sequence database. We have applied a preliminary version of the R3P method in combination with the 3D profile method for the prediction of several sequences to be folded as β/α barrels by sequence database searches with a number of β/α barrel profiles.[6]

Program Use and Availability

As described above there are a large number of ways to generate a profile. The method one uses depends on the application. For fold identification, there is no absolute favorite. No profile works best in all cases, but best results are generally obtained with continuous profiles or an R3P profile. For structure verification, the clear choice is continuous profiles. For alignment of an identified structure with the sequence, the clear choice is the R3P method. The following programs are available from the authors.

ENVIRON: Calculates area buried and fraction polar for each residue in a structure

3D_PROFILER: Reads the output file from ENVIRON and generates a discrete profile

PROFGEN: Reads the output file from ENVIRON and generates a continuous profile of the structure

MAKER3P: Produces an R3P profile directly from a coordinate file

ALIGNR3P: Produces a sequence–R3P profile alignment by an iterative refinement procedure

PROFILESEARCH: Determines optimal alignment scores for a database of sequences with a 3D profile; the alignment scores are converted to Z scores, and the output is a sorted list of sequences and alignment scores

VERIFY_3D: Determines the overall 3D profile score for a structure and the average score in regions of sequence along the structure

[36] SSAP: Sequential Structure Alignment Program for Protein Structure Comparison

By CHRISTINE A. ORENGO and WILLIAM R. TAYLOR

Introduction

Since the 1970s, protein structure comparison methods have become increasingly sophisticated. Early rigid-body techniques[1] have been used to study different mutant and ligand-bound forms. They are fast and extremely efficient for superposing very similar structures, but as structures diverge these methods cannot always identify equivalent positions because of insertions and deletions (indels).

A major incentive to developing more robust methods has been the need to analyze protein fold families, extracting information that can improve structure prediction and modeling. Because there are nearly 30 times more known sequences than structures, this is an important consideration. During evolution, the sequence of a protein may change, but the overall fold is much more conserved, remaining the same even if 70% of the sequence changes.[2,3] This gives rise to families of related structures and means that a new structure can be modeled on a known one if the proteins have similar sequences.

Analyses of protein families can help in protein structure modeling by setting tolerances on variability at different positions in the fold. Similarly, for structure prediction, information from protein families can improve template or profile-based methods by incorporating residue preferences in specific structural locations in the fold.[4,5]

In some protein families (e.g., dinucleotide binding proteins), very low sequence similarities (<10%) have been found and there are often very extensive indels, usually in the loops, whereas the core of the fold is much more conserved. For these much broader fold families, very sensitive structure comparison methods are needed. In particular these should be able to identify conserved structural regions that may be associated with specific sequence patterns. Such regions might be important for the folding pathway or for stabilizing the fold. To meet this challenge, a wealth of new compari-

[1] B. W. Matthews and M. G. Rossmann, this series, Vol. 11, p. 397.

[redacted]

son methods have been developed, some able to cope with very distantly related proteins.

There are now over 30 methods for comparing structures. This chapter discusses those flexible enough to align distantly related structures and therefore most suitable for identifying and analyzing protein fold families. To illustrate ways of overcoming the various difficulties encountered, we have focused on our method, sequential structure alignment program (SSAP), and describe the various modifications that have been required to handle more complex similarities. In particular, the need to identify similar motifs between proteins and the development of a multiple comparison method that can identify the consensus structure for a family of related proteins are discussed.

Different Approaches

There are two main approaches to structure alignment, both based on comparing the global protein geometry (reviewed in Ref. 6). Rigid-body techniques superpose structures in a common external frame of reference and measure distances between equivalent positions. Alternatively, the internal geometry of two proteins can be compared, that is, distances or vectors between residues in the same protein. Both types need strategies for coping with insertions and deletions. Some include information about local residue features, such as torsional angles, to enhance accuracy.

Although early rigid-body methods[1] had problems with indels, more recent solutions[7-9] include using dynamic programming to locate equivalent positions for superposition. Initial alignments are often obtained by compar-

(e.g., graph theory,[12] geometric hashing,[13] distance plot comparison[14,15]) along with various ways of coping with indels. In some graph theoretical approaches, graphs are based on distances and angles between secondary structure elements. Use of subgraph isomorphism algorithms[12] allows partial matches between proteins, thereby accommodating indels. Some very interesting similarities have been identified using these techniques, and they are fast allowing new structures to be scanned against all those known to search for matches.

