Visualization of structural similarity in proteins

Friedrich Rippmann* and William R. Taylor

Laboratory of Mathematical Biology, The National Institute for Medical Research, London, UK

Two new methods for the visualization of structural similarity in proteins with known three-dimensional structures are presented. They are based on the degree of equivalence of α-carbon pairs in two proteins. The quantitative measure for residue equivalence is the comparison score generated using the sequence and structure alignment method of Taylor and Orengo, which is based on the comparison of interatomic distances (and other properties that can be defined on a residue basis).

The first method uses information on corresponding α-carbon positions to display vectors joining these structurally equivalent residues. These vectors can be defined as target constraints, and their minimization "bends" the two proteins toward a common average structure. In the average structure the corresponding residues virtually superpose, while insertions and deletions become clearly visible.

The second method uses the comparison scores to perform a weighted least-squares fit of the two structures. It is further used to color code the two structures according to the score value, i.e., their similarity, on a continuous scale from red to blue. Examples of the methods for the comparison of flavodoxin, chemotaxis Y protein and L-arabinose-binding protein are given.

Keywords: similarity of proteins, structure comparison, structure alignment

INTRODUCTION

The comparison of protein structures is of great importance for understanding the principles of protein structure and function (which is often correlated with structure). Computer graphics are an invaluable tool in structural comparison once the correspondence in the two structures has been defined. In proteins with more than ~30% sequence homology this has usually been done by aligning two or more sequences using dynamic programming techniques. (See, for example, Taylor.[1])

*On leave from German Cancer Research Center, Heidelberg, Germany

Address reprint requests to Dr. Rippmann at the Laboratory of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

Received 6 November 1990; accepted 15 January 1991

In cases with no detectable sequence homology the definition of correspondence is more difficult and is often done on a rather subjective basis, although systematic procedures have been described.[2–4] The method of Taylor and Orengo[4] uses all interatomic vectors (and other properties that can be defined on a residue basis) to define quantitatively the correspondence and degree of similarity between the residues of two proteins with known three-dimensional structures.

A least-squares fit is normally used for the superposition of two structures (i.e., the sum of the squared distances is minimized). The root-mean-square (rms) deviation, defined as

$$\left(1/n \sum_{i=1}^{n} |X_i - Y_i|^2\right)^{1/2}$$

and the average distance, defined as

$$1/n \sum_{i=1}^{n} |X_i - Y_i|$$

of corresponding α-carbon positions are used as a quantitative measure for the overall similarity of the two proteins.

METHODS

The procedure for visualizing the structural equivalence of proteins used structural alignment based on interatomic distances generated by an adapted version of SSAP.[4,5] The comparison scores generated by the program for each corresponding residue pair were used to perform a weighted superposition of the two structures. An adapted version of a FORTRAN subroutine by Van Gunsteren using the method of McLachlan[6] was used for the weighted least-squares fit. The rms deviation and the average distance of corresponding α-carbon positions were used as measures for the overall similarity of the two proteins. Corresponding residues were color coded on a graphical display according to their comparison score value (i.e., to their structural similarity as defined by the structural alignment program SSAP) in a range from red (high structural similarity) to blue (low or no structural similarity).

Specifically, two Protein Data Bank (PDB)-like sets[7] of corresponding α-carbon atoms containing comparison score values in the BVALUE field (which usually holds the tem-

perature factors in X-ray structures) were translated to their respective centers of gravity and superimposed using the comparison score values as weights in the least-squares fit. The resulting rotation matrix and translation vector were then used to transform the coordinates of an all-atom PDB file and the comparison score values were written to the BVALUE field.

The full range of comparison scores was divided by the number of colors available in the molecular modeling program used. Version 3.0 of QUANTA (Polygen Corporation, Waltham, MA, USA) was used on a Silicon Graphics IRIS 4D/80GT workstation to display the two superimposed structures. They were colored according to their comparison score value in a continuous range from red to blue. The ranges for the fourteen display colors were read from the file created by the program.

Alpha-carbon traces proved most useful for visualization purposes, but any of QUANTA's display options can be used. To distinguish the α-carbon traces of the two structures, one trace was marked by small spheres around the α-carbon atoms (using the SURFACE options with appropriate parameters, e.g., 8% van der Waals radius).

