PROTEINS: Structure, Function, and Genetics 14:139–167 (1992)

Fast Structure Alignment for Protein Databank Searching

Christine A. Orengo, Nigel P. Brown, and William R. Taylor
National Institute for Medical Research, London NW7 1AA, England

ABSTRACT **A fast method is described for searching and analyzing the protein structure databank. It uses secondary structure followed by residue matching to compare protein structures and is developed from a previous structural alignment method based on dynamic programming.**

Linear representations of secondary structures are derived and their features compared to identify equivalent elements in two proteins. The secondary structure alignment then constrains the residue alignment, which compares only residues within aligned secondary structures and with similar buried areas and torsional angles. The initial secondary structure alignment improves accuracy and provides a means of filtering out unrelated proteins before the slower residue alignment stage.

It is possible to search or sort the protein structure databank very quickly using just secondary structure comparisons. A search through 720 structures with a probe protein of 10 secondary structures required 1.7 CPU hours on a Sun 4/280. Alternatively, combined secondary structure and residue alignments, with a cutoff on the secondary structure score to remove pairs of unrelated proteins from further analysis, took 10.1 CPU hours. The method was applied in searches on different classes of proteins and to cluster a subset of the databank into structurally related groups. Relationships were consistent with known families of protein structure.

Key words: protein structure, structure comparison, structure alignment, dynamic programming, databank search

INTRODUCTION

Dynamic Programming in Structure Comparison

Some recent methods of protein structure comparison employ dynamic programming techniques to align proteins. These techniques were developed by Needleman and Wunsch[1] for aligning amino acid sequences, and are extremely well suited to handling insertions and deletions.

In the method of Taylor and Orengo,[2] a local structural environment is defined for each residue in a protein, by the set of vectors from the C_β of the residue to the C_β of all other residues in the same structure. Residues in two proteins are then matched by comparing their structural environments and an alignment of the proteins generated by dynamic programming techniques. Other properties of the residues such as torsional angles, accessibility, or hydrogen bonding patterns can also be matched to improve alignments.[3] A more recent and faster version matches only subsets of residues from the two proteins. These are located in structurally similar regions and their selection improves the accuracy of the final alignment.[4] The method is completely automatic, requiring only residue coordinates and residue properties.

Zuker and Somorjai[5] use a method which combines both dynamic programming and superposition. Their algorithm finds a set of nonoverlapping fragments from one protein and, from a second protein, a set of equal sized fragments, which can be optimally superimposed on the first set. Alternatively, the approach of Šali and Blundell[6] adopts a multilevel representation of protein organization. The protein is described by elements at several hierarchical levels: residue, secondary structure, supersecondary structure, motif, domain, or globular structure. Each element can then be allocated features which relate to the protein fold. These features can be individual properties of the element or they can be associated with relationships between elements. Property comparisons are stored in a two-dimensional similarity matrix while relationships are compared using simulated annealing and information regarding equivalences stored in the same similarity matrix. Alignment of the proteins is then obtained from this matrix by dynamic programming. Although the method uses a very flexible definition of topological equivalence which should help in the alignment of remote structures, it is less automatic in its application.

Received July 10, 1991; revision accepted October 25, 1991.
Address reprint requests to Dr. Christine A. Orengo at her present address: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England.

Secondary Structure Comparisons

Protein structural similarity can also be recognized by comparing secondary structures. Structural relationships can be more clearly seen at this level, as changes which occur during the course of evolution tend to be restricted to insertions and deletions in the surface loops. Within families of proteins, secondary structure topology is relatively well conserved allowing proteins to be classified into different structural classes.[7] Furthermore, the reduced number of elements to compare means that the methods can be used for fast searches of the databank.

Richards and Kundrot[8] have presented an extension of distance plot methods already established for protein comparison at the residue level.[9] Two-dimensional matrices are used to indicate relationships between interacting secondary structures in the protein, where cells reflect the properties and geometric relationships between segments. Two proteins can be compared by looking for similar patterns in their distance plots. To search for a structural motif a distance matrix mask of the motif is generated and compared to that of the protein, and the overall fit calculated as the root mean square difference between the two matrices. However, since every residue position within the mask is considered, the sensitivity can be affected by insertions and deletions.

Abagyan and Maiorov[10] use simplified representations of secondary structures to search for similar fragments of structure in the protein data base. Each α-helix or β-strand is represented by a vector along its axis and vectors are used as connections between secondary structures. The topology of the structure can be described by lengths of the vectors and the planar and dihedral angles between them. Superposition of two vector representations is performed using McLachlan's algorithm.[11]

A related approach, developed by Murthy,[12] incorporates the techniques of dynamic programming. In the first step a similar vector representation of helices and strands is generated. One molecule is then rotated relative to the other and the probability that two elements are equivalent is related to the angle between them after superposition. Probability values for a given orientation are placed in a two-dimensional matrix which is evaluated using dynamic programming. The score derived from the matrix is weighted depending on the similarity of vectors between equivalent elements and plotted as a function of the 3 eulerian angles. Structural similarity is revealed by the presence of a significant peak in the score map at the correct orientation.

With the aim of producing a rapid method for databank searches, Mitchell et al.[13] use graph theory to compare secondary structure geometry between proteins. Linear representations of secondary structures are determined and then the angles between pairs of axes, together with distances between midpoints and closest approach distances, are evaluated and stored. Each protein or motif can then be represented by a graph, the nodes of which correspond to the linear secondary structures, and the edges to the interline angles and distances. It is possible to check whether a query graph is contained within one or more of the stored graphs of databank structures by application of subgraph isomorphism algorithms. The approach is very fast (a search through 326 structures with a motif containing 15 secondary structures taking roughly 30 min).

Outline

In this paper we present a method for searching and analyzing the protein structure databank, which aligns both secondary structures and residues between proteins. The motivation for using secondary structures was to generate a rapid initial structural alignment, which can be used to filter out pairs of proteins not sufficiently related to merit a more detailed residue comparison. An earlier method which uses dynamic programming techniques[2,3] has been modified to align both secondary structure and residue elements (Combined Method). Linear representations of secondary structures are derived and their properties and relationships compared to identify equivalent elements in two proteins. The secondary structure alignment then guides the residue alignment (which is further confined to residues having similar buried areas and torsional angles), as only residues within aligned secondary structures are compared. Accuracy is thus improved as this set contains a very high proportion of equivalent residues.

It is important not to confuse our approach with sequence alignment methods which modify the gap penalty in regions of secondary structure or equate secondary structures of like type (e.g., Barton and Sternberg or Fischel-Ghodsian et al.[14,15]). The biases introduced in such methods apply generally to *all* pairs of like secondary structure, whereas our method identifies specific pairs by considering global structure and then performs a true structure comparison rather than a sequence alignment.

With our method it is possible to search or sort the protein structure databank very quickly using just secondary structure comparisons. Additionally, residue alignments can be generated, with a cutoff on the secondary structure score to remove pairs of unrelated proteins from further analysis. Suitable cutoffs were determined using sets of known structurally related and unrelated proteins from the databank.

METHODS

The method of structure comparison described below is an extension of the previously published

method of Taylor and Orengo.[2–4] The original residue environment method, which forms an integral part of the new work, is summarized briefly before covering the new method of secondary structure matching. The full working method operates as a multistage process starting with secondary structure matching and finishing with residue alignment as shown in Figure 1.

Residue Structural Alignment Method (SSAP)

In the structural alignment method of Taylor and Orengo,[2] the alignment between two protein chains was found from a comparison of residue structural environments using dynamic programming techniques, where a residue's structural environment was given by the set of vectors from its C_β to all other C_βs in the same chain, defined in terms of a local frame of reference centered on the C_β of that residue. Two environments i, j, were compared using dynamic programming to find the best path through a lower or distance level matrix, which scored the difference between the respective vector sets. Vector sets were compared using the following function to score the cells of the lower level matrix:

$$S_{\text{res}}(k,l) = \frac{w}{a+\delta}, \quad \delta = |\vec{\mathbf{v}}_{\text{res}}(\text{i,k}) - \vec{\mathbf{v}}_{\text{res}}(\text{j,l})| \qquad (1)$$

where $\vec{\mathbf{v}}_{\text{res}}(i,k)$, is the vector from residue i to residue k in the first chain and similarly for $\vec{\mathbf{v}}_{\text{res}}(j,l)$ in the second chain so that δ is a difference function. The constants w, a were set to 500, 10 respectively.

For each pair of residues i, j (one residue from each protein chain) the optimal paths through the associated distance level matrices were accumulated in a corresponding higher or residue level matrix. Thus, scores for similar regions of structure in the two proteins were reinforced by many distance level pathways passing through the residue matrix cells. Finally, dynamic programming was also applied to identify the best pathway through the residue matrix, giving the alignment of the two proteins.

Secondary Structure Matching

The secondary structure comparison method is essentially the same as for residue matching, except that secondary structure elements replace residues in the higher and lower level matrices, so that different definitions for reference frames and properties for inclusion in the lower level scoring function are required.

Secondary structure assignments were taken from DSSP (Dictionary of Secondary Structure in Proteins[16]), giving three types of linear secondary structure: α-helix (**H**), 3_{10}-helix (**G**), extended- or β-strand (**E**). Extended strands were further classified as parallel, antiparallel, or ambiguous (participating in both parallel and antiparallel ladders) on the basis of DSSP β-ladder definitions, making five types of secondary structure in all.

A selection of features, both geometric and physicochemical, was examined for single structures and for pairwise interactions, to characterize structural environments in the lower level matrices. These included vector sets, angles, buried areas, and overlap between elements.

Pruning of the number of secondary structure comparisons performed was effected by only comparing like with like, using the five secondary structure types distinguished above.

Derivation of Axial Vectors

For each protein, the secondary structure assignments of DSSP were used to define helical (**G, H**), and β-sheet (**E**) elements. However, DSSP types **G** and **H** were modified as follows. Where DSSP generated assignments of the form $\mathbf{G}_{[l]}\mathbf{H}_{[m]}$ or $\mathbf{H}_{[m]}\mathbf{G}_{[n]}$ corresponding to α-helices with contiguous 3_{10}-helix caps, the **H** designation was "grown" at the expense of the **G** region to form a single helix. Rarely, a contiguous region of form $\mathbf{H}_{[l]}\mathbf{G}_{[m]}\mathbf{H}_{[n]}$ was found, in which case the **G** region was divided evenly between the two **H** regions. Isolated **G** regions were left intact.

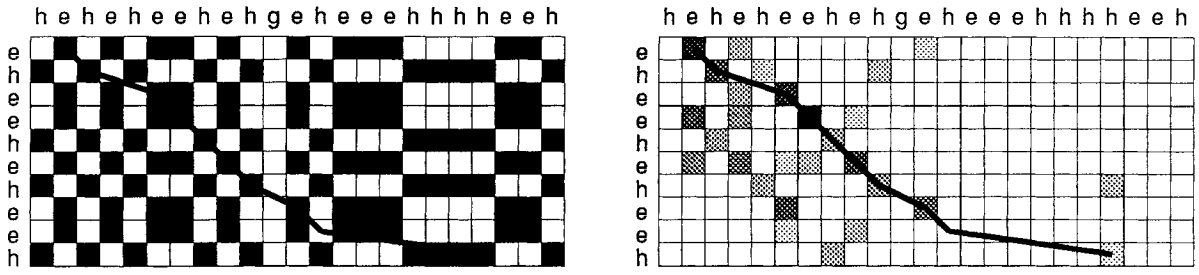
Using the coordinates of the main chain C_α, C, N, and O atoms of each secondary structure and treating each atom as a unit mass, the principal inertial axes for the structure were found.[17] The axial vector $\vec{\mathbf{a}}$, was taken as the inertial axis with the smallest moment and assigned direction from N-terminus to C-terminus. The terminal C_α atoms of the element were projected onto the axial vector to define endpoints.

Derivation of a Local Reference Frame

Relative positional information regarding a structure pair was defined in terms of a local cartesian frame of reference for the first member of the pair in a way similar to that of Taylor and Orengo,[2] which defined frames for residues.

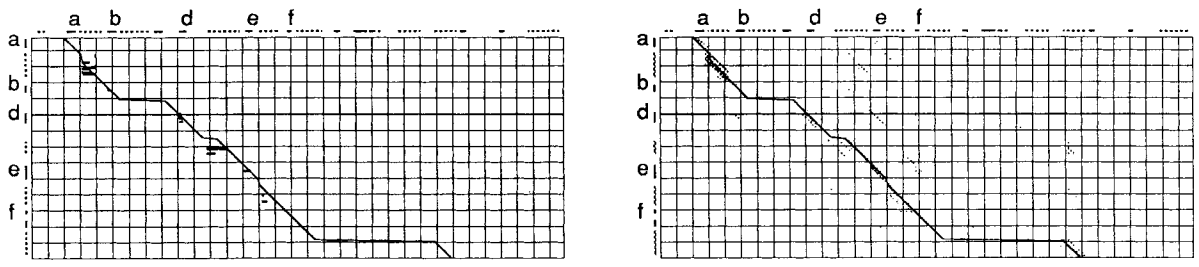
Three axes for a structure were defined centered on the midpoint of the axial vector. The axial vector was taken as the first axis, while the second and third axes were defined by vector cross-products with respect to a second constructor vector. Different constructor vectors were used to define several experimental frames for analysis. These were the cylindrical and spherical hydrophobic moments ($\vec{\boldsymbol{\mu}}_{\text{cyl}}$, $\vec{\boldsymbol{\mu}}_{\text{sph}}$, respectively) described below and also the vector $\vec{\mathbf{v}}_{\text{mm}}$, joining the midpoint of the given structure to the midpoint of the next consecutive axial structure vector along the polypeptide chain.

Various pairwise features relating the secondary structure pair, some defined in this local frame and some independent of the frame, were then used to characterize the interaction.



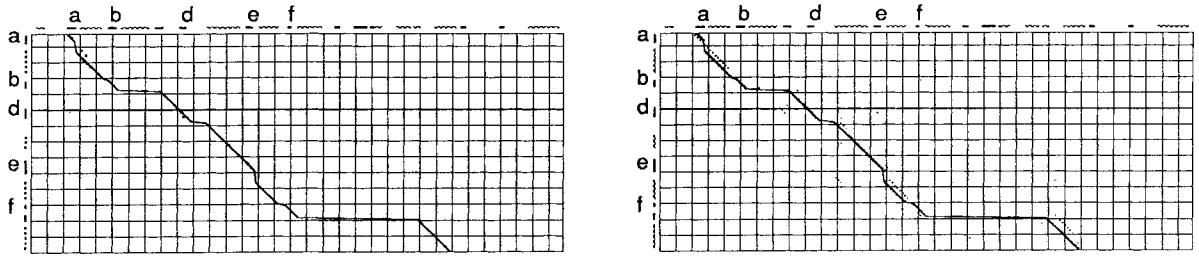
(a) (b)

Stage 1 Secondary structures. Only element pairs of the same type of secondary structure are compared (e.g. both α) in (a) so that a large proportion of possible comparisons is excluded for the alignment in (b).



(c) (d)

Stage 2 Selected residues. Residue pairs within aligned secondary structures and with similar accessible areas and torsional angles are selected in (c) and aligned in (d). The alignment pathway is largely correct (although the **b** strand is shifted slightly) and superposes well on the selected subset.



(e) (f)

Stage 3 Final alignment. The 20 highest scoring residue pairs (e) from stage 2 are recompared to generate the final pathway (f), in which all β-strands are correctly aligned. A high proportion of residues selected at this stage are along the correct alignment path, as are high scoring cells.

Fig. 1. Operation of the combined secondary structure and residue alignment method (SSAPc). At successive stages, the selection of elements for comparison is refined to contain a higher proportion of equivalences and to reduce "noise." Element pairs selected for comparison appear in the left matrices (black) and resultant score matrices (shaded by relative score) and alignment paths on the right. The paths are superimposed on the left matrices for visual comparison with the selected elements. Symbols: (**a**), (**b**) **h** = α-helix, **e** = β-strand; (**c**)–(**f**) hashed edge bars = α-regions, solid bars = β-regions **a**–**f**.

