J. Mol. Biol. (1994) **235**, 625–634

Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins

Vladimir N. Maiorov and Gordon M. Crippen

College of Pharmacy, University of Michigan
Ann Arbor, MI 48109, U.S.A.

In the study of globular protein conformations, one customarily measures the similarity in three-dimensional structure by the root-mean-square deviation (RMSD) of the C^{α} atomic coordinates after optimal rigid body superposition. Even when the two protein structures each consist of a single chain having the same number of residues so that the matching of C^{α} atoms is obvious, it is not clear how to interpret the RMSD. A very large value means they are dissimilar, and zero means they are identical in conformation, but at what intermediate values are they particularly similar or clearly dissimilar? While many workers in the field have chosen arbitrary cutoffs, and others have judged values of RMSD according to the observed distribution of RMSD for random structures, we propose a self-referential, non-statistical standard. We take two conformers to be intrinsically similar if their RMSD is smaller than that when one of them is mirror inverted. Because the structures considered here are not arbitrary configurations of point atoms, but are compact, globular, polypeptide chains, our definition is closely related to similarity in radius of gyration and overall chain folding patterns. Being strongly similar in our sense implies that the radii of gyration must be nearly identical, the root-mean-square deviation in interatomic distances is linearly related to RMSD, and the two chains must have the same general fold. Only when the RMSD exceeds this level can parts of the polypeptide chain undergo nontrivial rearrangements while remaining globular. This enables us to judge when a prediction of a protein's conformation is "correct except for minor perturbations", or when the ensemble of protein structures deduced from NMR experiments are "basically in mutual agreement".

Keywords: globular proteins; protein structure comparison; optimal rigid body superposition; three-dimensional structural motif; enantiomorphous relationships

1. Introduction

Suppose we have two alternative conformations of some globular protein, and we want to decide how similar they are. Typical situations where this might arise are the ensemble of structures calculated from NMR studies on a protein's conformation, or the comparison of the X-ray crystal structure with structures proposed from homology modeling or more ambitious tertiary structure prediction methods. In any case, the alternatives are reasonably compact and globular, and we are concerned with the general folding of the backbone. Hence, the protein is typically represented by its virtual C^{α} atom chain of n residues or points. For a quantitative single-number measure of structural similarity between structures A and B, one generally uses either the "distance RMSD"† (also called the "distance matrix error") (Nishikawa & Ooi, 1972; Levitt, 1976):

$$D^2_{\mathrm{dis}}(\mathrm{A}, \mathrm{B}) = (n(n-1)/2)^{-1} \sum_{i<j}^{n} (d_{\mathrm{A}ij} - d_{\mathrm{B}ij})^2 \quad (1)$$

(where $d_{\mathrm{A}ij}$ and $d_{\mathrm{B}ij}$ are the corresponding distances between the ith and jth atoms) or the "coordinate RMSD" after optimal rigid body superposition (Rao & Rossmann, 1973). In this work, we focus on the comparison of entire globular structures and the effect this has on assessing the significance of co-

† Abbreviations used: RMSD, root-mean-square deviation; CMD, correlation matrix determinant.

0022-2836/94/020625-10 $08.00/0

ordinate RMSD values. It is quite a different problem to judge the significance of spatial similarity when different proteins are compared, allowing insertions and deletions in the amino acid sequence (Rossmann & Argos, 1976, 1977; Sippl, 1982; Remington & Matthews, 1978, 1980; Abagyan & Maiorov, 1988, 1989; Zuker & Somorjai, 1989; Vriend & Sander, 1991; Aleksandrov *et al.*, 1992).

While distance RMSD is easy to calculate, it fails to distinguish between mirror images. Coordinate RMSD distinguishes mirror images, but the calculations are more complicated because first, both structures are translated so that their centroids are at the origin, and then one is rotated so that the squared deviation in corresponding coordinates is minimized. A number of different algorithms have been developed to carry out this optimal rigid body superposition of pairs or groups of structures (McLachlan, 1972, 1979, 1982; Diamond, 1976, 1988, 1992; Kabsch, 1976, 1978; Lesk, 1986; Mackay, 1984; Zuker & Somorjai, 1989; Kearsley, 1989, 1990, KenKnight, 1984; Gerber & Muller, 1987; Sutcliff *et al.*, 1987; Shapiro *et al.*, 1992). In what follows, we shall refer to coordinate RMSD as simply RMSD.

