3-D Lookup: Fast Protein Structure Database Searches at 90 % Reliability

Liisa Holm and Chris Sander

European Molecular Biology Laboratory
D-69012 Heidelberg, Germany
Holm@EMBL-Heidelberg.de

Abstract

There are far fewer classes of three-dimensional protein folds than sequence families but the problem of detecting three-dimensional similarities is NP-complete. We present a novel heuristic for identifying 3-D similarities between a query structure and the database of known protein structures. Many methods for structure alignment use a bottom-up approach, identifying first local matches and then solving a combinatorial problem in building up larger clusters of matching substructures. Here, the top-down approach is to start with the global comparison and select a rough superimposition using a fast 3-D lookup of secondary structure motifs. The superimposition is then extended to an alignment of C^{α} atoms by an iterative dynamic programming step. An all-against-all comparison of 385 representative proteins (150,000 pair comparisons) took 1 day of computer time on a single R8000 processor. In other words, one query structure is scanned against the database in a matter of minutes. The method is rated at 90 % reliability at capturing statistically significant similarities. It is useful as a rapid preprocessor to a comprehensive protein structure database search system.

Introduction

Protein families are traditionally identified by sequence database searches. In recent years, an increasing number of distant evolutionary relationships that are not evident by sequence comparison have been revealed by similarity of 3-D protein structures, both because of a rapid increase in the number of known structures and because of improved methods of detection.

The problem of structure comparison is much more complicated than sequence string comparison because a 3-D match requires co-operative similarity in the relative disposition of many parts of the structure. Structure alignment is an optimization problem that requires the transformation of intuitive notions of structural similarity into objective quantities, with suitable choice of parameters. (Meaningful distance measures, in our view, are difficult to construct in this context.) Whatever the measure and search variables, the search landscape contains very many local optima due to the recurrence of secondary structure elements (helices and strands) and small tertiary structural motifs, i.e., associations of two, three, four helices or strands. However, in practical applications it is not necessary to locate the absolute optimum of the object function in each pair comparison. This is because one is usually only interested in those matches that involve the folding pattern of an entire structural domain.

The algorithm described here is meant to be fast if not complete. Comparison with a classification of the protein structure database using the "slow, reliable" (Orengo, 1994) Dali algorithm (Holm and Sander, 1993) is used to calibrate the method's reliability in detecting statistically significant similarities. Although the present method will not find all "neighbours" of a query structure in the database, it saves time in the identification of easy-to-find hits. Some of the remaining similarities can be detected using knowledge of already classified folds, by means of consistency checks of family relations. In database searching, this quick prefilter catches a large fraction of the interesting similarities. More sensitive but slower search methods must be used to check all remaining areas of search space but they can discard regions that fall below

the already known level of similarity. This strategy of using multiple algorithmic approaches to the structure comparison problem makes sure that nothing is missed while the overall procedure becomes much more efficient.

Definitions

Objective Function

The objective is to find a rigid-body 3-D superimposition of two structures that yields the maximum number of equivalent residue centres. The equivalence relation is defined to require that the spatial separation of the C^{α} atoms is below 4.0 Å. In addition, the constraint of sequential alignment is imposed so that topographical rearrangements (cutting and pasting of loop connections) or chain reversal (cutting and flipping a segment) are excluded. Formally: let us label the n equivalent pairs as (a_i,b_i), i=1,...,n where a_i is the residue number in protein A and b_i is the residue number in protein B. If the residues in A are sorted as $a_1<a_2<...<a_n$, then $b_1<b_2<...<b_n$ is required.

Vector Description of Protein Architecture

Higher order correlations in the sequence of C^{α} positions can be exploited to produce simplified descriptions of protein structure. Globular proteins have a layered architecture. Chain direction is reversed by loops or sharp turns at the surface. The solvent inaccessible core is made up of essentially straight segments. These are called secondary structure elements (SSEs) and are of two types, i.e., helices and strands of sheets, which can be identified by regular patterns of backbone-backbone hydrogen bonding. To a first approximation, the geometry of helix and strand segments can be represented by vectors. The structural core of globular proteins tends to be well conserved while surface regions change more rapidly in evolution. Most similarities of interest can therefore be identified by focussing on the core elements.