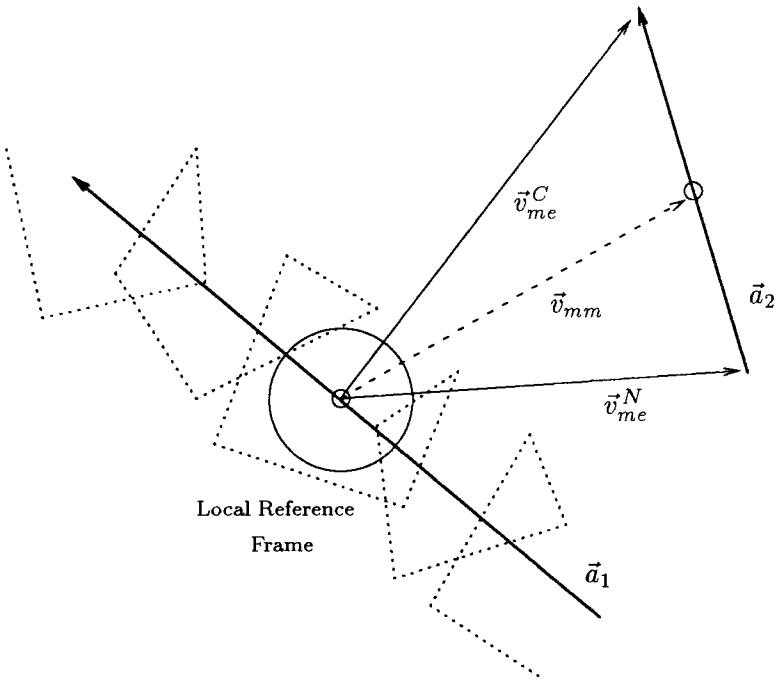


Fig. 2. Secondary structure environments. The environment of an axial structure vector $\vec{\mathbf{a}}_1$ is described by up to three vectors to other structures such as $\vec{\mathbf{a}}_2$ in the protein. These environment vectors are defined in the local reference frame of $\vec{\mathbf{a}}_1$ (see main text), and are the midpoint–midpoint vector $\vec{\mathbf{v}}_{mm}$, and the two midpoint–endpoint vectors, $\vec{\mathbf{v}}^{N,C}_{me}$.

Secondary Structure Features—Single

Vector length

The length of the secondary structure in Ångström defined by projections of the N- and C-terminal C_α atoms onto the axial vector.

Surface area

The solvent accessible area of an isolated secondary structure was estimated by the method of Lee and Richards[18] using a 1.4 Å radius probe, which is commonly used to represent a water molecule rolled over the molecular envelope.

Hydrophobic sum

The hydrophobicity values on the Kyte and Doolittle scale[19] were summed for the component residues to give an overall hydrophobicity H_{sum}, for each secondary structure.

Cylindrical and spherical hydrophobic moments

Hydrophobic moments[20] were defined for each secondary structure. Two moments, cylindrical and spherical, were used. The cylindrical moment is the conventional helical structural hydrophobic moment,[20] which sums residue moments radial to the axial vector and therefore takes no account of amphiphilicity along the axis. The spherical moment takes this into account, being the sum of moments radial to the centroid (which is also the midpoint of the axial vector $\vec{\mathbf{a}}$) of the structure.

The cylindrical moment of an N residue structure was defined as $\vec{\boldsymbol{\mu}}_{cyl} = \Sigma_{j=1}^{N} H_j \hat{v}_j$, where H_j is the hydrophobicity (see above) of the jth residue, and $\hat{v}_j$ is a unit vector perpendicular to and pointing from the axial vector to the C_α of the jth residue in the structure.

The spherical hydrophobic moment $\vec{\boldsymbol{\mu}}_{sph}$, was defined in a similar way, except that $\hat{v}_j$ was interpreted as a unit vector pointing from the midpoint of the axial vector to the C_α of the jth residue in the structure.

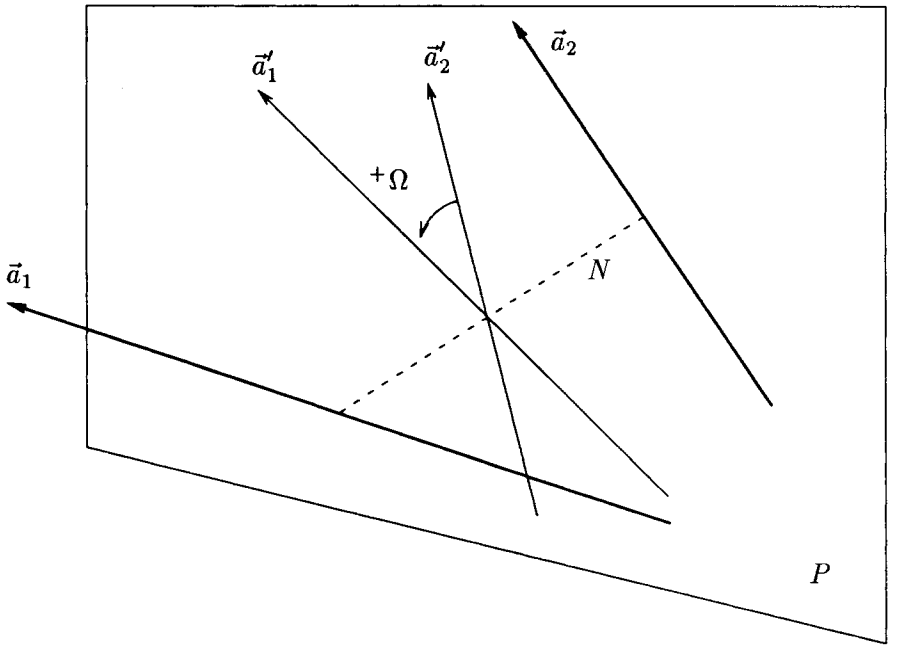
Secondary Structure Features—Pairwise

Midpoint–midpoint vector

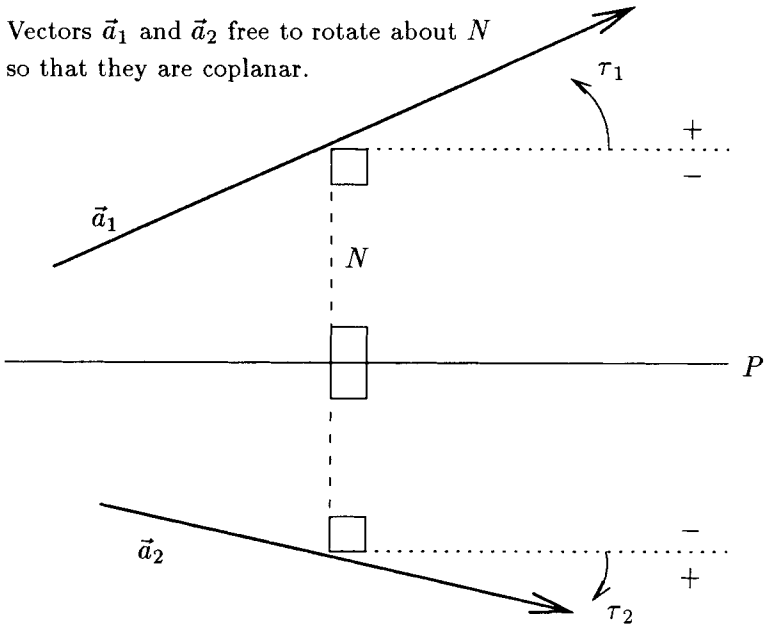
A vector was used to describe the distance and orientation of two secondary structures, treating the second structure as a point. Given a local reference frame centered on some secondary structure $\vec{\mathbf{a}}_1$, the vector from the origin of this frame to the midpoint of the axial vector $\vec{\mathbf{a}}_2$, corresponding to the other structure, was defined in the local coordinate system of the first structure, as the midpoint–midpoint vector $\vec{\mathbf{v}}_{mm}$ (see Fig. 2).

Midpoint–endpoint vector

Information concerning the length and orientation properties of the second structure vector was



(a)



(b)

Fig. 3. Secondary structure geometry. Illustration of a projection plane P and the normal N connecting a pair of axial vectors $\vec{a}_1$, $\vec{a}_2$, showing how choice of N determines the values of Ω and τ. (**a**) The interaxial angle Ω ($|\Omega| \leq 180°$) between projections $\vec{a}_1'$ and $\vec{a}_2'$. The interaxial distance is the length of N. (**b**) Side elevation showing the associated signed tilt angles for the vectors $\vec{a}_1$ and $\vec{a}_2$. The pairwise tilt angle τ, is the sum $\tau_1 + \tau_2$.

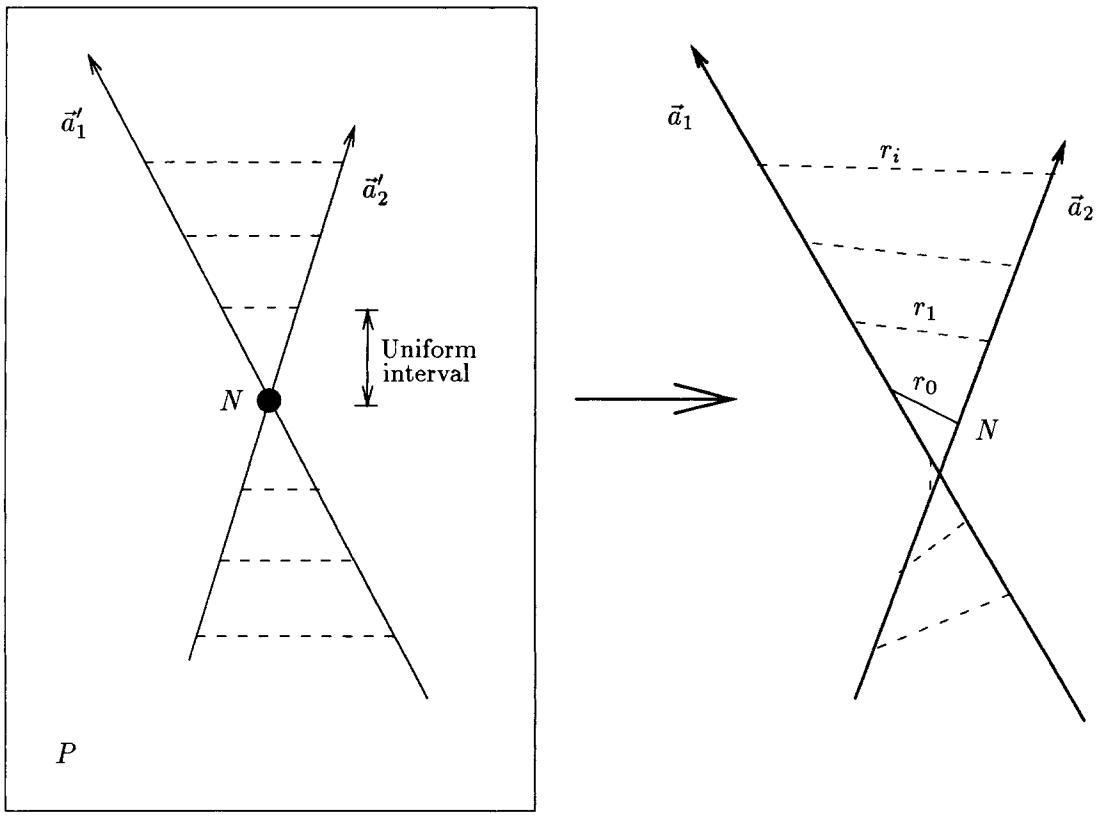


Fig. 4. The overlap measure. On the left, the projections $\vec{a}'_1$, $\vec{a}'_2$ of two axial vectors are shown in the projection plane *P*, as in Figure 3. A series of regularly spaced constructor lines proceed from the normal *N*, which passes up through the page. These constructors are projected back onto the original vectors $\vec{a}_1$, $\vec{a}_2$ on the right, to form a series of "rungs" whose lengths r_i, contribute to the overlap measure $olap = \Sigma_i 1/\text{length}(r_i)$.

added in a variant on the midpoint–midpoint vector. Two midpoint–endpoint vectors $\vec{v}_{me}^{N,C}$ were defined from the midpoint of the first structure to the N and C endpoints of the axial vector of the second structure (see Fig. 2).

Interaxial and tilt angles

Angles were measured in relation to a projection plane between a pair of axial vectors. Generally, a plane *P*, was defined normal to some constructor line *N*, connecting the two axial vectors so that the interaxial angle Ω, was defined by the angle of the vector projections in this plane. Interaxial angles were defined over ±180°, and used the sign convention such that the angle is negative if the near vector is rotated clockwise with respect to the far vector (see Fig. 3).

A second angle τ, also shown in Figure 3, described the tilt or divergence of the paired vectors out of the common plane. Tilt angles were defined as the sum of the tilts of the individual vectors with respect to the mutual projection plane, where each individual tilt was signed positive if the vector pointed outward from the common plane.

Alternative sets of interaxial and tilt angles were defined using three different constructor lines *N*: common perpendicular, closest approach distance, and midpoint–midpoint distance. The midpoint–midpoint distance gave the best results and the other projections will not be discussed further.

Buried interface area

The solvent accessible area lost on packing a secondary structure pair was estimated as the difference in areas between the isolated structures and the pair in situ, using the method described above.

Overlap

An experimental composite measure *olap* was designed to combine angular, distance, and length information for a pair of secondary structures. Given a projection plane and plane normal as illustrated in Figure 4, a set of lines was constructed linking adjacent arms of the vector images $\vec{a}'_1$, $\vec{a}'_2$ in the plane. The lines were evenly spaced (0.5 Å) and parallel so that, when projected back onto the axial vectors $\vec{a}_1$, $\vec{a}_2$, the projections formed the rungs of a ladder centered about the plane normal. The overlap measure was then calculated over all "rungs" r_i, as $olap = \Sigma_i 1/\text{length}(r_i)$. Increased separation along

TABLE I. Dataset for Secondary Structure Feature Parameter Optimizations*

Chain	Resolution (Å)	Description
α proteins		
1ECA	1.4	Hemoglobin (erythrocruorin, aquo Met), *Chironomus thummi thummi*
4HHB A B	1.74	Hemoglobin (deoxy), human
2LH4 A	2.0	Leghemoglobin (aquo, Met), yellow lupin
1MBN	2.0	Myoglobin (aquo, Met), sperm whale
2MHB A B	2.0	Hemoglobin (aquo, Met), horse
β proteins		
3FAB H L	2.0	λ-Immunoglobulin Fab', human
1FC1 H L	2.9	Fc fragment (Ig, g_1 class), human
Mixed α/β proteins		
4ADH	2.9	Apo-liver alcohol dehydrogenase, horse
4FXN	1.8	Flavodoxin (semiquinone form), *Clostridium mp.*
1GPD	2.9	D-Glyceraldehyde-3-phosphate dehydrogenase, lobster
5LDH	2.7	Lactate dehydrogenase, pig
4MDH	2.5	Cytoplasmic malate-dehydrogenase, pig

*Protein structures from the three classes of protein packing (α, β, α/β) taken from the Brookhaven Protein Databank (release January 1991) and used in optimizing the parameters which control the scores for secondary structure matching.

the connecting plane normal, shorter elements, and wide interaxial angle all operate to reduce the value of *olap*, which would be at a maximum for abutted, perfectly parallel/antiparallel, mutually long structures.

Optimization of Secondary Structure Parameters

In order to determine which properties and relationships are most useful in identifying equivalent secondary structures, the parameters for matching different features were optimized using a test set of proteins selected from the structure databank. These were chosen so as to represent the three main classes of protein composition and packing relationships: α, β, and mixed α/β (see Data section).

