PROTEINS: Structure, Function, and Genetics 11:52–58 (1991)

Detection of Common Three-Dimensional Substructures in Proteins

Gerrit Vriend and Chris Sander
European Molecular Biology Laboratory, D-6900 Heidelberg, Federal Republic of Germany

ABSTRACT **We present a fully automatic algorithm for three-dimensional alignment of protein structures and for the detection of common substructures and structural repeats. Given two proteins, the algorithm first identifies all pairs of structurally similar fragments and subsequently clusters into larger units pairs of fragments that are compatible in three dimensions. The detection of similar substructures is independent of insertion/deletion penalties and can be chosen to be independent of the topology of loop connections and to allow for reversal of chain direction. Using distance geometry filters and other approximations, the algorithm, implemented in the WHAT IF program, is so fast that structural comparison of a single protein with the entire database of known protein structures can be performed routinely on a workstation. The method reproduces known non-trivial superpositions such as plastocyanin on azurin. In addition, we report surprising structural similarity between ubiquitin and a (2Fe-2S) ferredoxin.**

Key words: protein structure comparison, superposition, clustering, folding units, sequence alignment

INTRODUCTION

Comparison of protein structures has many areas of application. Three-dimensional similarity can be used to produce protein alignments in cases where sequence similarity is so weak that sequence alignment programs fail.[1] Structure-based sequence alignment can reveal evolutionary relationships and provide the basis for the construction of phylogenetic trees.[2] Multiple alignment of structures naturally leads to the definition of a common structural core of a protein family,[3] to the identification of structurally important conserved contact regions,[4] and to the detailed study of residue replacements in conserved structural context.[5]

The principal difficulty in comparing three-dimensional protein structures is that of identifying structurally equivalent residues. Once a list of equivalent residues is known, elegant solutions to the problem of optimal superposition in 3-D[6] can be used to produce explicit coordinates of one protein in the framework of the other. Superficially, the equivalencing problem is similar to the problem of one-dimensional alignment of amino acid sequences. There is, however, an added complication in that clusters of residues locally similar in three-dimensional space may involve chain regions separated by many residues, i.e., arranged non-locally in sequence space. It is therefore not sufficient to compare one-dimensional neighborhoods in sequence space, but also necessary to compare three-dimensional neighborhoods in real space. For this reason, one-pass dynamic programming algorithms are not suitable for this problem.

Several authors have invented generalizations of sequence alignment algorithms in order to solve the 3-D equivalencing problem. For example, Taylor and Orengo[7] first define a local measure of similarity between any two sequence positions in two proteins by aligning the contact environments of each residue in protein A with that of each residue in protein B, using a dynamic programming algorithm. Subsequently, they solve the one-dimensional alignment problem in terms of new local similarities derived from the first step, again by dynamic programming. The algorithm can be thought of as solving the problem of aligning two contact maps (or distance plots), allowing insertions and deletions but adhering strictly to the sequential order of residues along the chain. This method is conceptionally neat and works well, but it is costly in computer time, as the algorithm is of order $N(A)^2N(B)^2$, where N(X) is the chain length of protein X. Sali and Blundell[8] use a Monte Carlo method, simulated annealing, to deal with the complexity of optimizing structural superposition, whereas Zuker[9] uses a dynamic programming algorithm.

Several other known methods for protein structure comparison are not based on generalizations of sequence alignment algorithms, but use a variety of iterative schemes to optimize superposition.[10–19] These methods have been extensively used for (closely) related structures. However, they each suffer from one or more of the following drawbacks: (1)

Received August 27, 1990; revision accepted February 1, 1991.

Address reprint requests to either Chris Sander or Gerrit Vriend, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 102209, D-6900 Heidelberg, Germany

large insertions and deletions are difficult to recognize; (2) only sequential alignments can be detected; (3) neither the occurrence of multiple copies of a motif nor spatial similarity in spite of different loop connections are detectable; (4) manual initial alignment is required; and (5) massive CPU resources are needed.

