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 $\beta\alpha$ barrels by sequence database searches with a number of $\beta\alpha$ 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.

ENVIROBN: Calculates area buried and fraction polar for each residue in a structure
3DPROFILER: 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
VERIFY3D: Determines the overall 3D profile score for a structure and the average score in regions of sequence along the structure

SSAP: Sequential Structure Alignment Program for Protein Structure Comparison

By Christine A. Orenge 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.\(^2\) 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.\(^4\)

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 comparis-
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 involve information about local residue features, such as torsional angles, to enhance accuracy.

Although early rigid-body methods had problems with indels, more recent solutions include using dynamic programming to locate equivalent positions for superposition. Initial alignments are often obtained by comparing sequences or torsional angles. The fit can then be optimized by iterating through cycles of superposition followed by alignment based on distances between superposed positions.

The effect of indels on superposition methods can also be minimized by removing the variable loops and comparing secondary structures. Different algorithms can 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 Sali and Blundell. Relationships are compared using simulated annealing, to identify possible equivalences between the two proteins. This information and that of similarities for a range of residue features (e.g., accessibility, torsional angles, and volume) are accumulated in a two-dimensional similarity matrix, except axes are labeled with positions from each protein. In a final step, corresponding residues are determined using single dynamic programming. Most families in the protein structure data bank have been compared and aligned using this approach. Contributions of scores from

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. 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 Ca atom. Vectors give more information on relative positions in a view than simple distances. Similarly, using Cβ atoms gives more information than Ca atoms, particularly for alternating positions along a β strand. Because views are defined in the coordinate frame of the Ca 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-2} in protein A with vector v_{k-2} in protein B.
protein B (see Fig. 1) is

\[ S_{\text{vect}} = a(l + \delta) \]  

where

\[ \delta = v_{l-1} - v_{k-1} \]  

In Eq. (1) \( \delta \) is the distance between the vectors, and \( a \) (=500) and \( \delta \) (=10) are parameters that were optimized using a large set of structure comparisons from the protein data bank. Parameter \( b \) softens the contribution of local distances and prevents residue pairs scoring too highly for similar local geometry (e.g., if they are both in helices) regardless of their relative positions in the structures.

The optimal pathway aligning the views is obtained by dynamic programming and added to a summary similarity matrix (Fig. 2). All residue views are compared in this way, and a final dynamic programming evaluation of the summary matrix gives the optimal alignment of the two structures. Use of dynamic programming at two levels, that is, to align all residue views and finally to align the residues, has been described as double dynamic programming.

The advantage of summing the whole path, rather than putting a single value in the residue pair cell, is that information about structurally similar regions tends to be reinforced with the addition of each path. Consequently, cells in the matrix corresponding to structurally equivalent positions have much higher scores than the average (typically 100-fold) depending on the sizes of the proteins being compared. This makes the method relatively insensitive to the value of the gap penalty and very robust, as equivalent regions can be easily recognized. A nominal gap penalty of 50 is chosen for all types of comparisons. To reduce noise caused by adding paths from nonequivalent residue positions, a cutoff of 10 on the path score prevents low scoring pathways from being accumulated.

An upper limit can be placed on the number of insertions or deletions expected between the proteins being compared. This is equivalent to placing a window on the vector and summary score matrices (Figs. 2 and 3). Only vector pairs or residue pairs within the window are compared and scored, reducing the time required for an alignment. Default window sizes are typically set to be the difference in the number of residues in the proteins being compared, plus 50.

Including Other Information besides Global Geometry

In addition to comparing structural environments, other residue information can be included29 (e.g., accessibility, torsional angles, volume). Differences in these properties are scored in the summary similarity matrix. Information about other structural relationships, such as hydrogen bonding patterns or disulfide bonds, can also be included in the residue views. Weights for the contribution of these different features and relationships were optimized on a set of comparisons in the immunoglobulin and dinucleotide-binding families of proteins. Vector views provide most information and are weighted highly, but for some of the more remote comparisons improvements can be obtained by including information about hydrogen bonding and accessibility.

Increasing Speed

Various modifications have been implemented to speed up the method. In the first,31 views are only compared if the residues appear to be in similar structural locations, that is, they have similar accessibilities and torsional angles. A parameter \( \delta_{\text{tot}} \) controls the allowed difference in these properties. The selection criterion expressed by Eq. (3) below has been found to give

\[ \text{compare all residues within window (14 CPU mins)} \]

\[ \text{compare residues having similar phi, psi angles and accessible areas (80 CPU secs)} \]

\[ \text{compare residues within aligned secondary structures and having similar phi, psi angles and accessible areas (16 CPU secs)} \]
Comparisons with both retrieval and retrieval structures established a robust significance of the independent variable. Extension tests involving

to generate a significant score for data bank starting the SSAP score

Normalizing SSAP score for data bank scoring

From the first pass are recorded in the second pass.

The second step is a scoring pass of the second passscoring performed.

