PROTEINS: Structure, Function, and Genetics 14:309–323 (1992)

Multiple Protein Sequence Alignment From Tertiary Structure Comparison: Assignment of Global and Residue Confidence Levels

Robert B. Russell and Geoffrey J. Barton
Laboratory of Molecular Biophysics, University of Oxford, Oxford OX1 3QU, England

ABSTRACT **An algorithm is presented for the accurate and rapid generation of multiple protein sequence alignments from tertiary structure comparisons. A preliminary multiple sequence alignment is performed using sequence information, which then determines an initial superposition of the structures. A structure comparison algorithm is applied to all pairs of proteins in the superimposed set and a similarity tree calculated. Multiple sequence alignments are then generated by following the tree from the branches to the root. At each branchpoint of the tree, a structure-based sequence alignment and coordinate transformations are output, with the multiple alignment of all structures output at the root. The algorithm encoded in STAMP (STructural Alignment of Multiple Proteins) is shown to give alignments in good agreement with published structural accounts within the dehydrogenase fold domains, globins, and serine proteinases.**

In order to reduce the need for visual verification, two similarity indices are introduced to determine the quality of each generated structural alignment. S_c quantifies the global structural similarity between pairs or groups of proteins, whereas P_{ij}' provides a normalized measure of the confidence in the alignment of each residue. STAMP alignments have the quality of each alignment characterized by Sc and P_{ij}' values and thus provide a reproducible resource for studies of residue conservation within structural motifs. © 1992 Wiley-Liss, Inc.

Key words: **protein structure comparison, sequence alignment, structure alignment, dynamic programming, dehydrogenase fold**

INTRODUCTION

There are now many protein families with more than one member having a crystallographically or NMR determined three-dimensional structure. Comparison of structures within these families aids our understanding of the evolutionary and thermodynamic constraints on the particular protein fold (e.g., see Bashford et al.[1]). Superimposition of similar protein structures and generation of a corresponding structure-based sequence alignment is an essential step in any such analysis. The resulting alignments provide reliable information that may be used as a benchmark to establish confidence levels for sequence alignment derived without structural information (e.g., see Barton and Sternberg,[2] Argos,[3]) for the derivation of environment-specific mutability matrices (e.g., see Overington et al.,[4]) as the basis for molecular modelling studies (e.g., see Blundell et al.[5]), and for the derivation of patterns or profiles that encapsulate the essential features of the protein fold (e.g., see Barton and Sternberg,[6] Bashford et al.,[1] Taylor,[7] Luthy et al.,[8] Bowie et al.,[9]).

A number of automated techniques for the comparison of two protein three-dimensional structures have been devised. Rossmann and Argos[10] compare two structures by first superimposing them on their centroids. A probability function consisting of distance and conformational terms is then applied to the comparison of all pairs of residues, and a path tracing routine used to identify the best set of equivalences between the two proteins. A search of rotational and translational space is performed to locate the orientation of the structures that yields the largest set of equivalences.

In contrast, the techniques of Remington and Matthews[11] and McLachlan[12] first divide each of the two proteins into a set of overlapping segments of a predetermined length. All pairs of segments are then fitted by a least-squares procedure and the distribution of r.m.s. (root mean square) values examined for unusual features. This procedure avoids the need to search rotational/translational space at the cost of ignoring the conformational contribution of residues outside the length of the fragments. In common with fragment based sequence comparison methods (e.g., see Fitch[13]), the procedure does not yield directly a sequence alignment and does not cope explicitly with the possibility of gaps (insertions and deletions).

Received September 27, 1991; revision accepted February 7, 1992.

Address reprint requests to Dr. Robert B. Russell, Laboratory of Molecular Biophysics, University of Oxford, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, England.

Recently, dynamic programming sequence comparison methods that are able to produce an alignment including gaps (e.g., see Needleman and Wunsch[14]), have been applied to protein three-dimensional structures.[15] Taylor and Orengo[16–18] built these ideas into an elegant "two-level dynamic programming" procedure for three-dimensional structure comparison. Their approach finds the optimal alignment of two protein structures that simultaneously takes into account several different features of the protein conformation (e.g., phi/psi angles, accessibility, interresidue vectors). Similarly, Šali and Blundell[19] make use of many protein features in their comprehensive comparison programme COMPARER. However, they mix dynamic programming with simulated annealing optimization as an alternative to the Taylor and Orengo two-level dynamic programming method.

The procedure of Šali and Blundell[19] is the only general method previously described that can systematically generate structure-derived multiple protein sequence alignments. Sutcliffe et al.,[20] present an algorithm that is able to generate a multiple sequence alignment and superposition for closely similar structures, but their method is sensitive to concerted shifts in secondary structure elements.[19] Alignments derived by visual inspection are always effected by subjectivity to some extent, making them difficult to use for comparison purposes. This is particularly manifest when alignments derived by different people disagree. For this reason alignments obtained by the method of Šali and Blundell (when available) provide ideal benchmarks against which to compare others.

In this study we describe a multiple structure alignment procedure that provides a series of alignments for structures or domains thought to have a similar fold by following a systematically derived hierarchy of structural similarity.

The method has been successfully applied to several protein structural families. These include closely related structures, such as aspartyl proteinases or immunoglobulin constant domains, as well as more distantly related structures with little sequence similarity, e.g., triose phosphate isomerase (TIM) barrels and viral coat proteins exhibiting the "jelly roll" fold (unpublished results). The procedure has also aligned the distantly related structures of azurin and plastocyanin in agreement with Adman[21] (unpublished results). Here we illustrate the method by the alignment of the globin and serine proteinase families for which it gives nearly identical results to the more complex Šali and Blundell technique and similar to other previously published structure-based alignments. The success of the algorithm is further demonstrated by the more stringent test of aligning domains exhibiting the "dehydrogenase fold."

Our method has the advantage over previous techniques of assigning both the overall quality of alignment at each stage of the hierarchy and of providing a confidence level for each aligned group of amino acids within each alignment, thus reducing the need for extensive visual inspection of structure superpositions prior to application of the alignment.

METHODS

The multiple structure alignment algorithm proceeds in three stages: (1) generation of an initial superposition and structure-derived tree based upon a multiple sequence alignment; (2) refinement of the superposition found in stage 1 and creation of multiple sequence alignments and structural trees derived from the structural equivalences found; and (3) assignment of reliability values to each region of the alignments. The algorithm makes extensive use of three techniques, least-squares fitting, hierarchical cluster analysis, and dynamic programming. For the sake of clarity, these are briefly reviewed here.

When given two structures, *A* and *B*, techniques for least-squares fitting[22, 23] take a set of *n* atoms (*x,y,z*) from *A* and *n* equivalent atoms from *B* and calculate the transformation (translation and rotation) that minimises the r.m.s. deviation as given by:

$$(\sum_{i=1}^{n} ((x_{A_i} - x_{B_i})^2 + (y_{A_i} - y_{B_i})^2 + (z_{A_i} - z_{B_i})^2)/n)^{1/2}.$$

In the following descriptions, "fitting" or "superposition" will refer to the calculation of an r.m.s. value and associated transformations for a pair of coordinate sets.

Hierarchical cluster analysis takes *N* objects and scores calculated for the comparison of each of the $N(N-1)/2$ possible pairs of objects. The method returns a dendrogram or tree that organizes the objects according to their similarity, with the most similar objects in the group clustered at the highest "branches" of the tree. This technique has been applied to protein sequence data to estimate phylogeny,[24] as a starting point for multiple sequence alignment[25] and as a starting point for multiple structure alignment.[19] The rationale for using a tree-like addition as opposed to a simple sequential addition is that the most similar structures are aligned first, and therefore aligned most accurately. More distantly related structures will be compared at later stages allowing the most accurate alignments to be maintained for as long as possible during the procedure. Any multiple comparison method that makes use of a tree-like addition will suffer from pairwise-order dependency: the results obtained may differ if the order in which individual elements are compared is altered. The experience of others[19, 25] and ourselves suggests that a tree-like addition is an effective way of minimizing order of dependency and generally produces superior alignments over a simple sequential addition.

Dynamic programming is a general technique that takes two sequences $A_{1\ ...\ P}$, $B_{1\ ...\ Q}$ and a matrix $M_{P,Q}$ where each value $M_{i,j}$ contains the score for the comparison of A_i with B_j. The method returns the best alignment of the two sequences including insertions/deletions and a score for the alignment. Since its introduction in molecular biology,[14] dynamic programing has been widely used to optimize the alignment of two or more protein sequences.[25] However, the method is generally applicable to sequential data of any sort.[26] Here, we extend the use of dynamic programming to sequences of three-dimensional coordinates as first suggested by Barton and Sternberg.[15]

Initial Superposition

Given the three-dimensional structures of N proteins that share similar folds, the amino acid sequences are first multiply aligned using the algorithm of Barton and Sternberg.[2] The C_α atoms of the M aligned positions where no gap occurs are then used for structure superposition. For each of the $N(N-1)/2$ pairs of proteins, the M equivalent atoms are compared by a least-squares fitting procedure,[12] giving r.m.s. values for each comparison. Single linkage cluster analysis is then applied to these r.m.s. values and a dendrogram calculated. A multiple structure superposition is generated by following the dendrogram from the tips of the branches to the root, superimposing the structures joined at each branchpoint or *node*. When more than two structures are being compared, averaged atomic coordinates are used in calculating the r.m.s. fit. Figure 1 illustrates an example derivation of a dendrogram from a similarity matrix derived from the pairwise comparison of four structures.