Here, an algorithm for protein structure comparison is proposed that overcomes these problems. The procedure consists of three steps. One, selection of sequence-local fragments that superpose well to create a diagonal plot. This is done efficiently using distance geometry criteria[20] followed by a fast algorithm for 3-D superposition.[6] Two, cluster analysis on pairs of fragments in order to identify larger structural units. Three, final optimization of the set of equivalent residues by minor trimming and extension and final optimal superposition of coordinates.

METHODS

From Sequence Alignment to 3-D Alignment

The point of departure of our algorithm is a so-called diagonal plot, a standard device for the graphical representation of sequence alignments. The two axes of such a rectangular plot are the sequences of the two proteins. A diagonal line segment, i.e., a trace inclined at 45° represents the similarity of a stretch of a certain length in protein A with a stretch of the same length in protein B without insertions or deletions, e.g., the similarity between two beta strands. The classical sequence alignment problem is that of finding an overall optimal path through the diagonal plot, connecting diagonal line segments, such that the overall similarity, i.e., residue similarity summed over all residues in all fragment pairs, is optimal, with sequential order strictly maintained. Sequential order is satisfied for two pairs of matching fragments, say, A1/B1 and A2/B2, if the fragments occur in the same order in both proteins, i.e., either A1<A2 and B1<B2 or A1>A2 and B1>B2, but not in mixed order like A1<A2 and B1>B2 or A1>A2 and B1<B2, where < means "comes before" in sequence and > means "comes after." See Figure 1A for an example.

In 3-D alignment, an optimal sum over fragment pair similarities alone does not guarantee that the matched segment pairs are part of similar substructures. For example, suppose that fragments A1 and A2 each are similar in shape to fragments B1 and B2, respectively. If, however, the spatial relationship of A1 and A2 in protein A differs from that of B1 and B2 in protein B, then the fused substructure A1 + A2 is not similar to B1 + B2; see Figure 1B for an example. The generalization of this argument from two to N fragment pairs leads to a clustering algorithm in which a new fragment pair Ai/Bi is added to an existing cluster of pairs if the spatial relationship between Ai and the A fragments in the cluster is similar to that of Bi and the B fragments already in the cluster. The requirement of strict sequential order of equivalenced fragments can be dropped. For example, one can allow A2/B2 to join in a cluster with A1/B1 even if A1 < A2 and B2 > B1 in sequence.

Technically, similarity of spatial relationship can be evaluated either in terms of explicit 3-D superposition or in terms of intrafragment distances. For example, one could simply determine the optimal superposition of A1 + A2 as one piece onto B1 + B2 and apply a cutoff in positional rms (root mean square) deviation as a criterion for joining these pairs into a common cluster. Alternatively, on could compare a set of alpha-carbon distances within A1 + A2 with an equivalent set within B1 + B2. A much more efficient test of spatial relationship can be made in terms of quantities already calculated in the production of the diagonal plot. This is our key technical point. The idea is to assess the similarity of spatial orientation between a pair A1/B1 and another pair A2/B2 by comparing the rotation operator attached to each pair comparison. The union of fragments A1 + A2 in protein A is similar to the union of fragments B1 + B2 if and only if the rotational transformation that best superposes A1 onto B1 is similar to the one that best superposes A2 onto B2. The comparison of operators can be performed by multiplying one operator by the inverse of the other and quantifying the departure of the result from the unit operator. Technical details of our method are given in the next three subsections.

Diagonal Plot

The first step in the creation of the diagonal plot is the comparison of fragments in the two proteins to be aligned. All fragments with a certain minimum length from one protein are compared with all fragments of the same length from another protein. Fragment pairs of similar structure are retained and reported in the diagonal plot as diagonal traces. In order to save computer time, a geometrical filter is applied to each pair of fragments in terms of intrafragment distances. If two fragments have the same structure, they will also have the same set of internal distances. So, if distance criteria are violated, the two fragments cannot have the same structure and need not be compared in more detail. However, the converse is not true. Even if distance sets are similar, the structures may be significantly different. Therefore the (fast) comparison of sets of distances has to be followed by a (slower) explicit three-dimensional superposition in order to eliminate false positives.

