[42] Phylogenetic Relationships from Three-Dimensional Protein Structures

By MARK S. JOHNSON, ANDREJ ŠALI, and TOM L. BLUNDELL

Introduction

As evolution progresses, accumulated changes in DNA and RNA lead to differences in amino acid sequences and corresponding alterations in the tertiary structure of proteins. Although these changes take place most often on the protein surface exposed to the solvent, mutations can be accepted within the inaccessible hydrophobic interior formed from packed secondary structure elements. Thus, relative translations and rotations of α helices and β strands do occur while their general spatial relationships remain highly conserved.[1,2] As a result, when comparisons are made among more distantly related structures, fewer topologically equivalent positions are found and lead to greater root mean square (RMS) deviations.[3,4]

Consequently, one should be able to chart the evolution of proteins from a comparison of their structures. Indeed, as protein structures are generally more conserved in evolution than are amino acid sequences,[5] they can be used to infer relationships among proteins where an alignment of their sequences is not statistically significant.[6] The first phylogenetic tree derived from structural information was based on the number of topologically equivalent positions in several dinucleotide- and mononucleotide-binding proteins[7] (for a review, see Matthews and Rossmann[8]). Following this work, Johnson *et al.*[6] compared six families of homologous structures and sequences (immunoglobulins, *c*-type cytochromes, globins, serine proteinases, eyelens γ-crystallins, and nucleotide-binding domains) and showed that trees based on sequence and structure are generally congruent.

Pattern matching and comparison methods for proteins can be employed to derive both the equivalences between residues and the overall similarity scores which can then be used to construct trees. It is important to recognize what information about the proteins is included in the comparison. This information generally focuses on only a few aspects, and there are two major groups of comparison methods: those using sequences of the amino acid residues and those employing three-dimensional structures of proteins.

The most familiar sequence comparison methods are the dynamic programming procedures based on the algorithm of Needleman and Wunsch,[9] which has recently been exploited to obtain multiple sequence alignments.[10-13] These methods usually consider the mutation rates of amino acid residues[14,15] to derive optimal comparison scores and corresponding alignments. The utility of many amino acid properties for the alignment of amino acid sequences was systematically explored by Argos[16] using his own optimization algorithm. Physical properties of amino acid residues are also considered in the pattern matching technique of Taylor,[17] which can align several protein sequences simultaneously. The hierarchical aspects of sequence and secondary structure organization were implemented in the program ARIADNE[18] for the matching of a given sequence pattern to a protein sequence and in the approach developed by Rawlings *et al.*[19] for reasoning about protein topology.

Alternatively, methods which compare tertiary structures of proteins are dominated by the rigid-body least-squares superposition of the α-carbon (C_α) positions (see Matthews and Rossmann[8] for a review). However, Rao and Rossmann[20] also included the main chain direction in their pairwise comparison procedure, which enabled the alignment of more divergent protein structures. The rigid-body approach was recently extended by Sutcliffe *et al.*[21] for the simultaneous comparison of several

[1] A. M. Lesk and C. Chothia, *J. Mol. Biol.* **136,** 225 (1980).
[2] A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160,** 325 (1982).
[3] C. Chothia and A. M. Lesk, *EMBO J.* **5,** 823 (1986).
[4] T. J. P. Hubbard and T. L. Blundell, *Protein Eng.* **1,** 159 (1987).
[5] M. Bajaj and T. L. Blundell, *Annu. Rev. Biophys. Bioeng.* **13,** 453 (1984).
[6] M. S. Johnson, M. J. Sutcliffe, and T. L. Blundell, *J. Mol. Evol.*, in press (1990).
[7] W. Eventoff and M. G. Rossmann, *CRC Crit. Rev. Biochem.* **3,** 111 (1975).
[8] B. W. Matthews and M. G. Rossmann, this series, Vol. 115, p. 397.
[9] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48,** 443 (1970).
[10] D.-F. Feng and R. F. Doolittle, *J. Mol. Evol.* **25,** 351 (1987).
[11] D.-F. Feng and R. F. Doolittle, this volume, [23].
[12] G. Barton and M. J. E. Sternberg, *J. Mol. Biol.* **198,** 327 (1987).
[13] G. J. Barton, this volume, [25].
[14] R. M. Schwartz and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, p. 353. National Biomedical Research Foundation, Washington, D.C., 1978.
[15] D.-F. Feng, M. S. Johnson, and R. F. Doolittle *J. Mol. Evol.* **21,** 112 (1985).
[16] P. Argos, *J. Mol. Biol.* **193,** 385 (1987).
[17] W. R. Taylor, *J. Mol. Biol.* **188,** 233 (1986).
[18] R. H. Lathrop, T. A. Webster, and T. F. Smith, *Commun. ACM* **30,** 909 (1988).
[19] C. J. Rawlings, W. R. Taylor, J. Nyakairu, J. Fox, and M. J. E. Sternberg, *J. Mol. Graph.* **3,** 151 (1985).
[20] S. T. Rao and M. G. Rossmann, *J. Mol. Biol.* **76,** 241 (1973).
[21] M. J. Sutcliffe, I. Haneef, D. Carney, and T. L. Blundell, *Prot. Engineer.* **1,** 377 (1987).

protein structures. Sippl[22] has used information from intramolecular C_α distance matrices to compare protein structures, while Murthy[23] applied dynamic programming to compare the secondary structure organization of proteins, taking into account the absolute angles and distances between the idealized secondary structure segments in approximately superposed proteins. Similarly, Richards and Kundrot[24] took into account the internal relationships between the elements of secondary structure to search for a given pattern in a protein structure database. Sheridan *et al.*[25] considered the residue secondary structure type to define the weights for each residue pair, which were later used in a dynamic programming procedure to find the alignment of two proteins. Recently, Barton and Sternberg[26] applied a dynamic programming procedure and an intermolecular distance matrix for roughly superposed loops to obtain the alignment of hypervariable regions.

In this chapter, we first describe a method for the multiple rigid-body superposition of structures and show the usefulness of the pairwise rigid-body superposition in determining relationships among homologous protein structures.[6] We continue with a description of a more flexible alignment procedure that compares a number of structural properties and relationships through simulated annealing and dynamic programming algorithms.[27] This latter technique escapes the limitations imposed by rigid-body alignments: it can taken into account deformations and translocations of secondary structure elements such as those illustrated in Fig. 1 between the two sets of aspartic proteinase domains and in Fig. 2 for the cytochromes *c*. From both approaches, the rigid-body and multifeature methods for the comparison of structures, phylogenetic trees are derived for the proteins listed in Table I. In general, these trees are isomorphous to those that may be obtained from the alignment of the corresponding amino acid sequences.[28]

[22] M. J. Sippl, *J. Mol. Biol.* **156**, 359 (1982).
[23] M. R. N. Murthy, *FEBS Lett.* **168**, 97 (1984).
[24] F. M. Richards and C. E. Kundrot, *Proteins* **3**, 7[illegible] (1988).
[25] R. P. Sheridan, J. S. Dixon, and R. Venkataraghavan, *Int. J. Peptide Protein Res.* **25**, 132 (1986).
[26] G. Barton and M. J. E. Sternberg, *J. Mol. Graph.* **6**, 190 (1988).
[27] A. Šali and T. L. Blundell, *J. Mol. Biol.*, in press (1990).
[28] The programs described here were written in the C and FORTRAN programming languages on a MICROVAX II computer (Digital Equipment Corporation) running the VMS operating system, version 4.5 or higher. These programs will be available from the authors in the near future.