Individual features *f*, for secondary structures *i, j* in two proteins are scored in the lower level matrix [cf. Eq. (1)] as follows:

$$S^{f}_{sec}(k,l) = \frac{w}{a+\delta}, \quad \delta = |f(i,k) - f(j,l)| \qquad (2)$$

where w is the weight controlling the contribution of the feature to the total score, a is a constant, and δ is the difference in the feature between the two proteins from the perspectives of elements i and j when compared with elements k, l, respectively. A cutoff c, on the score S^{f}_{sec}, prevents low scores from being accumulated in the matrix.

Scores for combined features are expressed as a sum of weighted terms, for example, the total score $S^{\Sigma f}_{sec}$, for combined midpoint–midpoint vectors, angles, and overlap between secondary structures is, using the same indices:

$$S^{\Sigma f}_{sec}(k,l) = \frac{w\,\vec{\mathbf{v}}_{mm}}{a\vec{\mathbf{v}}_{mm} + \delta\vec{\mathbf{v}}_{mm}} + \frac{w_{\Omega,\tau}}{a_{\Omega,\tau} + \delta_{\Omega,\tau}} + \frac{w_{olap}}{a_{olap} + \delta_{olap}}$$

where

$$\delta_{\vec{\mathbf{v}}_{mm}} = |\vec{\mathbf{v}}_{mm}(i,k) - \vec{\mathbf{v}}_{mm}(j,l)|$$
$$\delta_{\Omega,\tau} = |\Omega(i,k) - \Omega(j,l) + \tau(i,k) - \tau(j,l)|$$
$$\delta_{olap} = |olap(i,k) - olap(j,l)|.$$

Each group of structures has characteristic features and favored packing arrangements, which may be conserved to different extents, so that the distributions of alignment scores may vary between classes. Optimal parameters were therefore determined for each class separately as well as for the whole group of test structures.

Combined Secondary Structure and Residue Alignment (SSAPc)

In the combined method, structures are first aligned by secondary structure matching so that for subsequent residue matching only the small subset of residues within the aligned secondary structures and having similar accessible areas and torsional angles is compared. This is similar to the fast residue alignment method of Orengo and Taylor,[4] in which the structural environments of only a small percentage of residues having similar structural locations between proteins (buried areas and dihedral angles) were compared.

Any errors occuring at the secondary structure stage are generally corrected when residues are used. Moreover, selection of residues by the secondary structure alignment results in an initial subset containing many equivalent residue pairs, which could be very small (< 0.1% of the total number of

TABLE II. Dataset for Secondary Structure Score (ln G_{sec}) Cutoff*

Chain	Resolution (Å)	Description
Calcium binding		
3CLN	2.2	Calmodulin, rat
5TNC	2.0	Troponin-C, turkey
Lyzosymes		
1ALC	1.7	α-Lactalbumin, baboon
1LZ1	1.5	Lysozyme, human
2LZ2	2.2	Lysozyme, turkey
1LZT	1.97	Lysozyme, hen
Cytochromes		
351C	1.6	Cytochrome c_{551} (oxidized), *Pseudomonas aeruginosa*
155C	2.5	Cytochrome c_{550}, *Paracoccus denitrificans*
3C2C	1.68	Cytochrome c_2 (reduced), *Rhodospirillum rubrum*
1CC5	2.5	Cytochrome c_5 (oxidized), *Azotobacter vinelandii*
1CCR	1.5	Cytochrome *c*, rice
1CYC	2.3	Ferrocytochrome *c*, bonito
3CYT	1.8	Cytochrome *c* (oxidized), albacore tuna
Globins		
2DHB A B	2.8	Hemoglobin (deoxy), horse
1ECA	1.4	Hemoglobin (erythrocruorin, aquo, Met), *Chironomus thummi thummi*
1FDH G	2.5	Hemoglobin (deoxy), human fetal
1HBS G H	3.0	Hemoglobin s (deoxy), human
1HDS A C	1.98	Hemoglobin (sickle cell), virginia white-tailed deer
4HHB A B	1.74	Hemoglobin (deoxy), human
2LH2 A	2.0	Leghemoglobin (aquo, Met), yellow lupin
1MBN	2.0	Myoglobin (ferric iron—metmyoglobin), sperm whale
1MBS	2.5	Myoglobin (Met), common seal
2MHB A B	2.0	Hemoglobin (aquo, Met), horse
1PMB B	2.5	Myoglobin (aquo, Met, pH 7.1), porcine
Hemerythrins		
1HMQ	2.0	Hemerythrin (Met), sipunculid worm
1HMZ D	2.0	Hemerythrin (azido, Met), sipunculid worm
2MHR	1.7	Myohemerythrin, sipunculan worm
Hydrolases		
2ACT	1.7	Actinidin (sulfhydryl proteinase), chinese gooseberry
1ACX	2.0	Actinoxanthin, *Actinomyces globisporus*
Copper binding		
2AZA	1.8	Azurin (oxidized), *Alcaligenes denitrificans*
1AZU	2.7	Azurin, *Pseudomonas aeruginosa*
1PAZ	1.55	Pseudoazurin (oxidized), *Alcaligenes faecalis*
1PCY	1.6	Plastocyanin, poplar leaves
Immunoglobulins		
1F19 H L	2.8	Fab fragment, mouse
3FAB H L	2.0	λ-Immunoglobulin Fab', human
2FB4 H L	1.9	Immunoglobulin Fab, human
2FBJ H L	1.95	Ig a Fab fragment (j539, galactan-binding), mouse
2HFL H L	2.54	Ig g_1 Fab fragment (hy/hel-5), mouse
3HFM H L	3.0	Ig g_1 Fab fragment (hy/hel-10), mouse
2IG2 H L	3.0	Immunoglobulin g_1, human
2MCP H L	3.1	Immunoglobulin mc/pc603 Fab-phosphocholine complex, mouse
1REI B	2.0	Bence–Jones immunoglobulin (REI variable portion), human
2RHE	1.6	Bence–Jones protein (λ, variable domain), human
Aspartyl proteases		
4APE	2.1	Acid proteinase (endothiapepsin), chestnut blight fungus
2APP	1.8	Acid proteinase (penicillopepsin), *Penicillium janthinellum*
2APR	1.8	Acid proteinase (rhizopuspepsin), bread mold
1CMS	2.3	Chymosin B (rennin), bovine
3PEP	2.3	Pepsin, porcine
1PSG	1.65	Pepsinogen, porcine

(continued)

TABLE II. Dataset for Secondary Structure Score (ln G_{sec}) Cutoff* (Continued)

Chain	Resolution (Å)	Description
Jellyroll virus proteins		
2PLV 1 3	2.88	Poliovirus (type 1, mahoney strain), human
1RMU 1	3.0	Rhinovirus 14 (mutant with Cys 1 199 replaced by 2 Tyr), human
2RMU 1	3.0	Rhinovirus 14 (mutant with Val 1 188 replaced by 2 Leu), human
2RS3 1 3	3.0	Rhinovirus 14 (complex with antiviral agent), human
Serine proteases (β-barrels)		
2CGA B	1.8	Chymotrypsinogen A, bovine
2CHA	2.0	α-Chymotrypsin A (tosylated), bovine
2GCH	1.9	γ-Chymotrypsin A, bovine
2KAI B	2.5	Kallikrein A, bovine
1NTP	1.8	β-Trypsin, bovine
1PAD	2.8	Papain, papaya
1TON	1.8	Tonin, rat
2TRM	2.8	Asn_{102}–Typsin (mutant with Asp_{102} replaced by Asn), rat
Alternating β/α domains		
6ADH	2.9	Holo-liver alcohol dehydrogenase, horse
7ADH	3.2	Alcohol dehydrogenase, horse
3ADK	2.1	Adenylate kinase, porcine
8CAT	2.5	Catalase, bovine
4DFR	1.7	Dihydrofolate reductase, *Escherichia coli*
1FX1	2.0	Flavodoxin, *Desulfovibrio vulgaris*
3FXN	1.9	Flavodoxin (oxidized form), *Clostridium mp.*
1GD1	1.8	Holo-D-glyceraldehyde-3-phosphate dehydrogenase, *Bacillus stearothermophilus*
2GD1	2.5	Apo-D-glyceraldehyde-3-phosphate dehydrogenase, *Bacillus stearothermophilus*
1GPD	2.9	D-Gyceraldehyde-3-phosphate dehydrogenase, lobster
3GPD	3.5	D-Glyceraldehyde-3-phosphate dehydrogenase, human
3LDH	3.0	Lactate dehydrogenase, dogfish
5LDH	2.7	Lactate dehydrogenase, porcine
8LDH	2.8	Apo-lactate dehydrogenase, dogfish
2LDX	2.96	Apo-lactate dehydrogenase isoenzyme, mouse
4MDH	2.5	Cytoplasmic malate dehydrogenase, porcine
1PFK	2.4	Phosphofructokinase, *Escherichia coli*
3PGK	2.5	Phosphoglycerate kinase, bakers' yeast
Hexokinases		
1HKG	3.5	Hexokinase A, yeast
2YHX	2.1	Hexokinase B, bakers' yeast
Serine proteases (alternating β/α domains)		
2PRK	1.5	Proteinase K, *Tritirachium album limber*
1SBC	2.5	Subtilisin Carlsberg (subtilopeptidase A), *Bacillus subtilis*
2SBT	2.8	Subtilisin novo, probably *Bacillus amyloliquifaciens*
2SNI E	2.1	Subtilisin novo (complex with chymotrypsin inhibitor), *Bacillus amyloliquifaciens*
Serine protease inhibitors		
2SEC I	1.8	*N*-Acetyl eglin-c (complex with subtilisin Carlsberg), *Hirudo medicinalis*
2SNI I	2.1	Chymotrypsin inhibitor (complexed with subtilisin novo), barley
1TEC I	2.2	Eglin-c (complex with thermitase), *Hirudo medicinalis*
4TPI I	2.2	Pancreatic trypsin inhibitor (complex with trypsinogen), bovine
Carbonic anhydrases		
1CA2	2.0	Carbonic anhydrase II (carbonate dehydratase), human
2CAB	2.0	Carbonic anhydrase form B (carbonate dehydratase), human
Phospholipases		
2BP2	3.0	Prophospholipase A_2, bovine
1P2P	2.6	Phospholipase A_2, porcine
3P2P	2.1	Phospholipase A_2 (mutant), porcine
1PP2	2.5	Phospholipase A_2 (Ca free), diamondback rattlesnake

*Groups of related protein structures taken from the Brookhaven Protein Databank (release January 1991), and used to determine the cutoff on secondary structure score for excluding unrelated protein pairs from further comparison.

TABLE III. Additions to Dataset Used in Cluster Analysis and Limited Searches*

Chain	Resolution (Å)	Description
156B	2.5	Cytochrome b_{562} (oxidized), *Escherichia coli*
1ABP	2.4	L-Arabinose-binding protein, *Escherichia coli*
2ABX B	2.5	α-Bungarotoxin, braided krait
2ALP	1.7	α-Lytic protease, *Lysobacter enzymogenes*
2B5C	2.8	Cytochrome b_5 (oxidized), bovine
3BCL	1.9	Bacteriochlorophyll-a protein, *Prosthecochloris aestuarii*
2CDV	1.8	Cytochrome c_3, *Desulfovibrio vulgaris*
2CHY	2.7	CheY, chemotaxis protein Y, *Salmonella typhimurium*
2CNA	2.0	Concanavalin A, jack bean
2CPP	1.63	Cytochrome P-450_{cam}, *Pseudomonas putida*
1CRN	1.5	Crambin, abyssinian cabbage
1CTF	1.7	Ribosomal protein (C-terminal domain), *Escherichia coli*
1CTX	2.8	α-Cobratoxin, cobra
2CYP	1.7	Cytochrome *c* peroxidase (ferrocytochrome *c*), bakers' yeast
3EBX	1.4	Erabutoxin B, sea snake
3EST	1.6	Elastase, porcine
1ETU	2.9	Elongation Factor Tu, *Escherichia coli*
3FXC	2.5	Ferredoxin, *Spirulina platensis*
4FXN	1.8	Flavodoxin (semiquinone form), *Clostridium mp.*
3GAP A	2.9	Catabolite gene activator protein–cAMP complex, *Escherichia coli*
2GBP	1.9	D-Galactose/D-glucose binding protein, *Escherichia coli*
1GCN	3.0	Glucagon (pH 6–pH 7 form), porcine
1GCR	1.6	γ-II crystallin, calf
2GN5	2.3	Gene 5 DNA binding protein, filamentous bacteriophage (m13)
1GOX	2.0	Glycolate oxidase, spinach
1HHO A B	2.1	Hemoglobin A (oxy), human
3HLA A	2.6	Human class I histocompatibility antigen a2.1
2HLA A	2.6	Human class I histocompatibility antigen aw68.1
1HMG A	3.0	Hemagglutinin, influenza virus
3ICB	2.3	Calcium-binding protein, bovine
2LBP	2.4	Leucine-binding protein, *Escherichia coli*
2LHB	2.0	Hemoglobin (cyano, Met), sea lamprey
2LIV	2.4	Leucine/isoleucine/valine-binding protein, *Escherichia coli*
1NXB	1.38	Neurotoxin B, sea snake
2OVO	2.0	Ovomucoid third domain, silver pheasant
2PAB A	1.8	Prealbumin, human
3PGM	2.8	Phosphoglycerate mutase, bakers' yeast
2PKA A	2.05	Kallikrein A, *Sus scrofa*
1PPD	2.0	2-Hydroxyethylthiopapain, papaya
5PTI	1.0	Trypsin inhibitor, bovine
1RHD	2.5	Rhodanese, bovine
4RHV 1	3.0	Rhinovirus 14, human
1RN3	1.45	Ribonuclease A, bovine
3RP2 A	1.9	Mast cell protease II, rat
5RXN	1.2	Rubredoxin (oxidized, Fe_{III}), *Clostridium pasteurianum*
4SBV A	2.8	Southern bean mosaic virus coat protein, cow pea strain
2SGA	1.5	Proteinase A, *Streptomyces griseus*
3SGB E	1.8	Proteinase B, *Streptomyces griseus*
2SNS	1.5	Staphylococcal nuclease, *Staphylococcus aureus*
2SOD O	2.0	Cu,Zn superoxide dismutase, bovine
2STV	2.5	Coat protein, satellite tobacco necrosis virus
2TAA A	3.0	Taka-amylase A, *Aspergillus oryzae*
1TIM A	2.5	Triose phosphate isomerase, chicken
3TLN	1.6	Thermolysin, *Bacillus thermoproteolyticus*
1TPO	1.7	β-Trypsin (orthorhombic) at pH 5.0, bovine pancreas
1UBQ	1.8	Ubiquitin, human

*Structures from the Brookhaven Protein Databank (release January 1991) added to those listed in Table II. The combined set was used to generate structurally related groupings by single linkage cluster analysis on the residue alignment scores.

TABLE IV. Optimal Parameters* for Different Local Reference Frames†

Constructor	Class	w/10	a	c	lc	Correct (%)
Midpoint–	α	100	5	10	10	100
next midpoint	β	200	5	10	10	87
vector	α/β	80	5	5	10	96
Vector of	α	100	5	10	10	93
hydrophobic	β	140	5	10	10	69
moment 1: $\vec{\mu}_{cyl}$	α/β	100	5	10	10	82
Vector of	α	100	5	10	10	94
hydrophobic	β	150	8	8	10	72
moment 2: $\vec{\mu}_{sph}$	α/β	40	5	4	10	64

*The best alignments within all three protein classes were obtained using the vector $\vec{v}_{mm}$, from the element midpoint to the midpoint of the next element along the chain, to construct the local reference frame.
†Tables IV–VIII show the optimal parameters obtained for matching different secondary structure features between proteins. Results for the three classes of protein packing (α, β, α/β) are listed separately within the tables, along with the percentages of equivalent secondary structures correctly aligned.