The SSE vector descriptors were extracted from the all-atom protein coordinates as follows. Each residue was initially assigned to one of helix, sheet or loop states using the program DSSP (Kabsch and Sander, 1983). Helix and strand segments were then extended at the ends by including loop residues up to a minimum length of 6 residues (strands) or 8 residues (helices). If a segment hit the borders of other segments before reaching the prescribed length, the whole segment was removed (assigned as loop). The *midpoint* of an SSE vector was defined as the average of all C^{α} coordinates in the segment. The *direction* vector was defined as equal to the vector from the midpoint of the N-terminal (first) half of the SSE to the midpoint of the C-terminal (second) half of the SSE. More sophisticated descriptions of protein geometry have been proposed (e.g., Thomas, 1994; Mitchell et al., 1990) and their use will be explored later, elsewhere.

Algorithms

The method for structure comparison has two parts. The first part is a 3-D lookup using the vector descriptors of SSEs that refers to the objective function only implicitly. The second part extends the comparison to the level of C^{α} atoms by a dynamic programming algorithm that optimizes the objective function explicitly; this approach is commonly used in this context although it is known to have a rather narrow radius of convergence (e.g., Sali and Blundell, 1990; Vriend and Sander, 1991; Russell and Barton, 1992; Subbiah et al., 1993). Our method is therefore heuristic in nature and we make no claim of mathematically rigorous optimization.

3-D Lookup

Heuristic. In principle, the search for the optimal translation-rotation operators is a problem with six degrees of freedom. Our fast 3-D lookup circumvents this complication by making an educated guess for the optimal superimposition. The guess is based on the observation that an optimal superimposition in terms of residue centres typically produces a close spatial coincidence of SSE vectors in the two proteins. Due to the way in which amino acid mutations are accommodated in protein structure, the positions and directions of the SSEs can indeed be better conserved than the positions of the residue centres that define them. Turning this around leads to the expectation that superimposing a subset of such well matching SSEs is sufficient to *approximately* regenerate

