LOPAL and SCAMP: techniques for the comparison and display of protein structures

Geoffrey J. Barton* and Michael J. E. Sternberg

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet St., London WC1E 7HX, UK

This paper describes two computer programs designed to assist in the comparison of protein structures. LOPAL (LOoP ALignment) applies a dynamic programming algorithm to the comparison of regions of protein three-dimensional (3D) structure and gives a similarity score and suggested sequence alignment with that score. SCAMP (Structure Comparison and Alignment of Multiple Proteins) is an interactive graphics program for the Evans and Sutherland PS300 graphics terminal that allows the simultaneous display, manipulation and pairwise least-squares fitting of up to nine independent structures. Together, LOPAL and SCAMP provide an integrated system for characterizing structural similarities in proteins with the aim of improving the accuracy of predicted protein structures. An application of these programs to loop regions in the immunoglobulin constant domains is illustrated.

Keywords: *dynamic programming, PS300 graphics program, protein structure, least-squares fitting, automatic structure comparison, immunoglobulin*

Received 14 July 1988
Accepted 2 August 1988

INTRODUCTION

The applicability of an item of computer hardware to a particular problem is ultimately governed by the software available for the device. In the field of molecular graphics, a number of highly developed software packages can exploit the interactive display potential of the Evans and Sutherland PS300 family. For example, FRODO,[1] which was originally developed to aid in the determination of new protein structures by X-ray crystallography, has in its more recent PS300 implementations become a very flexible general display tool for proteins, allowing most of the conceivable coloring and selection options to be explored (e.g., color by atom, residue type, residue range or B-value). In keeping with its crystallographic heritage, FRODO also allows the manipulation of fragments of a protein, substitution of side-chain types, and rotation about torsion angles. HYDRA,[2] with flexibility implicit in its modular design, also allows for nearly all the types of display and manipulation commonly found useful to be performed. Chemical Design Ltd.'s widely used commercial package CHEM–X now supplies features required by both the protein modeler and the inorganic chemist. Evans and Sutherland's own MOGLI program can fully exploit the capabilities of the PS300; however, to allow manipulation of several large protein molecules, more than the minimum 1 megabyte of mass memory is required.

Although these programs can between them satisfy all the display requirements of the structural chemist or molecular biologist, their very flexibility can lead to a sacrifice in ease and speed of interactive use for any one particular function. In addition, the ability to perform pairwise least-squares fitting on the displayed structures, a feature vital for comparing protein structures, is available in some of these programs (e.g., HYDRA, CHEM–X), although it is not central to their design philosophy. As a consequence, successfully fitting a series of objects and maintaining their display can be difficult or impossible. The program FITZ,[3] which allows up to four independent objects to be manipulated and fitted pairwise, meets some of the requirements. Unfortunately, this program was written for the Evans and Sutherland PS2 with a PDP 11/60 host computer, and it is somewhat limited in the style of display and ease of selection possible; furthermore, no PS300 version is currently available.

The program SCAMP,[4] described here, was written to facilitate the comparison of similar protein structures by allowing a number of independent objects to be manipulated and also to allow pairwise fitting by a least-squares procedure. Although SCAMP does not offer the wide functionality of FRODO, HYDRA and CHEM–X, it does allow the combination of different objects currently displayed to be readily and quickly modified.

*To whom all communications should be addressed at: Biomedical Computing Unit, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK

The program LOPAL[4] was developed as a tool to assist in classifying common strutures in protein loop regions. It applies the Needleman and Wunsch[5] sequence alignment algorithm to the comparison of three-dimensional (3D) structures, thus giving a score for the overall similarity of the two structures even when they are of different lengths. Equivalent residues are also suggested, and these can be displayed using SCAMP. An example application is presented for loops in the immunoglobulin constant domains.

INTRODUCTION TO LOPAL

Structure-based alignments are now often used as a more comprehensive basis for comparative model-building studies than using a single homologous structure.[6,7] In such studies, much is to be gained from a detailed knowledge of the homologous loop conformations. Rules derived for a particular loop can be usefully applied when model-building to establish the most likely conformation for the corresponding region of chain in the model.[6] There may be several members of a protein family for which the 3D structures have been solved by X-ray crystallography. If these are closely similar in overall chain fold, then it is relatively simple through a consideration of hydrogen bonding patterns to locate a sufficient number of equivalent residues within the homologous secondary structures to allow the molecules to be superimposed. The superposition might be based on one member of the family using SCAMP, as illustrated for the immunoglobulin constant domains in Color Plate 1, or possibly on a consensus "framework" by applying the techniques developed by Sutcliffe *et al.*[7,8]

Hydrogen bonding patterns in conjunction with superimposed structures can suggest where the corresponding secondary structures (β-sheets and α-helices) begin and end. Given this information, the loop regions that join the regular secondary structures can be defined and examined in greater detail by using molecular graphics. Rules relating residue types, lengths of loops and insertions/deletions can then be developed for each homologous loop. These detailed data on loop conformations can aid in model-building another member of the protein family.