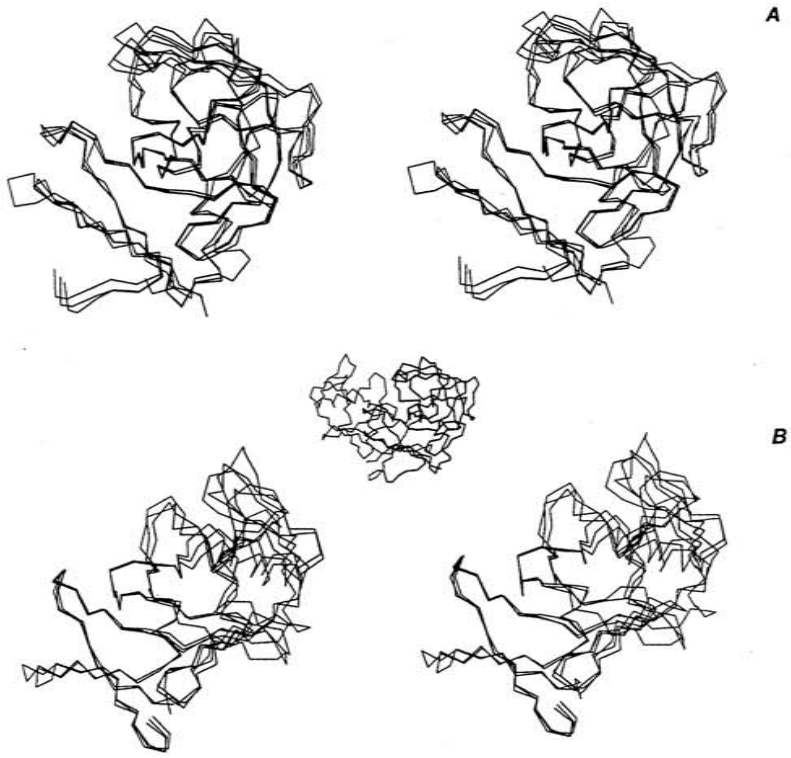

FIG. 1. Stereo views for the rigid-body superposed aspartic proteinase domains: (A) amino-terminal halves and (B) carboxy-terminal halves of endothiapepsin, penicillopepsin, and rhizopuspepsin (Table I). The amino- and carboxy-terminal domains are themselves a result of a duplication event and pack with C_2 symmetry. To show this, the C_α backbone for the entire two-domain structure of endothiapepsin (amino-terminal domain: thick line; carboxy-terminal: thin line) is inserted in the center; the C_2 rotation axis is located at the domain interface, approximately vertical and in the plane of the paper. Endothiapepsin, penicillopepsin, and rhizopuspepsin were split into domains at residues 174–175, 174–175, and 178–179 (crystallographic numbering), respectively.

Protein Structure Comparison by Rigid-Body Superposition

Topological Equivalence

The optimal superposition of two sets of coordinates is a common problem where the goal is to obtain the "best" fit of an object A to an object B over some set of coordinates said to be topologically equivalent for

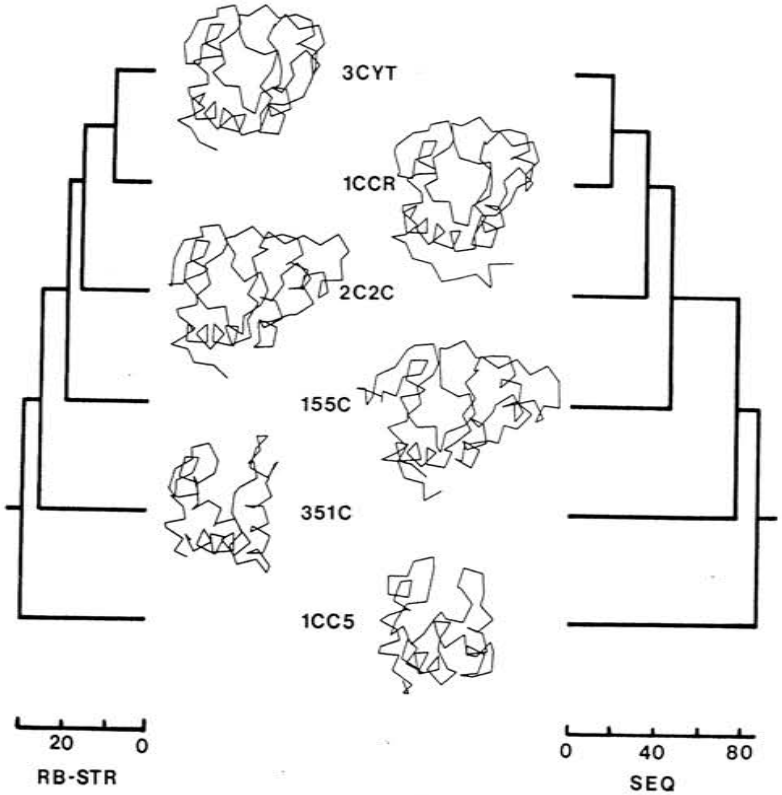

FIG. 2. Cladograms derived for the *c*-type cytochromes from both pairwise rigid-body comparisons (RB-STR) and multiple alignment of sequences (SEQ). C_α backbones are depicted for each of the structures after they have been fit with the multiple rigid-body superposition procedure; the structures have been translated only within the plane of the page. See Table I for identification of structures and sequences. The C_α coordinates belonging to the "unknown" (UNK) residues, as listed in the Brookhaven file for 155C, were excluded from the comparisons.

the two objects.[7,8,20,29,30] For proteins, these topologically equivalent positions can be defined as those C_α atoms in the two superposed structures that lie within a specified distance of each other, provided that these equivalences are colinear and hence obey the "no-knot" constraint.

[29] D. R. Ferro and J. Hermans, *Acta Crystallogr.* **A33,** 345 (1977).
[30] A. D. McLachlan, *Acta Crystallogr.* **A34,** 871 (1982).

