The normalizer of G is a supergroup with index j of the normalizer of V. Then $N(G)$ may be subdivided into j cosets of $N(V)$. Each of these cosets [except $N(V)$ itself] maps V onto another supergroup of G (*cf.* Engel, 1983). If $N(G)$ is the Euclidean normalizer and V is a space group, all equivalent supergroups are space groups again. If $N(G)$ is the affine normalizer, the supergroups equivalent to space group V may be affine groups (*cf.* example above).

Example (i) $G = F23 \qquad N(G) = Im3m(\frac{1}{2}\mathbf{a})$

$V = Fd3 \qquad N(V) = Pn3m(\frac{1}{2}\mathbf{a}), \quad j = 2.$

There exist two supergroups $Fd3$ which are mapped onto each other, *e.g.* by the centering translation of $N(G)$ with vector $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

Example (ii)

$G = Pmm2 \qquad N_E(G) = Z^1mmm(\frac{1}{2}\mathbf{a}, \frac{1}{2}\mathbf{b}, \mu\mathbf{c})$

$V = Pmmm \qquad N_E(V) = Pmmm(\frac{1}{2}\mathbf{a}, \frac{1}{2}\mathbf{b}, \frac{1}{2}\mathbf{c}).$

The index of $N_E(G)$ in $N_E(V)$ is infinite. Accordingly, there exist an infinite number of different supergroups $Pmmm$ for each group $Pmm2$. The mirror planes perpendicular to the **c** axis may be inserted at any height z within the unit cell of $Pmm2$.

(4) $N(G) \not\subset N(V)$ *and* $N(G) \not\supset N(V)$

There exists a largest common subgroup M of $N(G)$ and $N(V)$. The index of M in $N(G)$ gives the number of supergroups equivalent to V.

Example (*cf.* § 3):

$G = P\bar{3}m1 \qquad N(G) = P6/mmm(\mathbf{a}, \mathbf{b}, \frac{1}{2}\mathbf{c})$

$V = R\bar{3}m \qquad N(V) = R\bar{3}m(-\mathbf{a}, -\mathbf{b}, \frac{1}{2}\mathbf{c}).$

The common subgroup M is $P\bar{3}m1(\mathbf{a}, \mathbf{b}, \frac{1}{2}\mathbf{c})$. The index of M in $N(G)$ is 2. Each group $P\bar{3}m1$, therefore, has two supergroups $R\bar{3}m$ which differ with respect to the setting of the rhombohedral lattice (reverse and obverse).

For the classification of crystal structures it is necessary to derive for each crystal structure the corresponding idealized structure type with highest possible symmetry. This has been called aristotype by Bärnighausen (1980). In this context the knowledge of all different but Euclidean- (or affine-) equivalent supergroups of a space group is of special interest, because such different supergroups may result in different aristotypes for a given crystal structure.

I would like to thank Professor Dr Werner Fischer for many stimulating discussions and Professor Dr Hans Wondratschek for helpful remarks.

References

AHSBAHS, H. (1979). *Z Kristallogr.* **149**, 151–152.

BÄRNIGHAUSEN, H. (1980). *Math. Chem.* **9**, 139–175.

BIEBERBACH, L. (1912). *Math. Ann.* **72**, 400–412.

BILLIET, Y. (1981). *Acta Cryst.* **A37**, 649–652.

BILLIET, Y., BURZLAFF, H. & ZIMMERMANN, H. (1982). *Z. Kristallogr.* **160**, 155–157.

BURZLAFF, H. & ZIMMERMANN, H. (1980). *Z. Kristallogr.* **153**, 151–179.

ENGEL, P. (1983). *Z. Kristallogr.* **163**, 243–249.

FISCHER, W. & KOCH, E. (1983). *Acta Cryst.* **A39**, 907–915.

GUBLER, M. G. (1982*a*). Dissertation, Zürich.

GUBLER, M. G. (1982*b*). *Z. Kristallogr.* **158**, 1–26.

HIRSHFELD, F. L. (1968). *Z. Kristallogr.* **113**, 142–154.

International Tables for Crystallography. (1983). Vol. A. Dordrecht, Boston: Reidel.

JARRATT, J. D. & SCHWARZENBERGER, R. L. E. (1980). *Acta Cryst.* **A36**, 884–888.

KOCH, E. (1983). *Z. Kristallogr.* **182**, 144–145.

KOCH, E. & FISCHER, W. (1975). *Acta Cryst.* **A31**, 88–95.

NABONNAND, P. & BILLIET, Y. (1983). Abstr. Eighth Eur. Crystallogr. Meet. Liège, p. 291.

SENECHAL, M. (1983). *Acta Cryst.* **A39**, 505–511.

WONDRATSCHEK, H. (1983). *Introduction to Space-Group Symmetry.* In *International Tables for Crystallography*, Vol. A. Dordrecht, Boston: Reidel.

Acta Cryst. (1984). A**40**, 600–610

A Method for the Systematic Comparison of the Three-Dimensional Structures of Proteins and Some Results

BY MICHAEL LEVINE,* DAVID STUART† AND JOHN WILLIAMS

Department of Biochemistry, University of Bristol, Bristol BS8 1TD, England

(*Received* 2 *December* 1983; *accepted* 1 *May* 1984)

Abstract

A new rapid method of comparing three-dimensional protein structures using the sequence of dihedral angles is described. Systematic screening of protein structures by this method followed by detailed analysis reveals in particular that the calcium-binding protein carp parvalbumin is similar to cytochrome C2 from *Rhodospirillum rubrum*, cytochrome C is similar to hen lysozyme, carboxypeptidase A is similar to phage lysozyme. These results are completely unexpected and show interesting correlation with the

* Present address: Department of Physiology, University of Bristol, England.

† Present address: Laboratory of Molecular Biophysics, Zoology Department, University of Oxford, England.

0108-7673/84/050600-11$01.50

recently determined intron–exon pattern in the genes coding for some of them.

Introduction

A number of striking similarities between different protein structures have been reported. Up to now, however, the available methods have not been well suited to systematic comparison of all pairs of known protein structures. In this paper we describe a new approach which makes a systematic statistical screening feasible and we report the results of its application.

In comparing protein structures simple visual inspection of molecular models is an obvious first step. However, to make a detailed comparison of two proteins we need, first, to describe each protein as a simple set of sequential elements and then to apply a quantitative method for comparing each element of one structure with every element of the other. We can choose to represent the proteins in various ways (*e.g.* as a series of amino acids, α-carbon coordinates, secondary structure elements or dihedral angles) and this choice will influence the detail of the subsequent comparison. Computationally the comparison can be carried out by constructing a matrix whose row numbers i refer to the elements of protein 1 and column numbers j to those of protein 2. The ijth element of the matrix is the value of a measure of similarity between the protein elements i and j. The overall degree of similarity is defined in terms of a statistical analysis of the matrix.