Multiple Structure Alignment

The initial superposition is derived from a preselected set of atoms. Although the superposition provides the best three-dimensional fit of the chosen atoms according to the hierarchy, the fit will also reflect any errors in the initial sequence alignment. In order to refine the fit and to obtain a multiple sequence alignment based upon structural criteria, we have adapted the probability function of Rossmann and Argos.[10]

When comparing two structures, this function expresses the probability, P_{ij}, that residue i in one structure and residue j in another are equivalent as two terms:

$$\mathrm{P_{ij}} = \left\{\exp - \frac{\mathrm{d_{ij}^2}}{2\mathrm{E_1^2}}\right\}\left\{\exp - \frac{\mathrm{s_{ij}^2}}{2\mathrm{E_2^2}}\right\}$$

where E_1 and E_2 are constants, d_{ij} is the distance between C_α atoms, and s_{ij} is determined by:

a.

	A	*B*	*C*	*D*
A	–	3.0	1.0	4.0
B	–	–	3.5	0.5
C	–	–	–	3.4
D	–	–	–	–

b.

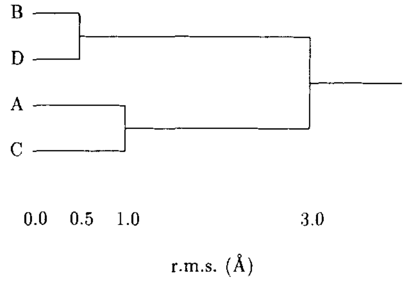

Fig. 1. Given four proteins A,B,C, and D, pairwise structure fitting may show rms values as shown in **(a)**. Single linkage cluster analysis yields the dendrogram shown in **(b)**. The initial superposition for these proteins would be obtained by fitting B to D (similarity of 0.5 Å), then A to C (1.0 Å), then fitting AC to BD using average coordinates for the AC and BD pairs to obtain the necessary transformations.

$$\begin{aligned}\mathrm{s_{ij}^2} = \{&(\Delta \mathrm{x_{ij}} - \Delta \mathrm{x_{i-1,j-1}})^2 + (\Delta \mathrm{y_{ij}} - \Delta \mathrm{y_{i-1,j-1}})^2 \\ &+ (\Delta \mathrm{z_{ij}} - \Delta \mathrm{z_{i-1,j-1}})^2 + (\Delta \mathrm{x_{ij}} - \Delta \mathrm{x_{i+1,j+1}})^2 \\ &+ (\Delta \mathrm{y_{ij}} - \Delta \mathrm{y_{i+1,j+1}})^2 + (\Delta \mathrm{z_{ij}} - \Delta \mathrm{z_{i+1,j+1}})^2\}.\end{aligned}$$

The first term gives a measure of the proximity in space of the two residues, whereas the second provides a measure of their conformational similarity. The values of E_1 and E_2 effect to weight the terms. For example, if $E_1 > E_2$, then the probability will be more indicative of the conformational similarity rather than the proximity of the two residues. This function may be readily expanded to include terms incorporating aspects of protein structure such as hydrogen bonding, distance to centroid, or solvent accessibility. The value of P_{ij} is calculated for every possible pair of residues, resulting in an m by n matrix of probabilities, where m and n are the number of residues in each structure being compared.

The path through this matrix with the highest sum of values (the *score*) corresponds to the best possible set of equivalences. This best path is found using a modified Smith-Waterman dynamic programming algorithm.[27] In our experience, no gap penalty is required to obtain good alignments and superpositions.

Once a set of equivalent residues has been found, the probability associated with each equivalence is examined. If a probability is greater than the threshold value, T (see Parameters section, following), then the corresponding pair of residues is used to perform a least-squares fit. Having superimposed the two structures, a new probability matrix is calculated, and the process is repeated iteratively until a minimum score difference (one-tenth of a percent of the score from the previous iteration) is obtained.

Multiple structure alignment follows a procedure similar to tree-based multiple sequence alignment methods (e.g., see Barton[25]). Each possible pairwise comparison for the group of proteins to be aligned is performed. Structural similarity scores, Sc (see Alignment Quality below) are used to derive a similarity matrix and corresponding dendrogram. The dendrogram is then followed from the branches to the root superimposing structures in order of their similarity.

When two groups of structures are to be aligned, the previously aligned sequences remain fixed relative to each other, and the probability matrix is calculated as follows. All pairs of comparisons between the two groups of structures are performed (e.g., if the alignment of A and B is to be aligned with the alignment of C and D, then the possible comparisons are A–C, A–D, B–C, and B–D). The average P_{ij} values for these comparisons at each position are used to create the probability matrix. In instances where a residue in one protein is to be compared to a gap in another, the neutral value of zero is added to the probability matrix at that position. The best path through the probability matrix, and the set of equivalent residues are found as for the pairwise comparisons.

In order to obtain an appropriate set of coordinates with which to perform a fit, an average set of equivalenced C_α atoms is calculated for both groups of structures being superimposed. Having superimposed the two groups of structures, the process is repeated iteratively as for the pairwise comparisons. The obtained transformations and corresponding sequence alignments are output at each node of the dendrogram so that each sub-alignment may be examined separately.

Normalization

In order that the relative quality of alignments within different families of proteins may be assessed, the P_{ij} values must be expressed on a standard scale. Pairwise alignments of different families of proteins show that equivalent P_{ij} values are observed for similarly aligned regions of secondary structure. However, the averaged P_{ij} values when comparing two structures decrease as the number of structures aligned increase, as do the mean and standard deviation (s.d.) This behavior is shown for 68 globins in Figure 2a. The mean and s.d. are also functions of the length of the structures being compared (Fig. 2b).

The dependency of mean and s.d. on length may be corrected by selecting a typical mean ($\bar{x}_t = 0.020$) and s.d. ($\sigma_t = 0.10$) to give a standard background measure for any comparison. These values were chosen as they are representative of typical protein lengths (i.e., 100–200 residues). The dependency of P_{ij} values on the number of structures being compared is overcome by applying a correction to these typical values for each multiple alignment as follows. During multiple alignment the mean and s.d. for the obtained averaged P_{ij} values are calculated and the corresponding pairwise values (i.e., for a single sequence of a similar length) are determined from the exponential relationship shown in Figure 2b:

$$\bar{x}_{\text{pairwise}} = \exp(-0.950 \log(L) + 0.686)$$
$$\sigma_{\text{pairwise}} = \exp(-0.474 \log(L) + 0.0152)$$

where L is the average length of the two alignments or single sequences being aligned. The ratio of multiple to pairwise values are used to correct the typical value:

$$\bar{x}_c = \bar{x}_t \left(\frac{\bar{x}_{\text{multiple}}}{\bar{x}_{\text{pairwise}}}\right) \qquad \sigma_c = \sigma_t \left(\frac{\sigma_{\text{multiple}}}{\sigma_{\text{pairwise}}}\right).$$

P'_{ij} values are calculated from:

$$P'_{ij} = \left(\frac{P_{ij} - \bar{x}_c}{\sigma_c}\right).$$

The P'_{ij} values follow similar trends over a wide range of protein structural families. Using P'_{ij} instead of P_{ij} enables standards for individual residue accuracy and overall alignment quality to be established (see Alignment Quality).

Parameters

Setting $E_1 = E_2 = 3.8$, after Rossmann and Argos[10] leads to good superpositions and alignments, provided that the initial superposition based on sequence alignment is reasonable. Setting $E_1 < E_2$ (making the probability of equivalence more dependent on distance than local conformation) leads to alignments based to heavily on the actual distance between atoms and does not accommodate concerted shifts in secondary structure. Setting $E_1 > E_2$ (making the probability of equivalence more dependent on local conformation) can be useful if it is thought that the initial superposition based on sequence alignment is poor. This relaxes the proximity requirement and avoids equivalencing those residues that the initial superposition has placed fortuitously near to each other. In such instances it is necessary to apply the method twice, once with $E_1 > E_2$, and again with $E_1 = E_2$. In this way the poor initial superposition can be corrected before being finally refined (e.g., see Dehydrogenase Fold Domains below).