The comparison of fragment geometry in terms of internal alpha-carbon distances is done by a method similar to that of Jones and Thirup,[20] except that only the distances from the first alpha-carbon atom to the last five alpha-carbon atoms in the same frag-

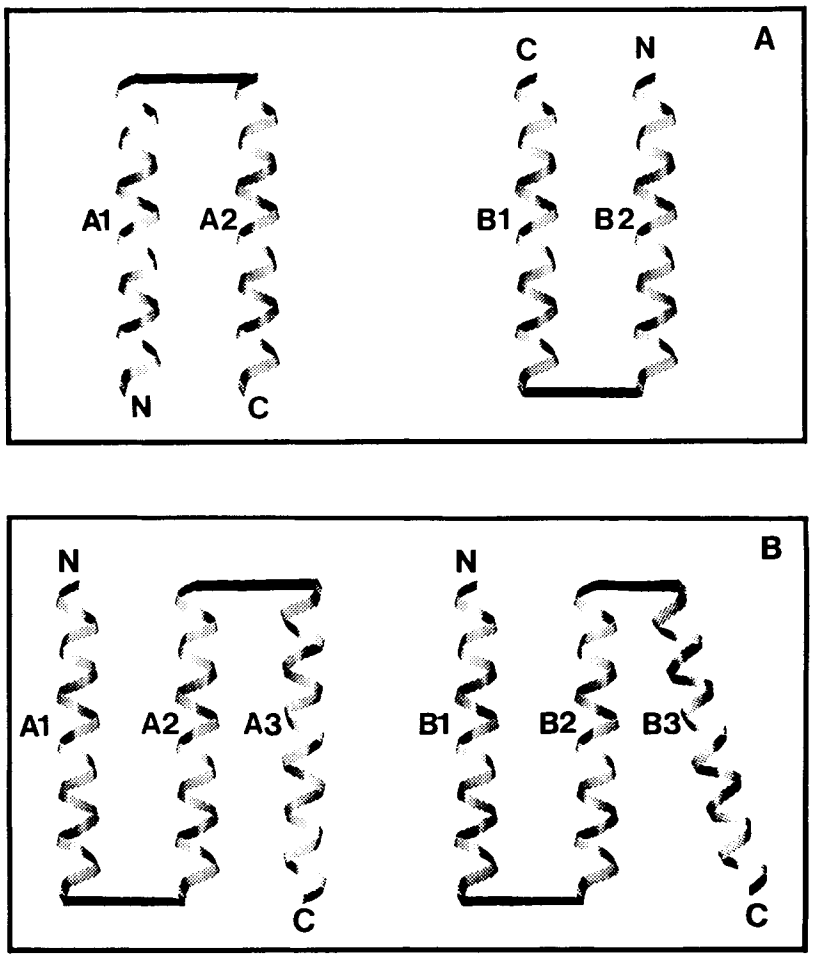

Fig. 1 Schematic example of comparing two alpha-helical protein structures, A and B, in which helix A1 is similar to helix B1, A2 to B2 and A3 to B3. **A.** The cores of the two structures superpose well, even though the interhelical connections are different. **B.** Helices A1 + A2 superpose well on B1 + B2. The third fragment pair, A3/B3, cannot join the cluster made by the other two helix pairs because the third helix has a different orientation in the two proteins.

ment are used. Using more than five distances did not appreciably improve the selectiveness of the filter. A pair of fragments is rejected if two equivalent alpha-carbon distances differ by more than a specified cutoff. To be sure that no pair will be rejected spuriously, this cutoff should be set at two times the maximum acceptable coordinate error after 3-D superposition of the fragments. The length of the shortest fragments compared is normally set at 10 to 15 residues. Using shorter fragments gives rise to an excessive number of matched fragments; using a longer minimum length tends to reduce the number of hits unacceptably.