Distance plot-based methods compare all the interresidue distances in one protein, to equivalent distances in another. Distance plots were originally suggested by Phillips[16] in 1970, and are two-dimensional matrices whose axes correspond to all the residue positions. Cells can be shaded according to distances between residues. Nearly identical structures can be compared simply by overlaying their distance plots. Early strategies for indels included chopping out apparently nonequivalent residue positions in order to be able to generate conformant plots that could be overlaid.[17]

In the method of Holm and Sander,[15,18] proteins are chopped into hexapeptide fragments to limit the effect of indels, and the distance plots are compared to find matching fragments. These are then recombined using simulated annealing. This flexible approach has been used to cluster all the known structures into protein families automatically. It can also be used to search for structural matches with different topologies, although this can be very time-consuming because all possibilities are explored. Few such instances have been identified to date.

A method that includes information about other relationships between residue positions (e.g., hydrogen bonding patterns) has been developed by

comparing different features and relationships have to be carefully weighted, and this makes the approach less suited to data bank searching and automatic clustering of protein families.

Sequential Structure Alignment Program: A Distance Plot-Based Method for Comparing Protein Structures

Some of the problems encountered in comparing distantly related proteins and ways of overcoming them can be illustrated for the SSAP method of Taylor and Orengo.[14] This compares internal geometry between proteins using the Needleman–Wunsch dynamic programming algorithms developed for sequence alignment. Instead of residue identities or physicochemical properties, three-dimensional geometry is compared to identify equivalent positions.

This is done by describing a structural environment or view for each residue which is the set of vectors from the Cβ atom to Cβ atoms of all other residues in the protein (Fig. 1). The view is defined within a common frame of reference for each residue based on the tetrahedral geometry of the Cα atom. Vectors give more information on relative positions in a view than simple distances. Similarly, using Cβ atoms gives more information

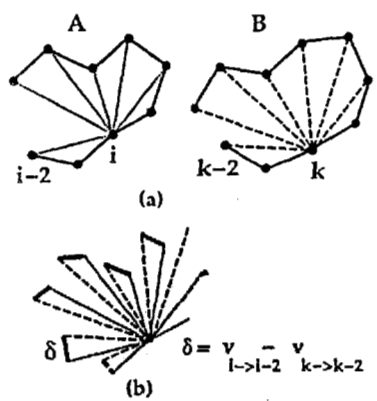


FIG. 1. Schematic representation of residue structural environments or views that are compared in the SSAP method.[14] In (a), A and B represent fragments of protein structure, and the dotted lines are vectors from residues *i* (in A) and *k* (in B). Common frames of reference are used to derive these vectors, based on Cα geometry. This means that views from *i* and *k* can simply be compared by calculating the difference between equivalent vectors.

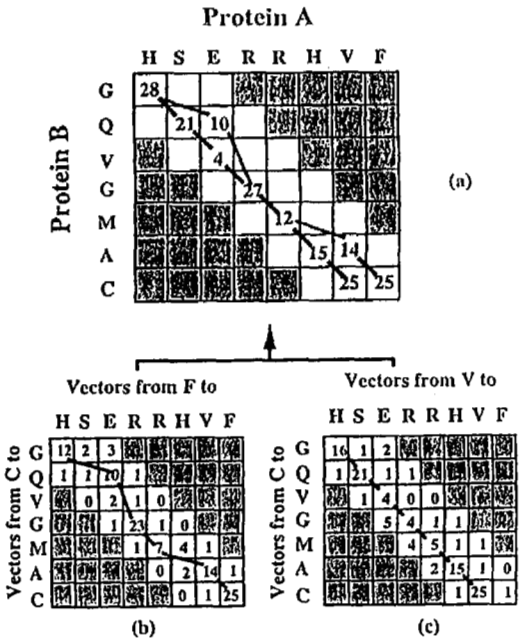


FIG. 2. Double dynamic programming algorithm used by SSAP. Residue views are scored in lower level, vector matrices (b, c). The alignment paths through these matrices, found by dynamic programming, are added to an upper level, summary matrix (a). Once views from all selected residue pairs have been compared, the optimal path through the summary matrix, again identified by dynamic programming, gives the alignment of residues in the two structures. Shaded cells are outside the window of selected residue pairs.

than Cα atoms, particularly for alternating positions along a β strand. Because views are defined in the coordinate frame of the Cα atom, they are rotationally invariant, which makes their comparison insensitive to the displacement of substructures.