Gibbs & McIntyre (1970) compared amino-acid sequences. The ijth element of the matrix is 1 if amino-acid residues i and j are identical and 0 otherwise. A sequence of residues identical in the two proteins will result in a diagonal line of 1's in the matrix. Insertions and deletions result in the line being parallel to the main diagonal but shifted to one side. Rossmann & Argos (1975, 1976, 1977) compare α-carbon coordinates X. The protein structures are rotated and translated relative to one another in a systematic way through all angles and a comparison matrix is constructed for each orientation. The ijth element of the matrix is expressed as a probability of equivalence P_{ij}:

$$P_{ij} = \exp(-d_{ij}^2/E_1^2)\exp(-S_{ij}^2/E_2^2), \qquad (1)$$

where $d_{ij} = |X_i - X_j|$,

$$S_{ij}^2 = (d_{ij} - d_{i+1,j+1})^2 + (d_{ij} - d_{i-1,j-1})^2$$

and E_1 and E_2 are empirical weighting factors. A run of equivalent residues results in a diagonal ridge of high probabilities. This method, while providing a detailed comparison of three-dimensional structures is very time consuming and not suited to a comprehensive search for similarities among all possible pairs of proteins.

More recently an alternative method has been proposed by Remington & Matthews (1978) [and improved by McLachlan (1979)] which abandons the attempt to choose equivalent portions by analysis of the matrix but rather returns to a method used for comparing protein sequences (Fitch, 1966). A major drawback of this technique is its complete inability to allow for insertion or deletion of residues in one structure with respect to the other.

We have devised a rapid statistical method based on dihedral angles for ranking the pairs of proteins according to the probability that they possess some structural similarity. The method is intended to be a screening procedure only and thus the highest ranking pairs must then be investigated further using the Rossmann & Argos method, for instance.

The course of the polypeptide chain of a protein can be specified by rotational (dihedral) angles φ and ψ which define the orientation of each peptide plane relative to the adjacent peptide planes (De Santis, Giglio, Liquori & Ripamonti, 1962; Ramachandran, Ramakrishnan & Sesisakharan, 1963). Each residue is represented by a point in a two-dimensional graph in which one axis represents φ and the other ψ. In this way the two angles are represented by a single point. This diagram shows all the information on the distribution of pairs of dihedral angles for the protein, but the three-dimensional course of the chain cannot easily be included. Balasubramanian (1977) introduced the idea of treating the dihedral angles as an ordered sequence. However, his method of plotting the angles on a diagram was not easy to use for quantitative comparisons.

The three-dimensional Ramachandran diagram

If the Ramachandran diagram is drawn in three dimensions (φ, ψ, n) with the third axis n representing residue number and with points for successive residues joined by straight lines we get a three-dimensional Ramachandran plot (TDRP), which has some useful properties. The most important of these is that, unlike a plot of atomic coordinates, the TDRP does not depend on the direction from which the molecule is viewed or on the choice of origin of coordinates. Any defined molecular structure will always appear the same on this plot apart from translation along the n axis. Similarly, repeated substructures within one molecule will be identical in the TDRP (apart from the translation) regardless of the relative orientations of the two parts of the molecule. Fig. 1 shows such a plot for the ferredoxin of *Peptococcus aerogenes* and the internal duplication in the structure is clearly seen (Adman, Sieker & Jensen, 1973).

Comparisons of pairs of molecules may be carried out very rapidly using the TDRP as a single rotationally and translationally invariant representation

of each molecule. We chose to set the ijth element of the matrix to the value

$$\Delta_{ij} = (\Delta\varphi + \Delta\psi)_{ij} = |\varphi_i - \varphi_j| + |\psi_i - \psi_j|.$$

A run of similar dihedral angles in the two proteins will result in a valley in the matrix. If there is an insertion of dissimilar structure in one of the proteins the valley will be displaced sideways. For a valley to denote matching conformations in the same direction along the chains, i and j must always increase along it.

Statistical evaluation of the matrix

The occurrence of low valleys in a particular matrix will depend upon the numbers of residues in α-helices or β-pleated sheets in each protein. This makes the comparison of different matrices difficult. We have developed the empirical statistical methods presented in the methods section. Analysis of the matrices was made using two kinds of algorithm. Firstly, we followed Rossman & Argos in re-defining the problem as that of finding the 'best' path through the matrix, defined here as that path which includes the greatest number of elements with values below a preset maximum and for which i and j always increase. The measure of similarity (S_1) was derived from the number of points in this path after application of the normalization procedure described below. In the second type of algorithm we followed Gibbs & McIntyre (1970) in using statistics based upon χ^2; in this way (detailed in the methods section) we obtained four related measures of similarity (S_2 to S_5) which were averaged and added to S_1 to give a single overall estimate (S) of the similarity of the two proteins.

In order to make a visual assessment of the matrix it was plotted as a two-dimensional array of points using a digital plotter linked to the computer. A point was plotted for any element if a cut-off criterion ($\Delta_{ij} < 30°$) was satisfied. Fig. 2 shows the matrix for the comparison of ferredoxin with itself. A clear off-diagonal line reveals the similarity between the two halves of the chain.

It is well known that dihedral angles are sensitive to small changes in the atomic coordinates and conversely that a small angular change can result in sizeable shifts in the positions of distant atoms. Experimental results are therefore necessary to judge whether these derived quantities are in fact characteristic of particular chain folds.

Methods

Analysis of the comparison matrix

The first algorithm finds the 'best' path through the matrix, defined as that which includes the greatest number of elements below a maximum and for which i and j (the matrix subscripts) always increase. This provides an estimate of similarity, χ_1.

A fast algorithm for finding this best path has been devised (Stuart, 1979). This involves searching all paths through the matrix within 100 points of the principal diagonal and use the techniques of list processing to achieve maximal speed.

The second kind of algorithm uses statistics based upon χ^2. For a particular diagonal line, N, in the matrix, using the usual definitions, this is

$$\chi^2_N = [(\sum^{n} \Delta_{ij}) - (n\bar{\Delta})]^2/n,$$

where n is the number of points in the diagonal line

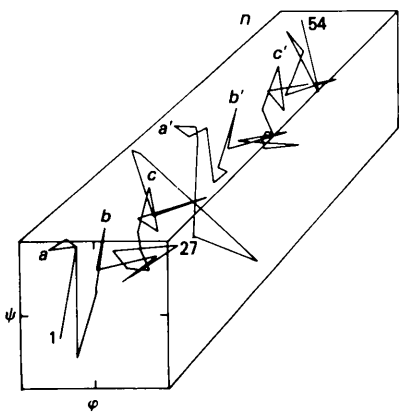

Fig. 1. Three-dimensional Ramachandran plot (TDRP) of the dihedral angles of ferredoxin *Peptococcus*. Axes φ, ψ, n are marked, where φ, ψ are dihedral angles and n is the amino-acid residue number. a and a', b and b', *etc.* are equivalent portions in the structure. The internal duplication of the structure is apparent.