TABLE V. Optimization of Parameters* for Midpoint–Midpoint Vector ($\vec{v}_{mm}$) Comparison†

	Parameter											
Class	w/10			a			c			lc		
α												
Value	60	**100**	140	2	**5**	8	5	**10**	15	5	**10**	25
%	99	100	98	99	100	100	99	100	100	99	100	100
β												
Value	120	**200**	280	2	**5**	8	4	**10**	15	6	**10**	50
%	62	87	54	82	87	74	62	87	69	87	87	85
α/β												
Value	30	**40**	60	2	**5**	6	3	**5**	15	6	**10**	12
%	N/A	95	84	78	95	91	89	96	84	91	96	95

*The starting set of parameters was $w/10 = 50$, $a = 5$, $c = 10$, and $lc = 10$. Parameters were varied one at a time in the sequence w, a, c, lc with the optimal values (shown in bold) replacing those in the above set. The vector to the midpoint of the next consecutive segment was used to define the local frame of reference.
†See Table IV for general notes.

TABLE VI. Optimal Parameters* for Interaxial (Ω) and Tilt (τ) Angles

y-Axis choice	Class	w/10	a	c	lc	Correct (%)
Tilt/interaxial	α	180	5	10	20	98
angles:	β	250	5	20	10	74
Ω, τ	α/β	60	5	8	10	75

*See Table IV for general notes.

residue comparisons). The optimal thresholds for differences in accessible areas and torsional angles were established by trials.

The structures are recompared using this small set of selected residue pairs. The residue score matrix from this alignment is further analyzed and the 20 highest scoring residue pairs are then recompared, giving the final alignment of the proteins (see Fig. 1).

Normalization of Alignment Scores

As described above, the final alignment is generated only for equivalent residue positions in the two structures. As well as removing noise caused by comparing nonequivalent positions, this also removes any amplification of local score occurring in secondary structure regions. This can be caused by the repetitive nature of secondary structures, which

TABLE VII. Optimal Parameters* for Combined Midpoint–Midpoint Vectors and Various Features†

Features matched	Class	$w/10$	a	c	lc	Correct (%)
Midpoint	α	100	5	10	10	100
vectors alone:	β	200	5	10	10	87
$\vec{v}_{mm}$	α/β	40	5	5	10	96
Vectors and	α	20	5	10	10	100
angles:	β	350	5	10	10	92
$\vec{v}_{mm}$, Ω, τ	α/β	24	5	10	10	98
Vectors and	α	12	5	6	10	100
buried areas:	β	10	5	10	10	87
$\vec{v}_{mm}$, *area*	α/β	8	8	10	15	98
Vectors and	α	10	5	10	10	100
overlap:	β	8	4	10	10	90
$\vec{v}_{mm}$, *olap*	α/β	6	5	10	10	96
Vectors and	α	2	5	5	10	100
hydrophobicities:	β	4	6	5	10	87
$\vec{v}_{mm}$, H_{sum}	α/β	10	5	0	10	95
Vectors and	α	6	6	10	10	100
hydrophobic moment:	β	6	5	5	10	87
$\vec{v}_{mm}$, $\vec{\mu}_{cyl}$	α/β	6	6	10	10	96
Vectors and	α	20	5	5	10	100
distances:	β	12	4	5	10	90
$\vec{v}_{mm}$, $\lvert\vec{v}_{mm}\rvert$	α/β	10	5	5	10	96
Vectors and	α	8	5	5	20	99
axis lengths:	β	10	5	5	20	90
$\vec{v}_{mm}$, $\lvert\vec{a}\rvert$	α/β	3	6	5	15	96

*Optimal parameters for midpoint vectors are given in the first row, while subsequent rows show the optimal parameters for different features tested in conjuction with vectors. Vector parameters were held constant while the parameters for other features were optimized.
†See Table IV for general notes.

results in good matches of local vectors, even between nonequivalent residues. Therefore, any magnification due to a local similarity component should be excluded from the final score used to express overall structural similarity.

Given the final residue alignment, the alignment path is noted and the upper matrix reset. A new upper matrix of scores is then calculated only for the residue pairs along the final alignment path by recomparing them and scoring and evaluating the corresponding lower level matrices. Scores on these lower level paths are accumulated in corresponding cells along the original upper path and the overall alignment score is calculated by summing along the latter, charging gap penalties for any insertions/deletions.

An overall final alignment score S_{elem} (either S_{sec} or S_{res}), is then normalized to take account of protein size in two ways. By dividing by the total number of element comparisons l, performed between the two proteins, an average alignment score A_{elem}, is obtained:

$$A_{elem} = S_{elem}/l.$$

Alternatively, the final score is divided by the square of the size of the smaller of the two proteins being compared (less the number of comparisons of elements with self since these are omitted) to give a better measure G_{elem}, of the global structural similarity:

$$G_{elem} = \frac{S_{elem}}{s(s-1)}, \quad s = \min(m,n)$$

where m, n are the sizes (number of elements) of the two protein chains.

The less structurally similar the proteins, the fewer the comparisons which would be performed between them, as selection criteria would ensure that only elements of similar secondary structure type or in similar structural environments are compared (see above). Therefore, the average score reflects the similarity between corresponding structural regions, regardless of the proportions these regions occupy in the whole structure, and may help to indicate local similarities between proteins. The global score, however, takes into account the overall structural similarity and is therefore a more useful general measure. If the smaller protein is identical to a fragment of the larger, the two scores will be the same.

TABLE VIII. Optimal Parameters* for Combined Features Across Protein Classes†

Features matched	Class	$w/10$	a	c	lc	Correct (%)
Midpoint	α					100
vectors alone:	β	80	6	10	10	71
$\vec{v}_{mm}$	α/β					84
Vectors and	α					99
angles:	β	60	6	10	10	82
$\vec{v}_{mm}$, Ω, τ	α/β					93
Vectors, angles	α					100
and distances:	β	10	5	10	10	80
$\vec{v}_{mm}$, Ω, τ, $\lvert\vec{v}_{mm}\rvert$	α/β					87
Vectors, angles	α					100
and overlap	β	8	5	10	10	90
$\vec{v}_{mm}$, Ω, τ, *olap*	α/β					90

*The table shows the effectiveness of parameters optimized over all classes when applied to the individual classes. Optimal parameters for midpoint vectors are given in the first row, while subsequent rows show the optimal parameters for features tested in conjunction with vectors. The best alignments were generated by combining midpoint vector matching with both angles and overlap matching.
†See Table IV for general notes.

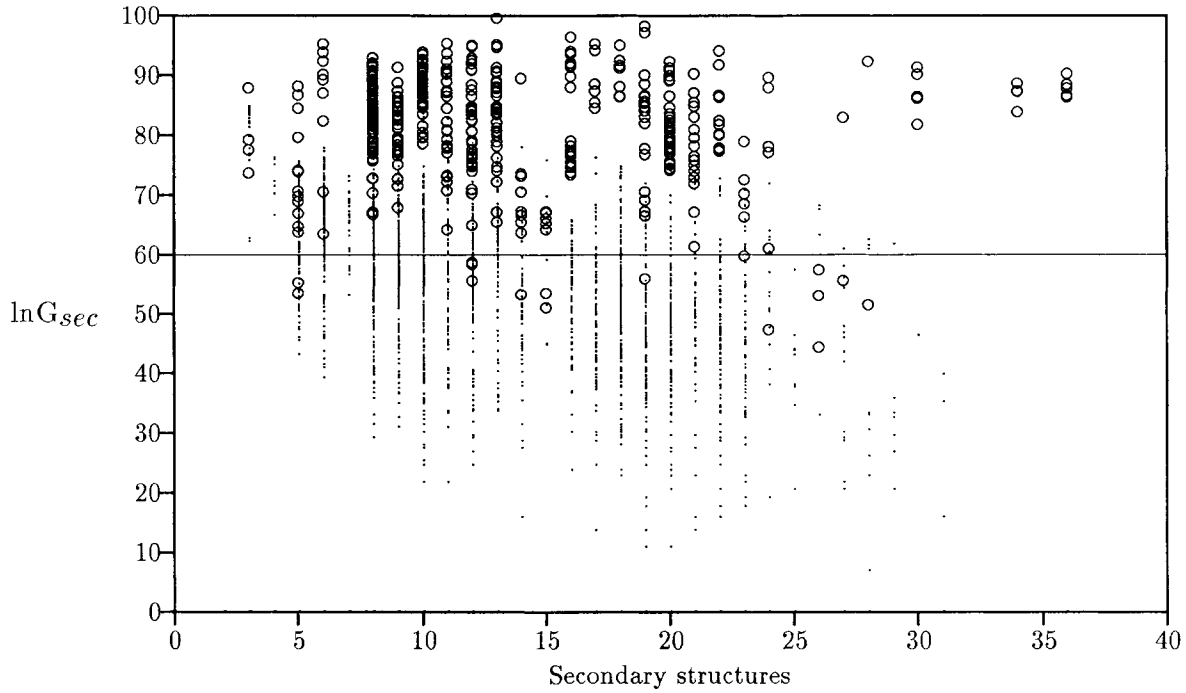


Fig. 5. Secondary structure alignment score (ln G_{sec}) against protein size (number of secondary structures) for related (○) and unrelated (·) protein structures (542 related, 5,151 unrelated comparisons). On the y-axis 100 indicates complete structural identity. The horizontal line drawn at ln G_{sec} = 60 marks the cutoff on the secondary structure score used for separating related and unrelated proteins.

Superposition of Aligned Structures

Once the residue alignment of a pair of proteins has been determined, superposition of the structures may be performed using the method of Rippmann and Taylor.[21] This generates a weighted superposition of the aligned residues, where weights are given by the individual residue pair alignment scores derived by SSAPc. After superposition, the weighted

root mean square (rms) deviation between the structures is calculated over the aligned residues using the same weights.

Programs

The combined secondary structure and residue alignment program (SSAPc) was written in C and runs under UNIX (SunOS 4.1 on Sun 4/280 and SS-1). The code used to generate solvent accessibilities was an implementation of Lee and Richards'[18] accessibility program from Birkbeck College, University of London. Originally written in Fortran 66 to run under VMS, it was ported with minor modifications to run under UNIX for use in this work.

Data

Protein coordinates were taken from the Brookhaven Protein Databank,[22] release January 1991.

Parameter optimization dataset

To test the secondary structure alignment method and optimize parameters, sets of structures were selected from the Brookhaven databank containing representative proteins from each of the three protein packing classes (all α, all β, and α/β). The correct alignments of these were known and some comparisons involved remote structures. Seven globin structures were taken from the α class, four immunoglobulins from the β class, and five structures containing the dinucleotide binding Rossmann-fold from the α/β class (see Table I for a listing).

Structure searching/sorting dataset

In order to establish lower limits on the secondary structure and residue alignment scores for related proteins, a subset of the databank was used, as indicated in Table II. For testing databank searches, all the structures in the Brookhaven databank of 3.0 Å resolution or better were used. A subset of 167 structures (Tables II and III) was used for testing the ability of the algorithm to sort the databank into related groups.

RESULTS AND DISCUSSION

Secondary Structure Alignment Method

For each of the secondary structure features described in the Methods section there are four parameters which adjust the scores. Three of these, *w, a, c,* control the values of scores in the lower level matrix, while the fourth parameter, lowcut (*lc*), is a threshold preventing low scoring pathways from being accumulated in the upper level matrix. Optimal values of the parameters used in matching each feature were determined.

Parameters were initially obtained for each protein class (α, β, α/β) separately, in order to establish if there were any differences in the degrees to which features were important and therefore conserved for a particular protein class.

As in earlier optimizations[3] parameters were varied systematically and, for each pair of structures compared, the percentage of secondary structures misaligned was recorded. Each parameter was varied independently and, once the optimum was found, the value was fixed during optimization of successive parameters. For each parameter, up to 10 values were tested, depending on the sensitivity of the alignments to the parameter. Total percentage misalignments within each protein class are listed in Tables IV–VIII for different parameters, together with ranges of the values tried and the optimal value obtained.

Preliminary tests showed that the information contained in individual properties of the secondary structures (e.g., accessibility, hydrophobicity, see Methods above) was not sufficient to give good alignments when these properties were considered alone. Similarly, relationships describing packing (buried areas and overlap) also performed poorly. On the other hand, geometric relationships, such as the vectors between midpoints/endpoints of segments and also angles between segments, contained sufficient spatial information to identify equivalent secondary structure segments.

Parameters for matching these relationships were therefore optimized first. Then further trials were performed, to establish the effect of using the most successful of the geometric relationships in combination with other features.

Optimal parameters for matching geometric features

For all classes, matching secondary structures by comparison of vectors between segment midpoints ($\vec{v}_{mm}$) gave the best alignments (Tables IV and V), and for the α set, correct alignments were obtained for all comparisons. When midpoint to endpoint vectors $\vec{v}_{me}^{N,C}$ were used, the alignments deteriorated, presumably due to variations in the lengths of the secondary structures in the different proteins.

Among the different local coordinate reference frames used to measure vectors between midpoints, the best alignments were for frames which used the vector to the midpoint of the next segment along the chain (see Table IV). Particularly for the α/β and the β class, frames using hydrophobic moments to construct the frame gave poor alignments. In these sets, buried β-strands have hydrophobic residues on both sides of the sheet, so that the moments generated are small and weakly conserved. There was little difference in alignments for frames generated using different hydrophobic moments.

The success of the vectors is in agreement with similarly good alignments obtained by matching vectors between residues.[2] Alignments in the β class were the least successful of the three classes. These structures are the most repetitive and

TABLE IX. Structurally Similar Pairs That Fall Below the ln G_{sec} Cutoff*

Pair		ln G_{sec}	Pair		ln G_{sec}
3LDH	1PFK B	59.8	3CYT O	351C	53.5
1FX1	3GPD G	58.7	2GD1 R	3ADK	53.5
1FX1	1PFK B	58.5	4DFR A	3ADK	53.3
1FX1	4DFR A	58.4	1GPD R	1PFK B	53.1
7ADH	1PFK B	57.4	2GD1 R	7ADH	51.5
2LDX	1PFK B	56.0	1GPD R	3ADK	51.1
1GPD R	7ADH	55.7	3GPD G	7ADH	47.4
1FX1	2GD1 R	55.7	2GD1 R	1PFK B	44.4
3C2C	351C	55.3			

*Pairs of proteins (Brookhaven codes and chain names) expected to show some structural relationship, but which score below the ln G_{sec} = 60 cutoff on the secondary structure alignment score (17 out of 542 total).