In a second step, pairs of fragments that are not rejected by the distance geometry filter are superposed using the least-square algorithm of Kabsch.[6] This is straightforward as the two fragments in a pair have the same length. For each fragment pair, the goodness of fit is evaluated in terms of the root mean square distances, Drms, and the largest distance, Dmax, between equivalent alpha-carbon atoms, after optimal superposition. Two fragments are considered to be sufficiently similar if these distance values are below specified upper limits, typically 2.0 Å for Drms and 3.8 Å for Dmax (tighter limits should be used for very similar proteins).

In a third step, accepted fragment pairs are elongated: one residue is added at the C-terminus of both fragments and the longer fragments are again superposed. This process is repeated until the next addition would lead to violation of the upper limits. Additional computer time is saved by avoiding the comparison of fragments that are entirely helical, as every helix always fits every other equally long or longer helix: fragments are only compared if they contain at least four non-helical residues. Secondary structure assignments are taken from the DSSP dictionary.[21] Also, fragment pairs are skipped that would be subsets of already stored fragments. Together, these three empirically developed steps produce very useful diagonal plots as input to the cluster analysis, with great economy of computer time.

Cluster Analysis

In order to assemble pairs of fragments into a pair of larger units, we use a simple incremental clustering procedure. In each step, one needs to determine if a pair of fragments can join a cluster. The most plausible way would be to simply add the fragment pair to the cluster and to perform a complete 3-D superposition on the entire set of equivalenced residues to see if the positional errors are less than some preset limit. This would, however, be a rather slow process, so a filter that determines if a pair of fragments could potentially join the cluster is needed. The filter we use assesses whether the fragment pairs A1/B1 and A2/B2 can be joined into a larger pair, i.e., if A1 + A2, taken as a rigid body, can be superposed well onto B1 + B2 (Fig. 1). This is done in two steps making use of quantities already calculated during the generation of the diagonal plot.

First, we check that the distance between the centers of mass of A1 and A2 is similar to the distance between the centers of mass of B1 and B2. If these intraprotein distances are too dissimilar, then there cannot be a good superposition of A1 + A2 onto B1 + B2. Second, the rotation matrix of the optimal superposition A1/B1 is compared with that of A2/B2. This is done by multiplying one superposition rotation matrix (R1) with the inverse of the other (R2) and quantifying the departure from the unit matrix in terms of the resulting net rotation angle δ given by

$$\cos \delta = \frac{1}{2}[trace(R_1 \cdot R_2^{-1}) - 1].$$

The discrepancy angle δ is equal to zero if R1 and R2 represent identical rotations. If δ is above a cutoff value, typically 0.2–0.3 radians, the two rotations are considered dissimilar and the fragments cannot be merged. The rationale behind this is as follows: if two proteins are perfectly superposable, then every pair of equivalent substructures is also perfectly superposable, with the same rotational component of the superposition transformation; deviations from perfect superposability can be measured in terms of deviations in the rotational component. Because the same reasoning does not hold for the translational component of the transformation, we use the vector between the centers of mass, as described above.

In order to determine the largest cluster(s) for a given protein pair, each pair of fragments should be used in turn to start a new clustering process. In general, this implies N^2 comparisons of pairs of fragment pairs, given N pairs of fragments in the diagonal plot. In practice, it is often satisfactory to terminate the search as soon as a sufficiently large cluster is found, say, exceeding the size of a minimal folding unit (> 40–50 residues) or, say, containing half of all residues in one of the proteins. If the two proteins have a measurable degree of similarity, it is likely that the fragment pairs near the main diagonal will provide the largest cluster. Therefore, in practice we search for clusters along the main diagonal first and terminate on cluster size, reducing the complexity of the clustering procedure from order N^2 to order N.