If proteins are nearly identical, residue views can be compared by simply subtracting equivalent vectors (Fig. 1). However, as with distance plots, insertions and deletions make it difficult to identify equivalent positions. This problem is solved by using dynamic programming to align residue views (Fig. 2). As with sequence alignment, a two-dimensional matrix is constructed (vector matrix). The axes are the vector sets of the two proteins, and cells are scored by subtracting the associated vectors. For example, the score for comparing vector $v_{i\to i-2}$ in protein A with vector $v_{k\to k-2}$ in

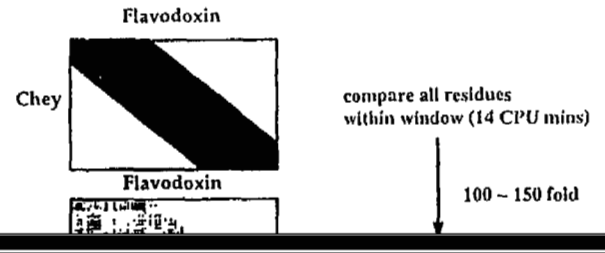
protein B (see Fig. 1) is

$$Svect_{i\to i-2,k\to k-2} = a/(b + \delta) \quad (1)$$

where

$$\delta = v_{i\to i-2} - v_{k\to k-2} \quad (2)$$

In Eq. (1) δ is the distance between the vectors, and a (=500) and b (=10) are parameters that were optimized using a large set of structure



pass is performed, recomparing the 20 highest scoring residues pairs to refine the alignment path.

[22] C. A. Orengo, N. P. Brown, and W. R. Taylor, *Proteins: Struct. Funct. Genet.* **14**, 139 (1992).

To generate a significant score for data bank searching, the SSAP score is normalized to be independent of protein sizes. Extensive trials[22] involving comparisons with both related and unrelated structures established a robust

scoring scheme for identifying related structures. This calculates the logarithm of the average similarity score for equivalent vectors. It is measured over all pairs of equivalent vectors from all equivalent residue pairs. Vectors to adjacent positions (±5 residues) are excluded to reduce the effect of high scores from local secondary structure similarity:

$$S_{SSAP} = \left(\sum_{i=1,i'=1}^{aln} \sum_{j=1,j'=1}^{aln} Svect_{i\to j,i'\to j'} \right) \Big/ (\text{maxequivs} * (\text{maxequivs} - 11)) \quad (4)$$

where $Svect_{i\to j,i'\to j'}$ is the vector similarity score for comparing vectors from equivalent residues i and i' in proteins A and B to equivalent residues j and j', respectively. In Eq. (4), *aln* is the number of aligned residue pairs, and maxequivs is the number of residues in the smallest protein and therefore the maximum possible number of equivalent positions between the proteins. Since the maximum score from comparing two identical vectors is 50, the final SSAP score is set to have a maximum value of 100 as follows:

$$S'_{SSAP} = \ln(S_{SSAP}) * 100/\ln(50) \quad (5)$$

Taking logarithms gave a better resolution of scores between structurally related and unrelated protein pairs when different scoring schemes were tested,[22] largely because differences between equivalent vectors increase exponentially as the similarity between proteins decreases.

SSAP scores above 80 are associated with highly similar structures. In many cases matching proteins have similar sequences or functions, suggesting they are homologous and have diverged from a common ancestor. Scores in the range of 70 to 80 indicate a similar fold but with more variation in the loops and larger shifts in secondary structure orientation. Often there is no sequence similarity or common function, and the relationships between the proteins are not clear. Either they have diverged a long way from the same ancestor or converged toward the same fold. As well as the similarity score, SSAP outputs the number of equivalent residue pairs between pro-

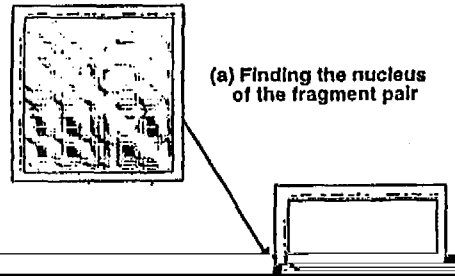
rest of the structure, with clearer sequence patterns, and hence easier to predict. Several such motifs have already been observed (e.g., β hairpins, $\beta\alpha\beta$ motifs), and templates expressing their recurrence within particular structures have improved prediction.[23] In the $\alpha+\beta$ class, folds are often asymmetric and complex, using motifs from all other classes which are then packed together in many different ways. Prediction of these structures will probably need an approach based more on recognizing motifs and understanding ways in which they prefer to pack.

For this reason, another version of SSAP was developed[24] (SSAPI) that finds conserved structural motifs between proteins. As for the original SSAP method, a summary matrix between two proteins is scored by comparing views from all residue pairs between the structures. Unlike the global method, where only the optimal vector alignment path is added to the summary matrix, in SSAPI the complete vector matrix is added to accumulate information about locally similar regions. Subsequently local paths within the matrix are extracted using a Smith–Waterman dynamic programming algorithm.