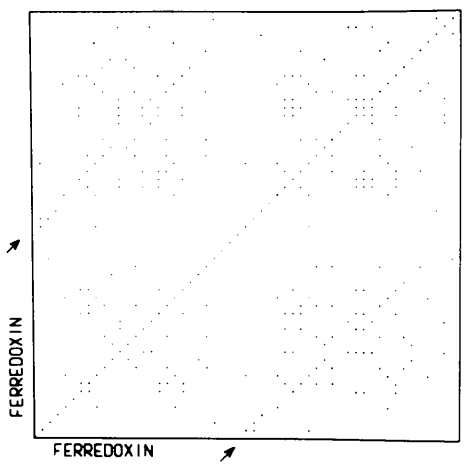

Fig. 2. A graphical representation of the dihedral-angle comparison matrix for ferredoxin with itself. The matrix is constructed by writing the residue numbers for the two polypeptide chains to be compared along the two sides of the matrix. The value to be assigned to the ijth element of the matrix is $\Delta_{ij} = (\varphi_i - \psi_j) + (\psi_i - \psi_j)$. A dot is placed in the diagram at the intersection of the ith row and the jth column if $\Delta_{ij} < 30°$. The off-diagonal lines indicated with arrows show the internal duplication clearly.

N, Δ_{ij} is the element ij of the matrix, $\bar{\Delta}$ is the average value of Δ_{ij} for the whole matrix. Summing the values of χ^2_N for all the diagonals in the matrix we get a second measure of similarity.

$$\chi_2 = \sum_N \chi^2_N.$$

This number gives equal weight to all the diagonals but it is clear that a low value of χ^2_N for a short diagonal (far from the principal diagonal) is less significant than would be the same value for a long diagonal. Thus we calculate another measure of similarity in which each χ^2_N is weighted according to the number of values, n, in the sum.

$$\chi_3 = \sum_N [(\sum^n \Delta_{ij}) - (n\bar{\Delta})]^2.$$

Since these χ^2 measures give equal weight both to runs of adjacent similar angles and to isolated similar angles two further measures (χ_4, χ_5) were obtained by repeating the calculations after removing all isolated points from the matrix. In this way five indices (χ_m) of similarity were obtained for each matrix.

Since the expected distributions of the five indices are unknown and may depend upon the particular structures being compared we used the following empirical method for normalizing the indices so that different pairs of proteins could be compared. Matrices $[\Delta_{ij}]$ were constructed in which the order of residues in one member of the pair was randomized. This was carried out ten times for each comparison. This gave ten values for each of the five indices and the expected mean ($\bar{\chi}_m$) and standard deviation (σ_m) could be obtained for each index. The actual values $\chi(\text{obs})_m$ could be compared with $\bar{\chi}_m$ using the following formula:

$$s_m = \frac{[\chi(\text{obs})_m - \bar{\chi}_m]}{\sigma_m}, \qquad m = 1, \ldots, 5.$$

Adding the average of the four s values derived from χ^2 statistics to the value s_1 derived from the 'best path' calculation provided a convenient overall estimate (S) of the similarity of two proteins. Other schemes could be devised.

Results

A magnetic tape was obtained from the Brookhaven Protein Data Bank (Koetzle *et al.*, 1977). Table 1 lists the proteins for which there was sufficient information for the dihedral angles to be calculated. Most of the repeated determinations of the same structures were omitted to save computer time. The 990 comparisons were carried out using programs written for a DEC PDP 11/45 computer, each comparison taking on average about one hour. The results were arranged according to the combined estimate of similarity (S) and Table 2 shows the first 50 of these.

Table 1. *Proteins compared*

The codes are those used by the Protein Data Bank (list obtained 20th July 1977).

No.	Label	Protein	Species
1	1REI	Part of Bence Jones protein (Immunoglobulin)	Man
2	1MBN	Myoglobin	Sperm Whale
3	1EST	Tosyl-elastase	Pig
4	1CAB	Carbonic anhydrase B	Man
5	1CAC	Carbonic anhydrase C	Man
6	1CPA	Carboxypeptidase A	Cow
7	1FXN	Flavodoxin	*Clostridium MP*
8	1FDX	Ferredoxin	*Peptococcus aerogenes*
9	1FDH	Fetal Deoxyhemoglobin	Man
10	155C	Cytochrome C550	*Paracoccus denitrificans*
11	2PGK	Phosphoglycerate kinase	Horse
12	1TIM	Triosephosphate isomerase	Hen
13	1HHB	Deoxyhemoglobin	Man
14	3CHA	α-Chymotrypsin	Cow
15	1FAB	γ-Immunoglobulin Fab fragment	Man
16	2CYT	Cytochrome C	Tuna
17	2SBT	Subtilisin novo	*Bacillus amylolique faciens*
18	1PAB	Prealbumin	Man
19	1CYC	Ferrocytochrome C	Tuna
20	3CNA	Concanavalin A	Jack Bean
21	2MBN	Myoglobin	Sperm Whale
22	1CHG	Chymotrypsinogen A	Cow
23	1HIP	High potential iron protein	*Chromatum vinosum*
24	1SOD	Cu, Zn superoxide dismutase	Cow
25	1C2C	Ferricytochrome C2	*Rhodospirillum rubrum*
26	3PTI	Trypsin inhibitor	Cow
27	2TLN	Thermolysin	*Bacillus thermoproteolyticus*
28	1CPV	Calcium-binding parvalbumin	Carp
29	1ADH	Alcohol dehydrogenase	Horse
30	1SBT	Subtilisin BPN	*Bacillus amylolique faciens*
31	1B5C	Cytochrome b5	Cow
32	1LHB	Hemoglobin	Sea lamprey
33	2ADK	Adenyl kinase	Pig
34	1GCH	γ-Chymotrypsin	Cow
35	2MHB	Methemoglobin	Horse
36	2DHB	Deoxyhemoglobin	Horse
37	2RXN	Rubredoxin	*Clostridium*
38	1SNS	Nuclease	*Staphylococcus*
39	1LYZ	Lysozyme	Hen egg
40	2PTB	β-Trypsin	Cow
41	1RNS	Ribonuclease	Cow
42	3LDH	Lactate dehydrogenase	Dogfish
43	1LZM	Lysozyme	Bacteriophage T4
44	1PAD	Papain	Papaya
45	1GPD	Glyceraldehyde 3-phosphate dehydrogenase	Lobster

We then examined in more detail some of the predictions of similarity which had not been reported before. Firstly, the diagonal plots themselves were examined. Secondly, structural comparisons similar to those of Rossmann & Argos (1975, 1976, 1977) were carried out to find the best rigid-body superimpositions using a specially written program (Stuart, 1979; Levine, Murihead, Stammers & Stuart, 1978). Table 3 gives the results of these calculations. Rossmann & Argos (1975, 1976, 1977) discuss the assessment for significance of such results. Thirdly, a comprehensive set of graphics programs (Stuart, 1979) enabled us to display pairs of proteins superimposed

Table 2. *The results of the dihedral angle comparisons*

S is defined in the text and is a measure of the probable similarity between two protein structures. The codes are those used in the protein data bank and are as defined in Table 1. The 50 pairs with the highest values are given.
i represent separate determinations of the same structure.
e represent structures which are known to be similar.
√ represent comparisons discussed in the text.