It is also useful to have lines connecting residues in the two structures to emphasize their equivalence. While this might easily be coded specifically, an interesting alternative is to specify a target constraint between the equivalent pairs of atoms. Using the graphical front end to the energy program CHARMM[8] in QUANTA such constraints were displayed suitably as dashed lines joining the atoms. This rather oblique method of displaying dashed lines has the amusing advantage that by setting the target constraint to zero distance and supressing the van der Waals repulsion and the electrostatic terms in the force calculation, the target separations between equivalent α-carbon atoms can be optmized by CHARMM. The effect is to bend and distort both structures toward a common average. The technique may be of limited use for quantitative analysis but is useful for clarifying superpositions where some components may have undergone small relative shifts and it avoids breaking the comparison problem into smaller substructures. (See Color Plates 1a and 1b and also Taylor et al.[9] for examples.)

RESULTS

This procedure was applied to two pairs of proteins having different biological functions and very low sequence homology that could not be aligned by conventional sequence alignment. Flavodoxin[10] and chemotaxis Y protein[11] are globular single-domain proteins of similar size. Arabinose-binding protein[12] is a two-domain protein of roughly double the size of the other two.

The coordinates of flavodoxin (PDB code 4FXN) and arabinose-binding protein (PDB code 1ABP) were taken from the Brookhaven Protein Data Bank.[7] Arabinose-binding protein was divided into its two domains, excluding the loop folding back from the second domain to the first. Residues 1–108 of domain P and residues 109–253 of domain Q were used; they contain all five strands of the β-sheet.

The α-carbon coordinates for chemotaxis Y protein (CHEY) were generated from a stereo picture,[11] except for the first two residues which were ill defined.

The structural alignment program SSAP was used for the pairs CHEY / 4FXN, 1ABP (P-domain) / CHEY and 1ABP (Q-domain) / CHEY. The parameters used were those published by Taylor and Orengo,[5] with the exception of window = 100 for the 1ABP (Q domain) / CHEY comparison. Tables 1a–c show the alignment of the three pairs and a simple graphical representation of the comparison scores of corresponding residue pairs. (The "darkness" of the symbols printed between the sequences corresponds to the value of the comparison score.) The secondary structure definitions generated by the DSSP program[13] are shown beside the aligned sequences. Only eight pairs of identical residues occur in the alignment of CHEY / 4FXN, fifteen pairs in 1ABP (P domain) / CHEY and six pairs in 1ABP (Q domain) / CHEY. The number of identical residues for the three comparisons, the number of aligned residues, the number of residues with comparison score > 0, the rms deviation and the average distance between the residues with comparison score > 0 are given in Table 2. The relatively high deviations in the range of 3–4 Å should not be interpreted

Table 1a. Alignment of chemotaxis Y protein (CHEY) with flavodoxin (4FXN). The symbols between the aligned sequences are a simple translation of the comparison score into shading. The secondary structure definitions of Kabsch and Sander are shown besides the sequence (in one-letter amino acid code). The phosphate-accepting residues in CHEY and those residues close to the flavin mononucleotide phosphate (distance <4Å) in 4FXN are emphasised with arrows