A threshold value of $P'_{ij} > T = 4.5$ was found to

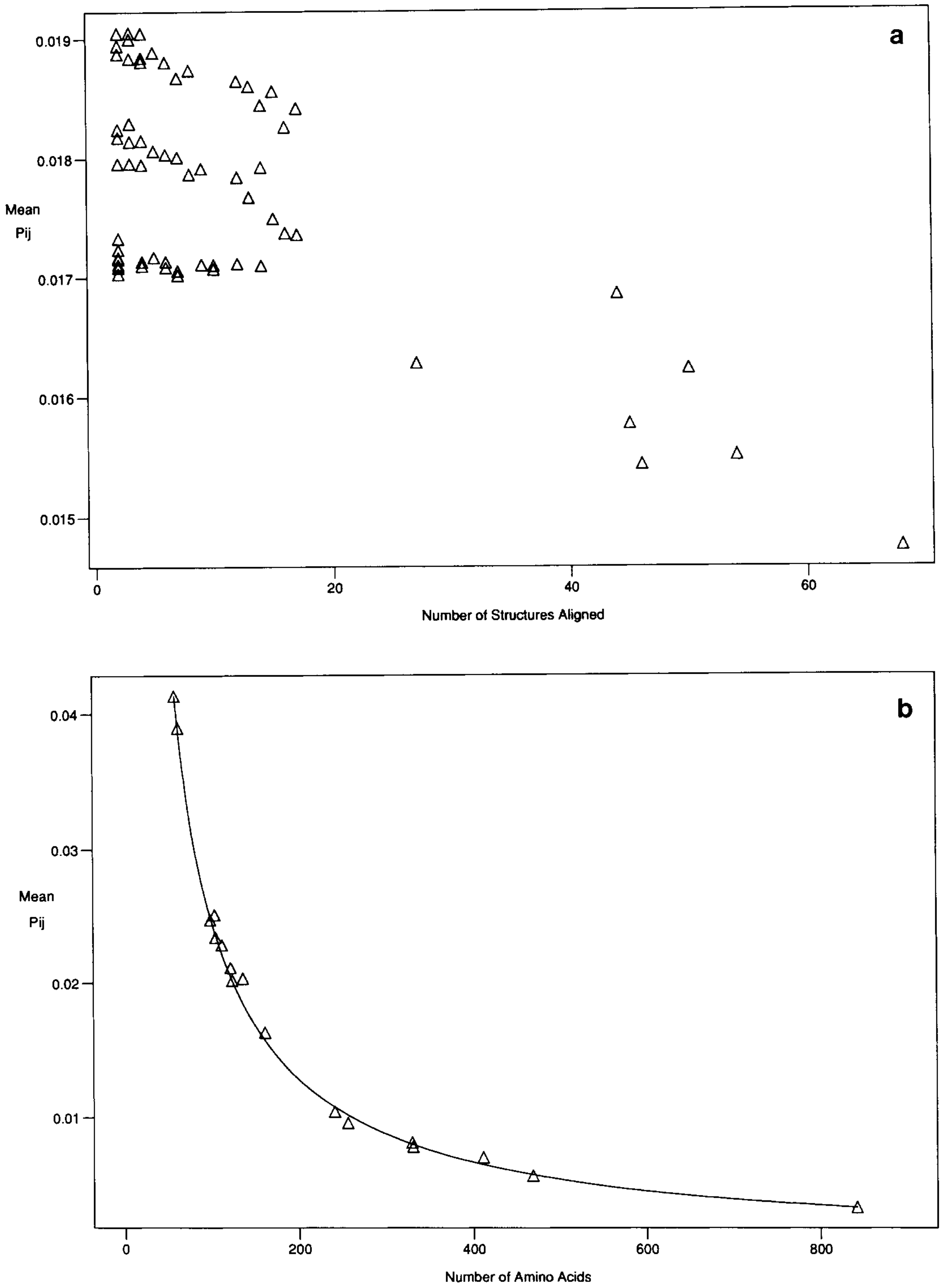

Fig. 2. **(a)** Plot of mean P_{ij} versus the number of structures aligned for the globins ($E_1 = E_2 = 3.8$). The plot of standard deviation vs. number of structures aligned shows a similar trend. **(b)** Plot of mean P_{ij} vs. number of amino acids for pairwise comparisons of proteins (of lengths ranging from 58 to 842) with themselves ($E_1 = E_2 = 3.8$). The plot of standard deviation vs. length also shows a similar trend.

yield good superpositions and alignments. Values less than 4.5 result in too many poor equivalences being chosen, leading to a poor fit, whereas values greater than 4.5 give too few equivalences and as a result the method is unable to correct an inaccurate initial superposition.

TABLE I. Examples of S_c Values Obtained for Pairwise Comparisons*

	Brookhaven codes (chain, range)			% reliability[†]	
Comparison	A	B	S_c	A	B
Globin compared to itself	1PMB (B, all)	—	9.8	100.0	100.0
Serine proteinase compared to itself	3SGB (E, all)	—	9.8	100.0	100.0
Immunoglobulin variable domains	2FB4 (L, 1–109)	2IG2 (L, 1–109)	9.6	98.2	98.2
	2FB4 (L, 1–109)	1MCP (H, 1–124)	5.9	64.9	58.1
Aspartyl proteinases	4APE (—, all)	2APR (—, all)	7.7	83.6	84.9
	4APE (—, all)	1PSG (—, all)	5.8	70.6	63.8
Snake toxins	1NXB (—, all)	1CTX (—, all)	5.5	62.9	54.9
Viral coat proteins	2MEV (2, all)	4SBV (A, all)	4.0	41.0	51.3
	2MEV (1, all)	2RS1 (3, all)	3.3	31.0	35.2
Immunoglobulin constant and variable domains	2IG2 (L, 1–109)	1MCP (L, 115–220)	3.8	39.6	41.5
	2IG2 (L, 1–109)	4FAB (H, 118–216)	3.4	32.4	36.4
Dehydrogenase fold domain and globin	8ADH (—, 191–320)	2HHB (B, all)	1.1	15.8	17.7
Globin and immunoglobulin variable domain	2HHB (B, all)	2IG2 (L, 1–109)	0.068	0.0	0.0

*Values quoted are for the comparison of structure A with B.

[†]Refers to percentage of each structure that lies in regions of high reliability (i.e., stretches of three residues or more having $P'_{ij} > 6.0$).

Alignment Quality

For a structure superposition and alignment to be useful without recourse to visual inspection of superpositions, it is necessary to have criteria for assessing the overall alignment quality, and for locating regions of high reliability within the alignment. These criteria were determined by aligning nine families of structures (immunoglobulin constant and variable domains, aspartyl proteinases, snake toxins, viral coat proteins, cytochrome c structures, serine proteinases, dehydrogenase fold domains, and globins) and analyzing the resulting superpositions using interactive graphics.

Individual equivalence accuracy

Examination of the superimposed families of proteins shows that regions within the alignments having P'_{ij} greater than 6.0 for stretches of three residues or more generally correspond to highly conserved elements of secondary structure when $E_1 = E_2 = 3.8$. Regions with values between 4.0 and 6.0 are normally structurally equivalent (>50% of the time). Such regions often correspond to alignment of similar, yet not identical regions of secondary structure, such as the alignment of a β-turn with one turn of an α-helix. Regions with values of less than 4.0 do not usually correspond to structurally conserved regions.

Overall quality

The scores from the dynamic programming (best path) routine provide a measure of alignment quality for both the pairwise and multiple comparisons. However, since these scores are a function of alignment length, it is necessary to normalize them so that alignments of unrelated families of proteins may be compared to one another. A more accurate measure of alignment quality is achieved by modifying the score, S_p (a sum of P'_{ij} values), from the best path routine as follows:

$$S_c = \left(\frac{S_p}{L_p}\right)\left(\frac{L_p - i_a}{L_a}\right)\left(\frac{L_p - i_b}{L_b}\right)$$

where S_c is the structural similarity score, L_p is the path length returned from the best path routine, L_a and L_b are the lengths of the two alignments (or single sequences) compared and i_a and i_b are the total length of gaps introduced into each alignment or sequence during the derivation of the new alignment. The first term removes the dependence on length, whereas the second two terms prevent short stretches of equivalences between unrelated proteins from yielding high scores.

Criteria for assessment of overall alignment quality were determined by examining a dendrogram for representatives of nine structural families. Comparison of any structure, such as a globin or serine proteinase (Table I), with itself yields a value of 9.8 (i.e., all P_{ij} values are equal to 1.0 for the best path). S_c values between 5.5 and 9.8 are obtained for the comparison of highly similar structures, such as immunoglobulin variable domains, or aspartyl proteinases, and generally imply that the alignments are reliable over >50% of their length. S_c values between 2.5 and 5.5 are obtained when less similar structures are compared, such as immunoglobulin constant and variable domains, snake toxins, or viral coat proteins. Values less than 2.5 indicate unrelated structures such as the comparison of a globin to an immunoglobulin variable domain. However,

TABLE II. Structures Used for Derivation of Structural Alignments of Globin and Serine Proteinase Families

Proteins	Brookhaven code (chain)	Length	Resolution	R-factor (Å)	Reference
Globins					
Human deoxy hemoglobin α-chain	2HHB(A)	141	1.7	0.16	42
Human deoxy hemoglobin β-chain	2HHB(B)	146	1.7	0.16	42
Sea lamprey hemoglobin V (cyano/met)	2LHB	149	2.0	0.14	43
Sperm whale deoxy myoglobin	4MBN	153	2.0	0.23	44
Chironomous thummi thummi erythrocruorin	1ECD	136	1.4	0.19	45
Lupinus luteus leghemoglobin	1LH1	153	2.0	—	46
Serine proteinases					
Bovine α-chymotrypsin	4CHA(A)	239	1.68	0.23	47
Porcine elastase	3EST	240	1.65	0.17	48
Bovine trypsin (orthorhombic)	2PTN	223	1.55	0.19	49
Rat tonin	1TON	227	1.80	0.20	50
Rat mast cell proteinase	3RP2(A)	224	1.90	0.19	51
Porcine kallikrein	2PKA(A,B)	232	2.05	0.22	52
Streptomyces griseus trypsin	1SGT	223	1.70	0.16	53
S. griseus proteinase A	2SGA	181	1.50	0.13	54
S. griseus proteinase B	3SGB(E)	185	1.80	0.13	55
L. enzymogenes α-lytic proteinase	2ALP	198	1.70	0.13	56

values greater than 1.0 may indicate the alignment of secondary structure elements, such as during the comparison of a dehydrogenase fold domain to a globin, where three helices are observed to be aligned.