Although SCAMP (described below) was specifically designed to assist in this type of study, the number of loop conformations that must be considered quickly becomes unmanageable as the number of proteins increases. Thus, for six proteins there are 15 pairwise comparisons to perform for each loop. LOPAL provides a systematic screening procedure that leads to an initial ranking of similarity between unequal-length segments of polypeptide chain.

AUTOMATIC METHODS FOR THE COMPARISON OF PROTEIN 3D STRUCTURES

Automated methods of comparing protein structure in three dimensions have aimed at locating and quantifying significant similarities between whole proteins or domains, usually with a view of inferring evolutionary relationships.

Rossmann and Argos[9-11] developed an elaborate procedure for identifying the structural and topological similarities between two proteins. They defined a probability function P_{ij} that had two factors; the first related the spatial proximity of residues i and j, while the second indicated the relative orientation of successive residues. The relative contribution of the two factors could be altered by adjusting weighting factors, thus allowing P_{ij} to reflect similar topologies, similar spatial equivalences or both. For any given orientation of the two molecules, the total number of equivalent residues was determined by first locating tentative equivalences as the highest P_{ij} values in the P_{ij} matrix, then extending the equivalences for increasing i,j. In order to avoid the need to provide an initial fit of the two proteins, one protein was systematically rotated through the three Eulerian angles. At each increment, (1) the number N of sequential equivalences was determined, (2) a linear least-squares procedure was used to locate the translation vector required to superimpose the equivalenced residues, (3) stages 1 and 2 were repeated, and (4) the final number of equivalences was recorded at the grid point corresponding to the current values of the Eulerian angles. Once complete, the significance of the largest values on the grid was estimated by comparing them to the background of values corresponding to random orientations. Rossmann and Argos applied this technique to many protein systems, including the identification of previously observed common structural features in lactate and glyceraldehyde-3-phosphate dehydrogenase, as well as suggesting weaker structural equivalences between hen egg white and phage lysozyme.

Remington and Matthews[12] took an alternative approach to determining structural similarity. First, they divided the two proteins into a set of overlapping segments of predetermined length. All pairs of chain segments were then fitted by a least-squares procedure, and the distribution of root mean square (r.m.s) deviations was plotted as a contoured comparison matrix. The statistical significance of the observed low r.m.s. values was assessed by looking for deviations from the normal distribution on cumulative probability plots (c.f. Fitch[13]). This technique located the expected significant similarities between hemoglobin and myoglobin, while a 60 residue similarity between T4 lysozyme and carp-calcium binding protein appeared not to be statistically significant. McLachlan[14] developed a faster fitting algorithm but applied the same segment comparison technique to the domains of chymotrypsin. However, since the distribution of r.m.s. values is not always normal, McLachlan used an empirical measure of significance obtained from the comparison of sample proteins representative of the main structural classes. His study suggested that the observed similarities between the domains of chymotrypsin were highly significant and could have resulted from gene duplication.

While these techniques can help to identify significant structural similarities between complete proteins, they are not ideal for comparing protein loops where the

aim is to identify a set of residues that are performing equivalent roles. Loops may be of *different lengths*, and a method that explicitly copes with this feature is required. In order to meet this requirement, the Needleman and Wunsch[5] algorithm that was originally developed for the comparison of amino-acid sequences was adapted to the comparison of 3D objects.

NEEDLEMAN AND WUNSCH ALGORITHM

This technique belongs to the class of *dynamic programming* algorithms (reviewed by Sankoff and Kruskal, 1983)[15], which produce a score for the best alignment of two sequences, given a scoring function for the alignment of each type of sequence element and a penalty function for the insertion of gaps. Briefly:

(1) Two *sequences* are defined as A, B with lengths m, n, respectively, A_i denotes the ith element of A, while the partial sequence $A^i = (A_i, A_{i+1}, A_{i+2}, \ldots, A_m)$.
(2) A matrix $R_{m,n}$ is generated in which each element, $R_{i,j}$ represents the score for A_i versus B_j.
(3) $R_{m,n}$ is acted on to generate $S_{m,n}$ where each element $S_{i,j}$ holds the maximum score for an alignment of A^i with B^j in which A_i is aligned with B_j and not with a gap.
(4) The matrix R can be used to obtain the best score and an alignment of the sequences. (For further details of this algorithm, see Reference 5).

It is important to bear in mind that although the Needleman and Wunsch algorithm was developed for the comparison of amino-acid sequences, it can be applied generally to the comparison of *any* two sequences for which the alignment of two sequence elements can be represented by a similarity scoring scheme. (See Reference 15 for further applications of dynamic programming.)

By applying this principle, LOPAL operates on protein Cα atoms expressed as *sequences* of x, y, z coordinates. The stages involved are summarized below and shown in Figure 1.

(1) The protein structures to be compared are first superimposed on regions of strong structural similarity; for example, the core α-helices or β-strands.
(2) A distance matrix D, where each element $D_{i,j}$ represents the distance between the $C\alpha$ atom of residue A_i and the $C\alpha$ atom of residue B_j, is constructed for the comparison of the segments of interest.
(3) The distance matrix is converted linearly to a similarity scale by subtracting all values from the largest distance in the matrix.
(4) The Needleman and Wunsch algorithm is applied to this similarity matrix to yield a best alignment of the two structures (no gap penalty was found necessary).
(5) The alignment calculated in (4) is stored as a list of vectors joining the equivalenced Cα atoms for subsequent display by SCAMP.