CHEY		4FXN				CHEY		4FXN				CHEY		4FXN		
		M						L		B		H	I = R		H
		A						N		T		H	A + D		H
		D						E				H	A + F		H
		K						D		S		H	A + E	100	H
		E						I		E		S 100	Q : E		H
		L + M				S		G - L		E			A : R		H
B		K # K		E		E		F @ I	50	E		S	G - M		H
		F @ I		E		E		I @ L		E			N		H
E		L @ V		E		E		I * G		E			G		H
E	10	V @ Y		E		E		S @ C		E			Y		T
E		V # W		E		E	-->	D # S	<--				G		B
	-->	D # S	<--			S		W + A		E		T	A - C		
S	-->	D = G	<--	S				N - M		E		E	S . V		E
		F + T	<--	S			60	M . G		T			V	110	E
H		S . G	10	S						D	T			V		
H		T . N	<--	H						E	T			E		S
H		M : T		H						V	60	E		T		
H		R + E		H						L		E	G : P			
H		R - K		H		B		P		E		E	Y = L			E
H	20	I . M		H		T		N . E		E			V = I			E
H		V - A		H		T		M . S		E			V = V			E
H		R + E		H		B		D . E		E		T	K - Q			E
H		N - L		H		T		G : F		H		T	110 P . N			S
H		L : I		H				L : E		H			E	120		
T		L : A	20	H		H		E = P		H			F . P			
T		K = K		H		H		L + F		H			D			G
T		E . G		H		H		L + I	70	H			E			G
T		L . I		H		H	70	K + E		H			T : A			G
		G I		H		H		T + E		H		H	A : E			H
	30	F E		H		H		I * I		H		H	A - Q			H
		S		T		H		R * S		H		H	T - D			H
		G		B		H		A = T		T		H	L = C			H
		K				S		D - K		T		H	E . I			H
		D				S		S				H	E . E	130		H
		N + V	30			G		A				H	K - F			H
B		N = N		E		G		M . I				H	120 L = G			H
		V * T		E		G		S - S		T		H	N . K			H
		E * I		E		G	80	A : G		T		H	K . K			H
E		E - N		E					K	80		H	I I			H
E		A + V		G					K		E	H	F A			H
S		E						L : V			E			N		T
S		D				E		P = A			E		H	E - I		
H		G				E		V = L			E		H K			
H	40	V				E		L * F			E		H L			
H		D				E		M : G			E		T G			
H		A				E		V + S			E		M			
H		L						T * Y			E					
H		N : S		G		T		A - G			S					
H		K . D		G		T		E - W		90	S					
T		L : V					90	A . G			S					
T		Q N						K D								
T		A . I	40	T		H		K . G			S					
T		G . D		B		H		E - K			H					
	50	G - E		T		H		N + W			H					
		F : L		T		H		I + M			H					

Table 1b. Alignment of L-arabinose binding protein domain P (1ABP"P") with chemotaxis Y protein (CHEY). The symbols between the aligned sequences are a simple translation of the comparison score into shading

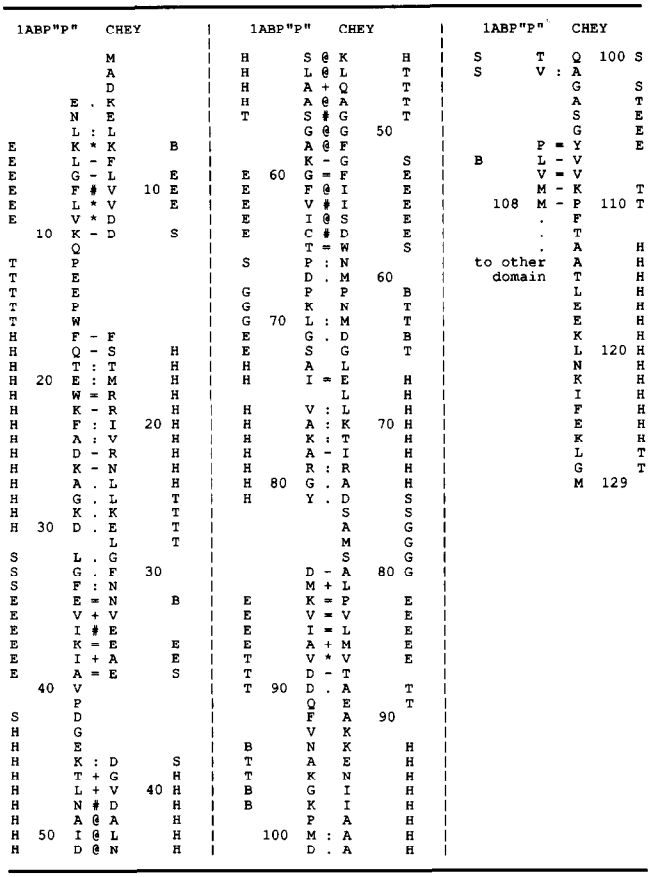
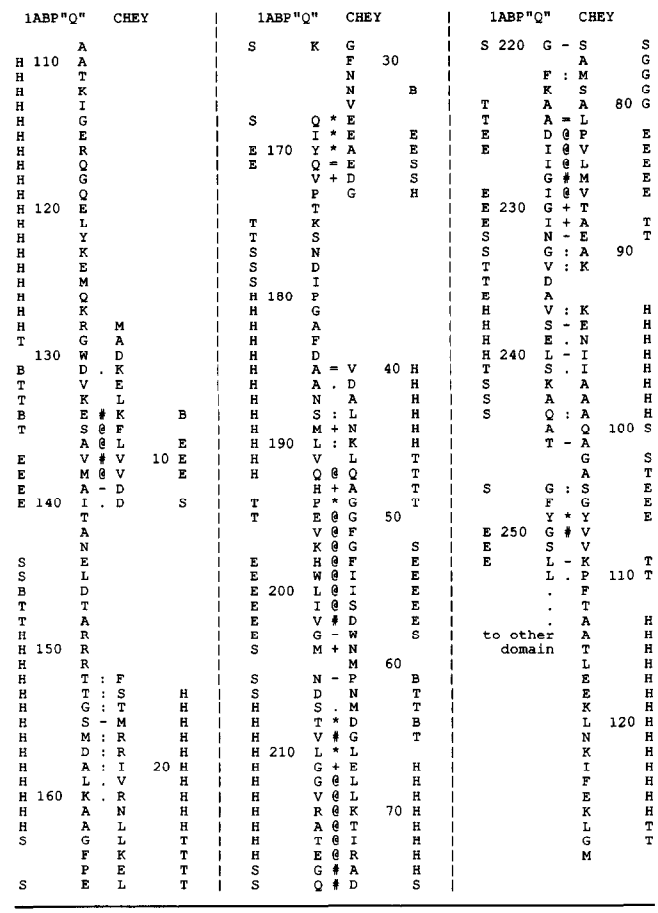