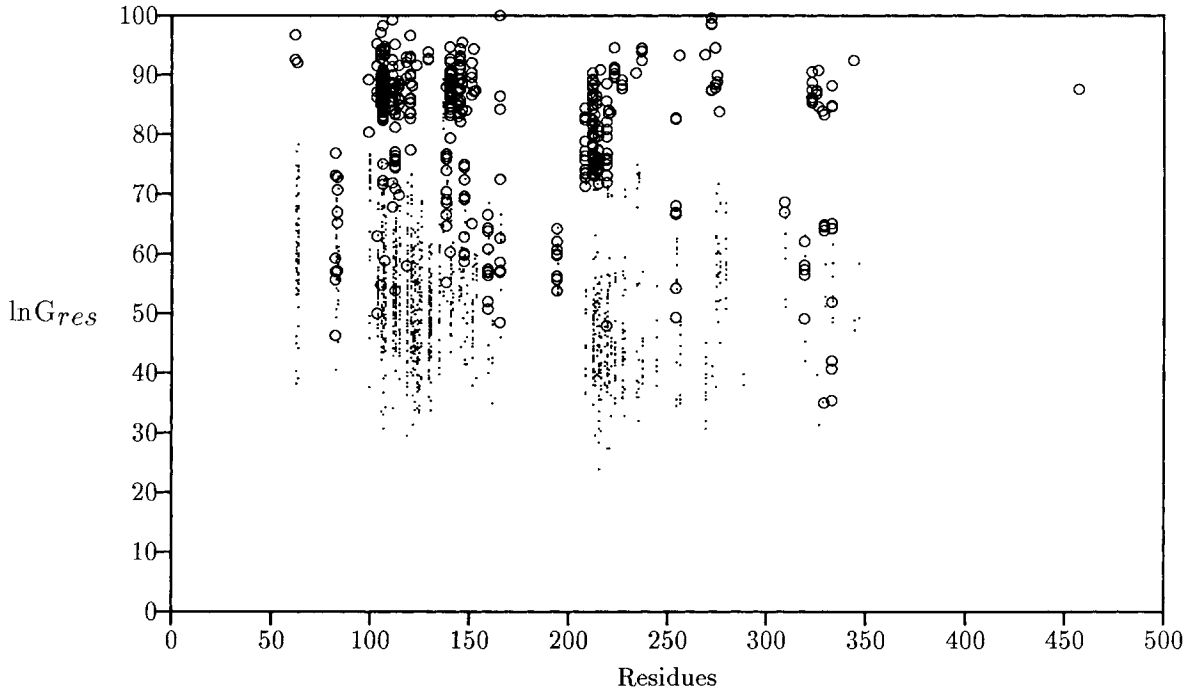


Fig. 6. Residue alignment scores (ln G_{res}) against protein size (number of residues) for comparisons between related (○) and unrelated (·) protein structures (542 related, 5,151 unrelated comparisons). On the *y*-axis 100 indicates complete structural identity. The slight decrease in score with protein size reflects the fact that larger proteins generally contain more than one domain, so that global similarity is often reduced.

contain many elements in fairly similar environments, which are hard to resolve using any of the matching options.

Comparing tilt and interaxial angles (Table VI) gave reasonable alignments for the α proteins. There are generally fewer secondary structure elements in this set and they occupy distinct positions in the structure. In the α/β set, strands in the β-sheet make similar angles with each other and with helices packing against the sheet, which confuses the alignment. Similarly for the all β proteins.

Improvements by combining vectors with other features

Optimal parameters for matching other secondary structure properties and relationships in combination with vectors between element midpoints are summarized in Table VII.

Including comparisons of tilt/interaxial angles between elements considerably improved the alignments of the β proteins. Angles contain extra information about orientations of the β-strands not provided by midpoint vectors alone.

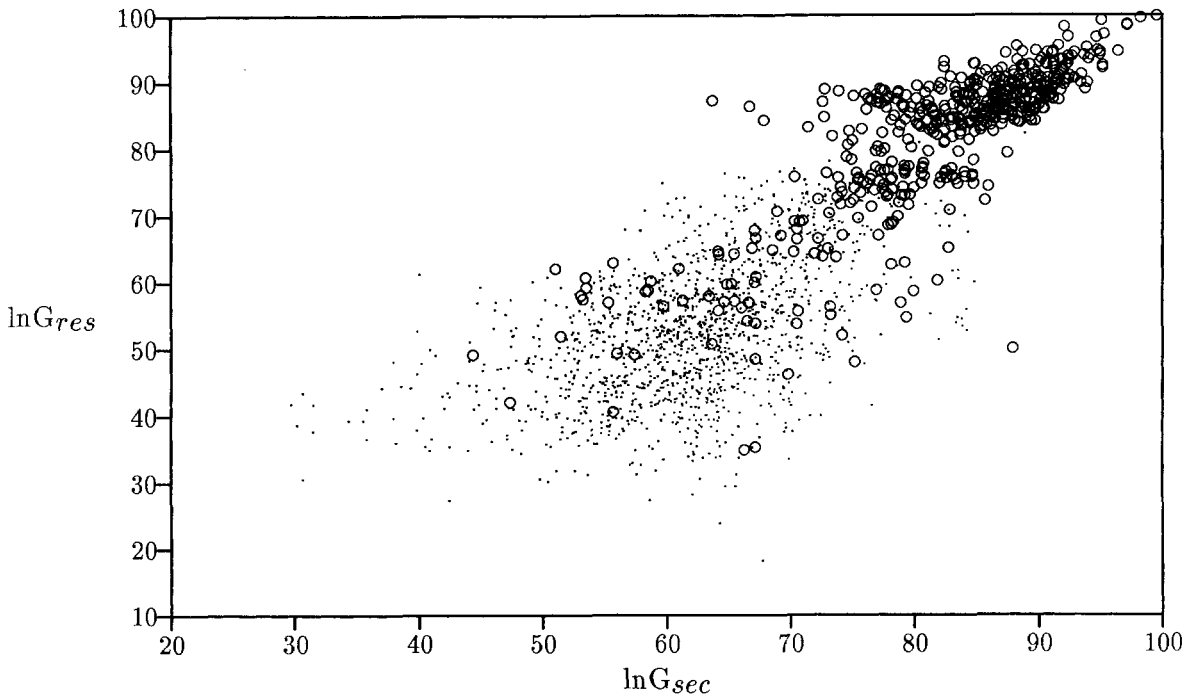


Fig. 7. Correlation of secondary structure and residue structural alignment score (ln G_{sec} and ln G_{res}, respectively) for related (○) and unrelated (·) protein structures (542 related, 5,151 unrelated comparisons). The correlation improves with greater structural similarity as shown by the correlation coefficients for the sets (related: 0.812 cf. unrelated: 0.538) and the relative displacement of the related set to the upper right.

Slight improvements were observed for the β class by also matching overlap of the elements, but not other features. It was thought that overlap, in particular, would give more substantial improvements. However, there may not be sufficient additional information in the geometry of the overlap measure to aid the comparison significantly.

Summarizing the results of Table VII, midpoint vectors together with tilt/interaxial angles are the features most powerful in identifying equivalent secondary structure elements and they give optimal alignments for all three classes of protein tested.

Optimal parameters for comparing the whole structure databank

Since alignments based on matching midpoint vectors and tilt/interaxial angles were successful for all classes, the parameters used to score the comparison of these features were reoptimized using all the structures in the test set as a single group. This gave a set of parameters for a databank wide comparison of protein structures.

Alignment scores and parameter values are shown in Table VIII. The contribution of vector matching and angle matching has to be balanced, as increasing the weight on the angles relative to midpoint vectors improves the alignments of the β class proteins, but impairs the α/β alignments. Further slight improvements were seen by including overlap. A final parameter set for using combined midpoint vectors/angles/overlap matching is shown in Table VIII.

Combined secondary structure and residue alignment method

In order to use the secondary structure alignment to guide a residue alignment, residue pairs are selected for comparison which are within the aligned secondary structures and have similar structural environments. Using the same set of test structures as above, trials showed that a cutoff = $(\delta_{area} + \delta_{\Omega,\tau})$ = 100, on the difference in structural environments, gave optimal residue alignments across the three protein classes considered.

Databank Search and Analysis

Cutoffs on secondary structure and residue scores for related proteins

As a first step to reducing databank search time, it is reasonable to exclude the comparison of predominantly α-helical structures with those which are predominantly β-type. Therefore, the compositions of residues in α and β secondary structures were calculated for each protein used in the search, for example:

$$\text{composition}_{\alpha} = \frac{\text{No. } \alpha\text{-helix residues}}{\text{No. secondary structure residues } (\alpha, \beta, 3_{10})} \times 100\%.$$

TABLE X. Search With 1PCY Showing the Highest Scoring* Comparisons†

Chain	Description	ln G_{sec}	ln G_{res}	Match	Match/ probe (%)	Match/ target (%)	Target/ probe
1PCY	Plastocyanin	100.0	**100.0**	99	100.0	100.0	1.0
2PCY	Plastocyanin	99.8	**99.2**	99	100.0	100.0	1.0
5PCY	Plastocyanin	94.2	**99.1**	99	100.0	100.0	1.0
6PCY	Plastocyanin	92.1	**99.0**	99	100.0	100.0	1.0
4PCY	Plastocyanin	92.7	**99.0**	99	100.0	100.0	1.0
3PCY	Plastocyanin	98.6	**98.7**	99	100.0	100.0	1.0
7PCY	Plastocyanin	91.8	**95.4**	97	99.0	98.0	1.0
1PAZ	Azurin	86.5	**89.6**	89	89.9	74.2	1.2
2PAZ	Azurin	87.6	**89.1**	89	89.9	72.4	1.2
2AZA B	Azurin	80.3	**82.3**	91	91.9	71.1	1.3
2AZA A	Azurin	82.1	**81.9**	91	91.9	70.5	1.3
1AZU	Azurin	82.2	**80.9**	90	90.9	72.6	1.3
3HFM H	Immunoglobulin	64.6	**76.6**	85	85.9	39.5	2.2
2FVB H	Immunoglobulin	67.8	**76.2**	81	81.8	68.6	1.2
2HFL H	Immunoglobulin	71.4	**75.9**	81	81.8	38.0	2.2
1FVB H	Immunoglobulin	72.8	**75.9**	81	81.8	68.6	1.2
2FBJ H	Immunoglobulin	71.1	**75.8**	81	81.8	37.0	2.2
2FVW H	Immunoglobulin	70.9	**75.6**	81	81.8	67.5	1.2
2MCP H	Immunoglobulin	66.2	**75.5**	87	87.9	39.4	2.2
1FVW H	Immunoglobulin	73.9	**75.5**	81	81.8	67.5	1.2
3FAB H	Immunoglobulin	66.4	**75.4**	86	86.9	39.3	2.2
2IG2 H	Immunoglobulin	62.9	**75.4**	87	87.9	36.7	2.4
2FB4 H	Immunoglobulin	59.6	**74.9**	86	86.9	37.7	2.3
1F19 H	Immunoglobulin	61.3	**74.9**	81	81.8	37.0	2.2
2FVB L	Immunoglobulin	64.7	**74.6**	86	86.9	81.1	1.1
1REI A	Immunoglobulin	75.9	**74.1**	85	85.9	79.4	1.1
3HFM L	Immunoglobulin	74.7	**74.0**	84	84.8	39.3	2.2
2FVW L	Immunoglobulin	73.8	**74.0**	85	85.9	75.9	1.1
2FBJ L	Immunoglobulin	69.1	**73.7**	83	83.8	39.0	2.2
1REI B	Immunoglobulin	74.4	**73.6**	84	84.8	79.2	1.1
2FB4 L	Immunoglobulin	69.8	**73.4**	88	88.9	40.7	2.2
3FAB L	Immunoglobulin	68.6	**73.2**	87	87.9	41.8	2.1
2RHE	Immunoglobulin	70.7	**73.2**	84	84.8	73.7	1.2
1FVW L	Immunoglobulin	75.0	**73.1**	82	82.8	73.2	1.1
4FAB L	Immunoglobulin	73.7	**72.7**	80	80.8	36.5	2.2
2IG2 L	Immunoglobulin	68.4	**72.7**	88	88.9	40.7	2.2
1TNF B	Lymphokine	67.8	**72.2**	91	91.9	59.9	1.5
1MCP L	Immunoglobulin	68.4	**71.9**	88	88.9	40.0	2.2
2HFL L	Immunoglobulin	71.0	**71.6**	86	86.9	40.6	2.1
1HFM L	Immunoglobulin	76.8	**71.6**	84	84.8	78.5	1.1
2MCP L	Immunoglobulin	72.9	**71.1**	88	88.9	40.0	2.2
2SOD Y	Superoxide dismutase	60.8	**71.0**	92	92.9	60.9	1.5
2HFM L	Immunoglobulin	73.6	**71.0**	84	84.8	78.5	1.1
2MCG 1	Immunoglobulin	64.3	**70.9**	88	88.9	40.7	2.2
2SOD G	Superoxide dismutase	62.8	**68.2**	92	92.9	62.2	1.5

*Only the first 45 comparisons are shown for brevity.
†Tables X–XIII list alignments of searches with different target proteins against the whole Brookhaven databank. The searches were ranked by ln G_{res} (in bold), and only list comparisons scoring ln $G_{res} \geq 60$ and where at least 80% of the probe protein has been aligned. Columns give the global log secondary structure score (ln G_{sec}), global log residue score (ln G_{res}), no. aligned residues (match), percent probe matched (match/probe), percent target matched (match/target), and ratio of target to probe sizes (target/probe).

Proteins with ≥ 90% residues in α are not compared with those containing ≥ 90% residues in β.

A fast databank search can be conducted by simply considering the secondary structures. Then any potentially related proteins giving high scores can be recompared by residue matching. Finally, high scoring residue alignments can be used to generate a superposition of the structures and calculate the root mean square deviation, if required. For this purpose cutoffs or lower limits were established on both the secondary structure and the residue alignment scores expected for related proteins.

Sets of related structures were identified using sequence alignments and information from the litera-

TABLE XI. Search With 3CYT Chain O*

Chain	Description	ln G_{sec}	ln G_{res}	Match	Match/ probe (%)	Match/ target (%)	Target/ probe
3CYT O	Cytochrome *c*	100.0	**100.0**	103	100.0	100.0	1.0
3CYT I	Cytochrome *c*	96.7	**97.6**	102	100.0	99.0	1.0
5CYT R	Cytochrome *c*	90.9	**97.2**	103	100.0	100.0	1.0
1CCR	Cytochrome c	94.6	**95.7**	103	100.0	92.8	1.1
2C2C	Cytochrome *c*	89.7	**92.3**	100	97.1	89.3	1.1
3C2C	Cytochrome *c*	92.5	**92.1**	100	97.1	89.3	1.1
155C	Cytochrome c_{550}	80.8	**87.9**	99	96.1	73.9	1.3
1CYC	Ferrocytochrome *c*	91.4	**85.9**	103	100.0	100.0	1.0
1COH C	Hemoglobin	77.1	**71.5**	87	84.5	61.7	1.4
2HHB C	Hemoglobin	71.3	**71.3**	87	84.5	61.7	1.4
2MHB A	Hemoglobin	77.1	**71.2**	84	81.6	59.6	1.4
2HCO A	Hemoglobin	76.9	**71.2**	87	84.5	61.7	1.4
4HHB A	Hemoglobin	78.5	**71.0**	87	84.5	61.7	1.4
3HHB A	Hemoglobin	77.1	**71.0**	87	84.5	61.7	1.4
1HCO A	Hemoglobin	77.3	**70.7**	86	83.5	61.0	1.4
1COH A	Hemoglobin	77.2	**70.3**	87	84.5	61.7	1.4
1HBS B	Hemoglobin	73.2	**69.7**	91	88.3	62.3	1.4
2HHB A	Hemoglobin	72.0	**69.0**	92	89.3	65.2	1.4
2MHB B	Hemoglobin	66.5	**68.8**	91	88.3	62.8	1.4
2DHB A	Hemoglobin	76.0	**67.9**	86	83.5	61.0	1.4
4HHB C	Hemoglobin	71.3	**67.7**	86	83.5	61.0	1.4
1HDS A	Hemoglobin	61.5	**67.7**	87	84.5	61.7	1.4
1HBS E	Hemoglobin	74.2	**67.6**	82	79.6	58.2	1.4
1HDS C	Hemoglobin	68.1	**67.1**	91	88.3	64.5	1.4
2CTS	Citrate synthase	64.0	**66.5**	94	91.3	21.5	4.2
2PRK	Pyruvate kinase	66.4	**64.8**	89	86.4	31.9	2.7
1PRC M	Photosynthetic reaction center	65.1	**64.2**	89	86.4	27.6	3.1
4TLN	Thermolysin	70.2	**63.8**	99	96.1	31.3	3.1
5TMN E	Thermolysin	70.2	**62.1**	92	89.3	29.1	3.1
2SNI E	Subtilisin	69.4	**60.9**	90	87.4	32.7	2.7
5TLN	Thermolysin	70.2	**60.6**	89	86.4	28.2	3.1
1HMG D	Hemagglutinin	66.3	**60.1**	83	80.6	47.4	1.7

*Refer to Table X for details.

ture regarding protein families (see Table II). Any proteins having greater than 30% sequence identity were taken to be related.