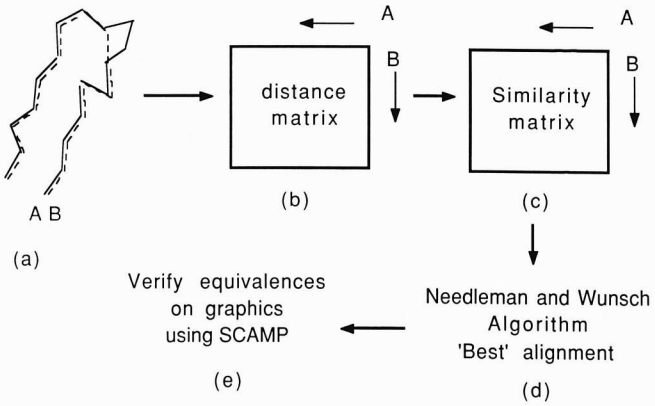


Figure 1. Sequence of operations performed by LOPAL. (a) LOPAL is supplied with two regions of protein structure that have been superimposed on the conserved cores. (b) A distance matrix is calculated. (c) The distance matrix is converted to a similarity matrix by subtracting all values from the largest entry in the matrix. (d) The Needleman and Wunsch algorithm is then applied to the similarity matrix yielding a "best" alignment and overall similarity score. (e) Residues suggested as equivalent can then be inspected using the graphics program SCAMP

This technique results in a single alignment that maximizes the similarity (i.e., minimizes the distance) between the two structures even when the segments are of different lengths. The score for the comparison gives a quantitative indication of how similar the two structures are to each other, while the alignment that may be inspected by using SCAMP can provide a starting point for further characterization.

LOPAL and SCAMP are currently being used as part of a strategy to systematically establish equivalent residues across *all* members of protein families for which X-ray structures are available. The results of this study will be described in detail elsewhere and form part of a "feature table" in the Birkbeck/Leeds relational database of protein structure. (S. Gardner, personal communication.) In order to illustrate the utility of LOPAL, an application to the A–B loops in the immunoglobulin constant domains is described here.

IMMUNOGLOBULIN CONSTANT DOMAINS

The immunoglobulin constant domain structure consists of one four- and one three-stranded antiparallel β-sheet that pack against each other and are linked by a single disulphide bridge. For the purposes of this study, the β-strands were labeled sequentially A–G, while start and end points of the loops were determined by reference to hydrogen bonding diagrams for each domain. Where one domain has a shorter β-strand, the start and end points of corresponding loops in all other domains were adjusted to give equal strand lengths. This procedure led to the definition of A–B loops from six constant domains shown in Table 1.

Four loops have 12 residues (MCPABL, FB4ABL, FB4ABH and FCCH3AB), one has 14 (FCCH2AB) and one has 11 residues (MCPABH). The core A and B

Table 1. Immunoglobulin constant domain A–B loops

Number	Brookhaven protein code	Residues	Loop code	Length (residues)
1	1MCP	L125 to L136	MCPABL	12
2	1MCP	H134 to H144	MCPABH	11
3	1FB4	H121 to H132	FB4ABL	12
4	1FB4	H130 to H142	FB4ABH	12
5	1FC1	A244 to A257	FCCH2AB	14
6	1FC1	A352 to A363	FCCH3AB	12

Table 2. Constant domain A–B loop comparisons ordered on distance

Comparison Loop X	Loop Y	Number of equivalences	Mean distance ±1.0 S.D. (Ångstroms)	
1	3	12	0.8	±0.4
1	6	12	1.4	±0.6
3	5	12	1.7	±0.9
1	5	12	1.7	±0.9
3	6	12	1.8	±0.7
5	6	12	2.4	±1.2
1	4	12	2.4	±1.7
3	4	12	2.7	±2.0
4	5	12	2.7	±1.7
4	6	12	2.9	±1.4
2	3	11	2.9	±1.1
2	6	11	3.1	±2.0
1	2	11	3.1	±1.4
2	5	11	3.4	±1.3
2	4	11	3.5	±1.8

β-strands (12 residues) for structures 2 to 6 were fitted to structure 1 using SCAMP on the basis of the 48 main-chain atoms. This fit, including the loop regions, is illustrated in Color Plate 2. Clearly there are some loops that are relatively similar to each other, and one in particular (2, MCPABH) that is quite different in conformation.

The equivalencing procedure was applied to the fitted loops using LOPAL and the mean and standard deviation of the distances between the equivalenced residues obtained was calculated for each comparison. Table 2 shows the results of the 15 pairwise comparisons in rank order of similarity, and Figure 2 illustrates a dendrogram produced from these data by applying single linkage cluster analysis.