Table 1c. Alignment of L-arabinose binding protein domain Q (1ABP"Q") with chemotaxis Y protein (CHEY). The symbols between the aligned sequences are a simple translation of the comparison score into shading



as a low similarity between the compared structures. Color Plates 2a–c show the color coded superposition of the respective structures and the similarities between them; the equivalence of the strands in the parallel sheets and corresponding helices are clearly visible, and the deviation in the region of the strands is much lower than the average deviation calculated with equal weights for all scoring residue pairs.

Figures 1a–c show the dependance of the average distance on the number of highest scoring residues. The average distance decreases almost monotonically with decreasing number of highest scoring residue pairs. This shows how critical it is to take the number of residue pairs properly into account when stating the rms deviation or the average distance as a criterion for the similarity of two structures. Such a plot, or alternatively, a cutoff for scoring residues could be used to define a core of greatest equivalence between proteins.

Figures 2a–c show the comparison scores of corresponding residues (triangles on dashed line; left axis, reversed) and the distance of corresponding α-carbon atoms (circles on solid line; right axis) along the sequence numbering of one protein. It can be seen that high comparison scores tend to be found in pairs of residues with small distances between their α-carbon positions. These positions are associated with the regions of defined secondary structure. To show this, the approximate positions of the β-strands were marked at the bottom of Figure 2 by solid boxes, and the α-helices by stars.

The minima in this comparison score profile can further

Table 2. Summary of the comparisons of chemotaxis Y protein (CHEY) to flavodoxin (4FXN) and to domains P and Q of L-arabinose binding protein (1ABP)

	CHEY/ 4FXN	1ABP P-domain/ CHEY	1ABP Q-domain/ CHEY
Total number of residues	129/ 138	108/ 129	144/ 129
Number of identical residues	8	15	6
Number of aligned residues	111	97	100
Number of residues with score > 0	110	89	89
Root mean square deviation (Å)	3.55	3.65	3.80
Average distance (Å)	3.28	3.03	3.29

