A fast method of comparing protein structures

M.R.N. Murthy

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, Karnataka, India

Received 14 December 1983

Comparative studies on protein structures form an integral part of protein crystallography. Here, a fast method of comparing protein structures is presented. Protein structures are represented as a set of secondary structural elements. The method also provides information regarding preferred packing arrangements and evolutionary dynamics of secondary structural elements. This information is not easily obtained from previous methods. In contrast to those methods, the present one can be used only for proteins with some secondary structure. The method is illustrated with globin folds, cytochromes and dehydrogenases as examples.

Protein structure *Homology* *Comparison* *Evolution* *Packing* *Secondary structure*

1. INTRODUCTION

Systematic comparisons of the 3-dimensional structures of proteins provide important information regarding their patterns of folding, evolutionary relationships and similarities of active site geometries. Due to the stringent restrictions on evolution of the 3-dimensional structures of proteins, distant evolutionary relationships are recognized only by comparative studies of structures. These studies also provide information on the mechanism by which proteins retain their 3-dimensional shapes during divergent evolution.

The problem of structural comparisons is that of finding the 3 rotational and 3 translational parameters which result in maximal superposition of the two structures. A divergent evolutionary relationship is suggested if the superposition follows a sequential order along the two polypeptide chains. However, significant but not sequential superposition suggests preferred packing arrangements or convergent evolution to a stable fold.

Two different approaches to this problem have been developed and extensively used. When two molecules with similar structures are properly oriented, the vectors between the structurally equivalent atoms will be nearly equal and parallel. A systematic search procedure using these vectors was developed in [1] (referred to as the RA procedure) to align the structures. It is possible to identify insertions and deletions and to evaluate the statistical significance of the similarity using the RA procedure. Authors in [2] (called the RM method) select segments of one molecule and fit it with every possible segment of the second molecule. If the structures are similar, a small number of comparisons will have unusually good and statistically significant fits. In this method, the segment length has to be chosen judiciously and insertions and deletions are not easily located. Both these methods require considerable computing time.

In contrast to these quantitative methods, authors in [3] developed a method recognizing the qualitative similarity of globular proteins by a simple diagrammatic representation of helices and sheets. In this method, the pattern of secondary structural elements in two molecules are compared and similarities are recognized on a qualitative basis. It is shown here that the ideas in [3] can easily be cast in a quantitative form. The resulting algorithm is fast and allows evaluation of the statistical significance of similarity. It also reveals certain features of protein folding not easily obtainable from the RA or RM method.

Published by Elsevier Science Publishers B V
00145793/84/$3 00

2. METHOD

2.1. Orientation search

Each protein molecule to be compared is considered to be made up of its secondary structural elements, the helices and sheets. The directions and centroids of these elements are evaluated. The direction cosines of the helix axes are evaluated by minimizing

$$\sum_{i=1}^{N-1} [l(x_i - x_{i+1}) + m(y_i - y_{i+1}) + n(z_i - z_{i+1}) - h]^2$$

where l, m and n are the direction sines, h is the height per residue of the helix, and N is the number of atoms within the helix. The direction cosines of sheet strands are evaluated as a simple average of the unit vectors along consecutive C_α positions. If two structures are similar, the equivalent structural elements will be parallel at an appropriate relative orientation. Hence, the secondary structural elements of the second molecule are rotated systematically through all possible eulerian angles and corresponding to each orientation, the angles θ_{ij} between the ith element of molecule 1 and jth element of molecule 2 are evaluated. The probability that these two elements are equivalent is set to

$$p_{ij} = w_{ij} f(\theta_{ij})$$

where $w_{ij} = 0$ if the structural elements are of different types. For elements of the same type (helices or sheets), the weight may be set to unity or to the number of atoms in the elements. It is easy to show that

$$f(\theta_{ij}) = \exp(-(2 \times 3.9 \times \sin\theta_{ij}/2)^2/2E^2)$$

corresponds to the expression used in the RA procedure for the scatter between vectors relating 3 adjacent residues. (E is the RMS value of the scatter). In practice, any expression which has large values for angles up to 20–30° is suitable. In the present method $P_{ij} = w_{ij} \cos^n \theta_{ij}$ where n is chosen to be between 5 and 10. At the appropriate orientation, the P_{ij} values between structurally equivalent elements will be large.