In order to decide whether two structures are similar, most investigators have simply chosen an arbitrary RMSD cutoff value, such as 3 Å. A few have constructed a population of alternative structures, observed the frequency distribution of RMSD among them, fitted the distribution usually to a Gaussian curve, and developed a statistically significant cutoff depending on n and the structural class of the protein (Cohen & Sternberg, 1980; McLachlan, 1979, 1984; Remington & Matthews, 1978, 1980; Aleksandrov *et al.*, 1992). Here, we develop an intrinsic RMSD cutoff for similarity that is not arbitrary or statistically based, but rather depends on special consequences of globular structures that have so far not been clearly recognized.

2. Methods

In order to understand the special properties of the RMSD for globular structures, we must first review McLachlan's analysis of optimal superposition (McLachlan, 1979). Given two structures A and B, each consisting of n points, first translate both of them so that their respective centroids are the origin of the coordinate system. Then we want to rotate the coordinate vectors $\mathbf{a}_k$ of A onto the corresponding coordinate vectors $\mathbf{b}_k$ of B by some proper rotation matrix $\mathbf{R}$ chosen to make the least-squares superposition:

$$D^2(\mathrm{A}, \mathrm{B}) = (1/n) \sum_{k=1}^{n} (\mathbf{R}\mathbf{a}_k - \mathbf{b}_k)^2, \tag{2}$$

which defines the coordinate RMSD, $D(\mathrm{A}, \mathrm{B})$. One can express this in terms of the radii of gyration (for coordinates referred to their centroid, $R_\mathrm{A}^2 = (1/n)\sum_{k=1}^{n} \mathbf{a}_k^2$, and similarly for B, independent of rotation) as:

$$D^2(\mathrm{A}, \mathrm{B}) = R_\mathrm{A}^2 + R_\mathrm{B}^2 - 2v, \tag{3}$$

where $v \geq 0$ is the optimal rotation correction. In terms of the original coordinates, the correlation matrix, $\mathbf{U} = (U_{ij})$, is defined by:

$$U_{ij} = (1/n) \sum_{k=1}^{n} a_{ik} b_{jk}, \quad i, j = 1, 2, 3, \tag{4}$$

which has $S = \mathrm{sign}\,(\det(\mathbf{U})) = \pm 1$. In what follows, we shall refer to the sign of the correlation matrix determinant (CMD). It turns out that the correction v from eqn (3) may be expressed as:

$$v = \lambda_1 + \lambda_2 + S\lambda_3, \tag{5}$$

in terms of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ of U. The worst case superposition occurs when $v = 0$ and, hence, $D^2 = R_\mathrm{A}^2 + R_\mathrm{B}^2$, but for globular structures, the radii of gyration are relatively small, and the distribution of D has a small upper bound.

The optimal rotation in the superposition procedure is a proper one, but sometimes mirror inverting one of the structures can lead to a lower RMSD. (See, for example, Kabsch, 1978; McLachlan, 1979; Diamond, 1990.) Denote by $\hat{D}(\mathrm{A}, \mathrm{B})$ (the "conjugated RMSD") the RMSD when either A or B is reflected. Then:

$$\hat{D}^2 - D^2 = 4S\lambda_3, \tag{6}$$

which is positive, negative or zero, depending on the CMD (Diamond, 1990).

3. Self-avoiding Configurations on a Simple Cubic Lattice

We have seen how D depends strongly on R_A and R_B, but we want to study the consequences of globularity on RMSD comparisons. An easy model system is the set of all configurations of a self-avoiding 27-center chain on a $3 \times 3 \times 3$ simple cubic lattice. Clearly, this is only the crudest approximation to real protein folds, but since the chain fills the whole space, every configuration has exactly the same radius of gyration, is roughly spherical and is very compact. Just to give it a little resemblance to proteins, we took the lattice spacing to be 3·8 Å. Because we can explicitly enumerate all 103,346 different configurations (excluding enantiomers) (Shakhnovich & Gutin, 1990), our sampling of conformation space is perfect. That means there are over 10^9 pairwise comparisons, so for that purpose, we used only every tenth configuration. The following results are unaffected by random choice and sample size.