Structure comparisons were performed both within and between groups of related structures. The distribution of alignment scores for different protein sizes suggested that the logarithm of the scores should be used. This was largely caused by using vectors to compare element separations. Local vectors appear more similar than remote vectors and the effect is amplified as the similarity between proteins is decreased. Therefore, scores for aligning proteins with similar sizes but varying similarity show greater variation than if distances between elements were compared.

In order to reduce this range of alignment scores, new logarithmic alignment scores were defined as follows:

$$\ln A_{elem} = \log_e(A_{elem})/\log_e(\max S_{elem}) \times 100$$

and

$$\ln G_{elem} = \log_e(G_{elem})/\log_e(\max S_{elem}) \times 100$$

where max S_{elem} is the maximum score [$\log_e(\max S_{sec}) = 10.12$, $\log_e(\max S_{res}) = 10.82$] observed for complete structural identity.

Figure 5 plots the log alignment score ln G_{elem}, against the sizes of proteins being compared. It can be seen that there is no dependence on the size. Also, the majority (90%) of comparisons between related proteins yield scores greater than 60. Those below this value were between remote members of the nucleotide binding family (e.g., 1FX1 and 3GPD, see Table IX). The scores showed no dependence on the classes of proteins being compared.

A cutoff of 60 on the secondary structure score prevented 75% of comparisons between unrelated proteins being pursued to residue alignments. Among the 25% remaining, a large number were between groups of structures with some degree of similarity, for example, the subtilisins and the nucleotide-binding domains, both of which contain a

TABLE XII. Search With 4FXN Showing the Highest Scoring* Comparisons†

Chain	Description	ln G_{sec}	ln G_{res}	Match	Match/ probe (%)	Match/ target (%)	Target/ probe
4FXN	Flavodoxin	100.0	**100.0**	138	100.0	100.0	1.0
3FXN	Flavodoxin	97.6	**97.1**	138	100.0	100.0	1.0
1FX1	Flavodoxin	88.0	**89.0**	137	99.3	93.2	1.1
8LDH	Lactate dehydrogenase	81.3	**78.5**	122	88.4	37.1	2.4
1LDM	Lactate dehydrogenase	81.4	**78.3**	122	88.4	37.1	2.4
6LDH	Lactate dehydrogenase	81.3	**78.2**	125	90.6	38.0	2.4
1LLC	Lactate dehydrogenase	80.9	**78.0**	124	89.9	38.8	2.3
2LDX	Lactate dehydrogenase	65.5	**77.9**	125	90.6	37.8	2.4
1LDB	Lactate dehydrogenase	82.2	**77.6**	122	88.4	41.5	2.1
2CHY	CheY chemotaxis protein	83.2	**77.4**	109	86.5	79.0	1.1
4MDH B	Malate dehydrogenase	81.9	**76.6**	122	88.4	36.7	2.4
2LDB	Lactate dehydrogenase	82.0	**76.5**	125	90.6	41.5	2.2
4MDH A	Malate dehydrogenase	81.7	**76.2**	125	90.6	37.5	2.4
1AT1 A	Aspartate transcarbamylase	72.3	**75.1**	116	84.1	37.4	2.2
2PFK C	Phosphofructokinase	77.6	**74.9**	125	90.6	41.5	2.2
7AT1 A	Aspartate transcarbamylase	72.7	**74.6**	124	89.9	40.0	2.2
1SBC	Subtilisin	77.2	**73.8**	122	88.4	44.5	2.0
5LDH	Lactate dehydrogenase	74.4	**73.3**	118	85.5	35.4	2.4
6AT1 C	Aspartate transcarbamylase	77.9	**73.2**	124	89.9	40.0	2.2
2SBT	Subtilisin	78.9	**72.7**	125	90.6	45.5	2.0
2AT1 A	Aspartate transcarbamylase	74.2	**72.0**	125	90.6	40.3	2.2
5ADH	Alcohol dehydrogenase	78.2	**71.8**	114	82.6	30.5	2.7
8ADH	Alcohol dehydrogenase	78.7	**71.6**	114	82.6	30.5	2.7
2LIV	Periplasmic binding protein	76.3	**71.4**	125	90.6	36.3	2.5
1ABP	Arabinose binding protein	79.6	**70.4**	120	87.0	39.2	2.2
3CPA	Carboxypeptidase	65.6	**70.3**	122	88.4	39.6	2.2
5CPA	Carboxypeptidase	65.6	**70.2**	127	92.0	41.4	2.2
1ETU	Elongation Factor Tu	74.4	**70.2**	129	93.5	72.9	1.3
3GPD G	D-Glyceraldehyde-3-phosphate dehydrogenase	77.6	**70.1**	114	82.6	34.2	2.4
3TS1	RNA synthetase	73.7	**70.0**	127	92.0	40.1	2.3
7AT1 C	Aspartate transcarbamylase	77.1	**69.8**	116	84.1	37.4	2.2
3LDH	Lactate dehydrogenase	84.0	**69.4**	110	79.7	33.4	2.4
2TS1	RNA synthetase	76.6	**68.8**	129	93.5	40.7	2.3
4GPD 4	D-Glyceraldehyde-3-phosphate dehydrogenase	74.4	**68.6**	112	81.2	33.9	2.4
4TS1 A	RNA synthetase	76.8	**68.4**	122	88.4	38.5	2.3
4GPD 2	D-Glyceraldehyde-3-phosphate dehydrogenase	77.8	**68.4**	112	81.2	33.6	2.4
2AT1 C	RNA synthetase	75.5	**67.9**	124	89.9	40.0	2.2
1GD1 O	D-Glyceraldehyde-3-phosphate dehydrogenase	70.1	**66.5**	127	92.0	38.0	2.4
2ATC A	Aspartate transcarbamylase	74.6	**66.4**	122	88.4	40.0	2.2
1TIM B	Triose phosphate isomerase	72.2	**66.4**	116	84.1	47.2	1.8
1RHD	Rhodanase	73.6	**66.3**	122	88.4	41.6	2.1
1GD1 R	D-Glyceraldehyde-3-phosphate dehydrogenase	70.1	**66.2**	129	93.4	38.9	2.4
1WSY A	Tryptophan synthase	74.2	**66.1**	131	94.9	52.8	1.8
4CPA	Carboxypeptidase	61.5	**65.4**	135	97.8	44.8	2.2
2GD1 P	D-Glyceraldehyde-3-phosphate dehydrogenase	71.6	**65.4**	120	86.9	35.9	2.4

*Only the first 45 comparisons are shown for brevity.
†Refer to Table X for details.

characteristic Rossmann-fold like structure. Comparisons between the immunoglobulins and the jellyroll virus structures sometimes gave scores above 60, indicating similarities in the geometries of the β-sheets.

Figure 6 shows the residue alignment scores ob-

TABLE XIII. Search With 2CHY Showing the Highest Scoring* Comparisons†

Chain	Description	ln G_{sec}	ln G_{res}	Match	Match/ probe (%)	Match/ target (%)	Target probe
2CHY	CheY chemotaxis protein	100.0	**100.0**	128	100.0	100.0	1.0
3FXN	Flavodoxin	83.2	**77.5**	109	86.5	79.0	1.1
4FXN	Flavodoxin	83.2	**77.4**	109	86.5	79.0	1.1
1LDM	Lactate dehydrogenase	79.5	**76.2**	120	95.2	36.5	2.6
8ADH	Alcohol dehydrogenase	83.2	**75.9**	117	92.9	31.3	3.0
2ATC A	Aspartate transcarbamylase	71.5	**75.9**	104	82.5	34.1	2.4
1SBC	Subtilisin	81.1	**75.9**	120	95.2	43.8	2.2
1FX1	Flavodoxin	80.1	**75.7**	106	84.1	72.1	1.2
2SEC E	Subtilisin	81.3	**75.6**	122	96.8	44.5	2.2
1PFK A	Phosphofructokinase	79.2	**75.4**	120	95.2	37.5	2.5
1SO1	Subtilisin	80.3	**75.3**	122	96.8	44.4	2.2
8LDH	Lactate dehydrogenase	79.4	**75.0**	117	92.9	35.6	2.6
1LLC	Lactate dehydrogenase	85.3	**74.7**	115	91.3	35.9	2.5
7AT1 C	Aspartate transcarbamylase	73.0	**74.5**	101	80.2	32.6	2.5
1ETU	Elongation Factor Tu	83.0	**74.4**	123	97.6	69.5	1.4
3PFK	Phosphofructokinase	79.4	**74.0**	120	95.2	37.6	2.5
4AT1 C	Aspartate transcarbamylase	75.7	**73.9**	101	80.2	32.6	2.5
2PFK C	Phosphofructokinase	80.1	**73.9**	123	97.6	40.9	2.4
1AT1 A	Aspartate transcarbamylase	74.5	**73.9**	101	80.2	32.6	2.5
8ATC C	Aspartate transcarbamylase	80.7	**73.7**	101	80.2	32.6	2.5

*Only the first 20 comparisons are shown for brevity.
†Refer to Table X for details.

tained for pairs of proteins having secondary structure scores of 60 and above. Residue scores are plotted against the sizes of protein being compared. There was no dependence on the packing class but there appeared to be a slight reduction in the score with increasing protein size. This may be related to the fact that larger proteins tend to contain more than one domain, so that global similarity is reduced when not all domains correspond. The distinction between scores for related and unrelated proteins was slightly less clear than for secondary structure matching. This is largely caused by the differences in the geometries of the loop regions, which can be quite large for less related structures and reduce the overall score considerably.

Residue scores for related proteins varied from 50 to 100 for complete identity. Examination of the alignments given by low scoring comparisons of between 50 and 70 showed that generally there was only partial similarity, for example, correspondence of only one domain between two multi-domain proteins. A large number of comparisons between remotely related proteins containing nucleotide binding domains fell into this group. Alternatively the proteins shared similar folds (e.g., the immunoglobulins and the azurins), but no strong sequence identity or functional similarity.

Scores of 80 or more were associated with significant structural similarity between the two proteins. The higher the score, the greater the similarity in terms of number of residues equivalent and correspondence of structural environments (see below for further discussion on the relationship between score and structural similarity).

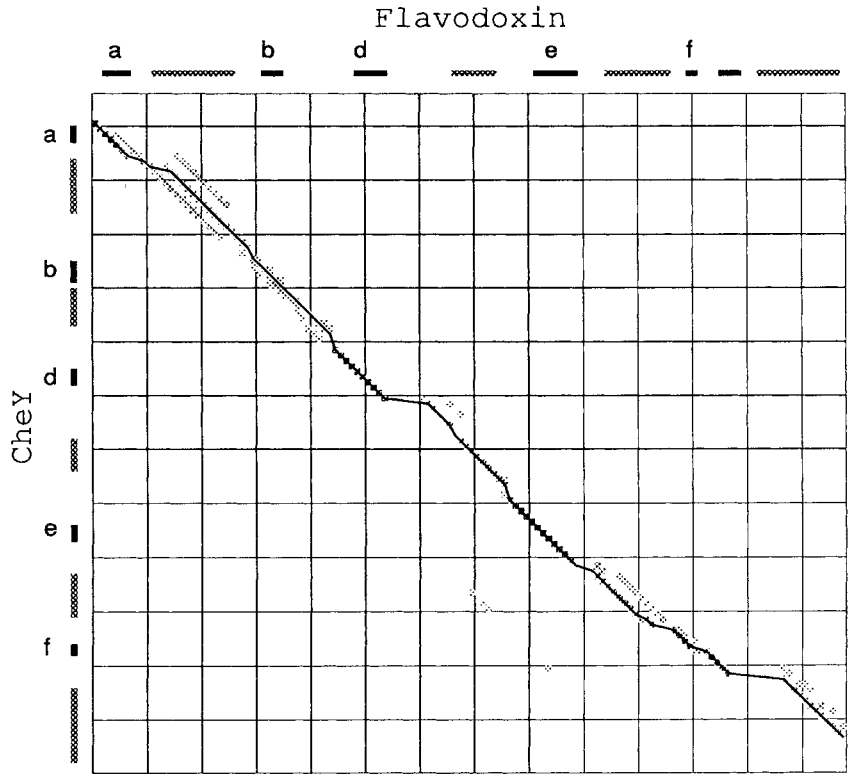
Low scores need to be treated with caution. In order to detect local similarities between two subdomains, it may help to consider both average (ln A_{res}) and global (ln G_{res}) residue scores. For similarity between subdomains of two proteins, the average score ln A_{res} will be high, reflecting the high scores in related regions identified by the residue selection procedure. However, the global score ln G_{res} will be low as only a small fraction of the proteins correspond structurally. In order to distinguish between more global or more local similarities, both average and global scores may be examined, together with the percentage of the smaller protein which has been equivalenced against the larger. The number of residues aligned between the two proteins together with the difference in their sizes, provides further information on the nature of the similarity between them (see Tables X–XIII). A modification of the method to search for all local similarities will be considered in a future publication.

Figure 7 shows a mainly linear correlation between secondary structure and residue scores for comparisons between related proteins. There is some variation due to the fact that not all the protein pairs have the same degree of structural similarity. In some cases the residue score will tend to be reduced due to a greater number of insertions and deletions while the secondary structure score may be relatively unaffected if these indels occur in loops between secondary structures.

```
      eeeee    hhhhhhhhhhhhhhh      eeee              eeeeee            hhhhh
     MKIVYWSGTGNTEKMAELIAKGIIESGK DVNTINVSDVNIDEL  LNEDILILGCSAMGDEVLEESEFEPFI
     @@@@@@@      -###@@@@@@+@@**+= -@*#@@@  @ +.+*  +@@@@@@@@@@@       = .=@=@@@
DKELKFLVVDDF      STMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGFGFIISDWNM      PNMDGLELL
      eee         hhhhhhhhhh              hhhhhhh       eee               hhh


hhh         eeeeeeee     hhhhhhhhhhhh   ee    eeee   hhhhhhhhhhhhhhh
EEISTK   ISGKKVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
@@@@@@  @@@@@@@@@@@@   :=@@@@@@@*    -#@@@  +@@@@         +.+=-@*+*=*.
KTIRADSAMSALPVLMVTAEA  KKENIIAAAQ    AGASG  YVVKP          FTAATLEEKLNKIFEKLGM
hhh          eee        hhhhhhhh        e  e                hhhhhhhhhhhhhh
```

(a)



(b)

Fig. 8. Structure alignment (**a**) of 4FXN (upper sequence MKIVY . . .) with cheY protein (lower sequence DKELK . . .). Symbols indicating structural similarity are ranked in increasing order of {blank . : − + * # @}. Shown in (**b**) is the final residue level alignment path for these proteins, symbols: hashed edge bars = α-helices, solid bars = β-strands **a–f.**

Databank searches

Test searches were conducted for each protein class. Cytochrome *c* (3CYT) was the test structure for the all α class, plastocyanin (1PCY) for the all β, and flavodoxin (4FXN) for the α/β. Searches were performed against all the structures from the Brookhaven databank[22] (release January 1991) with a resolution of 3 Å or better (a total of 740 structures). Tables X–XII list alignments scoring $\ln G_{res} \geq 60$, and where at least 80% of the probe protein has been aligned, ensuring that globally related proteins are identified. The tables show both

TABLE XIV. CPU Timings on a Sun 4/280 for Three Searches Against Different Size Datasets*

	Size		167 structures CPU time (min)		740 structures CPU time (hr)	
Probe	sec	res	sec	sec + res	sec	sec + res
4FXN	10	138	24	144	1.7	10.1
1PCY	10	99	25	96	1.9	5.0
3CYT	5	103	23	66	1.6	5.4

*The sizes of the probe proteins are given as number of secondary structures (sec) and residues (res), and search times are for secondary structure matching (sec) alone and for combined secondary structure and residue matching (sec + res).

secondary structure and residue alignment scores (ln G_{sec} and ln G_{res}), the percentages of probe and target proteins aligned, together with number of residues aligned, and the difference in protein sizes. All this information is useful in deciding the nature and degree of similarity.