Final Adjustment and Equivalencing

Creation of the diagonal plot and clustering of pairs of fragment in principle solves the problem of structural 3-D alignment. However, for practical reasons having to do with some of the time-saving approximations, a final pruning and fine-tuning of the largest clusters is performed. These reasons are: (1) pair comparison of fragments of length, say, less than 10 residues was avoided in the initial step, but in the final superposition, such short fragments could be interesting, provided they fit into the overall context; (2) in the clustering procedure, a new fragment pair was only compared with the starting member of the cluster, so one is not yet sure that the distance criteria are fulfilled for the entire cluster; and (3) it may be of interest to detect additional segments that are part of the cluster only if their chain direction is reversed.

The largest clusters of pairs of fragments are selected and fine-tuned with an iterative procedure similar in part to that of Rao and Rossmann.[17] First, the fragments are optimally superposed, such that Drms, the average positional deviation, is minimal. Subsequently, the list of equivalenced residues is re-examined and adjusted according to the criteria discussed below and a new overall transformation and Drms are determined. The process is iterated until no further adjustment is required. This termination condition is normally fulfilled within six to nine cycles. In the equivalencing pass of the final optimization a pair of residues is accepted: (1) if all equivalenced alpha-carbon positions are within Dmax of each other and; (2) if the pair of residues is part of two consecutive stretches of minimal length (say, 5 residues), acceptable according to (1). Optionally, fragments are allowed to run sequentially in opposite directions. The final cluster is reported after optimal superposition as a list of equivalenced residues, i.e., as the structure-derived sequence alignment.

RESULTS

As a test of the method, several well-known comparisons were redone: two hemoglobin chains, plastocyanin-azurin, and the two domains of rhodanese. In addition, we report discovery of an unexpected structural similarity: ubiquitin-ferredoxin. For the alpha and beta chain of hemoglobin[22] (Fig. 2), our alignment agrees with that of Lesk and Chothia.[3] For plastocyanin[23] azurin[24] (Fig. 3), our result agrees with the alignment by Adman.[25] A known

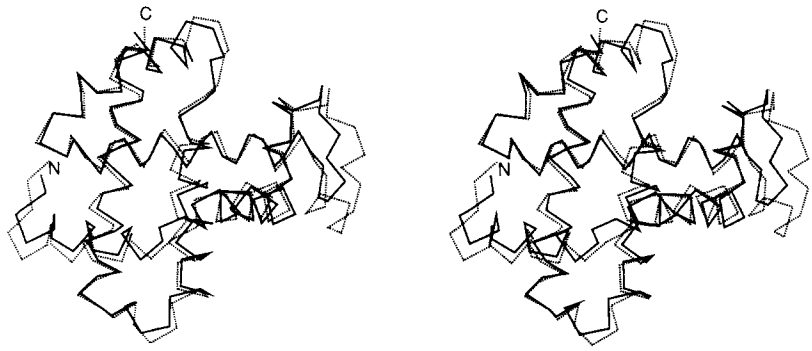

Fig. 2. Human deoxyhemoglobin beta chain (dashed lines) superposed on the alpha chain (solid lines). Stereo view, N terminus and C terminus labelled.