2.2. Selection of equivalences

The P_{ij} values constitute an equivalence probability matrix. The best sequential alignment of structural elements are derived from this matrix by an ingenious procedure developed in [5]. In this method, the matrix elements are manipulated such that the value of any cell i, j is incremented by the largest value of the cells with cell numbers $>i$ and $>j$. The cells are manipulated starting with the last row and last column. The maximum score thus obtained provides a criterion of fit. The maximum match pathway or the set of equivalences between structural elements that lead to the maximum score can also be obtained from the manipulated matrix. Authors in [4] developed a method for sequential alignment of the residues. The present method is more powerful since the whole matrix is used for the derivation of equivalences while in the RA procedure, only a limited number (usually 6) of largest probabilities are stored for each residue. The present procedure requires insignificant computing time since each structure contains only a few secondary structural elements.

2.3. Superposition of the structures

The score obtained by manipulating the equivalent probability matrix of the procedure in [5] depends on the relative orientation of the secondary structural elements and not on their positions. For achieving structural superposition, the score is now modified by a term that depends on the agreement between vectors relating two structural elements. Equality of these vectors ensures superposition of structures. The weighting factor is defined as

$$\sum_{ij} \exp(-|\Delta d_{ij}|^2/2E^2)$$

where $\Delta d_{ij} = d_{ij} - d_{i'j'}$, and i' and j' are elements of the second molecule which are structurally equivalent to elements i and j of molecule 1. These weighted scores are plotted as a function of the 3 eulerian angles. A significant peak in this map reveals structural similarity. The angles θ_{ij} and $|\Delta d_{ij}|$ values at the peak position reveal important details of packing of secondary structural elements. This information cannot easily be obtained from the RM method and is obtainable only after further analysis in the RA procedure.

Omission of the weighted factor and plotting only the scores obtained by the procedure in [5] might reveal other peaks which represent parallel orienta-

tion of secondary structural elements without achieving spatial superposition. Such data provide further insights into preferred packing arrangements of secondary structure of proteins.

The coordinates used here were taken from the Brookhaven National Laboratory Protein Data Bank [6]. Certain calculations were cross-checked with a structural comparison program kindly provided by Professor Rossman of Purdue University. Comparison of globin folds with 20° intervals in each eulerian angle takes about 1 min with the present procedure while it takes over 1 h with other methods [1,2] on comparable computers.

3. RESULTS AND DISCUSSION

Many pairs of proteins reported to possess significant similarity were selected and tested by the present method. The results are summarized in table 1. In most cases, one significant peak, corresponding to the eulerian angles reported in the literature, was obtained in the asymmetric unit of the search function map. The special features of some of the comparisons are discussed below.

3.1. Globin folds

In the comparison of myoglobin [7] and horse hemoglobin β chains [8], a single significant peak was obtained in the asymmetric unit. However, the signal-to-noise ratio was 8.3 when the distance criterion was suppressed and 18.3 when E was set to 10 Å. Also, the average background was 16 when $E = \infty$, and 5 when $E = 10$ Å. These variations reflect the special features of the globin fold which

Table 1

Results of comparing protein structures by the present method

Molecule 1	Molecule 2	n	E	Average weighted score	SD	Search interval	Maximum score	Eulerian angles corresponding to maximum score			Signal-to-noise ratio
Horse hemoglobin β chain	Horse hemoglobin α chain	10	∞	17	9	20	90	80	160	80	8.1
		10	5	4	13	20	238	80	160	80	18.0
	Myoglobin	10	∞	16	7	20	74	160	280	220	8.3
		10	10	5	10	20	186	160	280	220	18.3
Cytochrome b_5	Horse hemoglobin β chain	8	10	19	18	20	145	60	200	260	7.0
		2	5	23	13	20	133	60	210	270	8.4
Cytochrome c-551	Horse hemoglobin β chain	8	10	16	14	20	120	40	200	20	7.4 (θ_3 from 0–180°)
		8	10	16	14	20	156	80	100	280	10.0 (θ_3 from 180–360°)
		8	∞	9	4	20	24	40	200	20	3.8
		8	∞	9	4	20	27	80	100	280	4.5
Lactate dehydrogenase	Glyceraldehyde-3-phosphate dehydrogenase	6	10	17	25	20	222	40	20	340	8.2
NAD binding domain	NAD binding domain	6	10	7	16	20	144	40	20	340	8.6

is built on closely packed helices. Orientations other than the correct one result in the parallel alignment of a subset of helices, without spatial alignment.

