. J., Grindley, H. M., Park, J. E., Rice, D. W. & Willett, P. (1992). Three-dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph-theoretical techniques. *FEBS Letters*, **303**, 48–52.

Banner, D. W., Kokkinidis, M. & Tsernoglou, D. (1988). Structure of the ColE1 rop protein at 1·7 Å resolution. *J. Mol. Biol.* **196**, 657–675.

Barakat, M. T. & Dean, P. M. (1991). Molecular structure matching by simulated annealing. III. The incorporation of null correspondences into the matching problem. *J. Comp.-aided Mol. Design*, **5**, 107–117.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. *Nature (London)*, **206**, 757–761.

Bork, P., Sander, C. & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 7290–7294.

Chen, L., Rose, J. P., Breslow, E., Yang, D., Chang, W.-R., Furey, W. F., Jr, Sax, M. & Wang, B.-C. (1991). Crystal structure of a bovine neurophysin II dipeptide complex at 2·8 Å determined from the single-wavelength anomalous scattering signal of an incorporated iodine atom. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 4240–4244.

Chiche, L., Gaboriaud, C., Heitz, A., Mornon, J.-P., Castro, B. & Kollman, P. A. (1989). Use of restrained molecular dynamics in water to determine three-dimensional protein structure: prediction of the three-dimensional structure of *Ecballium elaterium* trypsin inhibitor II. *Proteins*, **6**, 405–417.

Dekker, N., Cox, M., Boelens, R., Verrijzer, C. P., van der Vliet, P. C. & Kaptein, R. (1993). The solution structure of the POU-specific DNA binding domain of Oct-1. *Nature (London)*, **362**, 852–855.

Djinovic, K., Gatti, F., Coda, A., Antolini, L., Pelosi, G., Desideri, A., Falconi, M., Marmocchi, F., Rotilio, G. & Bolognesi, M. (1991). Structure solution and molecular dynamics of the yeast Cu,Zn enzyme superoxide dismutase. *Acta Crystallogr. sect. B.* **47**, 918–927.

Eck, M. J. & Sprang, S. R. (1989). The structure of tumor necrosis factor-α at 2·6 Å resolution. Implications for receptor binding. *J. Biol. Chem.* **264**, 17595–17605.

Fischer, D., Bachar, O., Nussinov, R. & Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dynam.* **9**, 769–789.

Flaherty, K. M., DeLuca-Flaherty, C. R. & McKay, D. B. (1990). Three-dimensional structure of the ATPase fragment of a 70 K heat-shock cognate protein. *Nature (London)*, **346**, 623–628.

Fogh, R. H., Kem, W. R. & Norton, R. S. (1990). Solution structure of neurotoxin I from the sea anemone *Stichodactyla helianthus*. A nuclear magnetic resonance, distance geometry and restrained molecular dynamics study. *J. Biol. Chem.* **265**, 13016–13028.

Gribskov, M., McLachlan, M. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983). The theory and practice of distance geometry. *Bull. Math. Biol.* **45**, 665–720.

Hazes, B. & Hol, W. G. J. (1992). Comparison of the hemocyanin β-barrel with other Greek key β-barrels: possible importance of the "β-zipper" in protein structure and folding. *Proteins*, **12**, 278–298.

Hill, C. P., Yee, J., Selsted, M. E. & Eisenberg, D. (1991). Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. *Science*, **251**, 1481–1485.

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* **61**, 237–251.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Science*, **1**, 409–417.

Holm, L. (1986). Codon usage and gene expression. *Nucl. Acids Res.* **14**, 3075–3087.

Holm, L. & Sander, C. (1993*a*). Structural alignment of globins, phycocyanins and colicin A. *FEBS Letters*, **315**, 301–306.

Holm, L. & Sander, C. (1993*b*). Globin fold in a bacterial toxin. *Nature (London)*, **361**, 309.