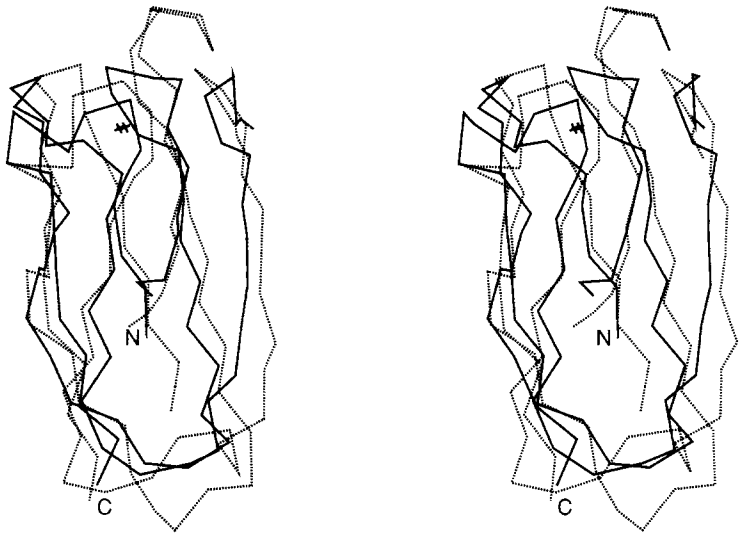

Fig. 3. Plastocyanin (dashed lines) superposed on azurin (solid lines). Two non-homologous loops (residues 43–62 in plastocyanin; residues 51–86 in azurin) are not shown, for clarity. The two bound copper ions are indicated by crosses. Stereo view.

example of internal duplication is bovine liver rhodanese.[1] This molecule is composed of two structurally similar domains, with no detectable sequence similarity between them. The cores of these two domains are almost identical, but the loop regions vary in length. The fragment match diagonal plot (Fig. 4), comparing the first with the second domain, has traces near the main diagonal that can be merged into one large cluster, corresponding to the superposition of the two domains (Fig. 5). The derived alignment is essentially identical to the one determined by the Ploegman et al.[1] with at most one residue more or less equivalenced at the ends of fragments.

A first database scan turned up several new structural similarities. One example is the pair ubiquitin[26] / ferredoxin[27] (Fig. 6) Ubiquitin, a 76 residue protein, is involved in protein breakdown via covalent conjugates, whereas ferredoxin, with 98 residues, functions as an electron carrier in the pho-

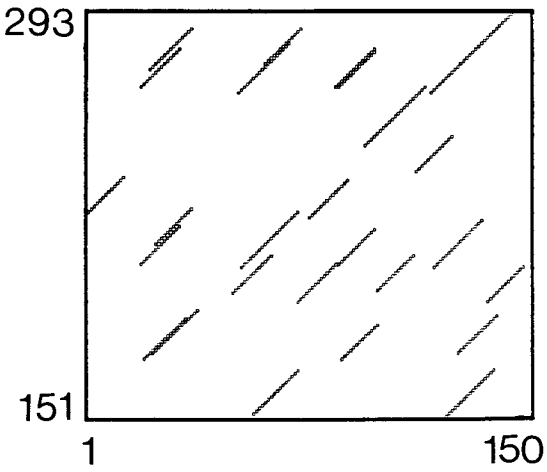

Fig. 4. Diagonal plot of fragment similarities in rhodanese, between the first domain (residues 1–150) and the second domain (residues 151–293), as used in the first step of the algorithm. Each trace corresponds to a fragment pair, which may or may not fit with the overall domain comparison.

Fig. 5. Rhodanese C-terminal domain (dashed lines) superposed on its N-terminal domain (solid lines). The terminal 4 (7) residues are not shown, for clarity. Stereo view.