3.2. Cytochromes

The structure of cytochrome *c*-551 [9] has 4 closely packed helices. Comparison of this structure with horse hemoglobin β chain [8] with 20° intervals in eulerian angles revealed 2 significant peaks in the asymmetric unit. These were at (40°, 200°, 20°) and (80°, 100°, 280°) with signal-to-noise ratios of 7.4 and 10.0, respectively (fig.1). Significant peaks at these positions were obtained both with and without the distance criterion. The equivalences of secondary structural elements and the angles between their axes are shown in table 2. The peak at (40°, 200°, 20°) of the map corresponds to the superposition reported by authors in [10]. The equivalences obtained were identical to those shown in table 2 for this position. They reported 49 equivalenced residues with an RMS distance between C_α atoms of 3.5 Å. As a check, the latter peak was examined by the program in [10]. This resulted in 46 equivalences with an RMS error of 3.5 Å. The latter peak was not recognized in [10] since those authors did not make a systematic search of all possible relationships. Their selection was based on a careful visual inspection of the structures. This observation underlines the significance of systematic comparisions. The relationship chosen in [10] also results in the superposition of heme normals and the position of heme iron atoms. However, the second peak reported here does not result in the superposition of active centers. Thus, it appears that the superposition suggested in [10] is the only functionally relevant matching. These results once again suggest the special nature of close packing of α-helices in the globin fold.

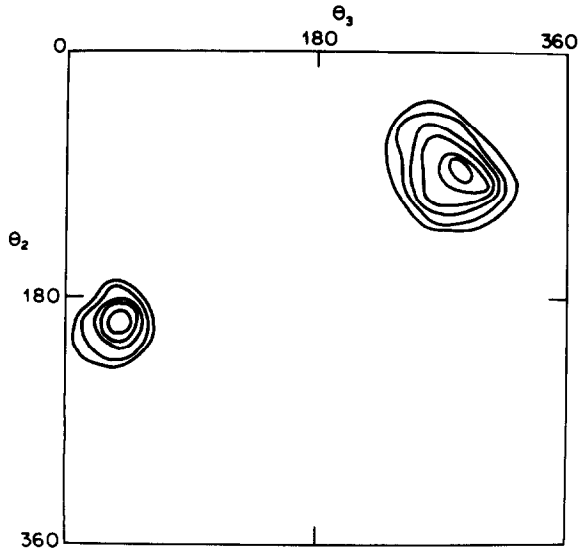

Fig.1. Comparison of cytochrome *c*-551 and horse hemoglobin β chain. The search interval was 20° in each eulerian angle. *n* and *E* were set to 8 and 10 Å, respectively. The map corresponds to $\theta_1 = 60°$.

Table 2

Comparison of cytochrome *c*-552 and horse hemoglobin β chain: structural equivalents and angle between equivalenced helices

(a) Peak at (40°, 200°, 20°)

Structural element in cytochrome *c*-551	Equivalent element in hemoglobin β chain	Angle (°)
α_1	C	18
α_2	E	8
α_3	F	68
α_4	G	0

(b) Peak at (80°, 100°, 280°)

Structural element of cytochrome *c*-551	Equivalent element of hemoglobin β chain	Angle (°)
α_1	A	22
α_2	FG	8
α_3	G	23
α_4	H	14

Comparison of the cytochrome b_5 [11] with the globin fold yielded a unique peak corresponding to the equivalence in [10]. The angles between the helical segments of horse hemoglobin β chain and the equivalenced helices of myoglobin, cytochrome *c*-551 and cytochrome b_5 are shown in table 3. These values are consistent with the findings in [12] that the geometry of helix packing is altered by as much as 30° to allow for changes in amino acid sequence during divergent evolution.