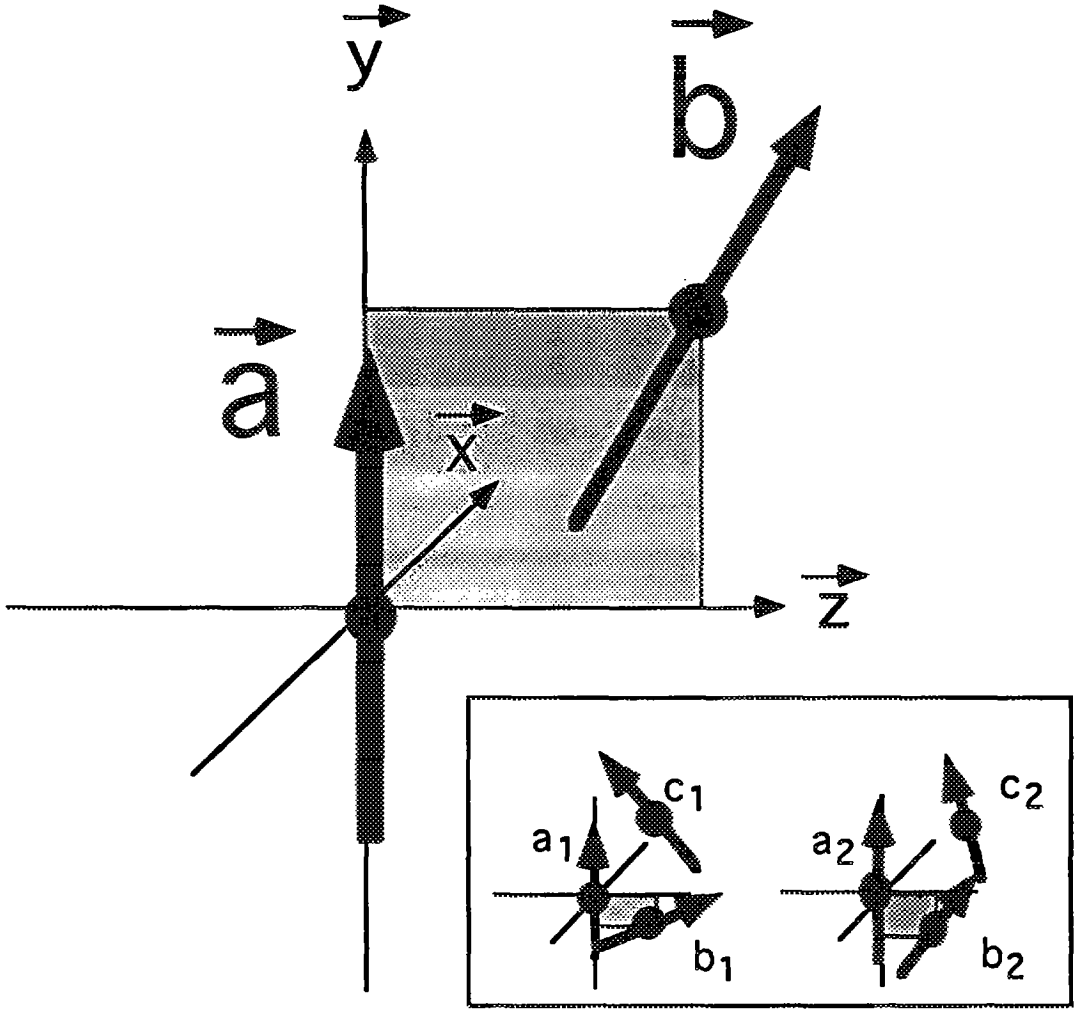

Figure 1: Coordinate system.
Protein structure is described as a set of *vectors* representing secondary structure elements (SSEs). An ordered pair of SSEs (*a* and *b*) defines a right-handed three-dimensional coordinate frame such that the midpoint of *a* is at the origin, the axis of *a* is along the positive y axis and the midpoint of *b* lies in the z-positive yz halfplane. It is required that the midpoint of *b* is not along the axis of *a*. (In practice, the singularity does not happen within machine precision.) The inset shows a comparison of the internal coordinate frames of two proteins (labelled 1 and 2): at the origin, the unit vectors $\widehat{a_1}=\widehat{a_2}$ are the same by definition and in this case the other SSEs match approximately in their directions ($\widehat{b_1}\approx\widehat{b_2}$ and $\widehat{c_1}\approx\widehat{c_2}$) and in the position of the segment midpoints (filled circles) relative to the origin.

the desired rigid-body transformation of the entire structures. In other words, the idea is to recover the whole from a comparison of the essential parts.

The key procedural step is to compare the spatial arrangement of SSEs in two proteins by superimposing appropriate internal coordinate frames, one for each protein. We define such internal coordinate frames in terms of the axis of one leading SSE and the direction to a second SSE (Figure 1). It is not known beforehand which frame to select in either protein. Fortunately, the number of possible coordinate frames is small enough to allow exhaustive testing of all frames for one structure against all frames for the other structure. For larger proteins, we further limit this number by excluding coordinate frames generated by pairs of SSEs that have a mutual distance larger than 12 Å.

Loading the target lookup grid. The number of matching SSEs between two proteins is counted efficiently using a 3-D lookup system that employs a storage and retrieval scheme reminiscent of hash tables. The idea is to precalculate all internal coordinate frames of a *target protein* and superimpose these on the axes of a 3-D grid. The SSE vectors are stored in this 3-D grid at the location of their midpoint coordinates in any particular frame. The grid cells have a size of 2 Å * 2 Å * 2 Å. Each cell contains a pointer to a linked list holding the explicitly transformed coordinates of the SSE vectors (midpoint and direction) together with identifiers for the generating coordinate frame (pair of SSEs) as well as sequential number and type (helix/strand) of the stored segment. Once loaded, the target protein can be probed by any number of query proteins.