No.	Proteins compared	*S*	No.	Proteins compared	*S*
1	1FDH–1HHB[e]	61·0	26	1B5C–3PTI[√]	7·3
2	1EST–2PTB[e]	25·9	27	1CPA–155C	7·3
3	1CHG–1GCH[e]	22·4	28	1SBT–1MBN[√]	7·2
4	2PTB–1CHG[e]	17·8	29	3CHA–2PTB[e]	7·1
5	1EST–1GCH[e]	16·4	30	155C–1SNS	7·1
6	2PTB–1GCH[e]	16·2	31	2CYT–1LYZ[√]	7·0
7	2CYT–1CYC[i]	15·1	32	1CPA–1LZM[√]	7·0
8	1HHB–1MHB[e]	14·3	33	1FDH–2MHB[e]	6·9
9	3CHA–1GCH[e]	13·2	34	2SBT–1SNS	6·9
10	1CAB–1CAC[e]	13·0	35	155C–1C2C[e]	6·8
11	2SBT–1SBT[i]	12·9	36	1FDH–1CHG	6·7
12	2CYT–1C2C[e]	12·6	37	2CYT–1PAB	6·5
13	155C–2CYT[e]	12·0	38	3LDH–1SBT[e]	6·5
14	3CHA–1CHG[e]	11·5	39	155C–1B5C[√]	6·3
15	1EST–1CHG[e]	11.3	40	1LBH–2MBN[e]	6·3
16	1MBN–2MBN[i]	10·7	41	1CPA–3CHA	6·3
17	1GPD–2MHB[√]	9.9	42	2CYT–1SNS	6·2
18	1CAB–2PTB[√]	9·8	43	1FXN–155C	6·1
19	1EST–3CHA[e]	9·6	44	1PAD–1GPD	6·1
20	1SBT–2MBN[√]	9·4	45	1CPA–1SBT[√]	6·1
21	1SBT–2DHB[√]	9·3	46	1B5C–3LDH	5·9
22	1B5C–1PAD[√]	8·3	47	1FXN–1TIM	5·8
23	1C2C–1CPV[√]	7·8	48	1B5C–2ADK	5·8
24	1FXN–2CYT	7·7	49	1FXN–1CYC	5.8
25	1LHB–1SNS	7·6	50	1RNS–1CHG	5·7

on each other with the relative rotations and translations indicated by the structural comparison calculation so that visual comparison could be carried out. As expected the method clearly is not exhaustive. The statistical nature of the dihedral angle comparison will lead to some similar protein structures being missed and also to some false positive indications of similarity. Secondary structure has a marked effect on the matrix; if there are α-helices in both proteins then rectangular blocks of low values will appear in the matrix. Large errors in calculated dihedral angles can result from quite small errors in model (*xyz*) coordinates and so the method requires accurate structures of consistent stereochemistry. However, it appears from the results that the dihedral angles are preserved in similar structures even though the amino-acid sequences may vary. Thus all the haemoglobin structures were picked out as being similar except for lamprey haemoglobin (1LHB). When the protein data bank coordinate file for lamprey haemoglobin was consulted it was found that the coordinates for this molecule were known to be much less accurate than those for the other haemoglobins. Since regularization and refinement of crystal structures is now normal this should not represent a major limitation of the method.

An improvement in the efficacy of the method could be brought about by basing the comparison upon clearly defined domains where these can be identified.

Table 3. *Results of comparison with the rigid-body superimposition program*

Columns 1, 2, 3, 4, 5 give the proteins compared and the number of residues in each. Columns 8 and 9 give the number of α-carbons which were found to be equivalent in the two structures and the r.m.s. α-carbon separations. The equivalent residues are defined by an iterative procedure. Firstly, a three-dimensional grid is defined whose points represent discrete values of the three Eulerian angles describing the relative orientations of the two molecules. For each orientation a comparison matrix is constructed whose elements are given in equation (1). Residues are equivalent if $P > 0{\cdot}05$. This criterion is in turn dependent on the chosen values of E_1 and E_2 (columns 6 and 7). The relative magnitudes of E_1 and E_2 determine the weight given to d_{ij} and S_{ij}, where d_{ij} is the distance between atoms i and j and S_{ij} depends on the similarity of the shapes of the polypeptide chain on either side of atoms i and j. The diagonal path through the matrix with the greatest number of equivalent residues is determined and this value is assigned to the appropriate point on the grid. The resultant three-dimensional map is examined for peaks defining possible matchings of the two structures. Starting with the highest peak a process of iterative least-squares refinement is carried out to maximize the number of equivalent residues.

Comparison number	Protein 1	Number of residues in 1	Protein 2	Number of residues in 2	E_1	E_2	Number of equivalent residues	R.m.s. distance (Å)	% residues equivalent in protein 1	% residues equivalent in protein 2
—	1CPV	34	1CPV	35	3	10	27	1·40	79	77
		(*NB* This is the comparison of the two halves of the calcium-binding fold)								
5	1EST	240	1GCH	236	2·7	6	225	1·80	94	95
17	2MHB	287	1GPD	333	3·8	3·8	97	4·02	40	29
18	2PTB	223	1CAB	258	3	10	60	3·13	27	23
20	2MBN	153	1SBT	275	3	10	58	2·86	38	21
22	1B5C	85	1PAD	428	3·8	5	46	2·65	54	11
23	1C2C	112	1CPV	109	3	10	52	3·27	46	48
24	2CYT	103	1FXN	138	3	10	57	3·11	55	41
26	1B5C	85	2CYT	103	3·8	5	35	4·13	41	34
27	1CPA	309	155C	134	3·8	5	78	4·63	25	58
32	1CPA	309	1LZM	164	3·8	5	70	3·83	22	43
36	1FDH	287	1CHG	226	3·8	5	86	4·22	30	38
37	2CYT	103	1PAB	228	3·8	5	63	4·41	61	28
38	3LDH	329	1SBT	275	3·8	5	131	3·53	40	48
39	155C	134	1B5C	85	3·8	5	50	4·7	37	59
41	1CPA	309	3CHA	236	3·8	5	87	4·07	28	37
45	1CPA	309	1SBT	275	3·8	5	122	3·93	39	44
46	1B5C	85	3LDH	329	3·8	5	42	3·90	49	13

Analysis of results

It can be seen that when the results are arranged in descending order of probable similarity (S), as in Table 2, the first 16 entries are pairs of proteins which are already known to have similar structures. We took this to be proof of the utility of the method.

We draw attention here to some particularly striking and unexpected similarities which were found on examining the rest of the table using the Rossmann & Argos method and the graphics.

Comparison of parvalbumin and cytochrome C2 (comparison 23 in Table 2)

Carp muscle parvalbumin contains 108 amino-acid residues and binds two calcium atoms. The three-dimensional structure has been determined by Kretsinger & Nockolds (1973) and Moewse & Kretsinger (1975) who showed that the two-calcium-binding loops (residues 40–74 and 75–109) are related by almost exact twofold symmetry. Cytochrome C2 from *Rhodospirillum rubrum* is a haem-containng respiratory protein. It comprises 112 amino-acid residues and its three-dimensional structure was determined by Salemme, Freer, Xuong, Alden & Krout (1973). The polypeptide chain encloses the single haem group without obvious symmetry between different parts of the molecule. Comparison of the amino-acid sequences of carp parvalbumin and *Rhodospirillum* cytochrome C2 by the method of Gibbs & McIntyre (1970) does not show any marked similarity between them. Despite these differences the TDRP method indicated a similar folding pattern in these two proteins (S = 7·8).