The distances shown in Table 2 and relationships suggested by Figure 2 are supported by inspecting the loops using molecular graphics. The 12-residue loops 1, 3 and 6 form a cluster of similar structures. The 14-residue loop 5 also belongs to this group, since, although it is two residues longer, it has a helical region in common with the other three loops. Loop 4, although it is also 12 residues long, is of a quite different conformation from 6, 1, 3 and 5, as indicated by Color Plate 2 and the complete pairwise data of Table 2. Similarly, the 11-residue loop, MCPABH, which has a clearly different conformation to all five other loops, is also shown as least similar by the LOPAL comparison data. It is therefore possible using this technique to obtain

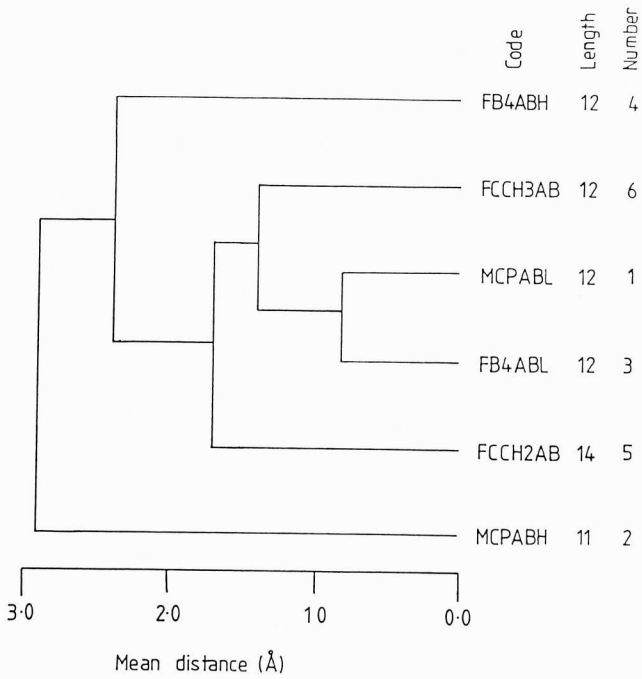


Figure 2. Dendrogram for A–B loops. Dendrogram resulting from the application of single linkage cluster analysis to the pairwise mean distances obtained from the application of LOPAL to A–B loops from six immunoglobulin constant domains. Code: loop codes defined in Table 1; Length: number of residues in loop; Number: loop identifying number from table 1

an objective indication of which loops are similar in conformation, even *when their lengths are different.* Furthermore, the equivalences suggested by LOPAL can be displayed to guide further analysis of the loop conformations.

Color Plate 3 illustrates the equivalence suggested for two loops deemed to be very similar (1 and 6), while Color Plate 4 shows the equivalences suggested for the pair of structures considered least similar (2 and 4). The alignment suggested for loops 2 and 4 indicates that Gly H138 (loop 4) is a single insertion. However, since the two loops are so different in overall conformation, a futher analysis that considered the positions and relative roles of side chains would be necessary before drawing any firm conclusions.

FURTHER DEVELOPMENTS TO THE ALIGNMENT ALGORITHM

The comparison method described here was initially developed for the study of short segments of polypeptide chain; however, larger structures are also amenable to the technique. Preliminary studies on the immunoglobulin domains and globins suggest that starting from an initially good fit, a large proportion of the regions identified as equivalent by detailed studies[16,17] can correctly be aligned automatically.

A potential drawback in applying the current implementation of LOPAL to the comparison of complete proteins is the need for the two proteins to be superimposed initially. However, the generality of the

Needleman and Wunsch algorithm has allowed several orientation-independent approaches to be investigated. By comparing *sequences* of Cα-Cα virtual bond angles, or φ/Ψ angles, a measure of the similarity of chain fold can be obtained. A more sophisticated approach involves the comparison of two *intra*molecular distance matrices,[18] and preliminary studies (A. Sali, personal communication) suggest that this method can yield alignments that agree well with expert structural comparisons.[16] A further refinement would be to combine similarity functions based on structural data (e.g., distance matrices, Cα angles) and amino-acid sequence data (e.g., comparison using Dayhoff's matrix), giving a scoring system that reflected both structural and amino-acid similarities.

The Needleman and Wunsch algorithm provides an elegant solution to the problem of aligning proteins on the basis of their 3D structures. However, it is important to remember that the method yields a best *global* alignment. If there are few or no regions of strong similarity, then, in common with amino-acid *sequence* alignments by this method,[19] the alignment is likely to exhibit few correctly equivalenced regions. LOPAL is therefore most useful when used to obtain an objective alignment of proteins with clearly similar chain folds, or for characterizing short variable regions bounded by structures of high similarity (e.g., loops). Bearing these limitations in mind, a logical extension to LOPAL would be to allow multiple protein alignment by adapting the algorithms of Barton and Sternberg[20] or Feng and Doolittle.[21] This would allow more than two structural segments to be simultaneously aligned. However, the more general problem of identifying common substructures in proteins that are not clearly of the same fold is probably better suited to the established Remington and Matthews[12] or Rossmann and Argos[10] approaches, which do not seek to identify a global best alignment.