TABLE I
PROTEIN STRUCTURES ALIGNED IN THIS STUDY

Brookhaven code[a]	Description	Resolution (Å)
Aspartic proteinases		
4APE	Endothiapepsin	2.1
2APP	Penicillopepsin	1.8
2APR	Rhizopuspepsin	1.8
Cytochromes *c*		
3CYT	Albacore tuna heart ferricytochrome *c* (oxidized)	1.8
1CCR	Rice embryo ferricytochrome *c*	1.5
2C2C	*Rhodospirillum rubrum* ferricytochrome c_2	2.0
155C	*Paracoccus denitrificans* cytochrome *c*-550	2.5
351C	*Pseudomonas aeruginosa* ferricytochrome *c*-551	1.6
1CC5	*Azotobacter vinelandii* ferricytochrome c_5	2.5
Globins		
2HHB	Human deoxyhemoglobin α and β chains	1.7
2HCO	Human carbonmonoxyhemoglobin α and β chains	2.7
1HHO	Human oxyhemoglobin α and β chains	2.1
1HBS	Human sickle cell hemoglobin	3.0
1FDH	Human deoxyhemoglobin α and γ fetal chains	2.5
2DHB	Horse deoxyhemoglobin α and β chains	2.8
1HDS	Deer sickle cell hemoglobin	2.0
2LHB	Sea lamprey hemoglobin V (cyano/met)	2.0
2MBN	Sperm whale metmyoglobin	2.0
3MBN	Sperm whale deoxymyoglobin	2.0
1ECD	*Chironomous thummi thummi* erythrocruorin	1.4
1LH1	*Lupinus luteus* leghemoglobin	2.0

[a] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).

Typically for proteins, one does not have *a priori* knowledge as to the extent of the topological equivalence, but only an idea of positions that are likely to be equivalent and which may be useful as a starting point for a comparison. This initial set of equivalences can be determined from an examination of atomic coordinates on a graphics device, from residues highly conserved in a sequence alignment, or from positions that are known to be crucial to the structural integrity, catalysis, or ligand binding of a protein. For the initial fit of two structures, three or more C_α positions must be specified to orient the structures. After the initial superposition of the structures, topological equivalences are redetermined, and the structures are again superposed based on this newly determined set of equiva-

lences. The procedure iterates until stability in both the topological equivalences and the RMS distance over the equivalent positions is obtained.

The Residual

The superposition of the structures hinges on the minimization of a function, δ (the residual), of the general form

$$\delta = \sum_{i=1}^{a} w_i(\vec{X}_i - \mathcal{R}\vec{Y}_i)^2 \tag{1}$$

where w_i is the weight for the ith pair of equivalenced positions $\vec{X}_i$ and $\vec{Y}_i$ of the two structures; $\mathcal{R}$ is the 3 by 3 rotation matrix that superposes structure B onto structure A. The use of this function presupposes that the coordinates for the two structures have been translated so that the centers of gravity for the two sets of equivalenced positions are located at the origin of the coordinate system. For the rigid-body comparison of structures in this chapter we have used a modification of a program MNYFIT[21] that was designed for the multiple alignment of structures by superposition.

Iterative Weighted Superposition

In the iterative weighted superposition procedure, one of the structures is chosen initially as the first approximation to the average of all structures (the framework), and each of the other structures is then fitted to it pairwise with the rapid procedure of McLachlan.[30] A new framework can then be calculated for b molecules from

$$\vec{F}_i^k = \frac{\sum_{j=1}^{b} w_{ij}^{k-1}\vec{Z}_{ij}}{\sum_{j=1}^{b} w_{ij}^{k-1}} \tag{2}$$

where $\vec{F}_i^k$ are the coordinates of point i on the framework at iteration k and $\vec{Z}_{ij} = \mathcal{R}_j\vec{Y}_{ij}$ are the coordinates of atom i from molecule j fitted to the previous framework F^{k-1}. The weight factors

$$w_{ij}^{k-1} = \frac{1}{\alpha_j^2 + \sigma_i d_{ij}} \tag{3}$$

include α_j as an estimate of the error in the coordinates of molecule j, σ_i, the standard deviation of the distance from the framework for the ith set of topologically equivalent positions a, and d_{ij}, the distance between atom i of molecule j to point i on F^{k-1}.

The residual is then calculated as

$$\delta^k = \frac{\sum_{j=1}^{b}\sum_{i=1}^{a} w_{ij}^{k-1}(\vec{F}_i^k - \vec{Z}_{ij})^2}{\sum_{j=1}^{b}\sum_{i=1}^{a} w_{ij}^{k-1}} \tag{4}$$

If the RMS distance between F^k and F^{k-1} is less than 10^{-5} Å and the difference between δ^k and δ^{k-1} is less than 10^{-5} Å then a minimum has been obtained and the equivalences can be updated.

These topologically equivalent atoms are determined from the optimal path through a matrix of Euclidean distances between all main-chain C_α positions from the proteins. To trace this path, a dynamic programming technique is used.[9,31] If the equivalences have not changed from the previous iteration, the superposition is complete. Otherwise, the molecules are fitted to F^k pairwise, a new framework determined, the residual calculated, and this process repeated until convergence is attained.

Distance Metric from Rigid-Body Superposition

From the pairwise rigid-body comparisons, two pieces of information can be obtained: the topological equivalences and the RMS distance over these equivalent C_α positions (Table II). The number of topologically equivalent positions is converted to a pairwise fractional topological equivalence (*PFTE*) by dividing by the length of the smaller structure. The RMS is a distance measure and is converted to a similarity score, the *SRMS*, calculated as 1 − RMS (Å)/3.5 (Å); a 3.5-Å cutoff is used in the definition of topological equivalence. The distance metric, D, employed in this study involves a weighted contribution [Eq. (5)] of these two parameters:[6]

$$D = -100 \ln(w_1 PFTE + w_2 SRMS) \tag{5}$$

where the weights, w_1 and w_2, are calculated from

$$w_1 = [(1 - PFTE) + (1 - SRMS)]/2 \tag{6}$$

$$w_2 = (PFTE + SRMS)/2 \tag{7}$$

with

$$w_1 + w_2 = 1 \tag{8}$$

For more closely related sequences, our experience has shown that the fraction of topologically equivalent positions may not differentiate be-

[31] M. L. Fredman, *Bull. Math. Biol.* **46**, 553 (1984).