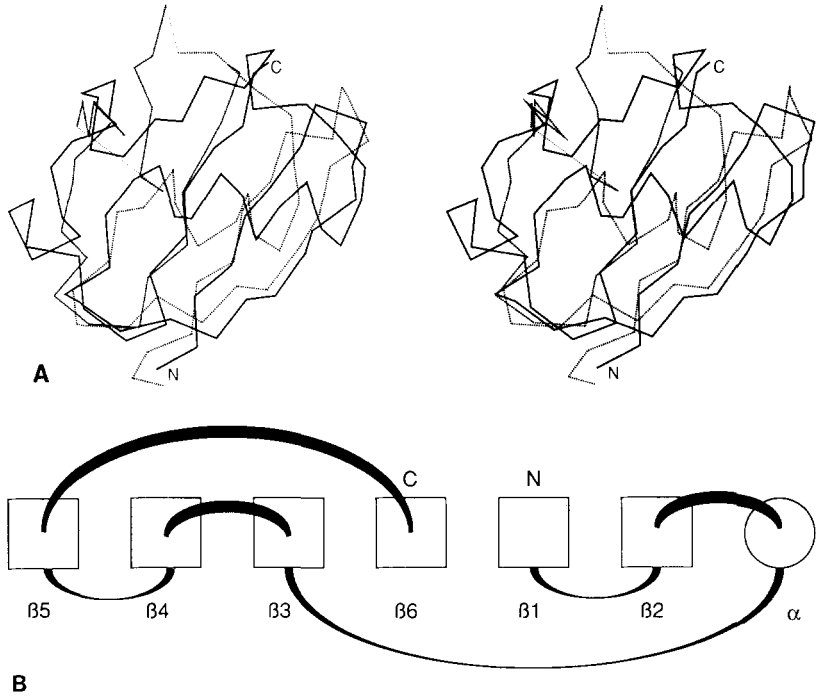

Fig. 6. **A.** Ferredoxin (dashed, 1UBQ) superposed on ubiquitin (solid lines, 3FXC). Stereo view. Two loops in ferredoxin for which no equivalent loops are present in ubiquitin are removed (top left), and replaced by thin dashed lines, for clarity. The superposition is generated by superposing the following fragments, equivalenced by the algorithm (1ubq range / 3fxc range): M1-L8/Y3-E10; T21-A28/N15-E31; G47-S57/G74-T84; E64-V70/D86-H92. **B.** Topology scheme for ferredoxin and ubiquitin. Circle: alpha helix; squares: extended strands. The alpha helix lies across the open hand formed by the beta strands 1,2,3,4, and 6. Strands 2, 1, 6, 3 and 4 form a sheet in which the irregular strand 5 does not participate. The two domains have exactly the same topology.

toreduction of cytochrome *c*. Surprisingly, the three-dimensional structures are remarkably similar. The overall rms deviation of 47 out of maximally 76 equivalenced alpha carbon atoms is 2.1 Å. Both structures can be described as a hand of five beta strands holding a short beta strand and an alpha helix in the center. There is no obvious analogy of protein function and, apparently, the structural similarity had gone undetected. Perhaps ubiquitin and ferredoxin do have a common ancestor. Alternatively, the ferredoxin and ubiquitin "beta-grasp" domain may be an energetically favored folding unit.

CONCLUSION

Our algorithm provides a novel tool for the comparison of protein structures with the options of allowing for altered loop topology and for reversal of chain direction. The entire procedure is fully automatic and can be used in a routine manner. The method is so fast that the comparison of one single structure with all known structures is possible with only a few hours CPU usage on a workstation. Large insertions or deletions or many insertions or deletions are no problem. The method can be used in any context where structural alignment is useful, e.g., to

determine reliable (structure based) sequence alignments, to aid in the definition of structural cores of protein families, and to find common three-dimensional folding units.

The method is implemented as an option in the molecular modeling and drug design program WHAT IF,[28] facilitating immediate visualization by computer graphics. WHAT IF is written in FORTRAN 77, with graphics drivers for Evans and Sutherland and Silicon Graphics computers. The program is available from G.V. for a minimal fee. Send electronic mail to VRIEND@EMBL-Heidelberg.DE on internet for information.