INTRODUCTION TO SCAMP

The initial impetus for developing SCAMP was to facilitate the detailed study of loop regions that interconnect regular secondary structural elements. When writing the program, the aim was to produce a system that would be an easy-to-use and *fast* interactive tool for comparing protein structures. The emphasis on speed is important, since when studying many similar structures considerable frustration can result if it takes more than a few seconds to alter the display — for example, from showing structures A, B, C to showing A, D, E. This is particularly true if all pairwise comparisons of a group are to be considered, as one might want to compare A with B, C, D, E in turn, then combine A and D to look at these versus C, and so on. In order to obtain sufficient speed of interaction, the program was designed to allow as much of the data as possible to reside on the PS300 and be controlled by the PS300 resident program rather than being repeatedly loaded from the host computer. However, due to restrictions on PS300 memory, this approach limits the complexity of both the display tree and function networks that can be used. As a result, SCAMP sacrifices the flexibility of coloring at the atom level, a feature of most other molecular graphics programs.

In order to perform quantitative and qualitative comparisons of sections of polypeptide chain, it is important to be able to superimpose quickly the regions of interest. For this reason, the ability to easily superimpose and obtain values for the r.m.s. deviation between two objects was a central consideration when writing SCAMP. The superposition or *least-squares fitting* problem has been considered by several authors (e.g., References 14, 22, 23). Given two molecules, A and B, the fit is calculated by first translating B so that its centroid coincides with that of A, then determining the orthogonal rotation matrix R that when applied to B minimizes:

$$\left(\sum_{i=1}^{n}(A_i - B_i)^2/n\right)^{\frac{1}{2}}$$

where A_i and B_i are the coordinates of the ith atoms in A and B, respectively. The routines to perform this operation in SCAMP implement McLachlan's method[14] and have been found to be both fast and reliable.

DESCRIPTION OF SCAMP FUNCTIONS

Up to nine independent objects can be displayed, and each object can be independently rotated, translated and colored and have residue names and numbers associated with it. The smallest object that can be displayed is a single amino-acid residue. In addition, each object has an attribute associated with it. This may take the form of another protein representation, a molecular surface as calculated by Connolly's program[24] or information about equivalent residues as supplied by the program LOPAL. Nine objects were chosen, since this allowed for a compact function network control structure. Nine of the 12 PS300 function keys are devoted to switching the dials between the transformation of each object individually. One key switches the dials to global transformations, and the remaining two keys switch two of the dials to allow the color and saturation of each object and its attribute to be changed.

On-Screen Menu

An on-screen menu lets the user toggle the display of an object, its attribute, residue numbers and residue names. The menu also allows any individual object to be reset to its original orientation and serves as the command center for the host program.

NEW initiates the loading of a new object or attribute that can be a protein structure in Brookhaven format, a join file from LOPAL or a Connolly surface file. If a protein structure is being selected, then either the whole structure or sections can be displayed. All atoms, Cα and main-chain-only representations are allowed. For speed of operation, connectivities are generated at the residue level, initially by reference to a lookup table for the standard amino acids. If the amino acid is unknown or has a nonstandard number of atoms, con-

nectivity is established by calculation on a bonding radius of 1.8 Å. Objects and attributes can be displayed using continuous or dashed lines as bonds (see Color Plate 5).

OLD allows a series of commands to be read from a previously written file.

FIT initiates the least-squares fitting procedures. The program asks which objects are to be fitted, whether on all, Cα or main-chain atoms and whether the whole molecule or only specified residues are to be used for the fit. The program consults the disk data files to obtain the coordinates used for fitting so that objects can be fitted on the basis of residues that have not been selected for display. For example, loops from a protein might be displayed, but fitting performed on the core secondary structures. The rotation and translation resulting from application of the least-squares procedure[14] is sent directly to the appropriate nodes in the PS300 display tree, thus instantly updating the current display.

CHAT gains access to a number of additional commands:

- PLOT allows a file containing the most recently generated vector list to be written. This can then be used via another program to produce stereo pairs on an x,y plotter.
- HISTORY toggles the writing of a file that records all commands given to the host computer either from the keyboard or the PS300. This file can be used later to reload the structures of interest via the OLD option.
- RX, RY, RZ and TR allow rotations and translations to be performed on the coordinates when no PS300 is available. These commands are useful for the production of plot files and transformed coordinates for further processing.
- SPAWN allows a subprocess to be generated under the VAX/VMS operating system so the user can inspect/edit files or run programs related to the graphics application.

In order to make user interaction easy, all commands entered at the keyboard are in free format, and error handling is performed so that if inappropriate data is supplied, the program returns to the central menu.

The SCAMP function networks and display tree consist of 2 000 lines of PS300 command language code that occupies 364 000 bytes of mass memory when loaded into the PS300. The FORTRAN host program is 3 500 lines long and makes use of a 3 000-line general-purpose subroutine library.

EXAMPLE DISPLAYS

The main advantages of SCAMP over other programs is its speed and ease of use when more than two objects are simultaneously required, and the ability to perform least-squares fitting on any pair of up to *nine* loaded objects. It is naturally difficult to convey the interactive advantages of SCAMP in the form of static pictures; however, the Color Plates discussed below illustrate the general display capabilities of the program.