TABLE II
FEATURES USED IN COMPARISON OF PROTEIN STRUCTURES[a]

Rigid-body structural comparisons:	
Number of topologically equivalent positions	
RMS distance over the topological equivalences	
Multifeature structural comparisons:	
Residues	Segments
Properties	
Identity	Secondary structure type
Physical properties	Amphipathicity
Local conformation	Improper dihedral angle
Distance from gravity center	Distance from gravity center
Side-chain orientation	Orientation relative to gravity center
Main-chain orientation	Side-chain accessibility
Side-chain accessibility	Main-chain accessibility
Main-chain accessibility	Position in space
Position in space	Global orientation
Global direction in space	
Main-chain dihedral angles	
Relations	
Hydrogen bond	
Distances to one or more nearest neighbors	Distances to one or more nearest neighbors
Disulfide bond	Relative orientation of two or more segments
Ionic bond	
Hydrophobic cluster	

[a] Structural features that are considered by the rigid-body and the multifeature approach to the comparison of protein structures and the determination of phylogenetic relationships are given. For the multifeature approach, various features are represented by rows and different levels of protein organization by columns. Only residue and secondary structure levels are shown here. The term property is used for all protein features that imply comparison of only one element from each protein. Conversely, the term relationship is used for a feature that implies comparison of at least two elements from each protein.

tween the structures.[4,6] Conversely, the RMS distance does provide a good measure of the difference between structures where the relationship is close. As a result, the weights, w_1 and w_2, are used to inversely modulate the contribution of the *PFTE* and the *SRMS* to the distance score [Eqs. (6)–(8)].[6] A matrix containing all pairwise distances can then be used directly by clustering or tree-generating techniques to display the relationships derived from the rigid-body structural comparisons.

Multifeature Comparison of Proteins

Although proteins within homologous families have the same tertiary folding, the elements of secondary structure undergo deformations, relative translations, and rotations to optimize packing of side chains and to adapt to evolutionary pressure. Thus, for two proteins with 30% sequence identity, the topologically equivalent residues defined by rigid-body superposition have a root mean square difference of approximately 1.5 Å and may comprise as few as one-third of the total number of residues.[4,6] This may not provide a sufficient basis for structural comparisons and emphasizes the requirement for a more flexible procedure for defining topological equivalence, one that can take into account relative movements and distortions of the secondary structure elements.

Structural Aspects of Proteins: Properties, Relationships, and Hierarchy

To achieve this flexibility, we include in the comparison method a number of protein features from several levels of the protein structure hierarchy (Table II). The protein is treated as a sequence of elements where each element is associated with a series of properties and may be engaged in a number of relationships with other elements. Additionally, these elements may exist at any level of the hierarchy of protein structure: residue, secondary structure, supersecondary structure, motif, domain, or globular protomer. For example, at the residue level, properties such as the local conformation and relationships like hydrogen bonds can be included. At the level of secondary structure, properties like segment solvent accessibility and relationships such as the relative spatial orientation of two segments can be incorporated.

Alignments from Dynamic Programming and Simulated Annealing

The comparison method is based on the dynamic programming technique generally used for sequence alignments.[9] In this method, one starts with the calculation of an N by M weight matrix $\mathcal{W}$ where N and M are the numbers of residues in the two compared proteins. This matrix is calculated in such a way that every element W_{ij} is proportional to the sum of the differences between various features of the residues i and j:

$$W_{ij} = \sum_l \left(\sum_p \rho^{pl} w_{ij}^{pl} + \sum_r \rho^{rl} w_{ij}^{rl} \right) \tag{9}$$

The contributions w_{ij} are the differences between individual features of the residues i and j, and factors ρ determine their relative weights. Superscript l runs over all levels of protein structure, superscript p stands for properties, and superscript r for relationships. When features at the second-

ary structure level are considered, the residues inherit the weights from the secondary structure segments involved, the corresponding α helices or β strands. It is trivial to define w_{ij}^{pl} for properties p; for example, the seventh property at the first level of structure, $w_{ij}^{7,1}$, describes the difference in the residue main-chain solvent accessibilities and is simply an absolute difference in the fractional main-chain accessibilities for the residues i and j from the first and second protein, respectively. In addition to properties, specific relationships such as hydrogen bonding interactions, which tend to be conserved in protein folds, can also be used in our comparison method. However, a relationship by its very nature affects more than one element in a sequence, and this precludes the simple procedure for the inclusion of this information into the residue-by-residue weight matrix $\mathcal{W}$.

To incorporate the information about relationships into the derivation of the final equivalences, we first use simulated annealing optimization[32] to obtain pairwise alignments based on relationships alone.[27] The underlying goal in the implementation of simulated annealing optimization is to maximize the number of equivalent relationships and minimize violations of a "no-knot" constraint. Since simulated annealing does not necessarily produce a global optimum, the optimization for every pair of structures is repeated several times. The fractional numbers of matching of residues i and j from proteins A and B, which can be obtained from several relationship alignments of proteins A and B in a straightforward way, are used to define the relationship weights w_{ij}^{rl}. These weights can be introduced directly into the residue by residue weight matrix $\mathcal{W}$ [Eq. (9)].

The dynamic programming algorithm then uses the matrix $\mathcal{W}$ to derive the most parsimonious alignment of the two structures. The overall distance score, which reflects the dissimilarity in selected features of the two proteins, is also obtained. A detailed description of these algorithms and their implementation in the program COMPARER may be found elsewhere.[27]

Multiple Structural Alignments

In the above description of the multifeature alignment method, we have assumed that the three-dimensional structures would be compared in a pairwise manner. However, such pairwise comparisons of several proteins may not be self-consistent, in the same way as pairwise sequence-based alignments may not be self-consistent. For this reason, we proceed by simultaneously aligning all structures. In COMPARER, we have adopted a strategy that employs a combination of the approaches by Feng and Doolittle[10,11] and Barton and Sternberg.[12,13] The procedure is divided

[32] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Science* **220**, 671, (1983).

into two parts. The first part is the construction of a dendrogram relating the homologous proteins, either *ad hoc* or from distance scores from pairwise comparisons.[10,11] The second part involves the gradual addition of new proteins, as imposed by the tree topology, into a growing multiple alignment. The most similar proteins and groups of proteins are structurally aligned first, and the gaps that are introduced do not change in later stages.[10,11] The weight matrix for the dynamic alignment of the two groups of previously aligned proteins is defined on the basis of the pairwise weight matrices relating the proteins from the two groups.

Distance Metric from Multifeature Comparison

With COMPARER, a pairwise distance score for each protein pair is obtained from the corresponding pairwise alignment implied by the multiple alignment. First, a sum of the weights W_{ij} that relate the residues equivalent in the pairwise comparison is found. This sum is then normalized via division by the number of equivalent residues in the pairwise comparison to give the intermediate score e. The final pairwise distance, E, that is used in the clustering procedure is then defined as

$$E = -100 \ln(1 - e/D_c) \tag{10}$$

D_c is a constant equal to the random value of the distance score e and is obtained for each protein pair by increasing the average of the weight matrix elements W_{ij} by three standard deviations of these elements divided by the square root of the number of equivalent residues. In addition to the distance score E, which does not incorporate information about gaps in the alignment, we considered a similar score that does incorporate gap penalties. Tree topologies for the two distance measures were the same in all cases.

Trees reflecting the evolution of different aspects of the proteins can be obtained by calculating the pairwise score E from the weights W_{ij} that were derived from different combinations of protein features. Thus, evolutionarily variable sequence features such as residue identity can be used for classification of similar proteins, and more conserved structural features, like hydrogen bonding, can be used for more divergent structures. Conversely, the clustering can also be used to infer the variability of a given protein feature in evolution.