Querying the grid. The search in the "3-D hash table" proceeds as follows. The grid is probed with a *query protein* by looping through each internal coordinate frame of the query protein. The given coordinate frame is superimposed on the grid axes. The query protein is now properly oriented to search for SSE matches by direct comparison with the 3-D coordinates that are stored in the grid for the target protein. To count a match between query and target SSEs, we require similar 3-D positions of the SSE midpoints (less than 4 Å distance), agreement of SSE type (helix-helix or strand-strand), a deviation of less than 30 degrees between the direction vectors, and similar sequential position (before-before or after-after) relative to the y-axis determining SSE (Figure 1). The 4 Å distance limit is chosen so that for most query segments there is at most one match with a target segment. The grid allows efficient pruning of search space, as a list of all possible candidate matches in the target protein is obtained through lookups in a few grid cells around the midpoint of any particular query SSE.

In comparing a query and target protein, the search algorithm keeps track of the number of SSE matches accumulated over each pair of coordinate frames. In database searching, the above comparisons are repeated for a large number of query proteins and the combination of coordinate frames which yields the highest number of matching SSEs is remembered for each pair of target/query proteins.

Refinement

The refinement step basically uses a textbook algorithm (Lesk, 1991, p. 132) which is repeated here for completeness. The previous 3-D lookup step yields preoriented coordinate sets *X, Y* of the two proteins which have n_X and n_Y residues. A sequential alignment (which maps every residue in the first protein either to null or a structurally equivalent residue in the second protein) is generated by the following iterative procedure:

Step 0: Initialize the "current" alignment with all nulls. Zero the iteration counter.

Step 1: Increment iteration counter by 1.

Step 2: Copy "current" alignment to "previous" alignment.

Step 3: Run a standard dynamic programming algorithm to maximize the sum of scores along a sequential path where the score *s* is a function of the Cartesian distance *r* of C^{α} atoms $i \in X$, $j \in Y$:

$s(i,j)=\max(0.0, 4.0\ \text{Å} - r(i,j))$, where $1 \le i \le n_X$ and $1 \le j \le n_Y$.

Step 4: The best trace returned by the dynamic programming algorithm is an alignment containing pairs (i,j) for each residue i in the first protein. Reset to (i,null) those pairs (i,j) which have a zero score, i.e., which are outside the 4 Å limit of similarity, as structurally non-equivalent.

Step 5: Compute the translation-rotation matrices that optimize the least-squares