3.3. Dehydrogenases

It is worth checking that the current procedure is able to show similarities of parts of larger struc-

Table 3

Angles between structural elements at the best superposition in a 20° interval search of horse hemoglobin β chain and the equivalent helices of structures with significant similarity to the globin fold

Helix of horse hemoglobin β chain	Molecule 2	Helix	Angle (°)
A	Myoglobin	A	0
B		B	11
C		C	16
D		D	23
E		E	16
FG		FG	14
G		G	8
H		H	0
A	Cytochrome b_5	α1	18
C		α2	18
E		α3	34
F		α4	24
G		α5	52
	Horse		
A	hemoglobin α	A	14
B		B	16
C		C	14
D		D	0
E		E	0
F		F	0
G		G	8
H		H	11

tures. Lactate dehydrogenase (LDH) [13] and glyceraldehyde-3-phosphate dehydrogenase (GPD) [14] have similar nucleotide binding domains while their catalytic domains are unrelated. Fig.2 shows the map for the comparison of the nucleotide binding domains alone while fig.3 is for the comparison between complete chains. The plots are very similar with essentially no change in the signal-to-noise ratio. The quality of these maps is comparable to those obtained in [1,15]. The angles between the β-strands at the peak position are shown in table 4. The structures of βA, βB and βC which constitute the adenine binding mononucleotide moiety seem to be conserved to a larger extent than the nicotinamide binding unit composed of βD, βE and βF. A similar trend was also observed using the RA procedure. After superposition, the RMS error for the atoms of βF was 2.7 Å, compared to the other strands which ranged from 1.0 to 2.0 Å. GPD does not contain a helix equivalent to LDH helix D. Although both structures have helix E, these helices are at a rather large angle of

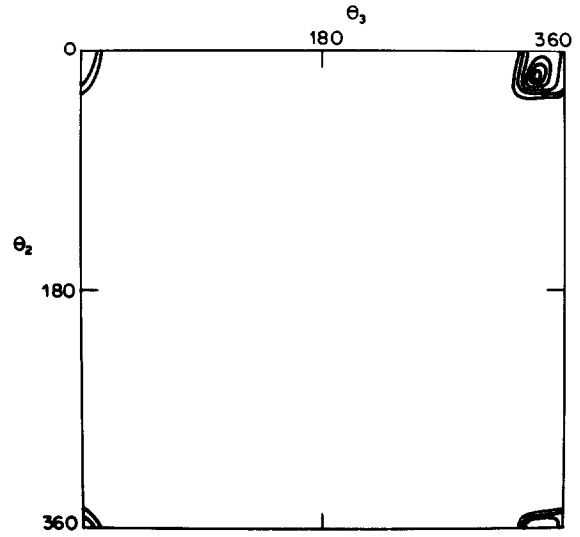

Fig.2. Comparison of the nucleotide binding domains of lactate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase. n and E were set to 6 and 10 Å, respectively. Search interval was 20°. $\theta_1 = 40°$.

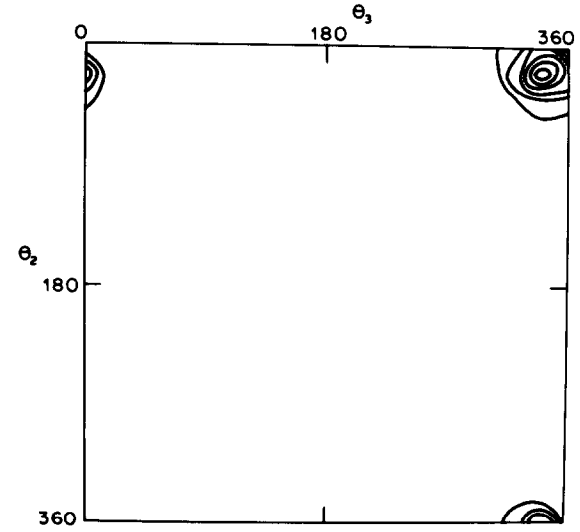

Fig.3. Comparison of the complete chains of lactate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase. The search conditions were identical to those for fig.2. The peak corresponds to the superposition of the nucleotide binding domains.