(c) select 20 highest scoring residues pairs

(d) residue structure

(e) selected residues

Pass Second

Structural alignment residues from initial

(f) identical initial

Pass First

Similarly, the alignment pair is passed to

(g) alignment similarity, residues from initial

SSAP motion structure comparison

651

1991

Three-dimensional considerations

optimal increases in speed and quality of alignment
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:

\[ S_{SSAP} = \left( \frac{\sum_{i} \sum_{j} \text{Svec}_{i-j} \cdot \text{Svec}_{i-j}}{(\text{maxequivs} \cdot (\text{maxequivs} - 11))} \right) \]

where \( \text{Svec}_{i-j} \) is the vector similarity score for comparing vectors from equivalent residues \( i \) and \( j \) in proteins A and B to equivalent residues \( j' \) and \( i' \), respectively. In Eq. (4), \( abn \) 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:

\[ S'_{SSAP} = \ln(S_{SSAP}) \cdot 100/\ln(50) \]  

(5)

Taking logarithms gave a better resolution of scores between structurally related and unrelated protein pairs when different scoring schemes were tested, 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 proteins. To ensure global similarity between folds, at least 60% of residue positions in the two proteins should be equivalent.

With scores between 60 and 70, proteins do not have the same global fold but usually belong to the same protein class and often have structural motifs in common. For example, comparisons of TIM barrels and Rossmann folds in the alternating \( \alpha/\beta \) class of proteins often return scores in this range because of matching \( \beta/\beta \) motifs.

**Identifying Domains and Local Structural Motifs**

Comparison methods also need to be able to identify common structural motifs. For some folds these may be more strongly determined than the rest of the structure, with clearer sequence patterns, and hence easier to predict. Several such motifs have already been observed (e.g., \( \beta \) hairpins, \( \beta/\beta \) motifs), and templates expressing their recurrence within particular structures have improved prediction. 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 (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:

\[ S_{run} = \frac{(S_{run} + S_{res})}{2} \]

where \( S_{res} \) 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 path by recomparing high scoring pairs improves alignments by reducing noise. Once the optimal fragment path has been identified, it is removed from the summary matrix together with adjacent cells by resetting scores to zero. This prevents subsequent paths crossing previously identified fragment pairs. Cells in secondary structures crossed by the path are also reset (Fig. 5). Alignment scores for each path are normalized to be independent of fragment size and in the range of 0–100.

The SSAPI program can also be used to search for a particular motif within a set of structures. Although slower than SSAP, requiring approxi-

I361

approximately 5 CPU minutes on an Indigo R4000 computer to identify all greek-key motifs in two immunoglobulin structures, it is feasible to use the method to search for common motifs among a set of proteins with similar architectures, for example, all the mainly α proteins which have predominantly orthogonal helix–helix interactions.

Using SSAP for Data Bank Searches and for Identifying Protein Families

Protein fold families can be identified by using fast SSAPc to compare protein structures. In the June 1994 release of the Protein Data Bank (PDB) there are over 3000 well-resolved structures (<3.0 Å) which cluster into 430 families on the basis of sequence similarity. 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. 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. 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 fold families.

A newly determined protein structure can be scanned against a set of representative structures from each of the 274 homologous fold families to check whether it is a novel fold or can be assigned to a particular family. The distribution of SSAP scores across the set enables structural matches to be expressed as standard deviations from the mean, so that their significance can be assessed (Fig. 6).

Multiple Structure Comparisons across Protein Family to Identify Conserved Positions

In the current structure data bank there are several fold families that are very highly populated with proteins having no sequence or functional similarity. Nearly 30% of known structures adopt one of these folds, and this increases to 50% if only nonhomologous structures are considered [where homologous proteins have >25% sequence identity or high SSAP scores (>80) and related functions]. Because of their ability to support a wide range of sequences and functions, they can be described as superfolds. Outside the superfolds, proteins in other fold families all have related functions. Most of the superfolds are very familiar structures (e.g., TIM barrels and globin folds) and have been known since the early days of crystallography. Relationships between the structures in the families are unclear. They may have diverged a long way from a common ancestor. Alternatively, the folds may be particularly stable arrangements of secondary structures toward which many structures converge. Whatever the cause, the existence of these superfold families has important implications for structure prediction and fold recognition. If a protein has no sequence or functional similarity to any known structure, then there is a 50% probability that it will either belong to one of the superfold families or have a unique fold.

The most recent version of SSAP (SSAPm) was developed\textsuperscript{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 vectors and information about their variability are calculated. All remaining structures in the set are then compared to the consensus and the highest scoring structure added next. This cycle of addition, construction of a new consensus, and pairwise comparisons against the consensus is repeated until all structures have been merged into the alignment. Information on variability is used to set weights for scoring vector comparisons. Similarly, a structure conservation measure, calculated at each position, weights vector comparison scores for that position and enhances the alignment of conserved core regions in the fold.