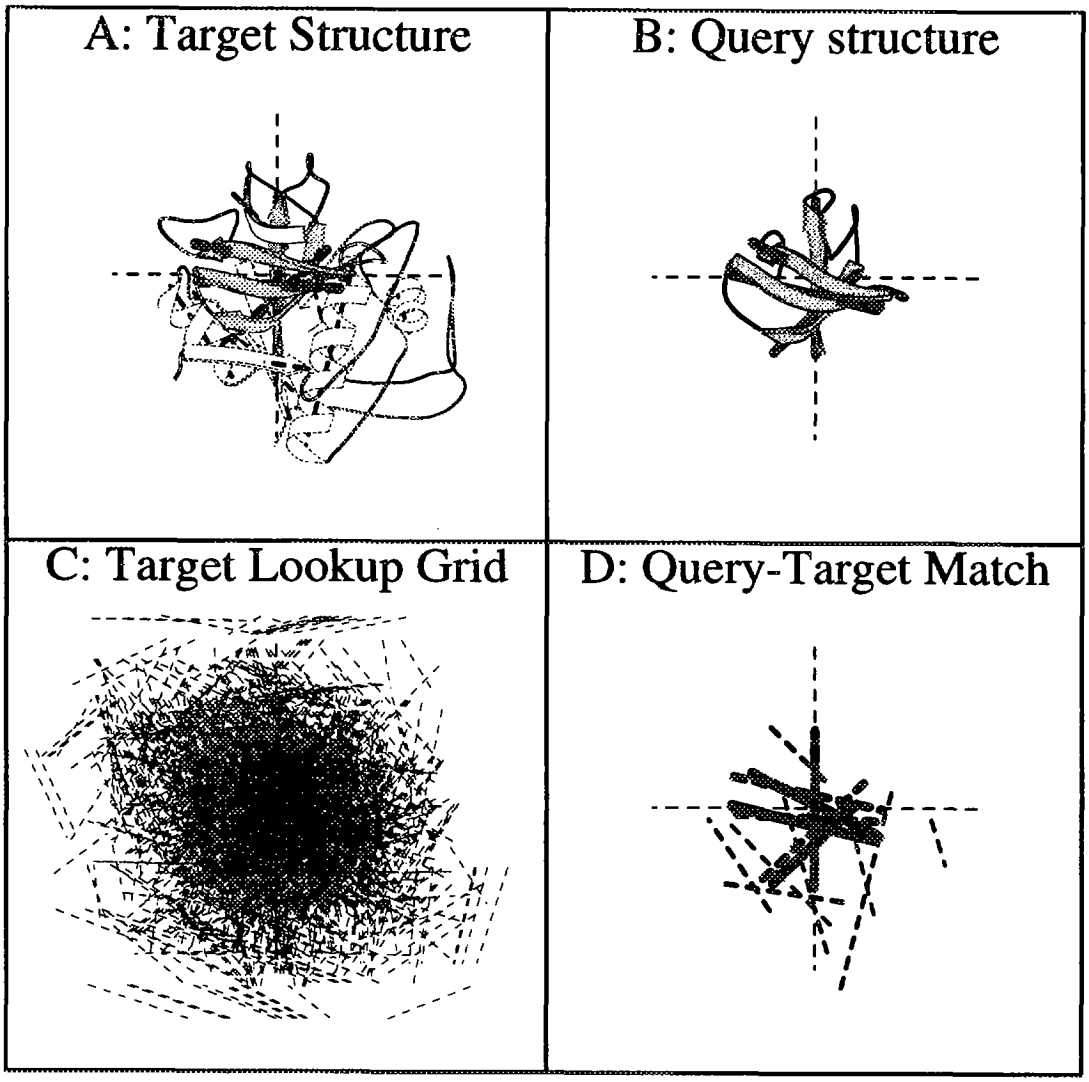

Figure 2: Principle of the heuristic.

A-B. Structure comparison of an SH3 domain from c-src kinase (1cskA, query structure) with the enzyme papain (1ppn, target structure) reveals similar domain folds (gray ribbons) although there is no sequence relationship between the proteins and one is much larger. The appropriate orientation of the molecules is found by exhaustive comparison of internal coordinate frames for each protein. **C.** The target structure, papain, loaded onto the grid. Each pair of SSEs where the segment midpoints are within 12 Å defines a coordinate frame relative to the grid axes. The figure shows the transformed positions of the 12 SSEs of papain (dotted lines) in each of the ~100 different coordinate frames defined by different pairs of SSEs. **D.** The target lookup grid is probed with the SH3 domain, which has 4 SSEs (thick continuous lines). The coordinate frames shown are the ones yielding the best three-dimensional match of four segments. These best matching frames are defined by SSEs (1,2) of the SH3 domain and SSEs (7,9) of papain. The equivalent SSE pairs are (1,2,3,4) in the SH3 domain with (7,9,10,11; thick dotted lines) in papain, respectively. Iterative extension of a residue-wise alignment starting from the preorientation defined by the SSE match shown here leads to equivalencing of 43 C^{α} atoms with 1.7 Å root-mean-square positional deviation on optimal least-squares superimposition. SSE vectors are here shown as lines centred at the midpoint of the segment and colinear with and twice the length of the direction vectors. Drawn with MolScript (Kraulis, 1991).

superimposition of the aligned C^{α} atoms in X onto their equivalent pairs in Y (Kabsch, 1978).

Step 6 Transform coordinate set *X'<-X* using the matrices from Step 5.

Step 7: Compare "current" alignment to "previous" alignment. If the two are not identical and the iteration counter is less than a limit (currently set to 20), then go to step 1.

Step 8: Return "current" alignment.