Program Details

The program for the initial superposition based on sequence alignment is written in Fortran 77, whereas the multiple structure alignment program is written in C. Both programs were developed on a Sun SPARCstation 1. Alignment of 6 globins takes ~ 4 minutes, whereas an alignment of 68 globins can be accomplished in ~ 9 hours.

Running the STAMP (Structural Alignment of Multiple Proteins) package on a family of protein structures results in a set of hierarchically grouped multiple sequence alignment files, one for each node in the structural similarity tree. Each file contains the transformations necessary to superimpose the native coordinates, the parameters used to generate the alignment, S_c value for the alignment, and P_{ij}' values for each aligned residue. Given a set of alignments for a family, it is therefore a simple matter to ask questions specific either to the most highly conserved (high P_{ij}') or variable (low P_{ij}') regions of the protein family. The hierarchical organisation of the structural alignments by S_c values also permits queries to be related to the overall similarity of the proteins.

The STAMP package is available from the authors.

Demonstration of the Method

The six globins, and the ten serine proteinases chosen by Šali and Blundell[19] (see Table IV) were selected to provide a direct comparison of the methods. Seven dehydrogenase fold domains were compared to demonstrate the effectiveness of the method on proteins with little sequence similarity. Tables II and III give summaries of the structures used for evaluation.

All protein structures were taken from the November 1990 release of the Brookhaven protein data bank,[28] except for glycogen phosphorylase b and 6-phosphogluconate dehydrogenase, which were kindly provided by Prof. L. N. Johnson and Dr. M. J. Adams, respectively. Secondary structure definitions were obtained from Kabsch and Sanders program DSSP [29].

RESULTS AND DISCUSSION

Globins

The globins are a family of all-α proteins that normally consist of eight helices arranged around a central heme group. The specific interactions of these helices have been studied by Lesk and Chothia[30] and Bashford et al.,[1] leading to the identification of 102 common residue sites between all globins. These studies provide a structurally based sequence alignment against which to compare the alignments obtained by our method.

The pairwise structural comparisons of these proteins give S_c values greater than 5.5 indicating a high degree of structural similarity between even the most distantly related globins (see Table IV for examples).

The final structurally derived alignment appears in Figure 3 (boxed, upper case positions in structural alignments correspond to those regions that the method shows to be reliably aligned—having P_{ij}' val-

TABLE III. Structures Used for the Derivation of Structural Alignments of the Dehydrogenase Fold Domain Family

Proteins	Brookhaven code (chain)	Range	Length	Resolution	R-factor (Å)	Reference
Dogfish lactate dehydrogenase	6LDH	20–163	142	2.0	0.202	57
Porcine lactate dehydrogenase	5LDH	20–163	142	2.7	0.196	58
Porcine malate dehydrogenase	4MDH(A)	14–160	160	2.5	0.167	59
Horse alcohol dehydrogenase	8ADH	191–320	130	2.4	0.190	60
Bacillus stearothermophilus glyceraldehyde 3-phosphate dehydrogenase	1GD1(O)	0–149	151	1.8	0.177	61
Rabbit skeletal muscle glycogen phosphorylase b	1GPB	562–711	150	1.9	0.191	36
Sheep 6-phosphogluconate dehydrogenase	PGDH*	1–128	128	2.5	0.185	62

*Code assigned in the absence of a Brookhaven code for 6-phosphogluconate dehydrogenase.

TABLE IV. Examples of S_c Values Obtained for Pairwise Comparisons Within Globin, Serine Proteinase, and Dehydrogenase Fold Domain Structural Families*

	Brookhaven codes (chain, range)			% reliability[†]	
Comparison	A	B	S_c	A	B
Globins	4MBN (—, all)	2HHB (B, all)	8.4	88.9	93.2
	4MBN (—, all)	1ECD (—, all)	7.4	80.4	90.4
	2HHB (A, all)	1LH1 (—, all)	5.8	61.0	56.2
Bacterial serine proteinases	3SGB (E, all)	2SGA (—, all)	8.4	84.9	86.7
Mammalian serine proteinases	2PTN (—, all)	1TON (—, all)	8.2	87.4	85.9
Mammalian and bacterial serine proteinases	3SGB (E, all)	1TON (—, all)	3.8	50.3	41.0
Dehydrogenase fold domains	5LDH (—, 20–168)	6LDH (—, 20–163)	7.6	88.4	90.3
	5LDH (—, 20–168)	4MDH (A, 1–160)	6.0	60.5	55.6
	8ADH (—, 191–329)	4MDH (A, 1–160)	3.9	50.0	40.6
	6LDH (—, 20–163)	1GD1 (O, 0–149)	3.4	36.8	35.1
	1GPB (—, 562–711)	1GD1 (O, 0–149)	2.6	22.0	21.9

*Values quoted are for the comparison of structure A with B.

[†]Refers to percentage of each structure (A and B) that lies in regions of high reliability (i.e., stretches of three residues or more having $P'_{ij} > 6.0$).

ues of greater than 6.0 over a continuous stretch of three residues or more). This alignment, considering both distance and local conformation agrees very closely with the alignments obtained by visual inspection,[1, 30] and with that obtained by the method of Šali and Blundell.[19] There are two differences within the regions identified by our algorithm to be reliably aligned (111 equivalences) and the alignment of Šali and Blundell. These lie in loop regions at the ends of helices where it is difficult to assign equivalences even by careful visual inspection. The first difference occurs at the C-terminal end of the C-helix (positions 54–56 in Fig. 3). In this region the Šali and Blundell alignment place an insertion of one residue in human hemoglobin α-chain (2HHBA; at position 56). The second difference occurs at the C-terminal end of the F helix (positions 111–114 in Fig. 3). Here our method places an insertion of one residue in Erythrocruorin (1ECD) and 1LH1 two residues before COMPARER does the same.

In contrast, the rigid body superposition method for Sutcliffe et al. fails to duplicate the alignment obtained by Lesk and Chothia, due to the variations helix packing angles between the most distantly related globins.[19]

Serine Proteinases

The serine proteinases are a family of β sheet proteins consisting of two similar β-barrel domains each of six antiparallel strands. A number of mammalian and microbial structures have been determined by X-ray crystallography and although there is <21% sequence identity between the mammalian and microbial serine proteinases, it has been observed that they adopt similar three dimensional structures.[31, 32]

Several attempts have been made to obtain a sequence alignment of mammalian and microbial serine proteases. James et al.[31] used a partially automated procedure and found that 60% of the α-carbon positions to be equivalent between these two groups; Craik et al.[33] also determined an alignment. These alignments disagree with each other in sev-

```
        0        10        20        30        40        50        60        70
1ECD               LSADQISTVQASFDKVkg    DPVGILYAVFKADPSIMAKFTQFAg k dlesiKGTAPFET
4MBN              vLSEGEWQLVLHVWAKVeadvagHGQDILIRLFKSHPETLEKFDRFKh lkteaemKASEDLKK
2HHBB           v hLTPEEKSAVTALWGKV  nvdeVGGEALGRLLVVYPWTQRFFESFG dlstpdavMGNPKVKA
2LHB     pivdtgsvapLSAAEKTKIRSAWAPVystyetSGVDILVKFFTSTPAAQEFFPKFKg lttadelKKSADVRW
1LH1             gaLTESQAALVKSSWEEFnanipkHTHRFFILVLEIAPAAKDLFSFLKg tsev  pQNNPELQA
2HHBA             vLSPADKTNVKAAWGKVgahageYGAEALERMFLSFPTTKTYFPHFD  l     sHGSAQVKG
Helix               aaaaaaaaaaaaaaa        bbbbbbbbbbbccccccccccc         ddeeeeee
                                                            ddddd

             80        90       100       110       120       130       140
1ECD    HANRIVGFFSKIIGEl    pn iea dVNTFVASHK PRGVTHDQLNNFRAGFVSYMKaht d fag aeAA
4MBN    HGVTVLTALGAILKKk    gh hea eLKPLAQSHAtKHKIPIKYLEFISEAIIHVLHsrhpgdfgadaqGA
2HHBB   HGKKVLGAFSDGLAHl    dn lkg tFATLSELHCdKLHVDPENFRLLGNVLVCVLAhhfgkeftppvqAA
2LHB    HAERIINAVDDAVASm dd tekmsm kLRNLSGKHAkSFQVDPEYFKVLAAVIADTVA      a gd  AG
1LH1    HAGKVFKLVYEAAIQlevtgvv vtdatLKNLGSVHV SKGVADAHFPVVKEAILKTIKevvgakwseelnSA
2HHBA   HGKKVADALTNAVAHv    dd mpn aLSALSDLHAhKLRVDPVNFKLLSHCLLVTLAahlpaeftpavhAS
Helix   eeeeeeeeeeeeeee             fffffffff ffffggggggggggggggggg            hh
                      f                       gggg

         150       160
1ECD    WGATLDTFFGMIFSKM
4MBN    MNKALELFRKDIAAKYkelgyqg
2HHBB   YQKVVAGVANALAHKYh
2LHB    FEKLMSMICILLRSAY
1LH1    WTIAYDELAIVIKKEMddaa
2HHBA   LDKFLASVSTVLTSKYr
Helix   hhhhhhhhhhhhhhhh
```

Fig. 3. Structurally derived sequence alignment of six globins ($E_1 = E_2 = 3.8$, $T = 4.5$). Boxed, upper case regions indicate regions of high reliability (i.e., $P'_{ij} > 6.0$). "Helix" denotes the labelling of globin helices after Bashford et al.,[1] for these regions. The last digit in the numbers above the alignment shows alignment position.

eral regions, hence neither provides an ideal benchmark to test the results of our method. Šali and Blundell applied their program COMPARER to this set of proteins and obtained a systematically derived alignment based on several features of protein structure. Though lacking in any definition of individual equivalence accuracy, this alignment is free from the subjectivity attached to manual alignments, hence provides an ideal template against which to compare the results obtained by our method.