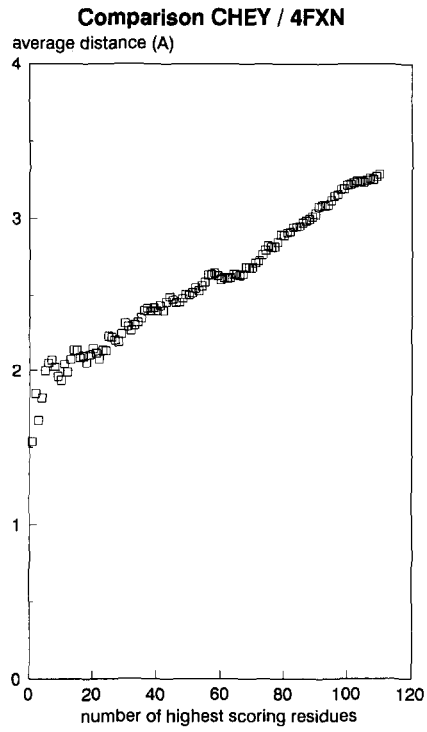


Figure 1a. Dependance of the average distance between corresponding α-carbon atoms of chemotaxis Y protein and flavodoxin on the number of highest scoring residues

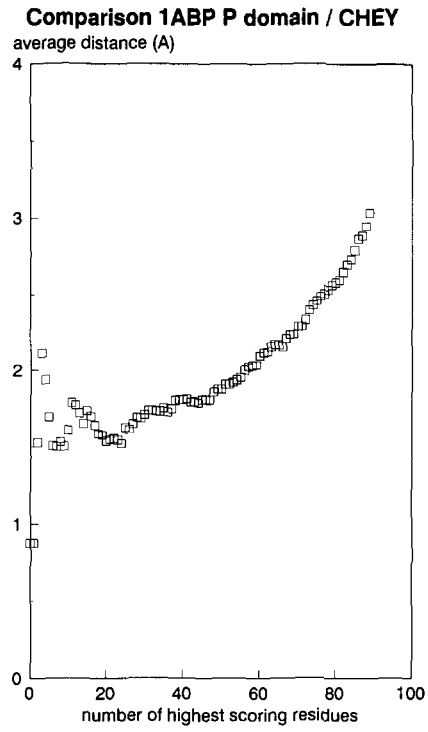


Figure 1b. Dependance of the average distance between corresponding α-carbon atoms of chemotaxis Y protein and L-arabinose binding protein domain P on the number of highest scoring residues

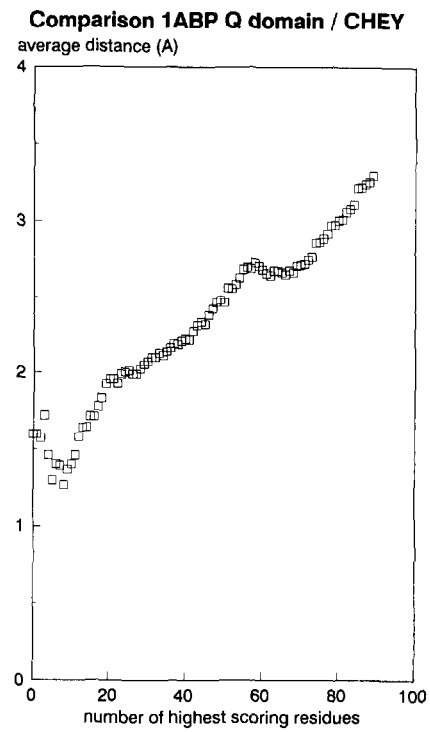


Figure 1c. Dependance of the average distance between corresponding α-carbon atoms of chemotaxis Y protein and L-arabinose binding protein domain Q on the number of highest scoring residues

be used to break the protein into smaller fragments for separate comparisons of the substructures with the highest similarity. Figures 3a–c show the correlation of the comparison score versus the corresponding α-carbon distance. The correlation is relatively weak, which is not surprising, because the comparison score is not a result of the distance between the two residues but of all the vectors of a residue within one structure compared to all vectors of the corresponding residue within the other structure.

It is, of course, tempting to use the purely structural superposition to compare the active sites of the three proteins. Flavodoxin is an electron carrier protein that accommodates the flavin mononucleotide as its prosthetic group. Chemotaxis Y protein is a cytoplasmic protein involved in signal transmission for the bacterial flagellar rotor.[14] It is assumed that it performs its regulatory function by accepting a phosphate group from chemotaxis A protein at ASP 12, 13 or 57.[11] Arabinose-binding protein is a periplasmic protein involved in the high affinity uptake of L-arabinose in *Escherichia coli*.