Cluster analysis

The method has been tested empirically by performing an all-against-all comparison in a set of 385 representative structures with less than 30 % pairwise sequence identity (Hobohm and Sander, 1992) and at least 3 secondary structure elements. This set was clustered into families by building an average linkage tree based on pairwise similarities (using the larger value for asymmetric X-Y and Y-X alignments). A similarly constructed tree using alignments by the Dali algorithm (distance matrix alignment by Monte Carlo optimization; Holm and Sander, 1993) was used as reference classification. Structural classes (families) were defined in the Dali reference tree using a Z-score cutoff of 2.0, where the Z-score is obtained from the original geometrical similarity score after normalization using a background distribution that takes into accout domain size (Holm and Sander, 1994). Clusters in the new tree were defined by either the number of matching SSEs or a Z-score calculated from the residue-level alignment (Table I).

Trees generated by Dali and by the new 3-D lookup methods were compared using a *split count* (*S*) for each family in the reference (Dali) tree. Perfect agreement (S=0) for a Dali family is obtained if there is a node in the new tree that encompasses all members of the family and no other proteins. Deviation from perfect agreement is measured by the number of separate clusters (nodes) needed to cover the Dali family minus one, i.e., by the number of splits. The Dali tree grouped the 385 proteins into 131 families. A relative *reliability index* (*R*) of the classification is given by $R=1-S/254$, where 254 is the number of nodes representing family relations (385 minus 131) and split counts S are summed over all Dali families.

Coordinates were retrieved from the Protein Data Bank (Bernstein et al., 1977). Coordinate entries are referred to as *codeX*, where *code* is the 4-letter Protein Data Bank identifier and *X* is the chain identifier. The C^{α} coordinates and segment definitions were read in from a preprocessed database.

Results and Discussion

We think that the present method works amazingly well considering the complexity of the structure alignment problem and the simplicity of the heuristic. An example from the 3-D lookup stage is shown in Figure 2 and another example of alignments after the refinement step in Figure 3.

Reliability Test

Benchmarks from three implementations of the heuristic are summarized in Table I. Already the simplest tree generated using only the 3-D lookup step to estimate the number of matching SSEs identifies more than two thirds of family relations (WOLF1). This search is blazingly fast, taking only a few minutes for the entire all-against-all comparison of 385 proteins. Rather few corrections were gained by testing more frames in the 3-D lookup (WOLF2). Although most families of large proteins are correctly identified in the trees, the number of matching SSEs is not a very sensitive measure for smaller proteins. A marked improvement is achieved by investing some time in the refinement step which extends the alignments to residue level (WOLF3). We found that assessing the quality of the 3-D matches using Dali scores gives better trees than using, e.g., the number of equivalenced residues.

Limitations and Future Improvements

The speed of the method comes at the cost of certain limitations. Structures with fewer than three segments are excluded. The coarse screening fails to detect a small but non-negligible fraction of strong similarities, which can of course be recovered in a slower (Dali) step. A frequent cause behind missed similarities between remote homologs is inconsistent definition of beginning and end of strand segments. Examples are the class of growth factors with an elongated sloppy β-fold that includes the platelet-derived growth factor BB, transforming growth factor-β2, and human