The S_c values for these comparisons (see Table IV for examples) show a high degree of similarity between the mammalian proteinases and *S. Griseus* trypsin, as well as between the other bacterial proteinases ($S_c > 6.8$). As expected, comparison of members from each of these sets of proteins indicates a much lower degree of similarity ($S_c < 3.9$).

The final structurally based multiple sequence alignment appears in Figure 4. There is very little similarity between the this and the conventional sequence-derived alignment (unpublished results). However, the alignment derived by our method shows only one difference to the alignment of Šali and Blundell[19] within the regions found to be reliably aligned: a shift of a four residue segment within a turn of *Streptomyces griseus* trypsin (1SGT; positions 165–170 in Fig. 4). Our alignment shifts this segment one residue towards the N-terminus relative to its position within the alignment of Šali and Blundell. The differences between our alignment and those obtained prior to Šali and Blundell are more pronounced, but have been discussed previously.[19]

Dehydrogenase Fold Domains

The Dehydrogenase, or Rossmann, fold consists of six strands of parallel β-sheet (βA, βB, βC, βD, βE, βF after Rossmann et al.[34]) with four or five helices (αB, αC, αD, αE, αF) running antiparallel to the sheet. It consists of two roughly identical units related by an ~ twofold axis running parallel to the strands between βA and βD. This fold has been observed in many dinucleotide binding domains including several dehydrogenases[35] and, more recently, in glycogen phosphorylase.[36] Although they exhibit almost identical folds, these structures have little sequence similarity.

Several attempts have been made to obtain structurally based alignments of these domains. Rossmann and coworkers[37, 35, 38] report structural

```
        0         10        20        30         40          50          60          70
3EST                  vvggteaqrnswps QISLQ y  r    sgsswah TCGG TLI r       qn WVMTAAHCV dr     e
4CHA  cg v paiqpvlsivngeeavpgswpw QVSLQ d  k     t    gfh FCGG SLI n       en WVVTAAHCG v      t
2PKA                  iiggrecek nshpw QVAIY h  y         ssf QCGG VLV n       pk WVLTAAHCK n      d
2PTN                  ivggytcgantvpy QVSLN s            gyh FCGG SLI n       sq WVVSAAHCY k      s
1TON                  ivggykceknsqpw QVAVI n             ey LCGG VLI d       ps WVITAAHCY s      n
3RP2                  iiggvesiphsrpy MAHLD i  vtekgl     rv ICGG FLI s       rq FVLTAAHCK g      r
1SGT                  vvggtraaqgefpf MVRLS                m GCGG ALY a       qd IVLTAAHCV sgsgnn
2ALP                        ani    vg GIEYS inn          as LCSV g FSV trgatk GFVTAGHCG t v na
3SGB                          i    sg GDAIY s  s         tg RCSL g FNV rsgsty YFLTAGHCT dg a  t
2SGA                          i    ag GEAIT t  g         gs RCSL g FNV svngva HALTAGHCT ni s  a
KS                                   bbbbb                  bbbb bbb          bbbbbbhhh

          80         90            100        110         120          130          140
3EST  l TFR vvvgehnlnq nngteq     yvg VQKIVV hpywntd       dvaa GYDIALLRLA qsv  tlnsyvql
4CHA  t SDV vvagefdqgsssek  iq    klk IAKVFK nskynsl       ti   NNDITLLKLS taa  sfsqtvsa
2PKA    NYE vwlgrhnlfe nentaq     ffg VTADFP hpgf nlsadgkdy     SHDLMLLRLQ spa  kitdavkv
2PTN    GIQ vrlgedninv vegneq     fis ASKSIV hpsynsn       tl   NNDIMLIKLK saa  slnsrvas
1TON    NYQ vllgrnnlfk depfaq     rrl VRQSFR hpdyiplp   vhdh    SNDLMLLHLS epa  ditggvkv
3RP2    EIT vilgahdvrk restqq     kik VEKQII hesynsv       pn   LHDIMLLKLE kkv  eltpavnv
1SGT  t SIT atggvvdlqs g a  av    kvr STKVLQ apgyng        t    GKDWALIKLA qpi  nq     pt
2ALP    TAR ig               gavv GTFAAR v f p                  GNDRAWVSLT  saqtl    lprv
3SGB    TWW an      s       arttvl GTTSGS s f p                 NNDYGIVRYT nttipk    dgtv
2SGA    SWS                      i GTRTGT s f p                 NNDYGIIRHS npa  aa   dgrv
KS      bbb                        bbbbbb                       bbbbbbbbbb

      150       160       170          180        190         200         210
3EST             gvlpragt ILAN n SPCYITGW gltrt  n gqlaqtlq QAYLPTVD  yaicssssywgst
4CHA             vclpsasd DFAA g TTCVTTGW gltr   y antpdrlq QASLPLLS  ntnck  kywgtk
2PKA             lelpt  q EPEL g STCEASGW gsiepgpddfefpdeiq CVQLTLLQ  ntfca  dahpdk
2PTN             islpt  s CASA g TQCLISGW gntks sg tsypdvlk CLKAPILS  dssck  saypgq
1TON             idlpt  k EPKV g STCLASGW gstnp se mvvshdlq CVNIHLLS  nekc      iet
3RP2             vplpspsd FIHP g AMCWAAGW gktgv  r dptsytlr EVELRIMD  ekacv  dyr y
1SGT             lkiat tt AYNQ   GTFTVAGW ganre  g gsqqryll KANVPFVS  daacr  saygne
2ALP  an g  ssfvtvrgs   t EAAV g AAVCRSGR  t           t    g YQCGTITA kn   v     tanya
3SGB  g        gqditsa  a NATV g MAVTRRGS  t           t    g THSGSVTA ln   a     tvnyg
2SGA  yl y ngsyqditta   g NAFV g QAVQRSGS  t           t    g LRSGSVTG ln   a     tvnygs
KS                        tttt   bbbbbbbb                     bbbbbbbb

             230          240       250          260       270          280      290
3EST      v      k  ns MVCA ggd  gvr SGCQGDSGGPLHC lvn gqy AVHGVTSFVS rlg      cn vt  rkp
4CHA      i      k  da MICA gas   gv SSCMGDSGGPLVC kkn gaw TLVGIVSWGS s t      cs ts   tp
2PKA      v      t  es MICA gyl pggk DTCMGDSGGPLIC n     g MWQGITSWGH t p      cg sa  nkp
2PTN      i      t  sn MFCA gyl eggk DSCQGDSGGPVVC s     g KLQGIVSWGS   g      ca qk  nkp
1TON      ykdnvt  d v MLCA gem eggk DTCAGDSGGPLIC d     g VLQGITSGGA t p      ca kp  ktp
3RP2      y       eykf QVCV gsp ttlr AAFMGDSGGPLLC a     g VAHGIVSYGH p d      a  k    pp
1SGT      l       vane EICA gypdtggv DTCQGDSGGPMFR kdnadew IQVGIVSWGY  g      ca rp  gyp
2ALP e gav       r  g LTQG n       a CMGRGDSGGSWIT sa    g QAQGVMSGGN v qsngnncgipasqrs
3SGB ggdvv       y  g MIRT n       v CAEPGDSGGPLYS g     t RAIGLTSGGS  g      ncssg   gt
2SGA s  giv      y  g MIQT n       v CAQPGDSGGSLFA g     s TALGLTSGGS  g      ncrtg   gt
KS                    bbbb           bbttttttbbbb          bbbbbbbbbb

          300          310
3EST t VFTRVSAYISWI nnviasn
4CHA g VYARVTALVNWV qqtlaan
2PKA s IYTKLIFYLDWI ddtitenp
2PTN g VYTKVCNYVSWI kqtiasn
1TON a IYAKLIKFTSWI kkvmkenp
3RP2 a IFTRVSTYVPWI navin
1SGT g VYTEVSTFASAI asaartl
2ALP s LFERLQPILSQY g   lslvtg
3SGB t FFQPVTEALVAY g   vsvy
2SGA t FYQPVTEALSAY g   atvl
KS     bbbbhhhhhhhh
```

Fig. 4. Structurally derived sequence alignment of 10 serine proteinases ($E_1 = E_2 = 3.8$, $T = 4.5$). Boxed, upper case regions indicate areas of high reliability (i.e, $P'_{ij} > 6.0$). "KS" denotes the predominant Kabsch and Sander (DSSP) secondary structure assignment[29] for each of these regions ("h" denotes helix, "b" denotes sheet, "t" denotes turn). The last digit in the numbers above the alignment shows alignment position.

alignments of lactate dehydrogenase (6LDH), glyceraldehyde-3- phosphate dehydrogenase (1GD1), and alcohol dehydrogenase (8ADH). These alignments are inconsistent with each other, particularly within the helical regions.