Although the three proteins have apparently unrelated biological functions, the spatial location of the active sites is very similar. A superposition of CHEY and 4FXN shows the similar orientation of the active residues in CHEY and of the residues accommodating the prosthetic group in 4FXN. It is interesting to note that the three phosphate-accepting aspartates of CHEY are structurally equivalent to those residues close to the flavin mononucleotide phosphate of 4FXN. (See Table 1a.)

Arabinose-binding protein binds its ligand between its two domains. Upon binding, the two domains change their positions relative to each other considerably,[15] but the structure within the domains remains almost unchanged. It is interesting to see that CHEY superimposes to both domains of 1ABP in an orientation where its active aspartic residues point to the arabinose binding cleft of 1ABP.

The active sites of these proteins are located in loops at the end of the parallel strands in the β-sheet. This location further seems to be common for most nucleotide-binding proteins and a detailed discussion will be published elsewhere.

DISCUSSION

The structural alignment of proteins using all interatomic vectors (and other properties that can be defined on a residue basis) is a fairly general method for the structural comparison of proteins with low or no sequence homology. It is relatively insensitive to insertions or deletions and it is able to detect equivalent substructures in different topological environments. The superposition method used here treats the two proteins as rigid bodies. However, this might be inadequate where a pair of equivalent substructures is found to be displaced relative to another pair of substructures. In this situation the proteins might easily be broken into smaller fragments, either at points of relative insertions and deletions of sequence or at a minimum in the comparison score profile. The fragments can then be superimposed by con-

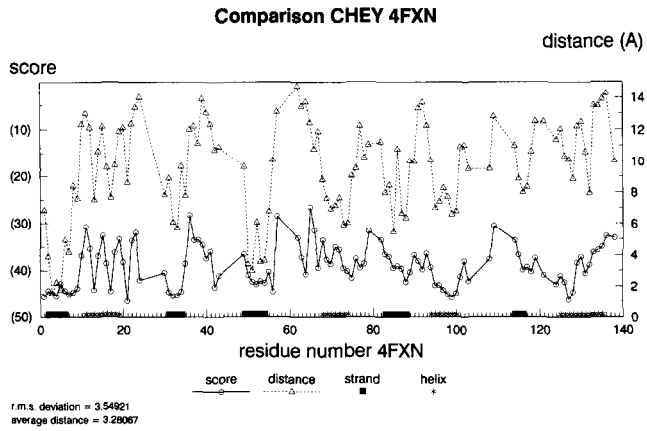


Figure 2a. Score (dashed line with triangles, left axis, reversed) and distance (solid lines with circles, right axis) of corresponding residues of chemotaxis Y protein and flavodoxin along the sequence numbering of flavodoxin (4FXN). Note that not all residues have a correspondence; only the marked positions have a defined value. The boxes at the bottom axis represent the approximate positions of strands (solid boxes) and helices (stars)

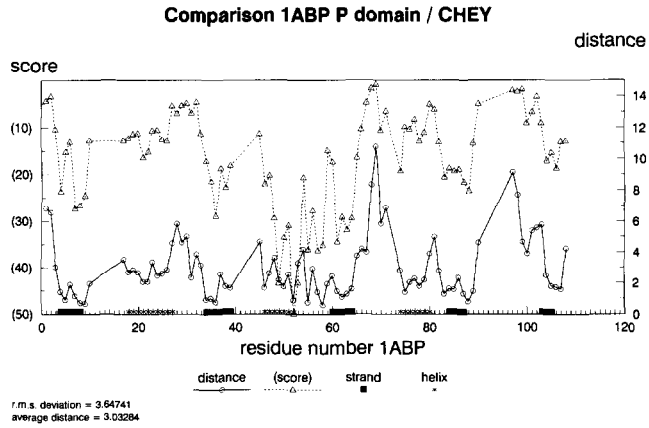


Figure 2b. Score (dashed line with triangles, left axis, reversed) and distance (solid lines with circles, right axis) of corresponding residues of chemotaxis Y protein and L-arabinose binding protein P-domain along the sequence number of L-arabinose binding protein (1ABP)