For each test structure, the search identified proteins of known structural similarity with the probe protein and the order of scores appears to agree with established relationships. For the plastocyanin search, the most structurally similar structures are the other plastocyanins in the databank and the azurins, while lower scores are associated with similarity of the plastocyanin fold to that observed in immunoglobulin domains (greek-key β-barrel[23]). In the cytochrome *c* search, the method correctly identifies other cytochromes known to be similar to the probe cytochrome (see Johnson et al.[24]).

Similarly, the flavodoxin search lists all other flavodoxins in the databank as being most structurally similar, while other proteins containing domains with the same Rossmann-fold type structure as flavodoxin are shown further down the list. Interestingly, the search also identifies the bacterial signal transduction protein, cheY (2CHY), as having a high structural similarity. This resemblance has been noted by Bowie et al.,[25] who matched hydrophobicity patterns contained in the sequence of cheY to the accessibilities of residues in the flavodoxin structure and also by Rippmann and Taylor,[21] who superimposed the proteins using equivalent residues identified by SSAP. Both proteins possess a β-sheet containing 5 strands, with alternating helices packed against the sheet on both sides.

Artymiuk et al.,[26] using graph theory techniques to search through the databank with cheY, suggested that its tertiary fold more closely resembled that of the GDP binding domain of *E. coli* elongation factor Tu (EF-Tu). However, a search with cheY, using the combined structural alignment method (see Table XIII), confirmed that while cheY scores well against both 4FXN and 1ETU, it has higher similarity to flavodoxin, both by score and percentage of the structures equivalenced (Table XIII, Fig. 8). Inspection of the superpositions of the structures, using molecular graphics, showed that, for both pairs (1ETU/2CHY) and (4FXN/2CHY), the β-sheets can be easily superimposed. However, EF-Tu possesses an additional β-strand over cheY and the arrangements of the helices are different and not easily superimposed on those of cheY.

Searches performed by secondary structure matching alone are very fast (Table XIV). Secondary structure scores are generally above the cutoff of 60 for related proteins. As discussed above, the residue scores show greater variation as they involve comparisons of loop and turn regions. The results further confirm that secondary structure score can be used as a reliable indicator of relatedness, and if the subsequent residue alignment yields low scores due to insertions/deletions, it may be better to repeat it using a local alignment algorithm.

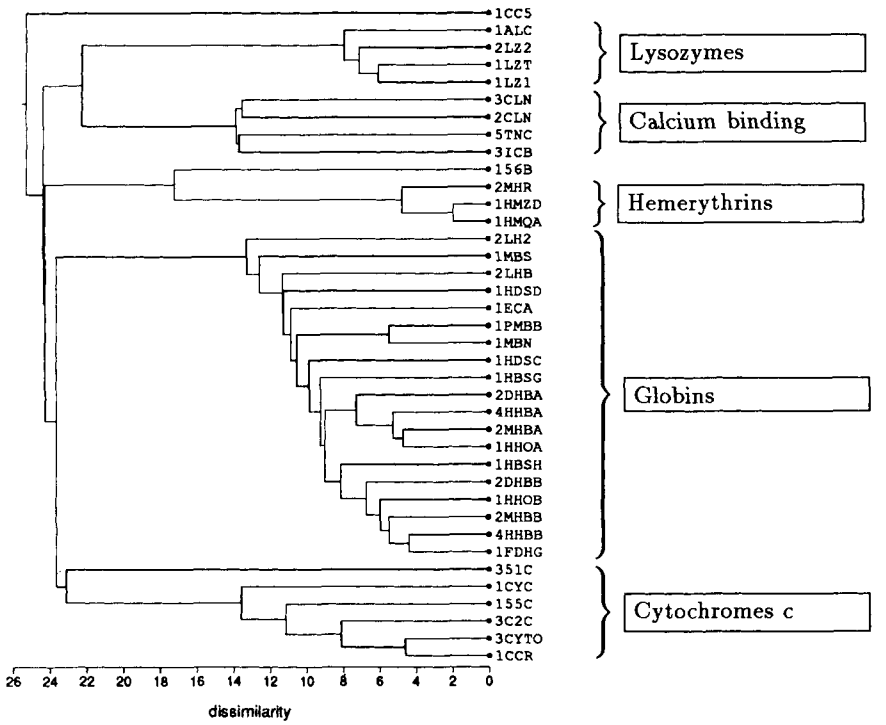
Sorting the Structure Databank Into Related Groups

Tables II and III list 167 high resolution structures from the Brookhaven databank which were used for a databank wide comparison. Restrictions developed above were employed, that is, no α proteins were compared with β proteins, and no pairs scoring less than ln G_{sec} = 60 by secondary structure matching were realigned by residue matching. Comparison of the whole databank with itself required 160 hr of CPU time running on a Sun 4/280.

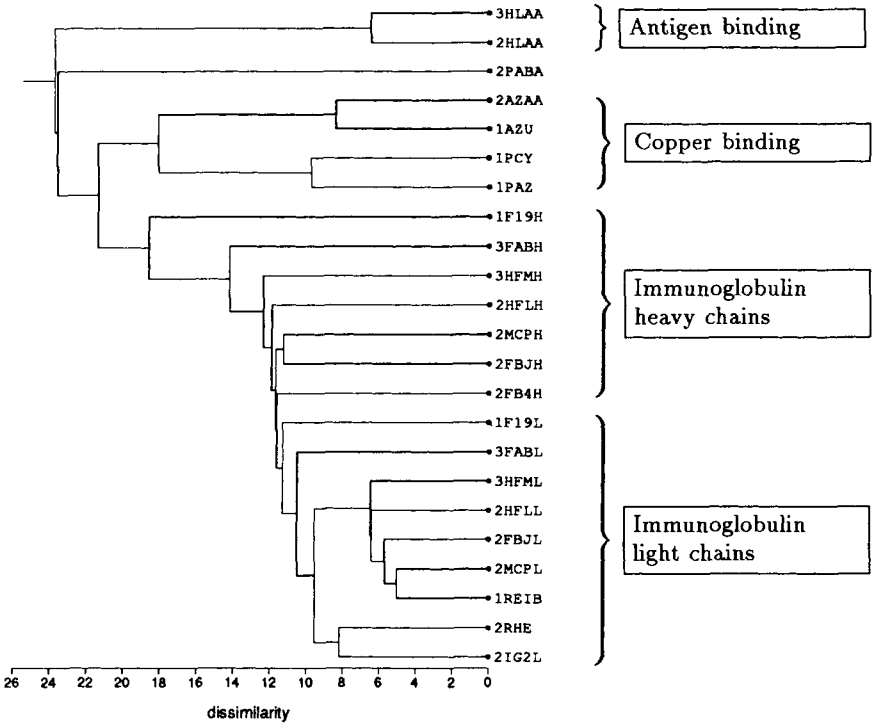
The resulting residue alignment scores were analyzed using single linkage clustering (see, for example, Krzanowski[27]) to generate a tree of related structures. Figures 9–11 show the groups identified by cutting across this tree at the 74.5 similarity level. All the groups can be related to known families of protein structures.

The tree in Figure 9a contains all α proteins. Within this, subtrees belonging to cytochromes, globins, hemerythrins, calcium-binding proteins, and lyzozymes are found.

All β proteins with aligned β-barrel (or simple greek-key) structures appear in Figure 9b, which separates out light and heavy immunoglobulin chains, antigen-binding proteins, and copper-binding proteins.

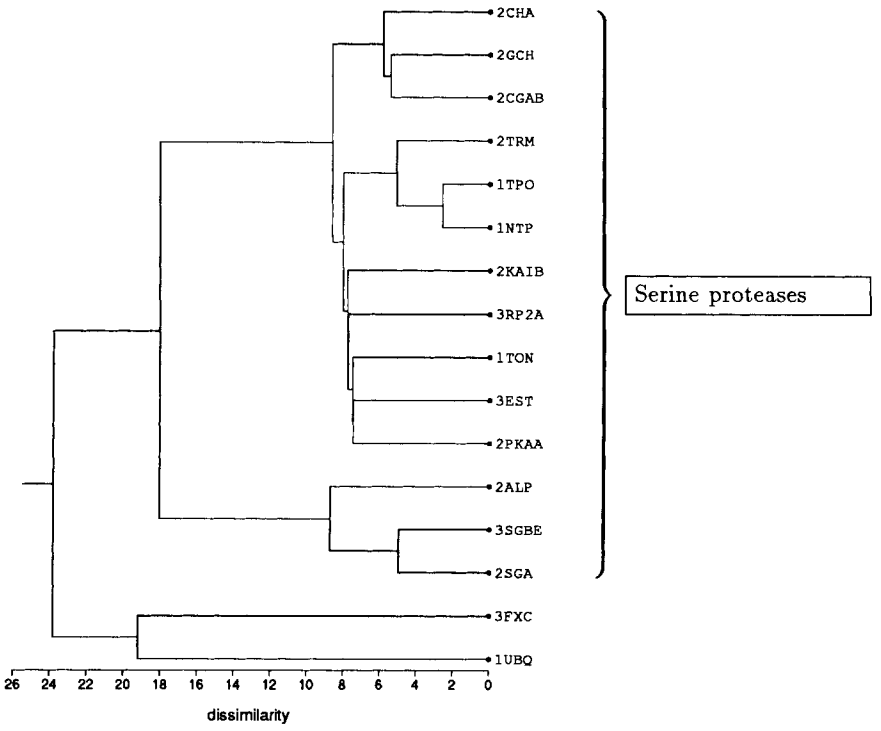


(a)

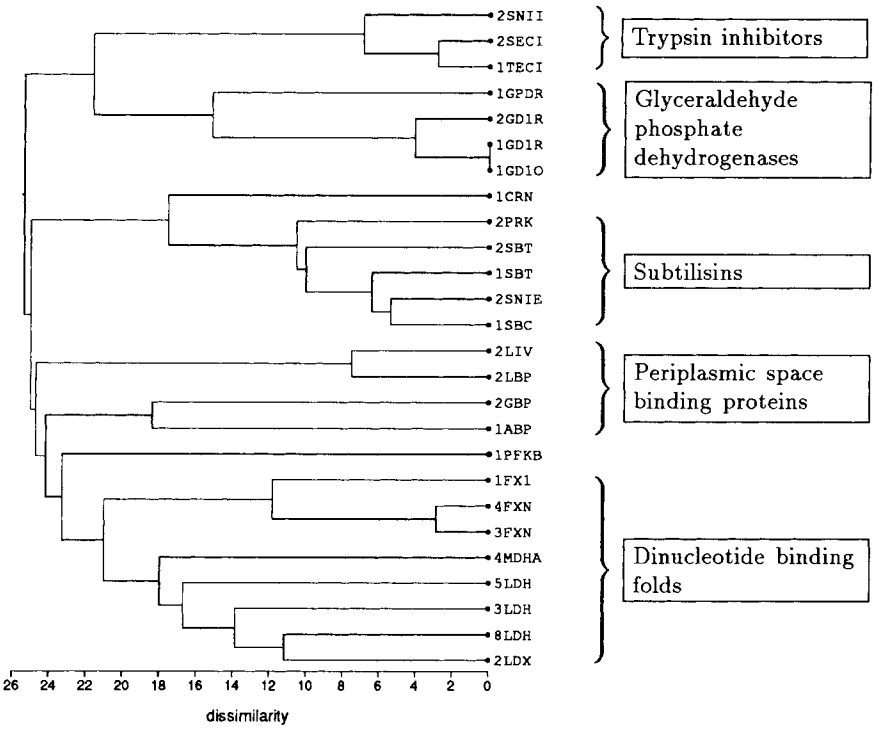


(b)

Fig. 9. Trees showing clustering of families. (**a**) Predominantly α proteins: cytochromes, globins, hemerythrins, lysozymes, and calcium-binding proteins. (**b**) Predominantly β proteins: immunoglobulins, antigen-binding proteins, and copper-binding proteins. The trees in Figures 9–11 were derived by single-link clustering of dissimilarities expressed as residue structural alignment scores ln G_{res}. All are subtrees of a single tree cut at a dissimilarity level of 25.5 (maximum dissimilarity = 100). Major groups are indicated in boxes.



(a)



(b)

Fig. 10. Trees showing clustering of families. (**a**) Serine proteases (trypsin-like). (**b**) Serine proteases (subtilisin-like), dinucleotide-binding folds and various other mixed α/β proteins. See Figure 9 for details of the method.

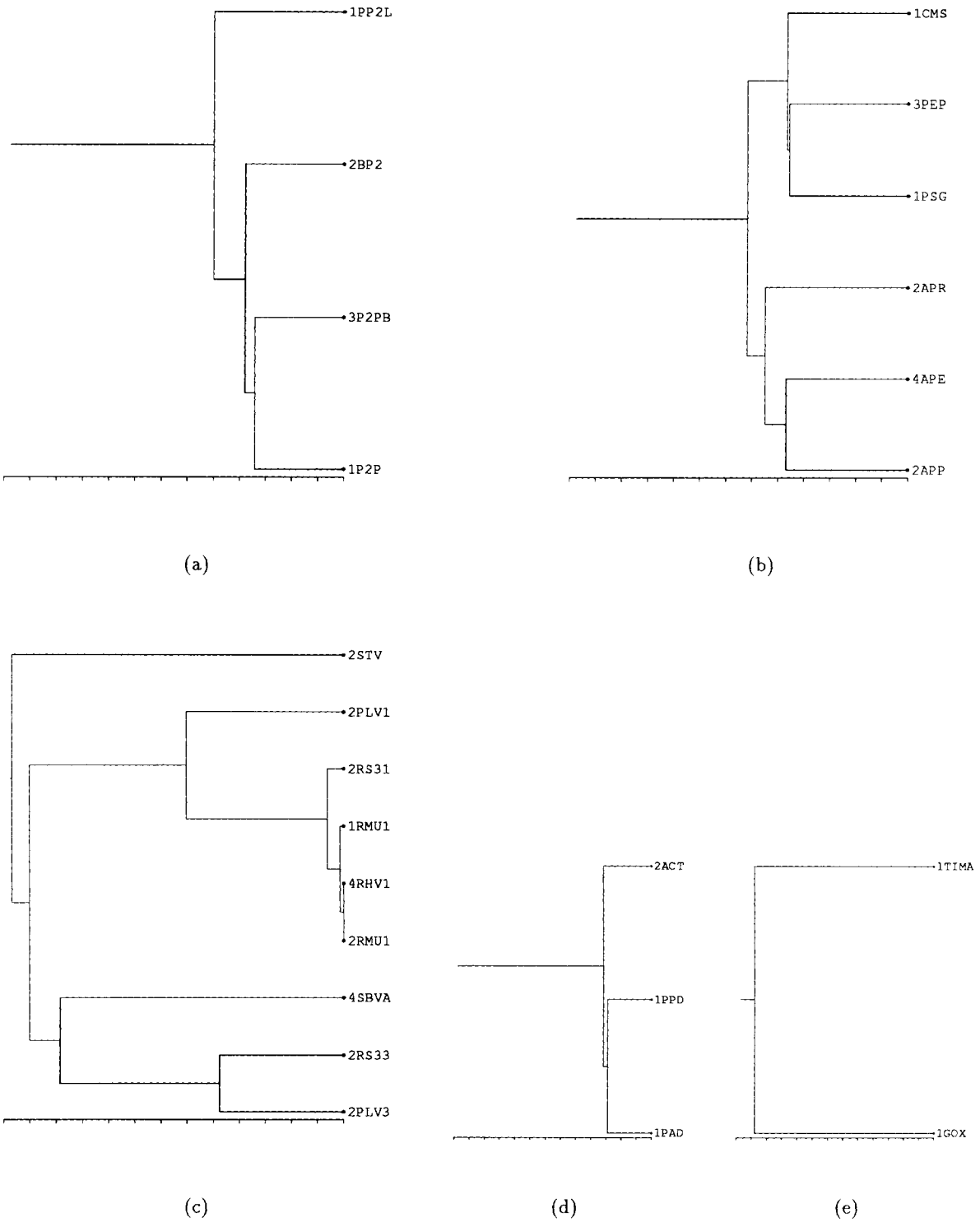


Fig. 11. Trees showing (**a**) phospholipases, (**b**) aspartyl proteases, (**c**) viral jellyrolls, (**d**) papain folds, and (**e**) Tim-barrels. See Figure 9 for details of the method and for the scale at the base of each figure.