Fig. 3 shows stereo drawings of carp parvalbumin and cytochrome C2 separately (*a* and *b*) and superimposed with the relative orientations given by the superimposition algorithm (*c*). Table 4 lists the equivalent residues and their C_α–C_α distances. It can be seen that from residue 33 in parvalbumin and residue 60 in cytochrome C2 onwards the polypeptide chain follows a similar course in the two proteins. There are two major differences between these parts of the two chains, which can be related to the presence of the haem group in cytochrome C2 and the calcium atoms in parvalbumin. (1) In parvalbumin an α-helix (residues 80–90) overlaps the haem position in cytochrome C2. The corresponding part of the chain in cytochrome C2 is in an extended conformation and is displaced 'behind' the 'right-hand' corner of the haem group in the orientation of Fig. 3. Since these two portions of chain must have different sequences of dihedral angles, they cannot be contributing to the high S value for this comparison. (2) In parvalbumin the two calcium-binding sites are formed from loops between adjacent helices. These loops are missing in cytochrome C2.

In contrast to the close matching of the courses of the C-terminal regions of the two proteins the N-terminal regions appear to be different in length and position. In cytochrome C2 the first 60 residues form a flap of mixed α and β structure across the 'front' and 'top' of the haem group. In parvalbumin, on the other hand, the N-terminal section is much shorter

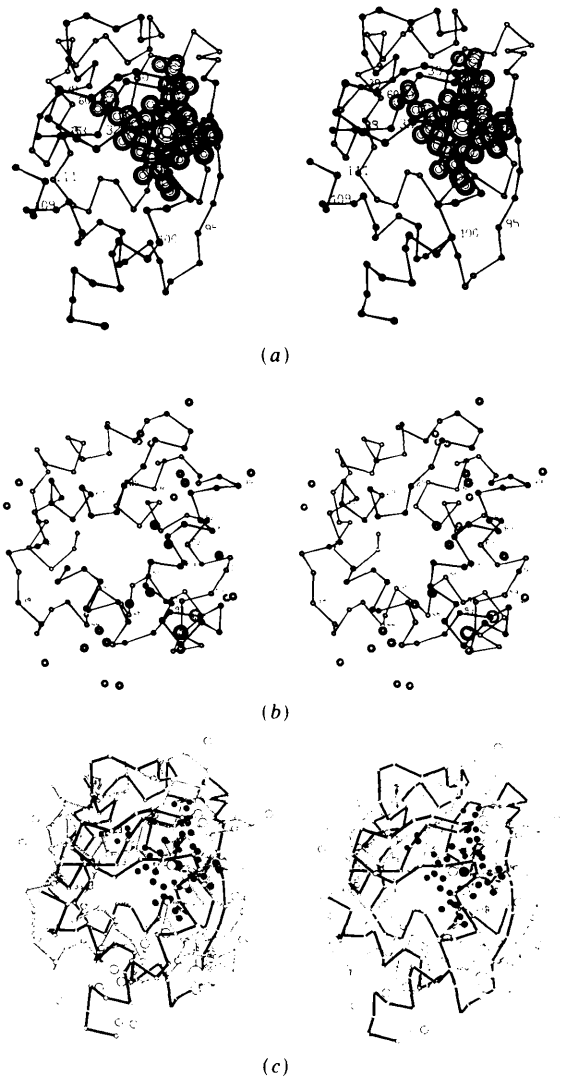

Fig. 3. Stereo diagrams showing the structure of: (*a*) cytochrome C, (*b*) carp parvalbumin, (*c*) (*a*) and (*b*) superimposed. α-carbon positions are joined by virtual bonds. In cytochrome C the haem group is shown [in (*a*) as shaded circles slightly smaller than the van der Waals radii of the atoms, in (*c*) the circles are smaller and completely filled]. For parvalbumin, the two bound calcium ions are shown as large unconnected circles and the water molecules as smaller unconnected circles [shaded in (*b*) and empty in (*c*)]. In (*c*) the virtual bonds for the cytochrome structure are filled in to distinguish them from the open bonds in the parvalbumin representation. The relative orientations between the molecules as illustrated here were determined by the rigid-body superimposition program. The residues at the beginnings and ends of the stretches of structural equivalence are numbered (note that the parvalbumin numbering scheme starts at residue 1 rather than 0 as in the Protein Data Bank).

Table 4. *Residues which were found to be equivalent in the comparison of cytochrome C2 (1C2C) and carp parvalbumin (1CPV)*

For each pair of equivalent residues the distance between their α-carbons is given and residues indicated in the Protein Data Bank as being in helices are marked. Note that the residue numbering used for the parvalbumin is increased by 1 compared to that used in the Protein Data Bank.

1C2C	1CPV		Δ (Å)	1C2C	1CPV		Δ (Å)
29	25		5·2	84	75		3·5
30	26		2·6	87	76		1·7
32	27		5·3	88	77		2·1
33	28		5·1	89	78		0·8
35	29	H_2	5·1	90	79		1·9
36	31		5·0	91	82		3·3
38	32		1·9	92	83	H_5	3·5
39	33		2·3	93	87		4·1
60	34		4·6	94	97		4·0
61	35		4·4	95	98		2·4
61	36		3·4	100	99		1·7
63	39		5·1	101	100		1·1
64	44		2·0	102	101		1·1
65	46		3·3	103	102		1·6
66	47		3·2	104	103		2·4
68	48	H_3	2·4	105	104	H_6	2·9
69	50		3·8	106	105		1·9
70	51		3·9	107	106		3·9
71	59		4·5	108	107		4·2
72	60		3·2	109	108		3·7
73	63		0·8	111	109		3·3
74	64		1·5				
75	65		2·9				
76	66		3·4				
77	67		3·2				
78	68		2·7				
79	69	H_4	2·4				
80	70		3·7				
81	71		1·8				
82	72		1·4				
83	74		2·2				

(30 residues) and is situated at the 'back' and 'top' of the molecule in the orientation of Fig. 3.

It may be significant, however, that in their description of the structure determination of carp parvalbumin, Kretsinger & Nockolds (1973) and Moewse & Kretsinger (1975) mention the difficulty of fitting this part of the structure to the electron density map. The N-terminal helix was poorly resolved and in their refinement large temperature factors were assigned to these atoms. On the 'front' of the parvalbumin molecule, in the position occupied by the first helix in cytochrome C, ordered water molecules were fitted to the density and these were assigned low temperature factors in the refinement. These water molecules are marked in Fig. 3 as open circles. We suggest that in view of the similarity of the other regions of the two molecules the interpretation of the parvalbumin electron density map should be reconsidered to confirm that the two structures are actually different here. Perhaps the molecule may be conformationally heterogeneous in this region and the helix could be twisted into the position occupied by the water molecules. In this case it would coincide with the N-terminal helix in cytochrome C. The structures would then be even more similar than present models.