Color Plate 5 shows the display of the main chain of Ribosomal protein L30 from *B. Stearothermophilus* (Mol__8) with the most highly conserved positions (conservation number $\geqslant 0.7$; see Zvelèbil *et al.* (1987)[25] for a description of conservation numbers) when aligned with L30 from *E. coli* displayed in all atom representation using dashed lines (Mol__7).

Color Plate 6 illustrates a main-chain representation of the CH1 domain from Immunoglobulin Kol (Brookhaven[26] code 1FB4). The β-strands are displayed as object 2; names and numbers for all residues have been loaded and selected for display. The loop regions of the domain have been loaded as the attribute of object 2 and are currently displayed in a different color.

Color Plate 1 demonstrates the fitting capability of the program. Six immunoglobulin constant domains are shown (Mol__1 to Mol__6) in a main-chain representation. Mol__2 to Mol__6 were each fitted pairwise to Mol__1 (FB4 CL) using the main-chain atoms from six residues in the cores of the proteins, including the two conserved Cys residues. It is clear from this Color Plate that the correspondence between the six structures is greatest in the β-strand regions and that there can be considerable variation in the loop regions.

SUMMARY

SCAMP was written with the display and study of homologous structures specifically in mind. However, the ability to have a large number of independent objects resident in memory is useful for more general graphic display. The structures loaded can be rapidly turned on and off or changed in color to highlight different points of interest simply by toggling a *single* menu item or turning a dial and without the need to reload or reissue commands from the host computer. For example, the six immunoglobulin domains, as well as the L30 main chain and conserved residues shown by Color Plates 1 and 5, were simultaneously present in the PS300 memory, and the same principle can also be applied to subsections of a single molecule. This feature can save considerable time when preparing illustrative photographs.

LOPAL introduces a novel application of the Needleman and Wunsch sequence-comparison algorithm[5] to the characterization of similarities in protein 3D structures. LOPAL has been used extensively to save time when studying loops in homologous proteins, since it gives a rapid first approximation to the "correct" alignment. Although LOPAL is useful in its current form, further developments to allow the comparison of intramolecular distance matrices and allow multiple alignment will further extend its utility.

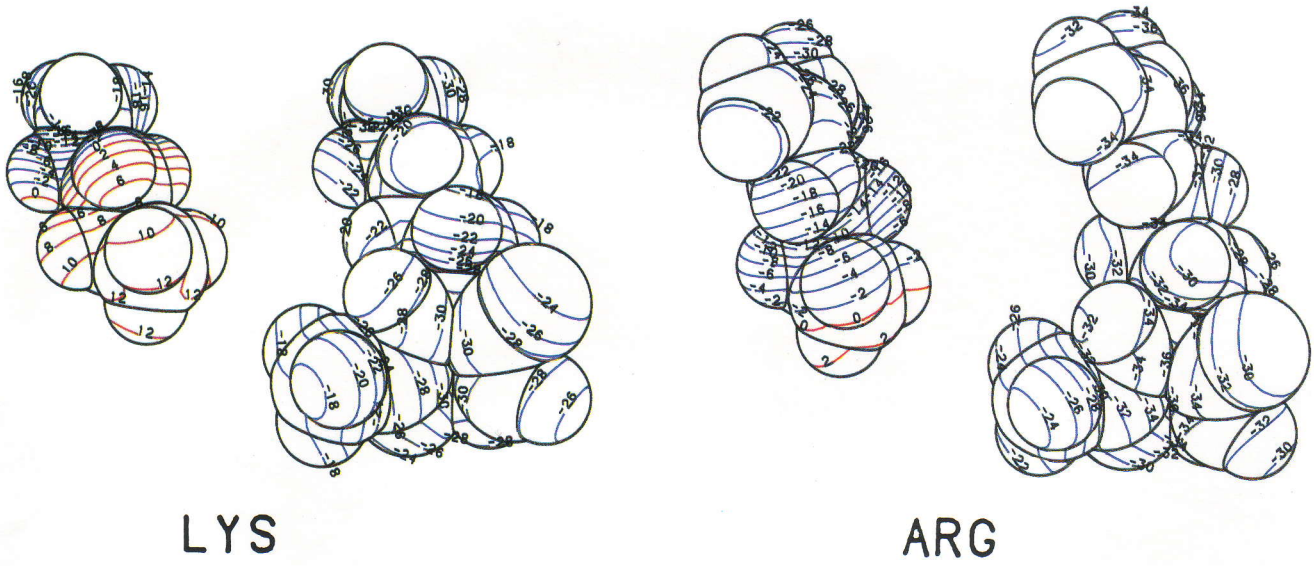
ACKNOWLEDGEMENTS

We would like to thank Dr. M. J. J. Zvelebil, Mr. S. Gardner and Mr. A. Sali for helpful discussions, Dr. S. White for supplying the L30 coordinates prior to publication, Dr. I. Tickle for the use of his program PSPARS and Professor T. L. Blundell. This work was

supported by the Science and Engineering Research Council, UK.