Holm, L., Ouzounis, C., Tuparev, G., Vriend, G. & Sander, C. (1992). A database of protein structure families with common folding motifs. *Protein Science*, **1**, 1691–1698.

Holmes, K. C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* **3**, 53–59.

Hyde, C. C., Ahmed, S. S., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* **263**, 17857–17871.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. (1990). Atomic structure of the actin: DNase I complex. *Nature (London)*, **347**, 37–41.

Kraulis, P. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.

Kraulis, P. J., Clore, G. M., Nilges, M., Jones, T. A., Pettersson, G., Knowles, J. & Gronenborn, A. M. (1989). Determination of the three-dimensional structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, **28**, 7241–7257.

Liebman, M. N. (1980). Quantitative analysis of structural domains in proteins. *Biophys. J.* **32**, 213–215.

Mathews, F. S., Bethge, P. H. & Czerwinski, E. W. (1979). The structure of cytochrome B562 from *Escherichia coli* at 2·5 Å resolution. *J. Biol. Chem.* **254**, 1699–1706.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Mitchell, E. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1989). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166.

Nishikawa, K. & Ooi, T. (1974). Comparison of homologous tertiary structure of proteins. *J. Theor. Biol.* **43**, 351–274.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J., Vershueren, K. H. G. & Goldman, A. (1992). The α/β hydrolase fold. *Protein Eng.* **5**, 197–211.

Ouzounis, C., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.* **232**, 805–825.

Phillips, D. C. (1970). Development of crystallographic enzymology. *Biochem. Soc. Symp.* **31**, 11–28.

Pletnev, V. Z., Kuzin, A. P. & Malinina, L. V. (1982). Actinoxanthin structure at the atomic level. *Bioorg. Khim.* **8**, 1637.

Rees, B., Bilwes, A., Samama, J. P. & Moras, D. (1990). Cardiotoxin V_{II}^4 from *Naja mossambica mossambica*: the refined crystal structure. *J. Mol. Biol.* **214**, 281–297.

Rees, D. C. & Lipscomb, W. N. (1982). Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2·5 Å resolution. *J. Mol. Biol.* **160**, 475–498.

Richardson, J. S. (1981). Anatomy and taxonomy of protein structures. *Advan. Protein Chem.* **34**, 167–339.

Rossmann, M. G. & Argos, P. (1976). Exploring structural homology in proteins. *J. Mol. Biol.* **105**, 75–95.

Sali, A. & Blundell, T. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**, 403–428.

Sander, C. (1990). Inverting the protein-folding problem. *Biochem. Soc. Symp.* **57**, 25–33.

Sippl, M. J. (1982). On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J. Mol. Biol.* **156**, 359–388.

Smith, J. L., Corfield, W. R., Hendrickson, W. A. & Low, B. W. (1988). Refinement at 1·4 Å resolution of a model of erabutoxin B. Treatment of ordered solvent and discrete disorder. *Acta Crystallogr. sect. A*, 44, 357–368.

Subbarao, N. & Haneef, I. (1991). Defining topological equivalences in macromolecules. *Protein Eng.* **4**, 877–884.

Taylor, W. & Orengo, C. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

Vellieux, F. M. D. & Hol, W. G. J. (1989). A new model for the pro-PQQ cofactor of quinoprotein methylamine dehydrogenase. *FEBS Letters*, **255**, 460–464.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins*, **11**, 52–58.

Weaver, L. H. & Matthews, B. W. (1987). Structure of bacteriophage T4 lysozyme refined at 1·7 Å resolution. *J. Mol. Biol.* **193**, 189–199.

Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K. & Jansonius, J. N. (1991). Structural conservation in parallel β/α-barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9169.

Wright, C. S. (1990). 2·2 Å resolution structure analysis of two refined *N*-acetylneuraminyl-lactose-wheat germ agglutinin isolectin complexes. *J. Mol. Biol.* **215**, 635–651.

Edited by F. Cohen