Methods

The alignment of sequence data was produced by the "historical" multiple alignment procedure of Feng and Doolittle;[10,11] trees were derived from the distance metric of Feng *et al.*[15] Structures and sequences

were obtained from the structure files of the Brookhaven Protein Data Bank[33] with the following exceptions: the sequence of cytochrome *c*-550 from *Paracoccus denitrificans* is as revised by Ambler *et al.*;[34] the globin sequences from *Parasponia*[35] and *V reoscilla*,[36] for which there are no structures, were obtained from the NBRVAT[37] sequence data bank.

Tree topologies and branch lengths were determined from the distance matrices using the program KITSCH from the phylogeny inference package (PHYLIP) of Felsenstein.[38] This procedure is a modification of the original Fitch–Margoliash[39] method and accounts for unequal rates of change among the proteins by adjusting distances so that the total branch lengths from the root of the tree to the tips of each of the leaves are equidistant. In addition, numerous topologies are explored by swapping branches locally. The "best" tree is defined as the one that minimizes the sum of the squared differences between the equivalent distances from the tree and the input matrix, where each squared difference is also normalized by the corresponding squared distance from the input matrix.

Phylogenetic Trees from Structural Comparisons

In this chapter, we concentrate on the phylogenies that can be inferred from distance scores obtained from the pairwise rigid-body superposition and the alignment of structures based on many features. The results stemming from these analyses of three-dimensional protein structures are then compared with those obtained from an alignment of the amino acid sequences. Three homologous families serve as examples; these include the amino- and carboxy-terminal domains of the aspartic proteinases, eukaryotic and microbial *c*-type cytochromes, and globins.

Aspartic Proteinases

Structures of three fungal aspartic proteinases (Table I) have been solved to high resolution: endothiapepsin (4APE), penicillopepsin (2APP),

[33] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).
[34] R. P. Ambler, T. E. Meyer, M. D. Kamen, S. A. Schichman, and L. Sawyer, *J. Mol. Biol.* **147,** 351 (1981).
[35] A. A. Kortt, J. E. Burns, M. J. Trinick, and C. A. Appleby, *FEBS Lett.* **180,** 55 (1985).
[36] S. Wakabayashi, H. Matsubara, and D. A. Webster, *Nature (London)* **322,** 481 (1986).
[37] R. F. Doolittle, *Science* **214,** 149 (1981).
[38] J. Felsenstein, *Evolution* **39,** 783 (1985).
[39] W. M. Fitch and E. Margoliash, *Science* **15,** 279 (1967).

and rhizopuspepsin (2APR). The sequence identity between any pair of these three proteins is roughly 40%. Additionally, Tang *et al.*[40] have shown that amino-terminal and carboxy-terminal domains, each comprising about one-half of the molecule, are related by C_2 symmetry.

The multifeature alignment of amino- and carboxy-terminal domains (Fig. 3a) gives equivalences that are generally identical to those obtained from careful inspection of the structures on a graphics terminal (Fig. 1). This is so even though fragments of secondary structure have undergone translations, rotations, distortions, and numerous changes in sequence. This alignment stands in sharp contrast to that obtained from either the rigid-body multiple structure superposition or from the multiple sequence alignment (Fig. 3). The multiple structure rigid-body technique locates only 43 topologically equivalent positions among the six domains (asterisked positions in Fig. 3a), each domain consisting of approximately 150 residues. These positions are in complete agreement with the COMPARER alignment (Fig. 3a). Pairwise rigid-body superposition of structures was used to derive the distances for tree construction and led to between 67 and 77 topologically equivalent positions between the two sets of domains. A comparison of the sequences of the amino-terminal domains with the carboxy-terminal domains of the aspartyl proteinases aligns the conserved active-site amino acid triad Asp-Thr-Gly that is present in both domains. Outside of the neighborhood surrounding the catalytic region, only one other section is aligned similarly to that obtained from the multifeature comparison method (Fig. 3).

The trees (Fig. 4) derived from the structural comparisons, either the rigid-body or the multifeature procedure, are congruent: the amino-terminal domains clearly branch apart from the carboxy-terminal domains. For the cluster of either domain, the shorter distance is between endothiapepsin and penicillopepsin (Fig. 4) and is consistent with an alignment of the three full-length sequences. By sequence, the branch order within each domain's cluster is well determined. However, the sequence similarity between the domains is not statistically significant at the level of 3σ; this illustrates the power of the structural comparison method where unequivocal relationships are found. The numerous structural features that are common to both sets of domains, most notably the hydrogen bonding patterns, are consistent with the notion that the domains result from gene duplication.[40]

[40] J. Tang, M. N. G. James, I. N. Hsu, J. A. Jenkins, and T. L. Blundell, *Nature (London)* **271,** 618 (1978).

(a) F-STR

```
          *****          ****          ***********  ******
4APE-N  ---STGSATTTPIDSLDDAYITPVQ-IGT-----PAQTLNLDFDTGSSDLWVFSSETTASEVDGQTIYTPSK
2APP-N  --AASGVATNTPTA-NDEEYITPVT-IG-------GTTLNLNFDTGSADLWVFSTELPASQQSGHSVYNPSA
2APR-N  --AGVGTVPMTDYG-NDIEYYGQVT-IGT-----PGKKFNLDFDTGSSDLWIASTLCT-NCGSGQTKYDPNQ

4APE-C  YTGSITYTAVSTKQ---GFWEWTSTGYAVGSGTFKSTSIDGIADTGTTLLYLPATVVSA---------YWAQ
2APP-C  YTGSLTYTGVDNSQ---GFWSFNVDSYTAGSQ-SGDG-FSGIADTGTTLLLLDDSVVSQ---------YYSQ
2APR-C  FKGSLTTVPIDNSR---GWWGITVDRATVGTSTVAS-SFDGILDTGTTLLILPNNIAAS---------VARA

                                    **** *
4APE-N  STTAKLLSGATWSISYGDGSSSSGD----VYTDTVSVGGLTVTGQ------------------AVESAKKVS
2APP-N  --TGKELSGYTWSISYGDGSSASGN----VFTDSVTVGGVTAHGQ------------------AVQAAQQIS
2APR-N  SSTYQAD-GRTWSISYGDGSSASGI----LAKDNVNLGGLLIKGQ------------------TIELAKREA

4APE-C  VSGAKSSSSV--------GGYVFPCSA-TLPSFTFGVGSARIVIPGDYIDFGPISTGSSSCFGGIQSSA---
4APE-C  VSGAQQDSNA--------GGYVFDCST-NLPDFSVSISGYTATVPGSLINYGPSGD-GSTCLGGIQSNS---
4APE-C  Y-GASDNGD---------GTYTISCDTSAFKPLVFSINGASFQVSPDSLVFEEF---QGQCIAGFGYG----

              ****                                    ****          ****
4APE-N  SSFTEDSTIDGLLGLAFSTLNTVSPTQQKTFFDNAKAS--LDSPVFTADLGY---HAPGTYNFGFIDTTA
2APC-N  AQFQQDTNNDGLLGLAFSSINTVQPQSQTTFFDTVKSS--LAQPLFAVALKH---QQPGVYDFGFIDSSK
2APR-N  ASFASG-PNDGLLGLGFDTITTVRG--VKTPMDNLISQGLISRPIFGVYLGKAKNGGGGEYIFGGYDSTK

4APE-C  ------GIGINIFGD--------------VALKAA----------FVVFNGA-----TTPTLGFASK----
2APP-C  ------GIGFSIFGD--------------IFLKSQ----------YVVFDSD-----G-PQLGFAPQA---
2APR-C  ------NWGFAIIGD--------------TFLKNY----------YVVFNQG-----V-PEVQIAPVA--E
```