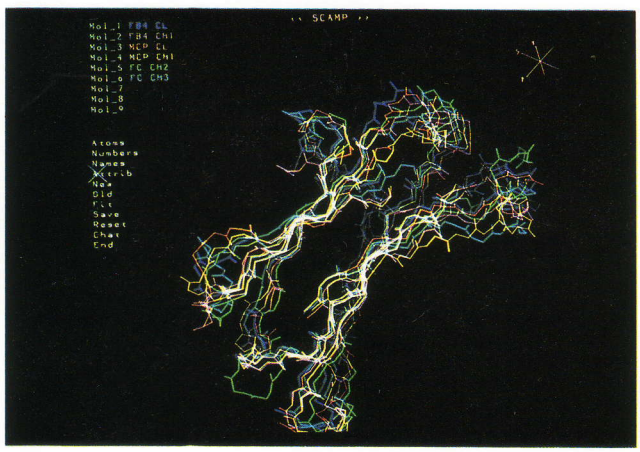
REFERENCES

1 Jones, T. A. *J. Appl. Crystallogr.* 1978, **11**, 268–272
2 Hubbard, R. E., in *Current Communications in Molecular Biology: Computer Graphics and Model Building,* Ed. Fletterick, R., and Zoller, M., 1986, Cold Spring Harbor Laboratory, 9–11
3 Taylor, G. *J. Mol. Graph.* 1983, **1**, 5–8
4 Barton, G. J. Computer analysis of protein sequence and structure, in PhD. Thesis, University of London, 1987, 223–257
5 Needleman, S. B., and Wunsch, C. D. A general method applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins *J. Mol. Biol.* 1970, **48**, 443–453
6 Greer, J. Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* 1981, **153**, 1027–1042
7 Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Prot. Eng.* 1987, **1**, 377–384
8 Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L. Knowledge based modelling of homologous proteins, part II: rules for the conformations of substituted sidechains. *Prot. Eng.* 1987, **1**, 385–392
9 Rossmann, M. G., and Argos, P. A comparison of the heme binding pocket in globins and cytochrome b5. *J. Biol. Chem.* 1975, **250**, 7525–7532
10 Rossmann, M. G., and Argos, P. Exploring structural homology of proteins *J. Mol. Biol.* 1976, **105**, 75–95
11 Rossmann, M. G., and Argos, P. Chemical and biological evolution of a nucleotide-binding protein. *J. Mol. Biol.* 1977, **109**, 99–129
12 Remington, S. J., and Matthews, B. W. A general method to assess similarity of protein structures, with application to T4 bacteriophage lysozyme. *Proc. Natl. Acad. Sci. USA* 1978, **75**, 2180–2184
13 Fitch, W. M. An improved method of testing for evolutionary homology. *J. Mol. Biol.* 1966, **16**, 9–16
14 McLachlan, A. D. Gene duplication in the structural evolution of chymotrypsin. *J. Mol. Biol.* 1979, **128**, 49–79
15 Sankoff, D., and Kruskal, J. B. (eds.). *Time Warps String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, 1983
16 Lesk, A. M., and Chothia, C. *J. Mol. Biol.* 1980, **136**, 225–270
17 Lesk, A. M., and Chothia, C. *J. Mol. Biol.* 1982, **160**, 325–342
18 Liebman, M. N. Correlation of structure and function in biologically active small molecules and macromolecules by distance matrix partitioning, in *Molecular Structure and Biological Activity,* Ed. Griffin, J. F., and Duax, W. L. Elsevier Science Publishing Co., 1982, 193–212
19 Barton, G. J., and Sternberg, M. J. E. Evaluation and improvements in the automatic alignment of protein sequences. *Prot. Eng.* 1987, **1**, 89–94
20 Barton, G. J., and Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.* 1987, **198**, 327–337
21 Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* 1987, **25**, 351–360
22 Ferro, D. R., and Hermanns, J. *Acta. Crystallogr.* 1977, **A33**, 345–347
23 Kabsch, W. *Acta Crystallogr.* 1978, **A32**, 922–923
24 Connolly, M. L. *Appl. Cryst.* 1983, **16**, 548–558
25 Zvelebil, M. J. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* 1987, **195**, 957–961
26 Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr, Meyer, D. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977 **112**, 535–542