Table 4

Angles between the β-strands of LDH and the structurally equivalent strands of GPD

Strand	Angle (°)
βA	8
βB	16
βC	12
βD	12
βE	22
βF	23

66°. However, the angle between helix D of LDH and helix E of GPD is close to 0°. Hence, when E is set to large values, helix D of LDH is chosen as equivalent to helix E of GPD. On setting $E = 5$ Å, helix E of LDH is chosen as equivalent to helix E of GPD. Thus, although helix D of LDH and helix E of GPD are similarly oriented, they are at different positions with respect to the dinucleotide binding fold. This part of apo-LDH is known to undergo a conformational transition during cofactor binding.

4. CONCLUSIONS

A new, fast method of comparing protein structures has been presented. This method works well with molecules rich in secondary structure. Authors in [12,16] and in [17] have shown that the 3-dimensional structures of globular proteins are retained by accommodating the effects of deletions and insertions in the loop regions and by small adjustments in the packing of secondary structural elements. Hence, the use of only the secondary structural elements for comparison is not as serious as might appear at the outset. With the increasing number of newer structures available, it becomes progressively more difficult to memorize the structures and hence fast methods of comparing a given structure with a large number of other known structures becomes important. The present method can be used to compare any newly determined structure with all the previously known structures without demanding excessive computing time.

The present method also has certain novel features not present in the earlier method. These include the data on the angles between superimposed structural elements and discrepancies in the vectors between pairs of equivalenced residues. These data provide information regarding preferred packing arrangements of secondary structural elements and also on the evolutionary dynamics of globular proteins.

ACKNOWLEDGEMENTS

I am grateful to Professor M.G. Rossmann of Purdue University for providing his homology program. I thank Professors M.G. Rossmann and V. Sasisekharan for critical reading of the manuscript. Thanks are due to the Department of Science and Technology, India, for financial support.

REFERENCES

[1] Rossmann, M.G. and Argos, P. (1976) J. Mol. Biol. 105, 75–95.
[2] Remington, S.J. and Matthews, B.W. (1978) Proc. Natl. Acad. Sci. USA 75, 2180–2184.
[3] Levitt, M. and Chotia, C. (1976) Nature 261, 552–558.
[4] Rossmann, M.G. and Argos, P. (1975) J. Biol. Chem. 250, 7525–7532.
[5] Needleman, S.B. and Wunch, C D. (1970) J. Mol. Biol. 48, 443–453.
[6] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shima Mouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[7] Takano, T. (1977) J. Mol. Biol. 110, 537–568.
[8] Ladner, R.C., Heidner, E.J. and Perutz, M.F. (1977) J. Mol. Biol. 114, 385–414.
[9] Almassy, R.J. and Dickerson, R.E. (1978) Proc. Natl. Acad. Sci. USA 75, 2674–2678.
[10] Argos, P. and Rossmann, M.G. (1979) Biochemistry 18, 4951–4960.
[11] Mathews, F.S., Argos, P. and Levine, M. (1971) Cold Spring Harbour Symp. Quant. Biol. 36, 387–395.
[12] Lesk, A.M. and Chotia, C. (1980) J. Mol. Biol. 136, 225–270.
[13] Holbrook, J.J., Liljas, A., Steindel, S.J. and Rossmann, M.G. (1975) in: The Enzymes, vol. 11 (Boyer, P.D. ed.) pp. 191–292, Academic Press, New York.
[14] Moras, D., Olsen, K.W., Sabesan, M.N., Buehner, M., Ford, G.C. and Rossmann, M.G. (1975) J. Biol. Chem. 250, 9137–9162.
[15] Rossmann, M.G. and Argos, P. (1977) J. Mol. Biol. 109, 99–129.
[16] Lesk, A.M. and Chotia, C. (1982) J. Mol. Biol. 160, 325–342.
[17] Chotia, C. and Lesk, A.M. (1982) J Mol. Biol. 160, 309–323.