The initial superposition of the dehydrogenase fold domains obtained from conventional sequence alignments is inaccurate. However, this orientation provides a good starting point for structural refinement. Using a sequence alignment to derive a preliminary superposition has the advantage of avoiding the exhaustive search of the rotational space proposed by Rossmann and Argos,[10] or the need to provide an initial fit by hand. In order to allow for the possible poor initial superpostion, the distance component of the probability equation was first weighted down by assigning $E_1 = 20.0$, $E_2 = 3.8$, and $T = 1.0$. The algorithm was then applied twice, first with this condition, then with the parameters set equal ($E_1 = E_2 = 3.8$ and $T = 4.5$) to give a similar refinement to that for the globins and serine proteinases.

The S_c values obtained for these domains (see Table IV) range in value from 7.6 for the most similar structures (6LDH and 5LDH) to 2.6 for the least similar structures (1GPB and 8ADH).

Figure 5 shows the final structurally derived sequence alignment for the dehydrogenase fold domains. The regions found to be reliable agree well with those obtained for 6LDH, 8ADH and 1GD1 by Rossmann and coworkers.[37, 35, 38] Some differences are observed within the alignments of the second and fourth helices (αC and αE; regions 71–78 and 181–187, respectively, in Fig. 5). However, previously published alignments are inconsistent within these regions making an objective comparison difficult. The alignment of αC differs from that of Otto et al.,[38] however, their alignment is derived on the basis of β-sheet superposition only[35] and is suspect within helical regions. The alignment of αE is identical to that reported by Rossmann et al.,[35] but it differs slightly from the alignment of Otto et al. (8ADH is shifted two residues along in our alignment). Our method aligns all six β-strands in agreement with both of these previous alignments.

The regions in the alignment found to be reliably equivalenced (upper case in Fig. 5) correspond the six stranded parallel β-sheet (βA to βF), and three of the helices (αB, αC and αE). The absence of helices αD and αF from the reliable regions is understandable as these helices are the least structurally conserved.[37] During the early stages of the procedure, when fewer, more similar structures are being aligned, both of these helices are found to be conserved (αD between 5LDH, 6LDH and 4MDH, and αF between 5LDH, 6LDH, 4MDH, and PGDH). Only when more distantly related structures are being compared do these helices no longer appear among the reliable regions.

Analysis of the reliably aligned regions reveals several interesting differences in the way the domains accommodate a similar fold. A striking example is the way in which αE (positions 181–187 in Fig. 5) packs against the sheet in malate dehydrogenase (4MDH), glycogen phosphorylase b (1GPB), and glyceraldehyde-3-phosphate dehydrogenase (1GD1). In all three structures one central residue on αE (position 184 in Fig. 5) interacts principally with residues on the sheet below; however, the actual residue-residue interactions are very different. In 4MDH, a leucine (LEU 115) residue on this helix interacts principally with two other hydrophobic residues (LEU 85 on βD and VAL 123 on βE, positions 148 and 195 in Fig. 5) on the β-sheet. In 1GPB, this position is occupied by a phenylalaine (PHE 681), which packs against several residues on the β-sheet. Finally, in 1GD1, this residue is a histidine (HIS 108) and forms a hydrogen bond with a glutamic acid residue (GLU 94 on βD, position 148 in Fig. 5) on the β-sheet ($N_{\epsilon2}$ on HIS 108 is 2.8 Å away from $O_{\epsilon2}$ on GLU 94). This helps to explain why conventional sequence alignment fails to align these domains, as several entirely different residues appear to accomplish the same interaction. A more detailed investigation of these interactions is in progress.

Bork and Grunwald[39] constructed consensus sequence patterns of several dehydrogenase NAD binding sites on the basis of 11 steric and physicochemical properties. They make use of these patterns to distinguish between nucleotide binding sites on the basis of sequence alone. They focus on the first β-α-β moiety, which contains the ribose binding loop sequence pattern GLY-X-GLY-X-X-GLY/ALA, which has been used to “fingerprint” nucleotide binding domains previously.[40] The alignment obtained by our method agrees with Bork and Grunwalds sequence-based alignment in this region (positions 1–58 in Fig. 5). It is interesting to note that the dehydrogenase fold domain of glycogen phosphorylase, which does not bind nucleotides in the same way as the other domains,[41] bears little structural or sequence similarity to the other domains in this region. An illustration of this difference is shown in Figure 6.

CONCLUSIONS

In this study, an algorithm is described for the generation of multiple protein sequence alignments from a comparison of their three-dimensional structures. The general conclusions are as follows.

1. The algorithm is efficient (typically taking less than 15 minutes of CPU time for groups of up to 10 structures) and automatic (no prior fitting is required).
2. Multiple structure alignments generated by the method agree with established alignments.

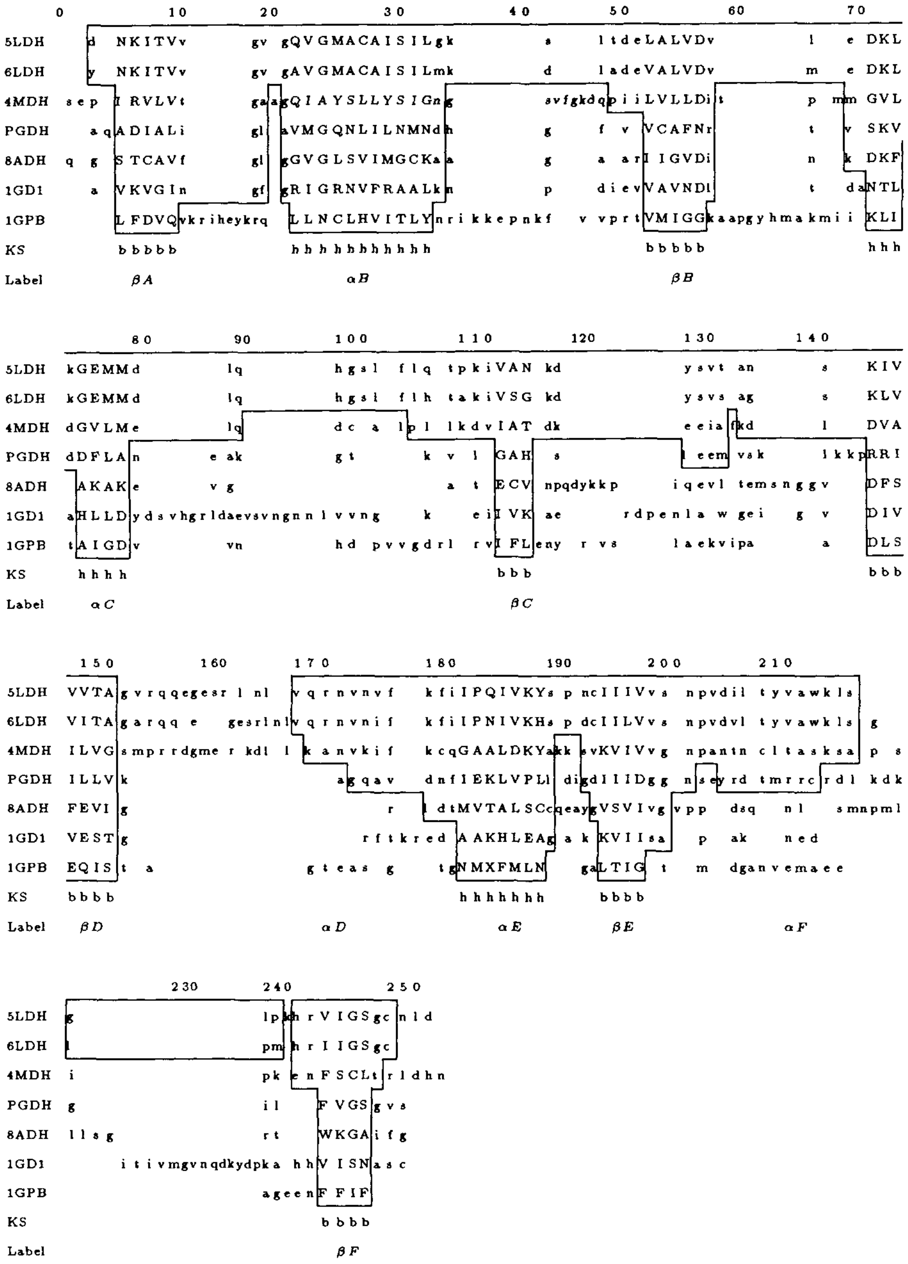
0 10 20 30 40 50 60 70
5LDH d NKITVv gv gQVGMACAISILgk s ltdeLALVDv l e DKL
6LDH y NKITVv gv gAVGMACAISILmk d ladeVALVDv m e DKL
4MDH sep I RVLVt gaag gQIAYSLLYSIGn g svfgkdqp i iLVLLDi t p mm GVL
PGDH aq ADIALi gl aVMGQNLILNMNd h g f v VCAFNr t v SKV
8ADH q g STCAVf gl gGVGLSVIMGCKa a g a ar I IGVDi n k DKF
1GD1 a VKVGIn gf gRIGRNVFRAALk n p diev VAVNDl t daNTL
1GPB L FDVQvkriheykrq LLNCLHVITLYnrikkepnkf v vprt VMIGGkaapgyhmakmii KLI
KS bbbbb hhhhhhhhhh bbbbb hhh
Label βA αB βB
80 90 100 110 120 130 140
5LDH kGEMMd lq hgsl flq tpkiVAN kd ysvt an s KIV
6LDH kGEMMd lq hgsl flh takiVSG kd ysvs ag s KLV
4MDH dGVLMe lq dc a lp l lkdvIAT dk eeia fkd l DVA
PGDH dDFLA n e ak gt k v l GAH s l eem vsk lkkp RRI
8ADH AKAK e v g a t ECV npqdykkp iqevl temsnggv DFS
1GD1 aHLLD ydsvhgrldaevsvngnnlvvng k ei I VK ae rdpenla w gei g v DIV
1GPB tAIGD v vn hd pvvgdrl rv I FL eny r vs laekvipa a DLS
KS hhhh bbb bbb
Label αC βC
150 160 170 180 190 200 210
5LDH VVTA gvrqqegesr l nl vqrnvnvf kfiIPQIVKYs pncIIIVvs npvdil tyvawkls
6LDH VITA garqq e gesrlnl vqrnvnif kfiIPNIVKHs pdcIILVvs npvdvl tyvawkls g
4MDH ILVG smprrdgme r kdl l kanvkif kcqGAALDKYakk svKVIVvg npantn cltasksa p s
PGDH ILLV k ag qav dnfIEKLVPLl digdlIIDgg n se y rd tmrrc rdl kdk
8ADH FEVI g r l dtMVTALSCcqeay gVSVIvg vpp dsq nl smnpml
1GD1 VEST g rftkred AAKHLEAg a k KVIIsa p ak ned
1GPB EQIS t a gteas g tg NMXFMLN ga LTIG t m dganvemaee
KS bbbb hhhhhhh bbbb
Label βD αD αE βE αF
230 240 250
5LDH g lp kh rVIGSgc nld
6LDH l pm h rIIGSgc
4MDH i pk e nFSCLt rldhn
PGDH g il F VGS gvs
8ADH llsg rt WKGA ifg
1GD1 itivmgvnqdkydpka hh V ISN asc
1GPB ageen F FIF
KS bbbb
Label βF