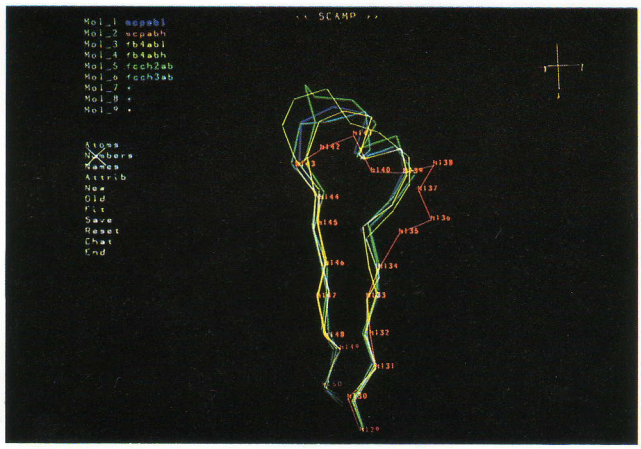


G.J. Barton and M.J. Sternberg

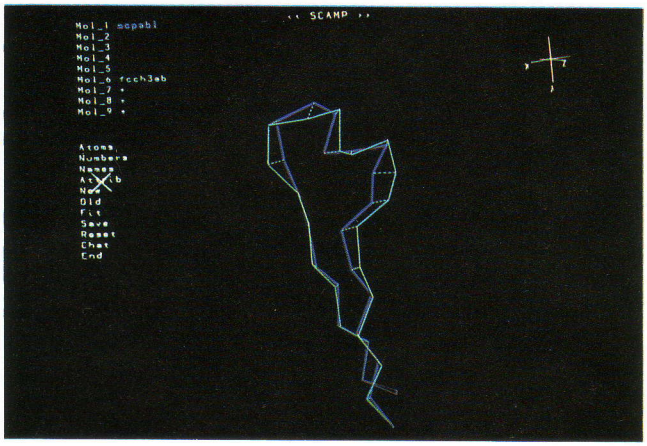
LOPAL and SCAMP: techniques for the comparison and display of protein structures



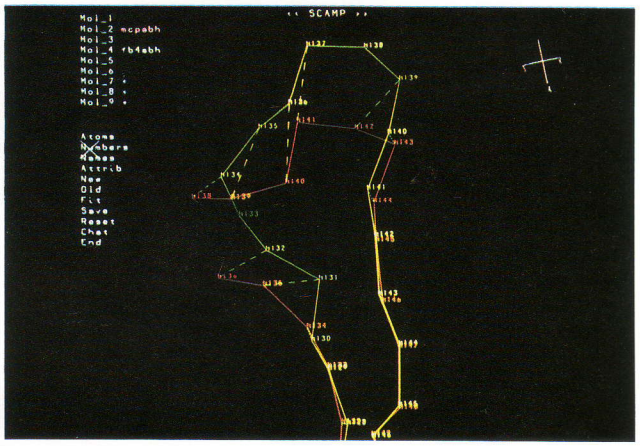
Color Plate 1. Six immunoglobulin constant domains displayed in mainchain only representation (Mol__1 to Mol__6) using SCAMP after pairwise least squares fitting to Mol__1 (see text)



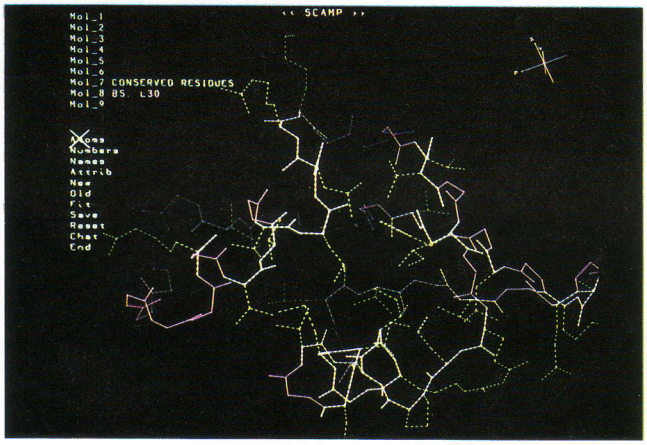
Color Plate 2. A-B loops from six immunoglobulin constant domains (see text for definition) after pairwise least squares fitting to Mol__1 on the core β-strand residues



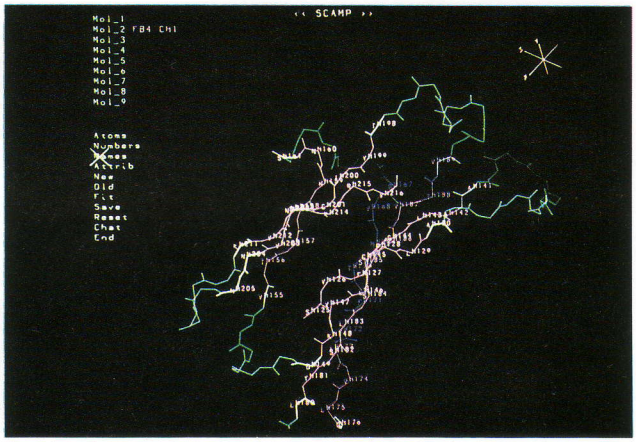
Color Plate 3. Equivalent residues suggested by LOPAL for two A-B loops automatically identified as very similar (MCPABL and FCCH3AB). The suggested equivalent residues are shown joined by dotted lines



Color Plate 4. Equivalent residues suggested by LOPAL for two A-B loops that are identified as very dissimilar (see text for further details)



Color Plate 5. Illustration of SCAMP display combinations. The complete mainchain for ribosomal protein L30 is shown in solid purple lines as Mol__8. The most highly conserved amino acids are shown in all atom representation with dashed lines in green (Mol__7). The *Objects* Mol__7 and Mol__8 may be independently rotated, translated or fitted



Color Plate 6. Illustration of SCAMP display including *Objects* and *Attributes* (see text for definitions). The β-strands from the CH1 domain of immunoglobulin Kol in mainchain only representation form the Mol__2 *Object* (purple). The residue numbers and names for this *Object* have been turned on by picking the NUMBERS and NAMES options. The loop regions (green) are displayed as the *Attribute* of Mol__2