Cytochrome C compared with hen egg white lysozyme (comparison 31 in Table 2)

Hen egg white lysozyme (Blake, Johnson, Mair, North, Phillips & Sarma, 1967) contains 129 residues and hydrolyses glycosidic bonds whereas tuna cytochrome C (Takano, Kallai, Swanson & Dickerson, 1973) contains 104 residues and is a haem-containing respiratory protein. Fig. 4 shows stereo drawings of the two proteins [(*a*) lysozyme, (*b*) cytochrome C with equivalent sections of lysozyme superimposed, r.m.s. distance 3·75 Å for 55 residues]. Table 5 gives the list of equivalent residues and their C_α–C_α distances.

Jung, Sippel, Grez & Gehutz (1980) have indicated the way in which the lysozyme gene is divided into separate exons. This scheme is marked on Table 5 (column 6). Also marked are the sections consisting of β-strand or α-helix in the two proteins (columns 2 and 4). Exon 2 in lysozyme which contains more than half of the total number of equivalences includes the catalytic residues and residues that bind the oligosaccharide substrate. It can be seen that the largest break in the list of equivalences is between exons 2 and 3. The bulk of the remaining equivalences is in exon 3 which is also involved in catalysis. The exon structure of cytochrome C is also known (Craik,

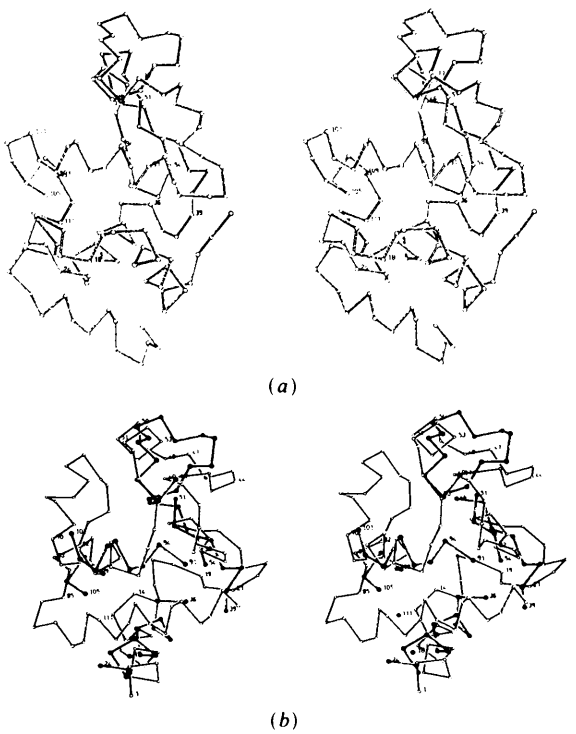

Fig. 4. Stereo diagrams showing the structures of: (*a*) hen egg white lysozyme (open lines and circles); (*b*) cytochrome C (open circles) with the equivalent residues of lysozyme superimposed (filled circles). The first and last residue of each stretch of structural equivalence is numbered.

Sprang, Fletterick & Rutter, 1982) and this is also shown in Table 5. It is interesting to note that the splice at residue 56 of cytochrome corresponds fairly well with the splice between the two principal exons of lysozyme involved in the similarity.

Comparison of carboxypeptidase A with phage T4 lysozyme (comparison 32)

Lysozyme from phage T4 (Matthews & Remington, 1974) has been shown by Rossman & Argos (1976) to have substantial sections which are similar to hen egg white lysozyme. It has 164 residues while carboxypeptidase has 309 residues.

Fig. 5 shows stereo drawings of the two proteins, (*a*) carboxypeptidase and (*b*) lysozyme with equivalent residues of carboxypeptidase superimposed. Table 6 shows the list of equivalent residues and their $C\alpha$–C_α distances (r.m.s. distance 3·83 Å for 70 residues). Also marked in Table 6 are the sections consisting of β-strands and α-helix. 54 of the equivalent residues in lysozyme correspond to an almost continuous stretch of chain starting near the N terminals and the remaining 16 equivalent residues comprise the C-terminal helix. The exon structure of carboxypeptidase has been reported by Craik *et al.* (1982) and this is shown in Table 6. Note that the junctions occur in all cases between stretches of equivalent structure, this is unexpected (the likelihood of it occurring by chance is about 10%).

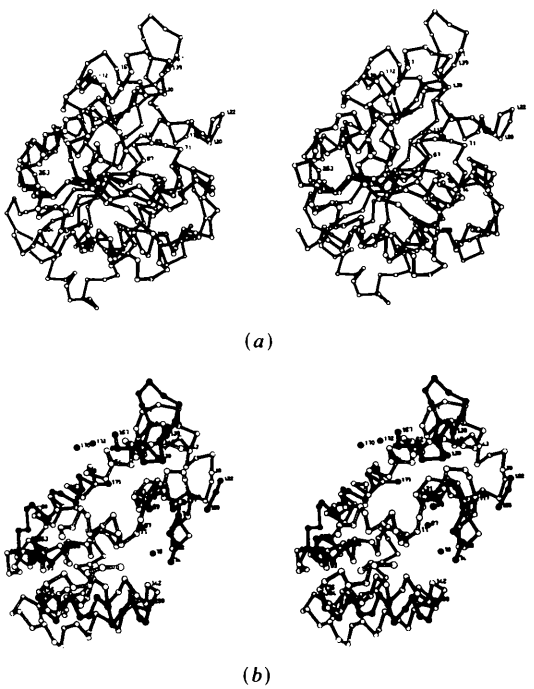

Fig. 5. Stereo diagrams showing the structures of: (*a*) carboxypeptidase; (*b*) phage T4 lysozyme (open circles) with the equivalent residues of carboxypeptidase superimposed (filled circles). The ends of the stretches of structurally equivalent residues are numbered.

Table 5. *Residues which were found to be equivalent in the comparison of cytochrome (2CYT), column 1, and hen egg white lysozyme (1LYZ), column 3*

For each pair of equivalent residues the distance between their α-carbons is given (column 5) and residues indicated in the Protein Data Bank as being in helices are marked (columns 2 and 4). The allocation of the residues to the different exons in the genes for lysozyme and cytochrome C is shown in columns 6 and 7, respectively.

1	2	3	4	5	6	7
2	CYT	1	LYZ	ΔÅ		
2		24		3·8		
3		25		3·8	1	
4		26		4·2		
5		27		4·3		
6		28		2·1		
7		29		2·0		
8	H_1	30	H_2	2·2		
9		31		3·9		
10		32		3·8		
11		33		3·2		
12		34		4·0		
14		35		3·0		
19		36		3·8		
21		39		3·4		
22		40		4·5		
23		41		3·4		
24		42		2·6		1
25		43		4·5		
26		44		4·0		
27		45		4·0	2	
28		46		3·6		
29		51		3·3		
30		52		2·4		
31		53		1·7		
32		54		2·0		
40		63		3·4		
41		64		2·5		
42		65		2·7		
43		66		3·4		
44		67		6·2		
47		68		3·8		
48		69		3·6		
50		70		3·3		
51		71		3·3		
53	H_2	72		1·7		
54		73		2·1		
55		74		4·6		
56		75		3·9		
57		76		3·8		
58		77		4·0		
59		91		6·2		
60		94		2·8		
61		95		4·4		
62		96		2·9		
63		97		0·7		
64		98		2·5	3	
65	H_3	99	H_4	0·7		
66		100		3·0		2
67		101		5·2		
68		105		4·0		
69		106		4·5		
70		107		5·0		
82		108		5·5		
83		109	H_5	5·8	4	
85		111		6·5		

Table 6. *Residues which were found to be equivalent in the comparison of carboxypeptidase (1CPA) and phage T_4 lysozyme (1LZM)*

For each pair of equivalent residues the distance between α-carbons is given and residues indicated in the Protein Data Bank as being in helices are marked. Column 6 shows allocation of the residues of carboxypeptidase gene into the exons of the gene.