Serine proteases which adopt orthogonal β-barrel structures are shown in Figure 10a, while aspartyl proteases with their twisted β-sheet structures appear in subtree Figure 11b. The subtree for the virus proteins which have a jellyroll greek-key β-barrel structure is given in Figure 11c.

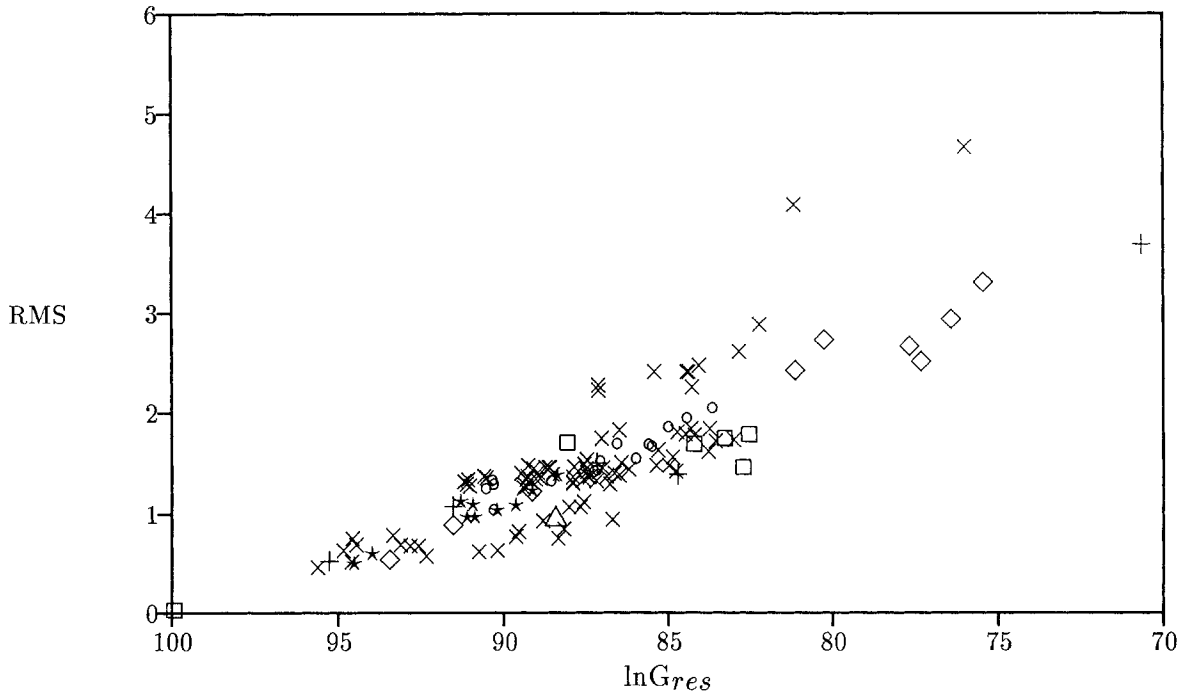


Fig. 12. Correlation of root mean square (rms) deviation obtained from structural superposition with global residue score (ln G_{res}) given by the combined structural alignment method. Structures were superposed using the method of Rippmann and Taylor[21] with residue equivalences from the combined structural alignment. Symbols in the plot refer to the following protein families: (○) aspartyl proteases, (◇) copper binding proteins, (+) cytochromes, (□) dinucleotide binding folds, (×) globins, (△) serine proteases (subtilisin-like), (*) serine proteases (trypsin-like).

Alternating α/β proteins form another tree (Fig. 10b). Within this, subgroups relating to subtilisins, nucleotide-binding proteins, and bacterial periplasmic space-binding proteins can be quite clearly seen. A small subgroup of trypsin inhibitor proteins is also present. These contain a few mainly helical secondary structures, which align moderately well with the long or double α-helices present in the large loops connecting β-strands in some α/β proteins.

Subtrees associated with some smaller groups containing only two proteins, such as the hexokinases and the carbonic anhydrases, are not shown. However, the method clearly separated these from the others.

Relationship between residue alignment score, rms, and sequence identity

The relationship between residue alignment score and rms deviation of the aligned structures was examined for groups of related structures identified by the cluster analysis. Structures were superimposed using the method of Rippmann and Taylor[21] and weighted rms values calculated, with the weights given by the residue pair scores of the structural alignment. Figure 12 shows a clear correlation between residue score and rms value: the higher the score the lower the rms value. Also, the correlation between sequence identity and both rms value and residue score (Figs. 13 and 14) agrees with that observed by Chothia and Lesk.[28] For a sequence identity greater than 30% and a corresponding residue score of 80–100, overall structural similarity between the two structures can be inferred.

A detailed analysis of structural relationships within each group is outside the scope of this paper. However, it can be seen that the method offers a powerful means not only of identifying related structures, but also of expressing the degree of similarity between them. Furthermore, because the alignment gives scores for each pair of equivalent residues,[2] regions of conserved structure between the proteins can be easily identified and linked to other features such as amino acid composition, accessibility, and hydrophobicity.

CONCLUSIONS

The combined secondary structure and residue method considerably improves the speed and accuracy of structural alignments by providing a more refined method of selecting initial residue pairs for comparison.

As secondary structure alignments are very fast, they can be used to rapidly search the structure databank for related proteins, with the advantage that any high scoring structures can be recompared by residue alignment, giving more detailed informa-

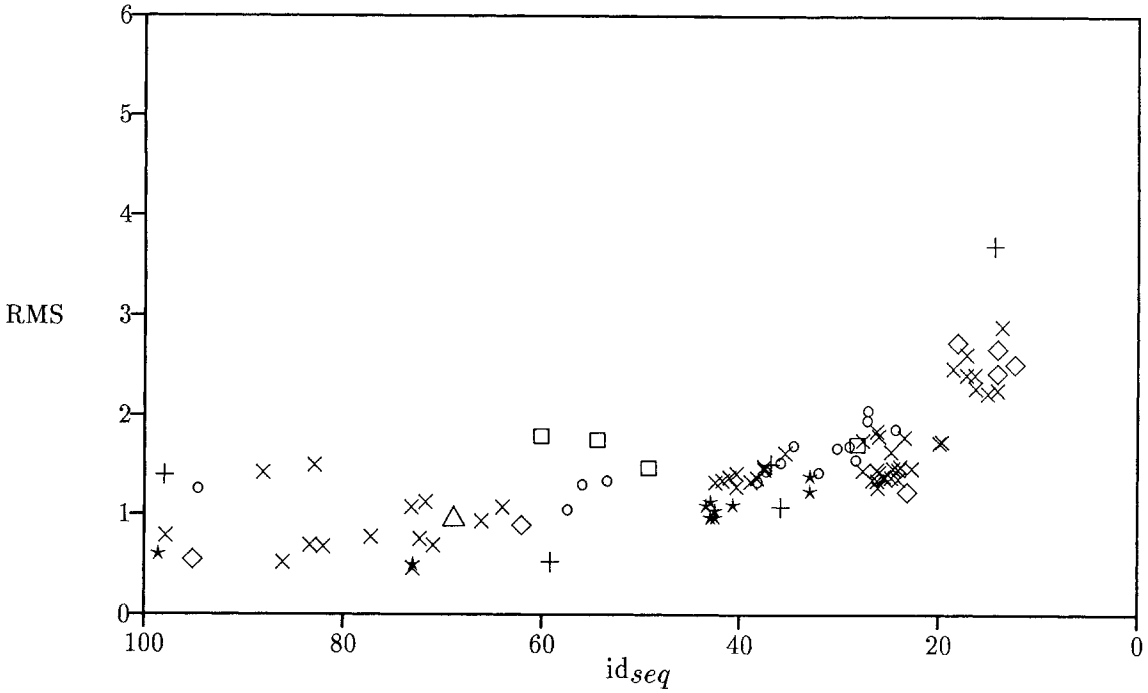


Fig. 13. Correlation of root mean square (rms) deviation with sequence identity (id_{seq}) for structurally related proteins. Structures were superposed using the method of Rippmann and Taylor[21] with equivalent residues identified by the combined structural alignment. Symbols as for Figure 12.

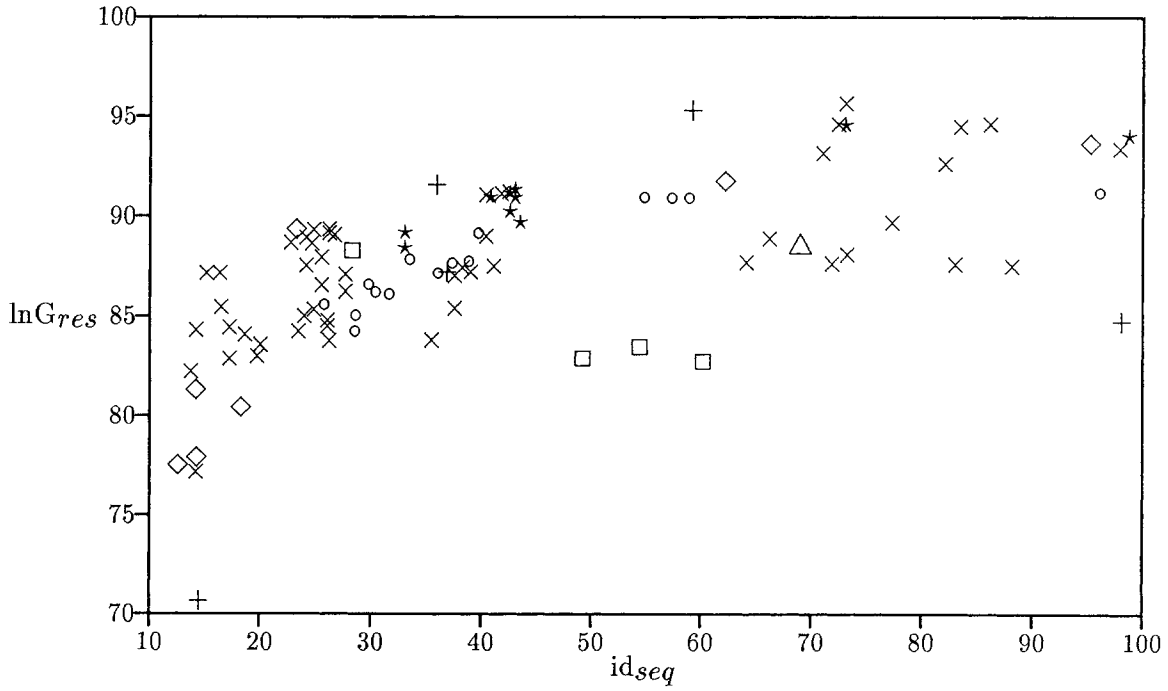


Fig. 14. Correlation of global residue alignment score ($\ln G_{res}$) with sequence identity (id_{seq}) for pairs of related proteins. Symbols as for Figure 12.

tion about the location of structurally similar regions within the proteins. When comparing all the structures in the databank, three quarters of the comparisons at the residue level can be prevented by filtering on the secondary structure score. This represents a saving of nearly 18 CPU days.

There is reasonable correlation between secondary structure and residue scores and tests using sets of related structures showed that both scores give a useful measure of structural similarity. Groups of related structures, identified both through searches through the databank and from comparisons involving the whole databank, are consistent with known similarities between proteins and with previously identified families.

The power and speed of the method lie first in its ability to use secondary structure comparisons to eliminate unrelated proteins from any further analysis. Second, once related proteins are identified, a detailed residue alignment is automatically generated, with improved accuracy, as mainly equivalent residue pairs are compared between proteins. Furthermore, the final score derived reflects the degree of overall structural similarity, while for each pair of aligned residues an individual score is known. This can be used both to determine the most highly structurally conserved regions between the proteins and to generate a weighted superposition of these regions.

NOTE ADDED IN PROOF

The weak similarity between ferredoxin (3FXC) and ubiquitin (IUBQ) shown in Figure 10a has also been noted and examined in detail in Vriend, G., Sander, C. Proteins: 11:52–58, 1991.

REFERENCES

1. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453, 1970.
2. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
3. Taylor, W.R., Orengo, C.A. A holistic approach to protein structure alignment. Protein Eng. 2(7):505–519, 1989.
4. Orengo, C.A., Taylor, W.R. A rapid method of protein structure alignment. J. Theor. Biol. 147:517–551, 1990.
5. Zuker, M., Somorjai, R.L. The alignment of protein structures in three dimensions. Bull. Math. Biol. 51(1):55–78, 1989.
6. Šali, A., Blundell, T.L. The 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.
7. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
8. Richards, F.M., Kundrot, C.E. Identification of structural motifs from protein coordinate data: Secondary structure and first level supersecondary structure. Proteins: Struct. Funct. Genet. 3:71–84, 1988.
9. Phillips, D.C. Development of crystallographic enzymology. Biochem. Soc. Symp. 31:11–28, 1970.
10. Abagyan, R.A., Maiorov, V.N. A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. J. Biomol. Struct. Dynam. 5(6):1267–1279, 1988.
11. McLachlan, A.D. Gene duplication in the structural evolution of chymotrypsin. J. Mol. Biol. 128:49–79, 1979.
12. Murthy, M.R.N. A fast method of comparing protein structures. FEBS Lett. 168:97–102, 1984.
13. Mitchell, E.M., Artymiuk, P.J., Rice, D.W., Willett, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J. Mol. Biol. 212:151–166, 1989.
14. Barton, G.J., Sternberg, M.J.E. Evaluation and improvements in the automatic alignment of protein sequences. Protein Eng. 1(2):89–94, 1987.
15. Fischel-Ghodsian, F., Mathiowitz, G., Smith, T.F. Alignment of protein sequences using secondary structure: A modified dynamic programming method. Protein Eng. 3(7):577–581, 1990.
16. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
17. Taylor, W.R., Thornton, J.M., Turnell, W.G. An ellipsoidal approximation of protein shape. J. Mol. Graph. 1(2):30–38, 1983.
18. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.
19. Kyte, J., Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
20. Eisenberg, D., Weiss, R.M., Terwilliger, T.C. The helical hydrophobic moment: A measure of the amphiphilicity of a helix. Nature (London) 299:371–374, 1982.
21. Rippmann, F., Taylor, W.R. Visualisation of structural similarity in proteins. J. Mol. Graph. 2:169–174, 1991.
22. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein databank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
23. Richardson, J.S. β-Sheet topology and the relatedness of proteins. Nature (London) 268:495–500, 1977.
24. Johnson, M.S., Šali, A., Blundell, T.L. Phylogenetic relationships from three dimensional protein structures. Methods Enzymol. 183:670–690, 1989.
25. Bowie, J.U., Clarke, N.D., Pabo, C.O., Sauer, R.T. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins: Struct. Funct. Genet. 7:257–264, 1990.
26. Artymiuk, P.J., Rice, D.W., Mitchell, E.M., Willett, P. Structural resemblance between the families of bacterial signal transduction proteins and of G proteins revealed by graph theoretical techniques. Protein Eng. 4(1):39–43, 1990.
27. Krzanowski, W.J. "Principles of Multivariate Analysis." Oxford Statistical Science Series-3. Oxford University Press, 1990.
28. Chothia, C., Lesk, A.M. The evolution of protein structures. Cold Spring Harbour Symp. Quant. Biol. LII:399–405, 1987.