(b) SEQ

```
4APE-N  -STGSATTTPIDSLD-------DAYITPVQIGT-P-AQTLNLDFDTGSSDL----------WVFSSETTAS
2APP-N  AASGVATNTPTAN-D-------EEYITPVTIG----GTTLNLNFDTGSADL----------WVFSTELPAS
2APR-N  AGVGTVPMTDYGN-D-------IEYYGQVTIGT-P-GKKFNLDFDTGSSDL----------WI-ASTLCTN

4APE-C  -YTGSITYTAVSTKQGFWEWTSTGY--AVGSGTFK-STSIDGIADTGTTLLYLPATVVSAYWAQVSGAKSS
2APP-C  -YTGSLTYTGVDNSQGFWSFNVDSYTAGSQSG-----DGFSGIADTGTTLLLLDDSVVSQYYSQVSGAQQD
2APR-C  -FKGSLTTVPIDNSRGWW----GITVDRATVGTSTVASSFDGILDTGTTLLILPNNIAASV-ARAYGASDN


4APE-N  EVDGQTIYT-PSKSTTAKLLSGATWSISYG-----DGSS---SSGDVYTD--TVSVGGLTVTGQAVESAKK
2APP-N  QQSGHSVYN-P--SATGKELSGYTWSISYG-----DGSS---ASGNVFTD--SVTVGGVTAHGQAVQAAQQ
2APR-N  CGSGQTKYD-PNQSSTYQA DGRTWSISYG-----DGSS---ASGILAKD--NVNLGGLLIKGQTIELAKR

4APE-C  SSVGG--YVFPC-SAT-LP------SFTFG-----VGSARIVIPGD-YIDFGPISTGSSSCFGGIQSSAGI
2APP-C  SNAGG--YVFDC-S-T-N-LPDFSVSIS-GYTATVPGSL--INYGP-SGD------G-STCLGGIQSNSGI
2APR-C  GD-GT--YTI---SCDTSAFKPLVFSI--------NGASFQVSPDSLVFEEFQ---G-QCIAG----F-GY


4APE-N  VSSSFTEDSTIDGLLGLAFSTLNTVSPTQQKTFFDNAKASLDSPVFTADL---GYHAPGTYNFGFIDTTA
2APP-N  ISAQFQQDTNNDGLLGLAFSSINTVQPQSQTTFFDTVKSSLAQPLFAVAL---KHQQPGVYDFGFIDSSK
2APR-N  EAASFASGPN-DGLLGLGFDTITTVRGVKTPMDNLISQGLISRPIFGVYLGKAKNGGGGEYIFGGYDSTK

4APE-C  GINIFG-----DVALKAAF----VVFNGA------------TTP----TL--------G---FASK--
2APP-C  GFSIFG-----DIFLKSQY----VVFD-S------------DGP----QL--------G---FAPQA-
2APR-C  GNWGFAIIG--DTFLKNNY----VVFN-Q------------GVP----------EVQIAPVAE
```

FIG. 3. Alignments of the aspartic proteinase domains from (a) multifeature (F-STR) and (b) multiple sequence (SEQ) comparisons; amino- and carboxy-terminal domains are labeled N and C, respectively (see Table I for identification of structures and sequences). The multifeature alignment was derived from the following properties and relationships: physical properties of amino acid residues (0.05), residue main-chain accessibilities (0.20), hydrogen bonding pattern (0.80), residue identities (0.10), Φ (0.05) and Ψ (0.10) dihedral angles, absolute C_α positions in space (0.15), and absolute main-chain direction in space (0.20). Numbers in parentheses represent the relative weights of the corresponding features. Asterisks indicate those positions that are considered topologically equivalent after multiple rigid-body superposition.

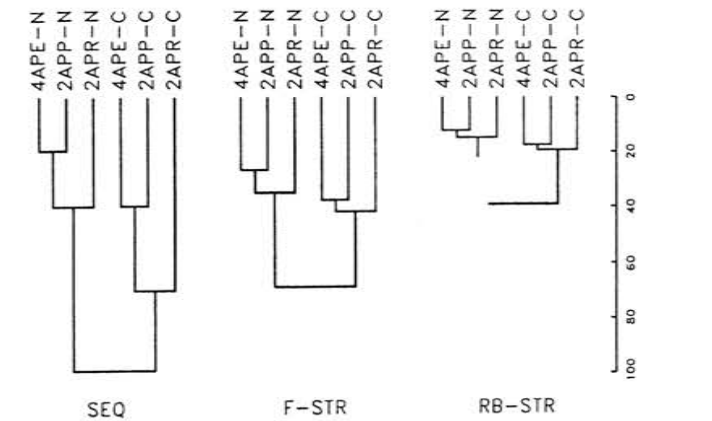

FIG. 4. Dendrograms derived for the domains of the aspartic proteinases from multiple sequence (SEQ), pairwise rigid-body (RB-STR), and multifeature (F-STR) alignments. For the multifeature tree, the same properties and relations were used as for the alignment in Fig. 3. See Table I for the identification of structures and sequences. The amino- and carboxy-terminal domains are labeled N and C, respectively.

Cytochromes c

Six *c*-type cytochromes that have either eukaryotic (bony fish and plant) or microbial origins are listed in Table I. These proteins cover a wide range of relationships; the percent identity obtained from the sequence alignment ranges between 62% for a comparison of albacore cytochrome *c* (3CYT) with rice embryo cytochrome *c* (1CCR), and 15% for a comparison of the rice cytochrome and the cytochrome c_5 from *Azotobacter vinelandii* (1CC5).[6] Trees constructed from pairwise rigid-body fits, the overall tree based on the COMPARER approach, and that from the multiple alignment of the corresponding sequences are in complete agreement with regard to topology (Figs. 2 and 5). In each case, the eukaryotic mitochondrial cytochromes *c* segregate from the microbial c_2, *c*-550, *c*-551, and c_5 cytochromes. The dendrograms agree with sequence-based trees described by others.[41]

A multiple structure rigid-body alignment identifies only 10 topologically equivalent residues out of the roughly 100 residues in each structure. The fitted structures, translated within the plane of the page, are shown in Fig. 2. It is clear from the view shown in Fig. 2 that there is more in

[41] R. M. Schwartz and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, p. 29. National Biomedical Research Foundation, Washington, D.C., 1978.