1	2	3	4	5	6
1	CPA	1	LZM	Δ(Å)	
71		18		3·6	
72		19		5·6	
73	H_2	20		3·2	3
74		21		3·1	
76		22		4·1	
112		31		4·4	
113		32		4·3	
114		33		3·5	
115		34		3·4	
116	H_4	35		4·4	4
117		36		3·1	
120		37		3·6	
121		38		2·8	
122		39		2·8	
130		43		5·6	
131		46	H_2	4·0	
132		47		2·6	
133		48		4·2	
134		51		3·8	5
135		52		4·9	
136		53		4·1	
137		54		4·5	
138		55		4·5	
139		56		5·8	
161		57		4·7	
162		58		1·7	
163		59		4·0	
164		60		2·4	
165		62		1·5	
166		64		3·6	
167		65		3·8	
170		68		5·7	
171		69		2·9	
175		70		4·9	
176		71	H_3	4·7	
177		72		5·3	
178		73		4·0	6
179		75		3·4	
181	H_5	76		2·1	
182		77		0·9	
183		78		3·0	
184		79		3·4	
185		80		3·0	
186		81		5·3	
187		85		4·4	
188		86	H_4	4·3	
189		88		2·1	
190		92		4·3	
263		93		3·2	
264		94		1·8	
265		95	H_5	2·6	
266		96		6·0	
267		97		6·2	
286		141		5·1	
290		142		3·2	
291		143		3·4	
292		144		4·1	
293		145		3·8	
294		146		3·4	8
295	H_8	147	H_{10}	3·7	
296		148		3·5	
297		149		3·1	
298		150		3·1	
299		151		2·0	
300		152		2·0	
301		153		2·3	
302		154		2·0	
303		155		2·5	
304		156		4·9	

Table 7. *Matching structural features giving high S values*

Comparison number	Protein 1	Protein 2	Comments
17	1GPD	2MHB	A pair of helices packed at the same angle, rest of structure dissimilar
18	1CAB	2PTB	β-sheet equivalent
20 and 28	1SBT	2MBN	A pair of helices packed at the same angle, the rest of the structure dissimilar
21 and 28	1SBT	2DHB / 1MBN	
22	1B5C	1PAD	β-sheet equivalent
26	1B5C	3PTI	β-sheet equivalent
39	155C	1B5C	Four helices in similar orientations

Comparison of carboxypeptidase with subtilisin (comparison 45)

Fig. 6 shows a stereo diagram of the 'nucleotide binding' folds of carboxypeptidase and subtilisin with the helices superimposed. The similarity of the folding in these two proteins has previously been pointed out by Rossman & Argos (1976). It is interesting to note that when the helices are superimposed in this way the strands of β-sheet in the two structures are staggered with respect to one another. Thus, although the two-dimensional topological diagrams (Levitt & Chothia, 1976) of these two structures are similar the packing arrangement is different.

Other comments on Table 2

We have attempted in Table 7 to identify some of the matching structural features which give rise to misleadingly high *S* values in Table 2. This table shows how portions of secondary structure packed similarly in two proteins in the absence of any other matching secondary structure gives rise to a high *S* value.

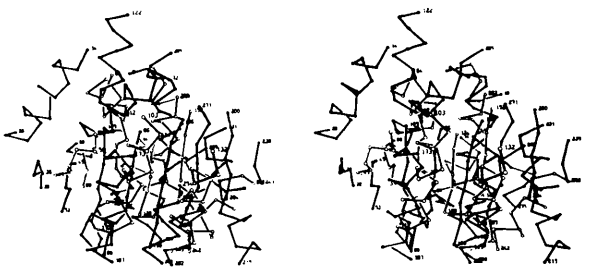

Fig. 6. Stereo diagrams showing α-helices and β-sheets of carboxypeptidase (solid lines) and subtilisin (open lines) superimposed. The numbers are given for the first and last residue in each piece of secondary structure.

Discussion

The results show that a systematic search is capable of revealing unexpected substructures in common between proteins which on functional grounds might have been thought to be entirely different. The structures of cytochrome C2 and carp parvalbumin have been known for many years and there are no functional or other similarities which would have led to their structures being compared with one another. In addition, a structural element which is part of the active site of lysozyme also appears in cytochrome C which has a different function. Thus, shared structural elements of proteins may evolve for reasons which have no obvious relationship to the protein's main function. It might have been thought that the structure of a haem protein would be quite different to that of a non-haem protein, but these examples contradict this reasonable expectation.

However, the results do support the view (Rossmann, 1974) that some of the diversity of protein structures is to be explained by chance recombination among whole sections of genetic material with subsequent divergence, rather than solely by accumulations of point mutations. The discovery that the bases coding for a single enzyme can be distributed in several sections over the chromosome, each corresponding to a separate stretch of polypeptide chain, provides a simple mechanism for this proposal and also an explanation of some of the results presented here. The known exon structures of the lysozyme, cytochrome and carboxypeptidase genes correspond well with the equivalences found here involving these proteins. It should be stressed with regard to these results that the structural comparisons were performed and the results analysed before any of the gene structures were determined. [Some of these results have been reported previously; Stuart (1979)]. Matthews, Grutter, Anderson & Remington (1981) and Artymiuk, Blake & Sippel (1981) have shown that stretches of equivalent residues in phage T4 lysozyme and hen lysozyme are concentrated in specific exons in the hen gene. The present results are the first demonstration that stretches of equivalent residues in functionally dissimilar proteins may also be concentrated in specific exons. In summary, this could be interpreted in terms of a possible mechanism in the evolution of these proteins, namely that new combinations of existing exons could give rise to new proteins with different functions. Although this idea has been suggested before (Gilbert, 1978), supporting data have been lacking. Alternative interpretations are that these sub-fragments are not homologous, in the strict evolutionary sense, but that they are either the result of convergent evolution to a common structure in order to perform a common function, or that they are coincidental similarities due to simple physical constraints. With regard to the latter, Craik, Rutter & Fletterick (1983) have noted that the boundaries between exons tend to occur at the surface of proteins and others (for example Sternberg & Thornton, 1977) have enumerated regularities in the ways in which α-helices and β-sheet structures are arranged in known protein structures. This is a very difficult and complex area. It is difficult, in general, in the absence of amino-acid similarities, to decide between these three alternatives. However, when there is both strong structural similarity and corresponding arrangement of exons the first theory seems to us to be less complicated and more economical in its assumptions and therefore to be preferred. It may also be possible to discover a series of homologous relationships linking two widely divergent molecules, thus indicating that these two molecules are indeed homologous [in morphology this is known as the 'serial criterion of homology' (Remane, 1952)].