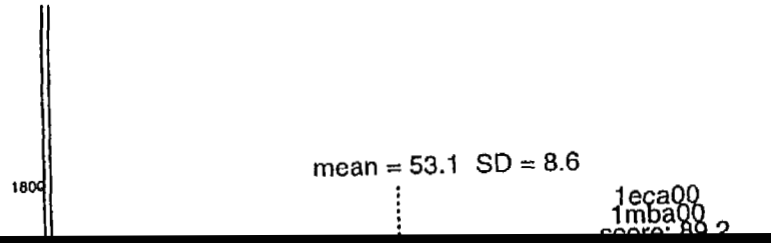
A fragment nucleus is first extracted from the summary matrix by tracing back a path from the highest scoring cell, until a running score (S_{run}) calculated at each position falls below a cutoff:

$$S_{run} = (S_{run} + Sres_{ij})/2 \quad (6)$$

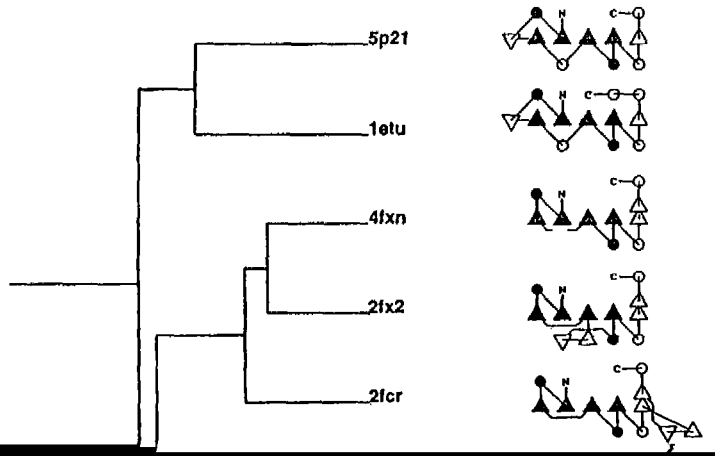
where $Sres_{ij}$ is the residue similarity score for cell ij in the summary matrix. The 20 highest scoring residue pairs from the fragment path are then recompared and their views aligned. Scores from along these alignment paths are accumulated in a separate fragment matrix (Fig. 5). The best local path through this fragment matrix is then sought by growing a path from the highest scoring cell and again truncating this once the running score falls below a cutoff. A softer cutoff is used to allow fragments to grow slightly. As with the global SSAP, the procedure of regenerating the

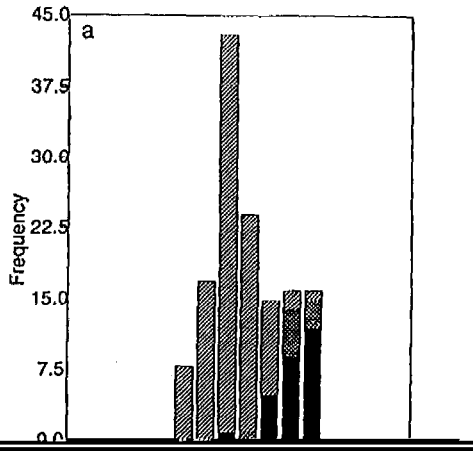


into 430 families on the basis of sequence similarity.[26] Sequence alignment is at least 10 times faster than structure comparison, and if more than 30% of the sequences correspond the folds will be the same.[2,3] Representatives from each family can then be structurally compared using SSAPc and families merged if their representatives match with SSAP scores above 80. This gives 274 homologous families. If the SSAP cutoff is softened to 70, 206 fold families are obtained.[26] Relationships between structures in these broader families is less clear, as functions and sequences can differ substantially. In view of this uncertainty, the families can be described as analogous



The most recent version of SSAP (SSAPm) was developed[28] with the aim of analyzing fold families and is particularly suited to broad structural families such as those of the superfolds. SSAPm multiply aligns a fold family to find conserved regions which can be used as structural fingerprints in prediction and recognition methods. All pairs of proteins in the family are compared using SSAP. The alignment of the highest scoring pair is then used to seed the multiple alignment, and a consensus structure is calculated consisting of average views at each residue position. Both average


5p21
1etu
4fxn
2fx2
2fcr
N
C


a
45.0
37.5
30.0
22.5
15.0
7.5
0.0
Frequency

[redacted]