ACKNOWLEDGMENTS

We thank Georg Tuparev, Anna Tramontano, and Ruben Abagyan for very helpful discussions; colleagues in the EMBL Biocomputing groups for providing useful test cases; W.L. Kabsch for use of the program U3B; Evans and Sutherland and Silicon Graphics for technical support; and many crystallographers for depositing their coordinates in the Protein Data Bank.[29]

REFERENCES

1. Ploegman, J.H., Drenth, J., Kalk, K. H., Hol, W.G.J. Structure of bovine liver rhodanese. J. Mol. Biol. 123:557–594, 1978.
2. Johnson, M.S., Sali, A., Blundell, T.L. Phylogenetic relationships from three-dimensional protein structures. Meth. Enzymol. 183:670–690, 1990.
3. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. J. Mol. Biol. 136:225–270, 1980.
4. Godzik, A., Sander, C. Conservation of residue interactions in a family of Ca-binding proteins. Prot Eng. 2:589–596, 1989.
5. Bordo, D., Argos, P. Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. J. Mol. Biol. 211:975–988, 1990.
6. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A 34:8274–828, 1978.
7. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1978.
8. Sali, A., Blundell, T.L. 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, 1990.
9. Zuker, M., Samorjai, R.L., The alignment of protein structures in three dimensions. Bull. Math. Biol. 51:55–78, 1989.
10. Levine, M., Stuart, D., Williams, J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. Acta Cryst. A40:600–610, 1984.
11. Remington, S.J., Matthews, B.W. A systematic approach to the comparison of protein structures. J. Mol. Biol. 140: 77–99, 1980.
12. Remington, S.J., Matthews, B.W. A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. Proc. Natl. Acad. Sci 175: 2180–2184, 1978.
13. Lesk, A.M. Detection of 3-D patterns of atoms in chemical structures. Comm. ACM 22:219–224, 1979.
14. Brint, A.T., Davies, H.M., Mitchell, E.M., Willet, P. Rapid geometric searches in protein structures. J. Mol. Graph. 7:48–53, 1989.
15. Barnton, G.J., Sternberg, M.J.E. LOPAL and SCAMP: techniques for the comparison and display of protein structures. J. Mol. Graph. 6:190–196, 1988.
16. Rossmann, M.G., Argos, P. Exploring structural homology of proteins. J. Mol. Biol. 105:75–95, 1976.
17. Rao, S.T., Rossmann, M.G. Comparison of super-secondary structures in proteins. J. Mol. Biol. 76:241–256, 1973.
18. Abagyan, R.A., Maiorov, V.N. A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. J. Biomol. Struc. Dyn. 5:1267–1279, 1988.
19. Abagyan, R.A., Maiorov, V.N. An automatic search for similar spatial arrangements of alpha helices and beta strands in globular proteins. J. Biomol. Struc. Dyn. 6: 1045–1060, 1989.
20. Jones, T.A., Thirup, S. Using known fragments in protein model building and crystallography. EMBO J. 5:819–822, 1986.
21. Kabsch, W., Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22:2577–2637, 1983.
22. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. 175:159–174, 1984.
23. Guss, J.M., Freeman, H.C., Structure of oxidized popular plastocyanin at 1.6 Å resolution. J. Mol. Biol. 169:521–563, 1983.
24. Baker, E.N., Structure of azurin from alcaligenes denitrificans. Refinement at 1.8 Å resolution and comparison of the two crystallographically independent molecules. J. Mol. Biol. 203:1071–1095, 1988.
25. Adman, E.T. Metalloproteins. P.M. Harrison, ed. Verlag Chemie Weinheim, 1985, Part 1, chapter 1, pp. 1–42.
26. Vijay-Kumar, S., Bugg, C.E., Cook, W.J. Structure of ubiquitin refined at 1.8 Å resolution, J. Mol. Biol. 194:531–544, 198.
27. Tsukihara, T., Fukuyama, K. Nakamura, M., Katsube, Y., Tanaka, N., Kakudo, M., Wada, K., Hase, T., Matsubara, T., Structure of a (2Fe-2S) ferredoxin from spirulina platensis. Main chain fold and location of side chains at 2.5 Å resolution. J. Biochem. (Tokyo) 90:1763–1773, 1981.
28. Vriend, G. WHAT IF: A Molecular Modeling and Drug Design Program. J. Mol. Graphics 8:52–56, 1990.
29. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. J. Mol. Biol. 112:535–542, 1977.