Fig. 5. Structurally derived sequence alignment of seven dehydrogenase fold domains ($E_1 = E_2 = 3.8$, $T = 4.5$). Upper case letters indicate areas of high reliability (i.e., $P'_{ij} > 6.0$) across all of the domains. The dendrogram derived from pairwise comparisons results in addition of structures to the alignment one at a time. Boxed regions indicate how the areas of high reliability change as more structures are aligned (i.e., from having nearly all residues reliably equivalenced in the alignment of 5LDH to 6LDH to having only short stretches of secondary structure reliably equivalenced in the final seven domain alignment). "KS" denotes the predominant Kabsch and Sander (DSSP) secondary structure assignment[29] for each of the upper case regions ("h" denotes helix, "b" denotes sheet). The "label" assigned to each region of secondary structure follows the nomenclature of Rossmann et al.[34] The last digit in the numbers above the alignment shows alignment position.

3. (a) A structural similarity score (S_c) is defined in order that overall alignment equality and structural similarity may be compared across a wide range of protein structural families. (b) A measure of individual residue accuracy (P'_{ij}) is defined in order that residue equivalences may be normalized with respect to both the number of structures in an alignment and the length of the structures being aligned.

4. Alignments having a structural similarity score (S_c) between 5.5 and 9.8 imply a high degree of structural similarity and reliability. Values be-

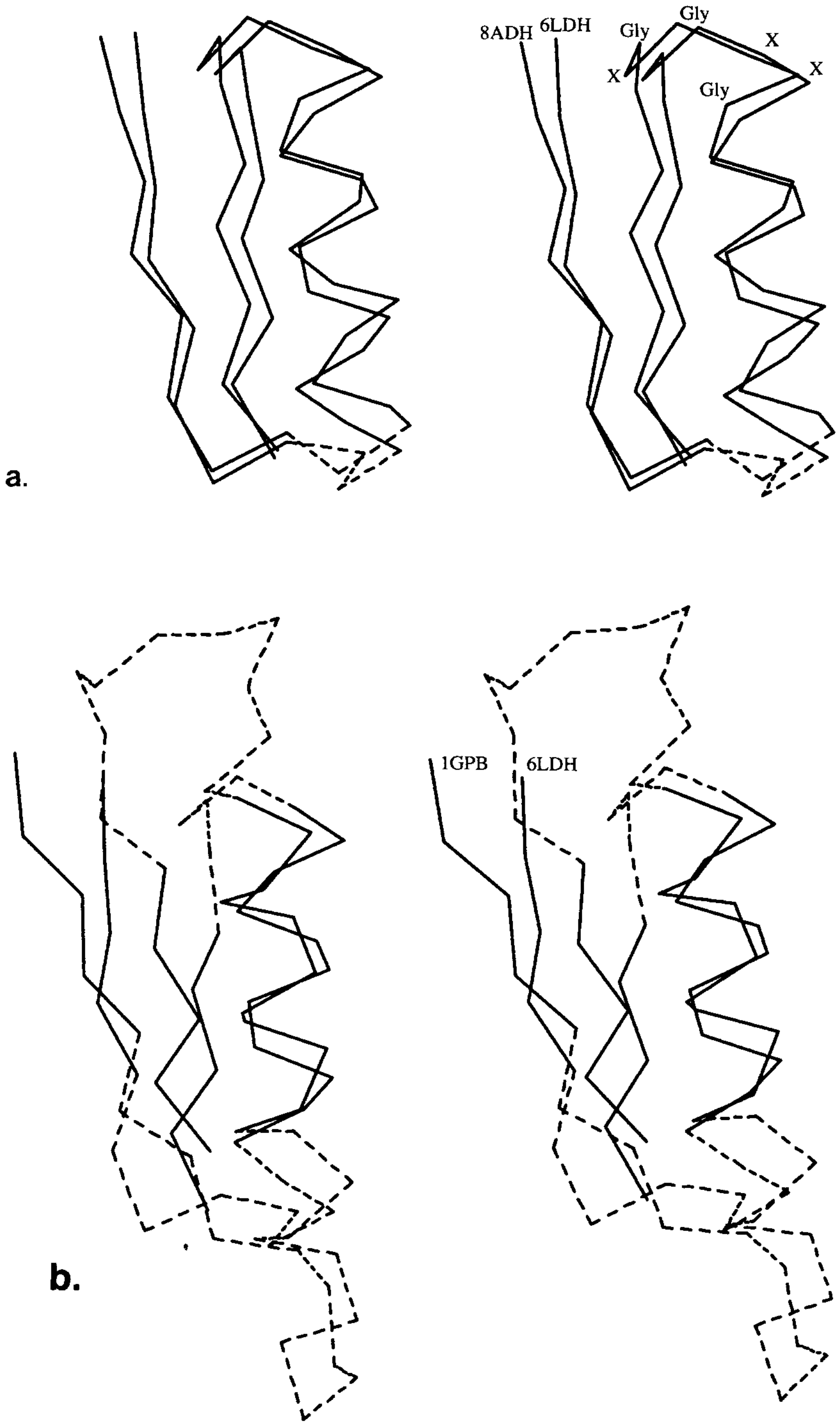

Fig. 6. Examples of structural superposition and equivalences as determined by STAMP. Solid lines represent regions found to be equivalent (i.e., $P_{ij}' > 6.0$), whereas broken lines indicate non-equivalent regions (i.e., $P_{ij}' < 6.0$). Structures are labelled at the C-terminal end. **(a)** Superposition of the first β-α-β moiety of 6LDH and 8ADH (aligned independently), around the ribose binding loop (6LDH: ASN 21 to ASP 52; 8ADH: SER 193 to ASP 223; positions 5 to 56 in Fig. 5). Strict structural equivalence is observed over the entire GLY-X-GLY-X-X-GLY pattern. **(b)** Superposition of 6LDH and 1GPB in the same region (1GPB: LEU 562 to GLY 607). Here the ribose binding loop is not conserved, and the total number of equivalences has dropped from 30 to 21.

tween 2.5 and 5.5 correspond to more distantly related structures, whereas values less than 2.5 generally indicate little overall structural similarity.

5. Stretches of three or more aligned positions with P'_{ij} values greater than 6.0 correspond to genuine topological equivalences, values between 4.0 and 6.0 are equivalent >50% of the time, whereas values less than 4.0 are generally not equivalent.

6. Regions defined in 5. having $P'_{ij} > 6.0$ generally correspond to regions of conserved secondary structure within a family of structures being compared.

ACKNOWLEDGMENTS

We thank Professor L. N. Johnson for her encouragement and support. R.B.R. is a Commonwealth Scholar, and a member of Keble College, Oxford. G.J.B. thanks the Royal Society for the support of a University Research Fellowship.