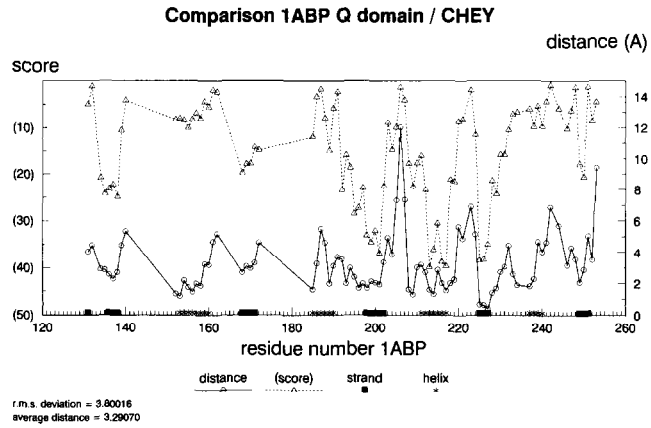


Figure 2c. Score (dashed line with triangles, left axis, reversed) and distance (solid lines with circles, right axis) of corresponding residues of chemotaxis Y protein and L-arabinose binding protein Q-domain along the sequence numbering of L-arabinose binding protein (1ABP)

ventional methods. This procedure is somewhat related to that of Remington and Matthews,[2] which superimposes a stretch of residues of fixed length (e.g., 20 or 40 residues) consecutively on all stretches of equal length in a protein. Our procedure has the advantage that the length of corresponding stretches is defined in a rational way. The method of Zuker and Somorjai[3] finds such corresponding stretches by aligning α-carbon distances using the dynamic programming technique, but their corresponding stretches cannot accommodate insertions or deletions. An alternative approach presented here, which avoids breaking the comparison problem into smaller substructures, is to bend the structures together using target constraint minimization. For the globular $\beta\alpha$-proteins compared here, the color coded superposition of two protein structures according to their degree of similarity (defined by the structural alignment method) is a useful tool for visualizing their overall structural similarity and especially for visualizing high similarity in certain parts. As structural similarity is often correlated with functional similarity, this will help also in the comparative functional analysis of proteins.

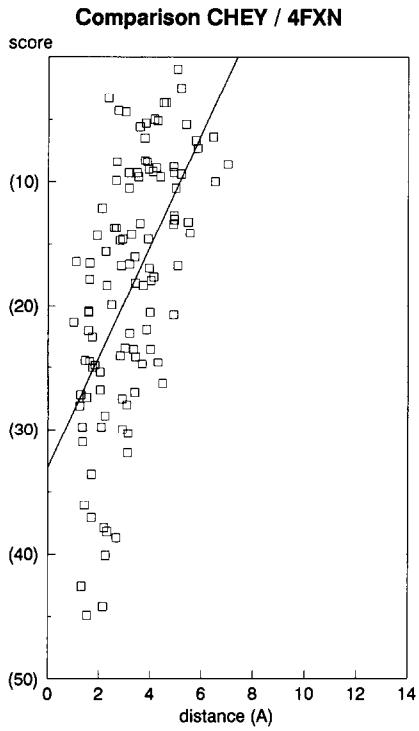


Figure 3a. Correlation of the comparison score and the distance of corresponding α-carbon atoms for the comparison of chemotaxis Y protein and flavodoxin. The regression equation is score = −4.47 distance + 33.02; r = 0.59; n = 110

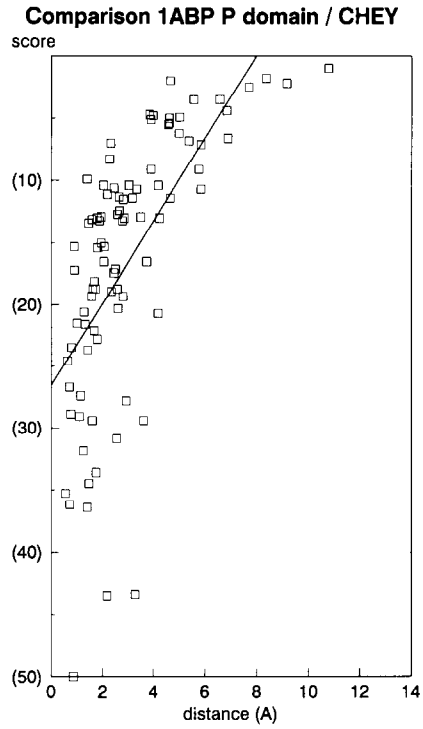


Figure 3b. Correlation of the comparison score and the distance of corresponding α-carbon atoms for the comparison of chemotaxis Y protein and L-arabinose binding protein domain P. The regression equation is score = −3.33 distance + 26.52; r = 0.64; n = 89