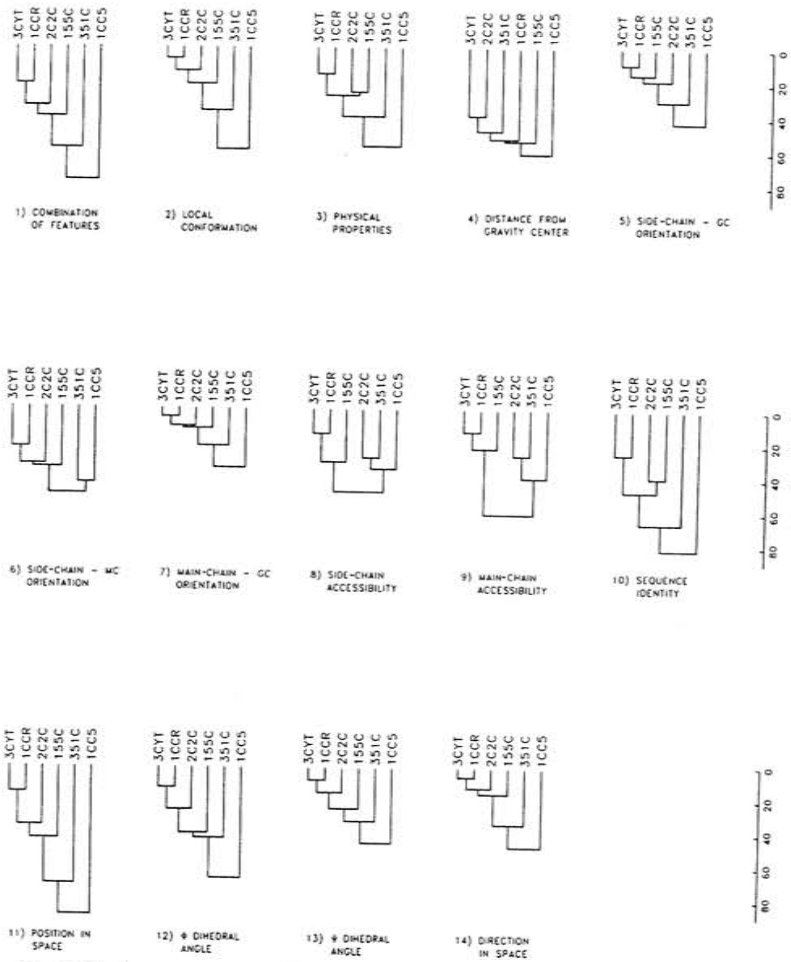

FIG. 5. Cladograms constructed from a multifeature alignment of the *c*-type cytochromes. Trees 2–14 were derived from the pairwise distances E [Eq. (10)] that were calculated using the individual features only. On the other hand, tree 1 was constructed from the same weight matrix elements W_{ij} that were used to derive the multiple alignment (data not shown): features included were physical properties of amino acid residues (0.20), distance of the C_α from the molecular gravity center (0.20), residue identities (0.10), absolute distance in space (0.30), and absolute main-chain directions in space (0.20). GC, Molecular gravity center; MC, main chain.

common among the structures than the small number of equivalences would suggest. The multifeature comparison method provides a technique for establishing a complete alignment of the structures (results not shown).

Trees reflecting various aspects of *c*-type cytochromes can be derived from their multiple alignment (Fig. 5). Trees 2 through 14 were obtained from the pairwise distances E that were calculated by considering only one feature in the derivation of weight matrix elements W_{ij} [Eq. (9)]. Tree 1 was constructed from the distances E obtained from the combination of features used to align the structures (Fig. 5). Of 14 trees, 7 have the same topology as the trees based on the multiple sequence and pairwise rigid-body comparison (Fig. 2); these include the trees derived from the local conformation of the main chain, orientation of the main chain relative to the molecular gravity center, absolute position of C_α atoms, main-chain direction in approximately superposed structures, Φ and Ψ dihedral angles, and the tree constructed from a combination of distances calculated from the features used to obtain the multiple alignment.

Trees that have a topology different from the most frequent one include the two trees derived from sequence information: the first of these two trees is based on the five physical characteristics of amino acid residues (such as hydrophobicity) that were found useful in construction of sequence alignments by Argos,[16] and the second tree is derived from a consideration of residue identities only. It may be noted that the clustering in these two trees is the same and corresponds to the subjective impression obtained from a consideration of the shape of the cytochrome structure in Fig. 2. The two trees reflecting similarities in the main-chain and side-chain accessibilities are also congruent with each other, but they are different from the two topologies mentioned above. In contrast, the unique and self-inconsistent topologies of the two trees that involve the orientation of side chains relative to the main chain and relative to the molecular gravity center imply that the orientation of the side chains is not a useful indicator for establishing relationships between divergent protein structures.

The three tree topologies (Fig. 5) based on the combination of features, sequence criteria, and solvent accessibility, demonstrate that evolutionary pressure does not act on all aspects of protein structure in the same way; thus, different criteria may be better for different purposes. For example, trees constructed from rigid-body superpositions are suitable in the selection of structures for determination of a framework[6] in homology-based protein modeling,[21] sequence-based trees are convenient for the description of evolutionary relationships among relatively similar proteins, while trees based both on pairwise rigid-body and multifeature comparison may be better for the analysis of more divergent structures.

Globins

The 19 globin structures listed in Table I were compared by the two structural procedures. Included within these comparisons are a number of globin structures that are identical in sequence (2HHB, 2HCO, and 1HHO; 2MBN and 3MBN) but differ in ligand binding [deoxy-, carbonmonoxy-, and oxyhemoglobins (human); deoxy- and metmyoglobins (sperm whale) as well as globins that are very distantly related to the vertebrate globins [the erythrocruorin from the larva of *Chironomus thummi thummi* (1ECD) and the leghemoglobin from the leguminous plant *Lupinus luteus*]. From the sequence alignments,[6] the vertebrate sequences are between 15 and 24% identical to the insect erythrocruorin and plant leghemoglobin. In contrast to the alignment of the aspartic proteinase domains (Fig. 3), the alignment of the globins using the multifeature approach is very similar to the multiple alignment constructed from the amino acid sequences (Fig. 6). On the other hand, the multistructure rigid-body comparison leads to only 15 topologically equivalent positions.