Application of methods such as those presented in this paper may assist in providing a comprehensive taxonomy of protein structures based upon the detailed course of the backbone rather than simplified topological relationships between secondary structural elements (Levitt & Chothia, 1976). This may help in the study of the principles of protein folding and relationship of these to protein function (*e.g.* Levine *et al.*, 1978).

We wish to express thanks to Dr Herman C. Watson for provision of computing facilities and also to the SERC for provision of post-doctoral fellowships (to ML and DS).

References

ADMAN, E. T., SIEKER, L. C. & JENSEN, L. H. (1973). *J. Biol. Chem.* **248**, 3987–3996.

ARTYMIUK, P. J., BLAKE, C. C. F. & SIPPEL, A. E. (1981). *Nature (London)*, **290**, 287–288.

BALASUBRAMANIAN, R. (1977). *Nature (London)*, **266**, 856–857.

BLAKE, C. C. F., JOHNSON, L. N., MAIR, G. A., NORTH, A. C. T., PHILLIPS, D. C. & SARMA, V. R. (1967). *Proc. R. Soc. London-.Ser. B*, **167**, 378–388.

CRAIK, C. S., RUTTER, W. J. & FLETTERICK, R. (1983). *Science*, **220**, 1125–1129.

CRAIK, C. S., SPRANG, S., FLETTERICK, R. & RUTTER, W. J. (1982). *Nature (London)*, **299**, 180–182.

DE SANTIS, P., GIGLIO, E., LIQUORI, A. M. & RIPAMONTI, A. (1962). *Nuovo Cimento*, **26**, 616–618.

FITCH, W. M. (1966). *J. Mol. Biol.* **16**, 1–7, 8–16, 17–27.

GIBBS, A. J. & MCINTYRE, G. A. (1970). *Eur. J. Biochem.* **16**, 1–11.

GILBERT, W. (1978). *Nature (London)*, **271**, 501.

JUNG, A., SIPPEL, A. E., GREZ, M. & GEHUTZ, G. (1980). *Proc. Natl. Acad. Sci. USA*, **77**, 5759.

KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F. JR, BRICE, M. D., BERNSTEIN, F. C., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

KRETSINGER, R. H. & NOCKOLDS, C. E. (1973). *J. Biol. Chem.* **248**, 3313–3326.

LEVINE, M., MURIHEAD, H., STAMMERS, D. K. & STUART, D. I. (1978). *Nature (London)*, **271**, 626–630.

LEVITT, M. & CHOTHIA, C. (1976). *Nature (London)*, **261**, 552–557.

MCLACHLAN, A. D. (1979). *J. Mol. Biol.* **128**, 49–79.

MATTHEWS, B. W., GRUTTER, M. G., ANDERSON, W. F. & REMINGTON, S. J. (1981). *Nature (London)*, **290**, 334–335.
MATTHEWS, B. W. & REMINGTON, S. J. (1974). *Proc. Natl. Acad. Sci. USA*, **71**, 4178–4182.
MOEWSE, P. C. & KRETSINGER, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.
RAMACHANDRAN, G. N., RAMAKRISHNAN, C. & SESISAKHARAN, V. (1963). *J. Mol. Biol.* **7**, 95–99.
REMANE, A. (1952). *Die Grundlagen des Natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik.* Leipzig: Akademische Verlagsgesellschaft.
REMINGTON, S. J. & MATTHEWS, B. W. (1978). *Proc. Natl. Acad. Sci. USA*, **75**, 2180–2184.
ROSSMANN, M. G. (1974). *New Sci.* **61**, 266–268.
ROSSMANN, M. G. & ARGOS, P. (1975). *J. Biol. Chem.* **250**, 7525–7532.
ROSSMANN, M. G. & ARGOS, P. (1976). *J. Mol. Biol.* **105**, 75–95.
ROSSMANN, M. G. & ARGOS, P. (1977). *J. Mol. Biol.* **109**, 99–129.
SALEMME, F. R., FREER. S. T., XUONG, N. H., ALDEN, R. A. & KROUT, J. (1973). *J. Biol. Chem.* **248**, 3910–3921.
STERNBERG, M. J. E. & THORNTON, J. M. (1977). *J. Mol. Biol.* **110**, 285–296.
STUART, D. (1979). PhD thesis. Univ. of Bristol.
TAKANO, T., KALLAI, O. B., SWANSON, R. & DICKERSON, R. E. (1973). *J. Biol. Chem.* **248**, 5234–5255.

Acta Cryst. (1984). A**40**, 610–616

The Influence of Rational Dependence on the Probability Distribution of Structure Factors

BY V. GRAMLICH

Institut für Kristallographie und Petrographie der ETH, CH-8092 Zürich, Switzerland

(*Received* 27 *January* 1984; *accepted* 1 *May* 1984)

Dedicated to Professor J. D. Dunitz on the occasion of his 60th birthday

Abstract

An index subgroup of strong main reflexions and cosets of weak reflexions are typical features of crystal structures with systematic rational dependence of the atom coordinates exhibiting a pseudotranslational symmetry. The mean squares of these normalized structure-factor sets which deviate significantly from unity are interpreted in terms of correlation coefficients of the atom coordinates. An asymptotic form of the von Mises distribution of a structure factor phase is derived which allows for rational dependence and makes explicit use of the $\overline{|E|^2}$ values of the different structure-factor sets. The formula provides a basis for the use of phase relationships of the type 'weak–strong–weak' proposed in the recent literature. The limits of the method are estimated. In particular, symmetry and homometry problems in superstructures are more complex than in usual cases and their careful consideration is essential for the success of procedures intending an automatic solution.

Introduction

The concept of rational dependence of atom coordinates in connexion with the statistics of normalized structure factors was introduced by Hauptman & Karle (1953). Renormalization was proposed in order to remove problems imposed by systematically strong and weak reflexion classes occurring in this context (Hauptman & Karle, 1959). No general statistical basis for this procedure was available. Despite some successful attempts at direct phase determination for superstructures, the method was not much further developed. Combined trial-and-error, Patterson and Fourier methods turned out to be a powerful tool (*cf.* Schulz, 1976, and references cited therein).

The application of direct methods to structures containing heavy atoms (Beurskens & Noordik, 1971) was successful even if the heavy atoms exhibited some subperiodicity (*cf. DIRDIF*: Beurskens, Bosman, Doesburg, Gould, van den Hark, Prick, Noordik, Beurskens & Parthasarathi, 1981, and references cited therein). In this context procedures using partial information (Main, 1976; Heinerman, Krabbendam & Kroon, 1977) may be mentioned. Giacovazzo (1983) developed a new theory for the use of *a priori* known partial structure information and compared his method with the difference structure factor (*DIRDIF*) approach. *DIRDIF* will fail if the input model consists of nearly all atoms in idealized positions (Beurskens & Bosman, 1982).

The main difference in the approach of this paper compared with others is the explicit use of the information $\overline{|E(\mathbf{h}_n)|^2}^{\mathbf{h}_n}$ (different numbers for different classes $n = 1, \ldots p$, if rational dependence is prominent). This is particularly interesting for those superstructures where a known 'average' model may explain the strong reflexions quite satisfactorily but

0108-7673/84/050610-07$01.50