We can view the total RMSD distribution (Fig. 1, solid line) as the sum of two skewed, equally populated distributions, corresponding to positive and negative CMDs, neglecting the 0·02% of the comparisons having zero CMD. This is in contrast to the usual assumed Gaussian shape (Remington & Mathews, 1978, 1980; Cohen & Sternberg, 1980) or the skew toward low RMSD for random, freely-jointed chains (McLachlan, 1984). The skew toward high RMSD may be a characteristic feature of comparison of compact, globular structures. The other important general features are: there is a minimal $D = D_0 = 1{\cdot}47$ Å > 0 that occurs strictly for positive CMD; only at a considerably higher $D = D_1 = 4{\cdot}47$ Å can there be negative CMD comparisons; and both distributions terminate at about the same $D = D_2$

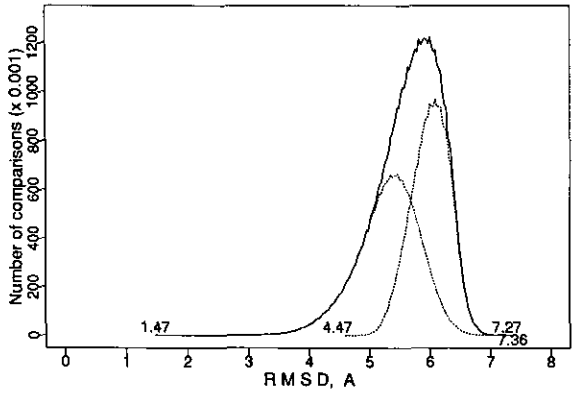

Figure 1. RMSD frequency distributions for 53,411,280 pairwise comparisons of 10,336 27-point self-avoiding configurations on a $3 \times 3 \times 3$ cubic lattice: total distribution (solid line) and two subdistributions corresponding to positive (left dotted) and negative (right dotted) CMD.

(7·27 Å for positive CMD and 7·36 Å for negative) because all configurations have the same small radius of gyration.

The conformational implications of D_0 are particularly important. For these lattice walks, this occurs when the two ends of the chain cooperatively exchange places, as in Figure 2. A smaller D could be achieved by moving just one end (necessarily outside the $3 \times 3 \times 3$ box), but these conformations are constrained to be extremely compact. The equivalent event in realistic protein structures would be for the N and C-terminal helices to be of similar length and lying next to each other in antiparallel orientation, so that they could exchange places without expanding the globule or significantly disturbing the rest of the chain.

Another useful way to display these cutoffs is to plot the conjugated RMSD *versus* RMSD, referred to as the $\hat{D}/D$ diagram, as in Figure 3. Because of the large number of points corresponding to 53,411,280 comparisons, we show only the envelope of minimal and maximal values of $\hat{D}$ for each D. Because we excluded enantiomers in our exhaustive list of configurations, we must conclude that the extreme symmetry of the Figure is due to our uniform coverage of the set of all comparisons. The angularity arises from the restriction to extraordinarily compact configurations having exactly the same radius of gyration. As a next step toward

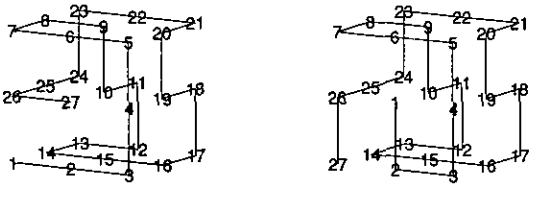

Figure 2. Example pair of 27-point self-avoiding cubic lattice configurations with smallest attainable RMSD for a $3 \times 3 \times 3$ lattice RMSD, 1·47 Å. The only difference between these configurations is the arrangement of the terminal segments 1 to 2 and 26 to 27.