REFERENCES

1. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold, unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
2. Barton, G. J., Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences: Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.
3. Argos, P. A sensitive procedure to compare amino acid sequences. J. Mol. Biol. 193:385–396, 1987.
4. Overington, J., Johnson, M.S., Šali, A., Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. Lond. B 241:132–145, 1990.
5. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326;347–352, 1987.
6. Barton, G. J., Sternberg, M. J. E. Flexible protein sequence patterns—a sensitive method to detect weak structural similarities. J. Mol. Biol. 212:389–402, 1990.
7. Taylor, W. R. Classification of amino acid conservation. J. Theor. Biol. 119:205–218, 1986.
8. Luthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins: 10:229–239, 1991.
9. Bowie, J.U., Luthy, R., Eisnenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
10. Argos, P., Rossmann, M. Exploring structural homology of proteins. J. Mol. Biol. 105:75–95, 1976.
11. Remmington, S.J., Matthews, B.W. A general method to assess similarity of protein structure with applications of t4 bacteriophage. Proc. Natl. Acad. Sci. 75:2180–2184, 1978.
12. McLachlan, A.D. Gene duplication in the structural evolution of chymotrypsin. J. Mol. Biol. 128:49–79, 1979.
13. Fitch, W.M. An improved method of testing for evolutionary homology. J. Mol. Biol. 16:9–16, 1966.
14. 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.
15. Barton, G. J., Sternberg, M. J. E. Lopal and scamp: Techniques for the comparison and display of protein structures. J. Molecular Graphics 6:190–196, 1988.
16. Taylor, W.R., Orengo, C.A. A holistic approach to protein structre alignment. Protein Engineering 2(7): 1989.
17. W.R. Taylor and C.A. Orengo. Protein structure alignment. J. Mol. Biol. 208:1–21, 1989.
18. Orengo, C.A., Taylor, W.R. A rapid method of protein structure alignment. J. Theor. Biol. 147:517–551, 1990.
19. Šali, A., Blundell, T.L. 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.
20. Sutcliffe, M.J., Haneef, I., Carney, D., Blundell, T.L. Knowledge based modelling of homologous proteins, part i: three dimensional frameworks derived from the simultaneous superpostion of multiple structures. Protein Engineering 1:377–384, 1987.
21. Adman, E.T. Structure and function of small blue copper proteins. In P.M. Harrison (ed.) "Metalloproteins: Metal Proteins with Redox Rules." Basel: Verlag Chemie, 1984: 1–42.
22. Mclachlan, A.D. A mathematical procedure for superimposing atomic coordinates of proteins. Acta Crystallographica A28:656–657, 1972.
23. Diamond, R. On the comparison of conformations using linear and quadratic transformations. Acta Crystallographica A32:1–10, 1976.
24. Blanken, R.L., Klotz, L.C., Hinnebusch, A.G. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. 19: 9–19, 1982.
25. Barton, G. J. Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol 183: 403–428, 1990.
26. Sankoff, D., Kruckal, J.B. (eds.). "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison." Reading, MA: Addison-Wesley, 1983.
27. Smith, T. F., Waterman, M. S. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197, 1981.
28. Berstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanovichi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
29. Kabsch, W., Sander, C. A dictionary of protein secondary structure. Biopolymers 22:2577–2637, 1983.
30. Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein strucutres: The structure and evolutionary dynmaics of globins. J. Mol. Biol. 136:225–270, 1980.
31. James, M.N.G., Delbaere, L.T.J., Brayer, G.D. Amino acid sequence alignment of bacterial and mammalian pancreatic serine proteases based on topological equivalences. Can. J. Biochem. 56:396–402, 1978.
32. Read, R.J., Brayer, G.D., Jurasek, L., James, M.N.G. Critical evaluation of comparative model building of streptomyces griseus trypsin. Biochemistry 23:6570–6575, 1984.
33. Craik, C.S., Rutter, W.J., Fletterick, R. Splice junctions: Association with variation in protein structure. Science 220:1125–1129, 1983.
34. Rossmann, M.G., Adams, M.J., Buehner, M., Ford, G.C., Hackert, M.L., Lentz jun., P.J., McPherson jun., A., Schevitz, R.W., Smiley, I.E. Structural constraints of possible mechanisms of lactate dehydrogenase as shown by high resolution studies of the apoenzyme and a variety of enzyme complexes. Cold Spring Habour Symp. Quant. Biol. 36:179–191, 1971.
35. Rossmann, M.G., Liljas, A., Branden, C.-I., Banaszak, L.J. Evolutionary and structural relationships among the dehydrogenases. The Enzymes 11:61–102, 1975.
36. Acharya, K.R., Stuart, D.I., Varvill, K.M., Johnson, L.N. "Glycogen Phosphorylase b; Description of the Protein Structure." London: World Scientific Publishing, 1991.
37. Rossmann, M.G., Moras, D., Olsen, K.W. Chemical and biological evolution of a nucleotide-binding protein. Nature 250:194–199, 1974.
38. Otto, J., Argos, P., Rossmann, M.G. Prediction of secondary structural elements in glycerol-3-phosphate dehydrogenase by comparison with other dehydrogenases. Eur. J. Biochem. 109:325–330, 1980.
39. Bork, P., Grunwald, C. Recognition of different nucleotide-binding sites in primary structures using a property-pattern approach. Eur. J. Biochem. 191:347–358, 1990.
40. Moller, W., Amons, R. Phosphate-binding sequences in nucleotide-binding proteins. FEBS Letters 186:1–7, 1985.
41. Stura, E.A., Zanotti, G., Babu, Y.S., Sansom, M.S.P., Stuart, D.I., Wilson, K.S., Johnson, L.N., Van De Werve, G.

Comparison of AMP and NADH binding to glycogen phosphorylase b. J. Mol. Biol. 170:529–565, 1983.

42. Perutz, M.F., Hasnain, S.S., Duke, P.J., Sessler, J.L., Hahn, J.E. Stereochemistry of iron in deoxyhaemoglobin. Nature 295:535, 1982.
43. Hendrickson, W.A., Love, W.E., Karle, J. Crystal structure analysis of sea lamprey hemoglobin at 2 Å resolution. J. Mol. Biol. 74:331, 1973.
44. T. Takano. Structure of myoglobin refined at 2.0 Å resolution. J. Mol. Biol. 110:569, 1977.
45. Weber, E., Steigemann, W., Jones, T.A., Huber, R. The structure of oxyerthyrocruorin at 1.4 Å resolution. J. Mol. Biol. 120:327, 1978.
46. Arutyunyan, E.G., Kuranova, I.P., Vainshtein, B.K., Steigemann, W. X-ray structural investigation of leghemoglobin VI. structure of acetate-ferrileghemoglobin at a resolution of 2.0 Å. Kristallografiya 25:80, 1980.
47. Blow, D.M. Structure and mechanism of chymotrypsin. Accounts of Chemical Research 9:145, 1976.
48. Radhakrishnan, R., Presta, L.G., Meyer Jr., E.F., Wildonger, R. Crystal structures of the complex of porcine pancreatic elastase with two valine-derived benzoxazinone inhibitors. J. Mol. Biol. 198:417, 1987.
49. Marquart, M., Walter, J., Deisenhofer, J., Bode, W., Huber, R. The geometry of the reactive site and of the peptide groups in trypsin and trypsinogen and its complexes with inhibitors. Acta Crystallographica 39:480, 1983.
50. Ashley, P.L., Macdonald, R.J. Kallikrein-related m-RNAs of the rat submaxillary gland. nucleotide sequences of four distinct types including tonin. Biochemistry 24:4512, 1985.
51. Reynolds, R.A., Remington, S.J., Weaver, L.H., Fisher, R.G., Anderson, W.F., Ammon, H.L., Matthews, B.W. Structure of a serine protease from rat mast cells determined from twinned crystals by isomorphous molecular replacement. Acta Crystallographica 41:139, 1985.
52. Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G., Bartunik, H. Refined 2 angstroms x-ray crystal structure of porcine pancreatic kallikrein Å and a specific trypsin-like serine proteinase. crystallization and structure determination and crystallographic refinement and structure and its comparison with bovine trypsin. J. Mol. Biol. 164:237, 1983.
53. Read, R.J., Brayer, G.D., Jurasek, L., James, M.N.G. Critical comparison of comparative model building of streptomyces griseus trypsin. Biochemistry 23:6570, 1984.
54. Moult, J., Sussman, F., James, M.N.G. Electron density calculations as an extension of protein structure refinement. streptomyces griseus protease at 1.5 Å resolution. J. Mol. Biol. 182:555, 1985.
55. Fujinaga, M., Sielecki, A.R., Read, R.J., Ardelt, W., Laskowski Jr., M., James, M.N.G. Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 Å resolution. J. Mol. Biol. 195:397, 1987.
56. Brayer, G.D., Delbaere, L.T.J., James, M.N.G. Molecular structure of the alphalytic protease from myxobacter 495 at 2.8 Å resolution. J. Mol. Biol. 131:743, 1979.
57. Abad.Zapatero, C., Griffith, J.P., Sussman, J.L., Rossmann, M.G. Refined crystal structure of dogfish apo-lactate dehydrogenase. J. Mol. Biol. 198:445, 1987.
58. Grau, U.M., Rossmann, M.G. The 2.7 Å x-ray structure of pig LDH a very close representation of the active ternary complex of lactate dehydrogenase. American Crystallographic Association, Abstracts Papers 8:39, 1980.
59. Birktoft, J.J., Rhodes, G., Banaszak, L.J. Refined crystal structure of cytoplasmic malate dehydrogenase at 2.5 Å resolution. Biochemistry 28:6065, 1989.
60. Branden, C.I., Tapi, O. Interdomain motion in liver alcohol dehydrogenase. structural and energetic analysis of the hinge bending mode. J. Biol. Chem. 261:15273, 1986.
61. Branlant, C., Oster, T., Branlan, G. Nucleotide sequence determination of the DNA region coding for bacillus stearothermophilus glyceraldehyde-3-phosphate dehydrogenase and of the flanking DNA regions required for its expression escherichia coli. Gene 75:145, 1989.
62. Adams, M.J., Gover, S., Leaback, R., Phillips, C., Somers, D.O'N. The structure of 6-phosphogluconate dehydrogenase refined at 2.5 Å resolution. Acta Crystallographica B B47:817–820, 1991.