The program CORA takes the multiple alignment given by SSAPm and generates a structural template for the family. Only positions with individual SSAP scores above a cutoff are written to the template, and average views to other selected positions are described together with vector variability information. The cutoff can be adjusted depending on the similarity of structures compared. Structural templates can be used both for fold recognition by threading type algorithms and in fold identification. Considerable improvements in resolution between related and unrelated structures can be obtained by searching against the structure database with a fold template (Figs. 7 and 8). The program CORALIGN is a modified version of SSAP that aligns a protein structure against a template generated by CORA. Again, individual SSAP conservation scores and information on vector variability are used to set weights for matching key conserved positions in the fold.

Conclusion

In summary, there are now several robust methods for comparing distantly related protein structures, some\textsuperscript{4,19,28} performing multiple comparisons with protein fold families to identify and characterize conserved structural features. Already, the data generated by these methods offer considerable hope for improvements in fold recognition.\textsuperscript{4,5}

This is timely as there are still far more protein sequences known (~100,000) than structures determined (~4000). By the year 2005, we may expect thousands more sequences from the various genome mapping projects. The ultimate goal is to understand the functions of these proteins, and knowledge of the protein structure is an essential step in this process.

Fig. 7. Dendrogram showing the structural relationships for a set of doubly wound α/β proteins. The x axis gives the SSAP score. On the right-hand side, schematic topology representations, drawn by the program TOPS, are shown for each structure. Helices are represented by circles and β strands by triangles. The common core of the fold, identified by SSAPm, is shaded and consists of a four-stranded β sheet with an α helix packed on each side.

Fig. 8. Improvements in fold identification obtained by scanning through the dataset of nonhomologous structures using a structural template based on the common core of the α/β doubly wound folds. (a) Distribution of SSAP scores obtained by scanning against the data set with a representative doubly wound α/β structure (Chey). (b) Distribution of SSAP scores using the core template. Solid bars are correct hits (other alternating βα type proteins with a similar fold), whereas crosshatched bars represent hits on TIM-barrel proteins (alternating βα type proteins with a different fold). Hatched bars are unrelated folds.
We can expect that analysis of protein structural families and identification of consensus structural templates will improve both template- and threading-based prediction algorithms for identifying the fold of a protein, particularly for the superfolds. For more complex and less frequently observed folds, characterization of common structural motifs and any associated sequence patterns would be expected to improve prediction.

Running SSAP and Generating Superposition of Structures

All versions of SSAP require an input WOLF file. This can be created by running the program WOLF, a modified version of DSSP\textsuperscript{29} that generates information about hydrogen bonding patterns and secondary structure assignments. The input file for WOLF is the Protein Data Bank\textsuperscript{25} file for the structure. WOLF files also contain residue accessibility and torsional angles, together with frames of reference for each residue, centered on the C\textbeta position.

Additional SEC files are needed for running SSAPs. These contain information about the line vectors representing secondary structures, together with pairwise distances and angles and overlap information for each secondary structure pair. They are generated by running the program SECLINE.

Besides returning the normalized global score, SSAP outputs an alignment of the two structures together with individual scores for each residue position. These are in the range of 0 to 100 and measure similarity in views for the residue pair. Scores are generally higher for secondary structure regions and around active sites.

SSAP also outputs two superposition (SUP) files, listing equivalent positions and scores. These are input for the program SUPRMS\textsuperscript{30}, which performs a weighted superposition using the SSAP residue scores as weights. SUPRMS is based on the McLachlan least squares method\textsuperscript{31} and returns the root-mean-square (rms) deviation between the structures together with the number of residue pairs superposed for a 3.0 Å cutoff. It also outputs a QUANTA graphics file for coloring the superposed structures according to similarity in views at each position.

For multiple comparisons across a family of structures, the program SSAPm generates a MULTISUP file listing equivalent positions and information about conservation of residue views. This can be input to the program MULTISUP\textsuperscript{32}, which performs weighted multiple superposition of the structures.

Availability

The SSAP package can be obtained by sending a request to orengo@bsm.bioc.ucl.ac.uk. Programs are distributed as binary files compiled for execution on a Silicon Graphics machine running UNIX.

References

\textsuperscript{29} W. Kabsch and C. Sander, \textit{Biopolymers} 22, 2577 (1983).

Introduction

The structure of a protein can elucidate its function, in both general and specific terms, and its evolutionary history. Extracting this information, however, requires a knowledge of the structure and its relationships with other proteins. These two aspects are not independent, for an understanding of the structure of a single protein requires a general knowledge of the folds that proteins adopt, while an understanding of relationships requires detailed information about the structures of many proteins.

Fortunately, this complex problem with its intertwined requirements is not insurmountable, for two reasons. First, protein structures can be fundamentally understood in ways that most of their sequences cannot. The comprehensibility of protein structures derives from the relatively few secondary structure elements in a given domain and the fact that the arrangement of these elements is greatly restricted by physics and probably by evolution. Second, resources are now available to aid recognition of the relationships between protein structures. The structural classification of proteins (scop) database hierarchically organizes proteins according to their structures and evolutionary origin.\textsuperscript{1} As such, it forms a resource that allows researchers to learn about the nature of protein folds, to focus their investigational efforts.