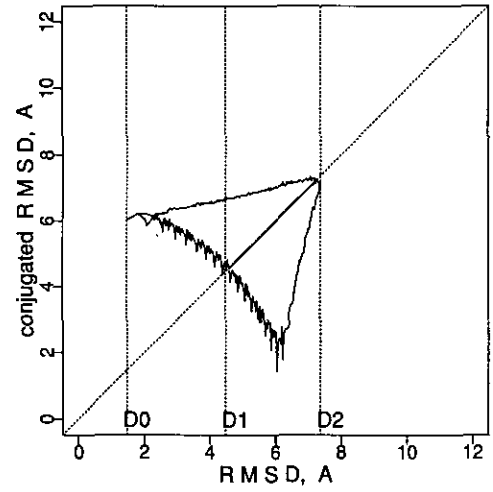

Figure 3. $\hat{D}/D$ diagram for 27-point self-avoiding configurations on the $3 \times 3 \times 3$ cubic lattice. Every 10th lattice configuration out of the exhaustive enumeration is included (10,336 configurations in the sample), resulting in 53,411,280 comparisons. Dotted lines mark the diagonal $\hat{D} = D$, as well as D_0, D_1 and D_2.

real proteins, we sampled 1336 configurations of 27-point, self-avoiding chains on a $4 \times 4 \times 4$ simple cubic lattice. As seen in Figure 4, the less uniform sampling of comparisons produces an asymmetric diagram, and the variation in radius of gyration rounds the corners. D_2 has increased because of the greater allowed radius of gyration, but D_1 remains the same. Although not clearly shown in the Figure, D_0 is decreased, because in some configurations the tail has room to be moved without affecting the rest of the chain.

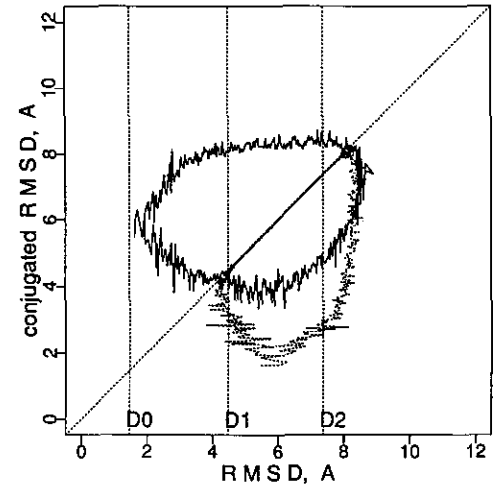

Figure 4. $\hat{D}/D$ diagram for 27-point self-avoiding configurations on $4 \times 4 \times 4$ lattice. Since their total number is so large that exhaustive enumeration can hardly be done in a reasonable time, only a limited sample of 1336 were used for the illustration, resulting in 891,780 comparisons. Dotted lines give the symmetrization by reflecting the upper half about the diagonal. For reference, the D_0, D_1 and D_2 values for Fig. 3 are marked.

4. Comparisons of Protein Structures

The next question is whether we can clearly identify D_0, D_1 and D_2 in real protein structures. First, we note that unless we restrict our survey to polypeptide conformations that are always so compact that the radius of gyration is small, D_2 will be large, and maximal dissimilarity is of limited interest in any event. Secondly, we want to determine D_1 as a function of the number of residues, n, so as to establish a general standard for "substantial dissimilarity" in protein conformational comparisons. Thirdly, since protein C^α coordinates are not confined to discrete lattice points, D_0 is nearly zero for very similar structures. We will show, however, there is a nontrivial minimal rearrangement threshold observed in comparisons of dissimilar proteins.

The total set of protein crystal structures we considered were taken from PDB (Abola *et al.*, 1987). We considered only the C^α coordinates of those proteins having resolution better than 3 Å, and without obvious chain breaks in the middle (Table 1). Disordered or unresolved residues at the N and/or C termini were not included in the polypeptide chains we consider here. For brevity, we refer to these chains by their PDB code and the chain identifier in the PDB file (e.g. 2pab.A is the A chain of prealbumin).