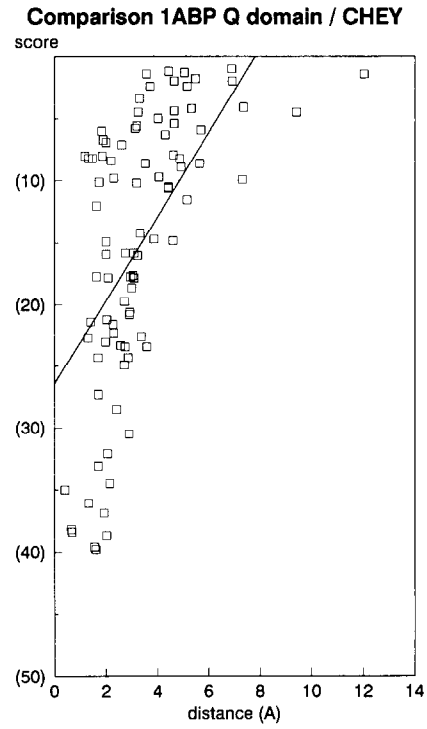


Figure 3c. Correlation of the comparison score and the distance of corresponding α-carbon atoms for the comparison of chemotaxis Y protein and L-arabinose binding protein domain Q. The regression equation is score = −3.39 distance + 26.39; r = 0.59; n = 89

REFERENCES

1 Taylor W.R. Pattern matching methods in protein sequence comparison and structure prediction. *Protein Eng.* 1988, **2** (2) 77–86

2 Remington, S.J. and Matthews, B.W. A systematic approach to the comparison of protein structure. *J. Mol. Biol.* 1980, **140**, 77–99

3 Zuker, M. and Somorjai, R.L. The alignment of protein structures in three dimensions, *Bull. Math. Biol.* 1989, **51**, 55–78

4 Taylor, W.R. and Orengo, C.A. Protein structure alignment. *J. Mol. Biol.* 1989, **208** (1) 1–22

5 Taylor, W.R. and Orengo, C.A. A holistic approach to protein structure alignment. *Protein Eng.* 1989, **2** (7) 505–19

6 McLachlan, A.D. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* 1979, **128**, 49–79

7 Bernstein, F.C., Koetzle, T.F., Williams, G., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanochi, T. and Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112**, 535–542

8 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* 1983, **4** (2) 187–217

9 Taylor, W.R., Orengo, C.A. and Pearl, L.H. Comparison of predicted and X-ray crystal structures of retroviral proteases. in *Protein Engineering* (M. Ikehara, Ed.) Japan Scientific Societies Press, Tokyo and Springer-Verlag, Berlin, 1990, 21–27

10 Smith, W.W., Burnett, R.M., Darling, G.D. and Ludwig, M.L. Structure of the semiquinone form of flavodoxin from Clostridium MP. *J. Mol. Biol.* 1977, **117**, 195–225

11 Stock, A.M., Mottonen, J.M., Stock, J.B. and Schutt, C.E. Three-dimensional structure of Che*Y*, the response regulator of bacterial chemotaxis. *Nature* 1989, **337**, 745–749

12 Gilliland, G.L. and Quiocho, F.A. Structure of the *L*-arabinose-binding protein from Escherichia coli at 2.4-Å resolution. *J. Mol. Biol.* 1981, **146**, 341–362

13 Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded geometrical features. *Biopolymers* 1983, **22**, 2577–2637

14 Stewart, R.C. and Dahlquist, F.W. Molecular components of bacterial chemotaxis. *Chem. Rev.* 1987, **87**, 997–1025

15 Newcomer, M.E., Gilliland, G.L. and Quiocho, F.A. *L*-Arabinose-binding protein–sugar complex at 2.4-Å resolution. *J. Biol. Chem.* 1981, **256** (24) 13213–13217