The sequence-derived tree (Fig. 7) agrees with trees based on sequence.[10,42,43] Overall, the trees constructed from either set of structural data are also in agreement (Fig. 7). The α chains cluster apart from the β/γ cluster; the sea lamprey hemoglobin clusters at about the same position as the human myoglobin, and each of these structures and sequences clusters apart from the erythrocruorin and leghemoglobins as one would expect on the basis of classic organismal evolution. COMPARER produces a tree that coincides completely with the sequence tree, but differentiation among the structures in cases where the sequences are identical also takes place. This also occurs with the pairwise rigid-body procedure, where all deoxy chains cluster together (2HHB, 1FDH, 1HBS, and 2DHB) and those subunits that have oxygen or carbon monoxide bound also cluster together; this occurs for both α and β chains and presumably reflects the conformational changes that occur on ligand binding. For all three trees, the deer sickle cell subunits are outliers to the α chains and the β/γ chains.

Concluding Remarks

About 30 new protein structures a : determined each year and added to the nearly 400 structures that have already been deposited in the Brookhaven Protein Data Bank.[33] While this is far from the number of known

[42] M. Goodman, G. W. Moore, and G. Masuda, *Nature (London)* **253,** 603 (1975).

[43] L. T. Hunt, S. Hurst-Calderone, and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.) Vol. 5, Suppl. 3, p. 229. National Biomedical Research Foundation, Washington, D.C., 1978.

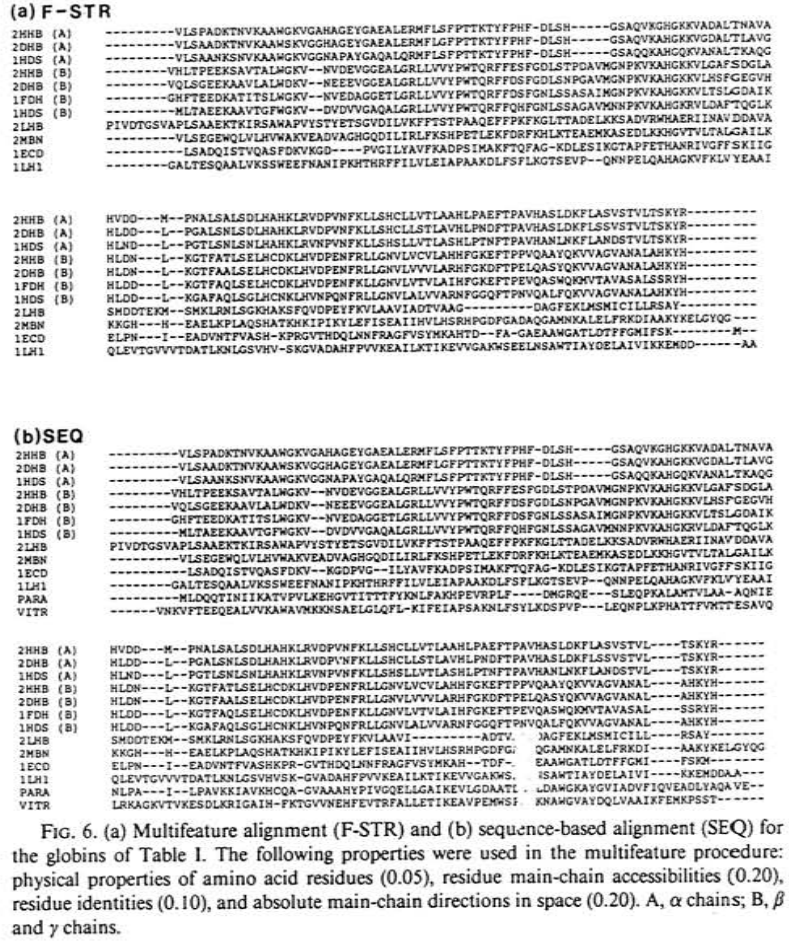

FIG. 6. (a) Multifeature alignment (F-STR) and (b) sequence-based alignment (SEQ) for the globins of Table I. The following properties were used in the multifeature procedure: physical properties of amino acid residues (0.05), residue main-chain accessibilities (0.20), residue identities (0.10), and absolute main-chain directions in space (0.20). A, α chains; B, β and γ chains.

primary sequences, the structural data contain considerably more information per protein than do the sequence data. This information may be used to resolve evolutionary relationships among proteins. The following statements summarize the major conclusions of this chapter.

(1) Phylogenetic trees that are generally in agreement with trees derived from amino acid sequences can be constructed from structural information.

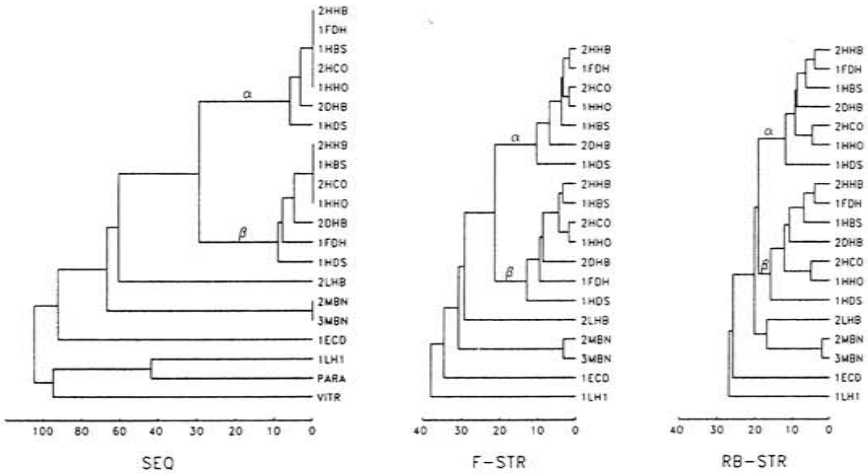

FIG. 7. Trees constructed for 19 globin structures (Table I) from pairwise rigid-body (RB-STR), multifeature (F-STR), and multiple sequence (SEQ) alignments. The tree for the multifeature approach was calculated from the same combination of properties that was used to derive the alignment in Fig. 6. α, α chains; β, β and γ chains.

(2) Trees that could not be obtained from the comparison of sequences alone may be established from structural comparisons: Evolutionary trees for divergent proteins, where the sequence relationships are not statistically significant, can be derived from structures because tertiary structure is more conserved in evolution than sequence. For example, even though the comparison of the amino- and carboxy-terminal domains of the aspartic proteinases reveals many common structural features, the sequence similarity is not statistically significant. Structural procedures can distinguish between different crystal structures for the same sequence. For example, perturbations of hemoglobin structures induced by ligand binding are seen in the tree constructed from rigid-body superpositions, where all deoxy chains cluster apart from those with bound ligands.

Acknowledgments

We thank our colleague John P. Overington for valuable discussions. We also thank Michael J. Sutcliffe. We thank the American Cancer Society (M.S.J.) for a fellowship and the ORS Awards Scheme and the Research Council of Slovenia for support (A.Š.).

Author Index

Numbers in parentheses are footnote reference numbers and indicate that an author's work is referred to although the name is not cited in the text.