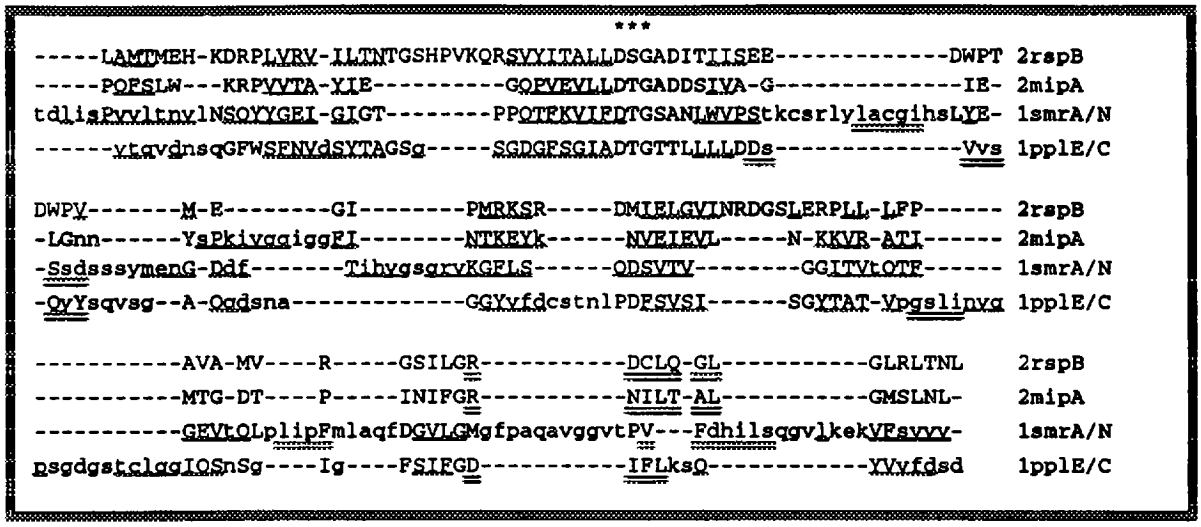

Figure 3: Example structural alignment: aspartic proteases.
Database search using the viral protease 2rspB as query structure identifies the known relatives from another virus (2mipA: 1.3 Å r.m.s. deviation over 87 C^{α} pairs after residue-level refinement), fungus (1pplE/N-terminal domain: 2.0 Å r.m.s. deviation over 74 C^{α} pairs after refinement) and mouse (1smrA/C-terminal domain: 2.0 Å r.m.s. deviation over 79 C^{α} pairs after refinement). Note the correct alignment of the active site signature D(S/T)G (***) although no sequence information is used in the search. Residues which are structurally equivalent with 2rspB are in uppercase, non-equivalent in lowercase, dashes indicate deletions. Strands are underlined, helices are doubly underlined. This figure combines three independent pairwise alignments with respect to 2rspB and extra gaps are used as padding characters to show the entire sequence of each of the matched domains.

Table I

method	description	clustering variable	CPU time*	time / pair	split count (S)	reliability (R)**
WOLF1	3-D lookup using only one frame (closest neighbour) per SSE, no refinement	number of matching SSEs	3.5 min	0.001 s	70	72 %
WOLF2	3-D lookup using all frames defined by SSE neighbours within 12 Å, no refinement	number of matching SSEs	40 min	0.02 s	68	73 %
WOLF3	3-D lookup using all frames defined by SSE neighbours within 12 Å, plus refinement	Z-score	1 d	0.6 s	24	91 %

* All-against-all comparison of 385 structures (150,000 comparisons) on an R8000 processor.
** Correctness of family classification compared to Dali reference tree.

chorionic gonadotropin (1pdgB, 2tgi, and 1hcnA/1hcnB), or the catalytic domains of verotoxin (1bovA) and enterotoxin (1ltsD). A possible remedy would be to define segments in terms of chain curvature rather than detailed atomic interactions, as in these cases the C^{α} traces are conserved although hydrogen bonding patterns (defined by the DSSP program) are not.

Another case of topological similarity missed by the current implementation of the fast 3-D lookup involves myoglobin and the membrane insertion domain of colicin A. These proteins are known to match over 6 helices or more than 100 residues with an rmsd of just over 3 Å (Holm and Sander, 1993a). However, in this "global" transformation there are no closely matching pairs of helix axes, so the current heuristic cannot work. It is conceivable that coordinate frames determined by some other set of reference points could be more sensitive, for example, using points of closest approach between triplets of SSEs.

The 3-D lookup is not guaranteed to give symmetric results for pair comparisons X-Y and Y-X. In addition, if folds have internal symmetry, it may happen that a suboptimal alignment is chosen at the 3-D lookup step, e.g. in the comparison of TIM [$(\alpha/\beta)_8$] barrels. The method could be extended to examine more than one initial superimposition as a starting point for refinement, or the sharp distance cutoff of the 3-D lookup could be replaced by a continuous function which might give better discrimination between the alternative frames.

The refinement step is the slow part of the algorithm. Database size being constant, the execution time scales linearly with the number of residues in the query structure. With some loss of sensitivity, speed could be gained by passing only those pairs to the extension step which have more SSEs in common than some cutoff. As the pairwise protein comparisons are independent, a substantial additional speedup would result from performing database searches in parallel.

Related Methods

The efficient search of protein structural databases is a vigorous area of research and development in computational molecular biology. Except for our definition of coordinate system, many of the basic concepts used in this work have been used before in one form or other. A number of iterative methods using dynamic programming enter the iteration cycle with an extensive prealignment at step 5 (e.g., Russell and Barton, 1992; Vriend and Sander, 1991; Subbiah et al., 1993), which in our experience is vulnerable to misassigned portions in the prealignment. Our innovation in defining the internal coordinate frames is using a *pair* of segments in each molecule as a trial unit rather than only one segment of backbone as in several clustering algorithms (e.g., Vriend and Sander, 1991; Alexandrov et al., 1992) or just one residue as at the bottom level of Taylor's double dynamic programming algorithm (Taylor and Orengo, 1989; Orengo et al., 1992). Using tertiary structural motifs apparently directs the search to accurate initial guesses of the global translation-rotation transformation, with less sensitivity to local deformations. Our method differs from algorithms working with pairwise relations of SSE vectors (e.g., Mitchell et al., 1990; Grindley et al., 1993) in the use of direct 3-D hashing rather than a tree search for a maximal common subgraph. The geometric hashing algorithm by Fischer et al. (1992) does lookups on interatomic distances for triplets of C^{α} atoms and then uses a complicated clustering procedure to work out the alignment. More related in spirit is the approach by Johnson et al. (1994) which uses a genetic algorithm for optimizing the translation-rotation matrices after superimposition of the centres of mass.

Conclusion

The new contribution of this work is the top down, coarse screening of structural similarity using vector descriptors of protein architecture. The speedup gained is orders of magnitude compared to our previous method (Dali), at the cost of some false negatives. The method will be part of a comprehensive structure database search system that uses multiple algorithmic levels and stored family information in order to efficiently determine the structural neighbours of a query protein.

References

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

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

Fischer, D.; Bachar, O.; Nussinov, R.; and Wolfson, H. 1992. An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn.* 9:769-789.

Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; and Willett, P. 1993. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* 229:707-721.

Hobohm, U.;Scharf, M.;Schneider, R.; and Sander, C. 1992. Selection of representative protein data sets. *Protein Sci.* 1:409-417.

Holm, L.; and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123-138.

Holm, L.; and Sander, C. 1993a. Globin fold in a bacterial toxin. *Nature* 361:309.

Holm, L.; and Sander, C. 1994. The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* 22:3600-3609.

Johnson, M. S.; Overington, J. P.; Edwards, Y.; May, A. C. W.; and Rodionov, M. A. 1994. The comparison of structures and sequences: alignment, searching and the detection of common folds. In L. Hunter (Ed.), 27th Hawaii International Conference on System Sciences, V: Biotechnology Computing (pp. 296-305). Los Alamitos, California: IEEE Computer Scociety Press.

Kabsch, W. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* 34:827-828.

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

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

Lesk, A. M. 1991. *Protein architecture. A practical approach.* Oxford: Oxford University Press.

Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; and Willett, P. 1990. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* 212:151-166.

Orengo, C. 1994. Classification of protein folds. *Current Opinion in Structural Biology* 4:429-440.

Orengo, C. A.; Brown, N. P.; and Taylor, W. T. 1992) Fast structure alignment for protein databank searching. *Proteins* 14:139-167.

Russell, R. B.; and Barton, G. J. 1992. Multiple protein sequence alignment from tertiary structure: assigment of global and residue confidence levels. *Proteins* 14:309-323.

Sali, A.; and Blundell, T. L. 1990. Definition of general topological equivalence in protein structures. *J. Mol. Biol.* 212:403-428.

Subbiah, S.; Laurents, D. V.; and Levitt, M. 1993. Structural similarity of DNA-binding domains of bacterophage repressors and the globin core. *Current Biology* 33:141-148.

Taylor, W. R.; and Orengo, C. A. 1989. Protein structure alignment. *J. Mol. Biol.* 208:1-22.

Thomas, D. J. 1994. The graduation of secondary structure elements. *J. Mol. Graphics* 12:146-152.

Vriend, G.; and Sander, C. 1991. Detection of common three-dimensional substructures in proteins. *Proteins* 11:52-58.