In order to determine D_0 and D_1, we carried out an "all-by-all" comparison, as in Remington & Matthews (1978, 1980), where all possible consecutive fragments consisting of L amino acid residues in one protein are compared to all those in the second by optimal superposition, resulting in $(N_1-L+1)\times(N_2-L+1)$ comparison, where N_1 and N_2 are the numbers of residues in the proteins, with the obvious assumption that the probe length $L \leq N_1, N_2$. This procedure is performed for all pairs of protein structures having adequate length. Only those segments passing our previously established criteria for compactness (Maiorov & Crippen, 1992) were included in the comparisons.

Since D_0 is anticipated to depend on the similarity of the proteins, we drew up a list of 66 "dissimilar protein pairs" (Table 2), based on the recommendations of Sander (Hobohm *et al.*, 1992; Holm *et al.*, 1992). In addition, we deleted the pairs of structures from apparent representatives of

Table 1

List of the proteins used in the work, sorted by PDB code

PDB code†			
155c	1abp	1bjl.1	1bjl.2
1cc5	1ccr	1cse.I	1cse.E
1ctf	1cts	1ecd	1est
1fx1	1hip	1hmg.B	1hmg.A
1hmq.A	1hoe	1hvp.A	1lh4
1lyz	1lz1	1mba	1mbd
1paz	1pfk.A	1pyp	1rei.A
1rhd	1sn3	1tim.A	1wrp.R
1ycc	2abx.A	2act	2alp
2aza.A	2b5c	2c2c	2cab
2cdv	2cyp	2fb4.L	2gn5
2ig2.L	2hhg.A	2hhb.B	2lhb
2lzm	2pab.A	2pka.A	2pka.B
2rhe	2sga	2sod	2ssi
2stv	2taa.A	351c	3adk
3ebx	3fab.L	3fab.H	3fxc
3fxn	3gap.A	3gpd.G	3icb
3pgk	3rp2.A	4ape	4dfr.A
4mdh.A	4rhv.3	4rhv.1	4tln
5cpa	5cpv	5cyt.R	6ldh
7api.A	8adh	9pap	

† In the case of more than one chain in a PDB file, the chain identifiers are given as a suffix.

Table 2

Protein structure pairs between which no observable structural similarities occur ("dissimilar protein pairs")

	Protein 1	Protein 2		Protein 1	Protein 2		Protein 1	Protein 2
1	1ccr	4ape	23	2act	2cyp	45	2stv	3pgk
2	1ccr	4tln	24	2act	3pgk	46	2stv	4ape
3	1ccr	7api.A	25	2alp	1hmg.A	47	351c	5cpa
4	1cse.I	1rei.A	26	2alp	1rhd	48	351c	7api.A
5	1cse.I	3fxn	27	2alp	1tim.A	49	3adk	1est
6	1ctf	1pfk.A	28	2cdv	4tln	50	3adk	2act
7	1ctf	3gap.A	29	2cdv	6ldh	51	3adk	4tln
8	1hip	1cse.E	30	2gn5	1est	52	3ebx	2sga
9	1hip	4tln	31	2gn5	3pgk	53	3ebx	3fxc
10	1hoe	1paz	32	2pab.A	1abp	54	3fab.L	1abp
11	1hoe	2alp	33	2pab.A	1pfk.A	55	3fab.L	3pgk
12	1hvp.A	3gap.A	34	2pab.A	3pgk	56	3fab.L	8adh
13	1hvp.A	4fdr.A	35	2pka.A	7api.A	57	3fxc	1abp
14	1rei.A	1paz	36	2pka.A	8adh	58	3fxc	1lz1
15	1rei.A	1pfk.A	37	2sga	1hmg.A	59	3icb	1hmq.A
16	1sn3	2aza.A	38	2sga	1tim.A	60	3icb	1lyz
17	1sn3	3fab.H	39	2sga	3pgk	61	5cpv	4ape
18	1wrp.R	2lhb	40	2ssi	1pyp	62	5cpv	4tln
19	1wrp.R	2sga	41	2ssi	2cyp	63	5cpv	7api.A
20	2abx.A	2cyp	42	2ssi	3adk	64	9pap	1hmg.A
21	2abx.A	2taa.A	43	2ssi	4rhv.3	65	9pap	1rhd
22	2act	1rhd	44	2stv	2cab	66	9pap	4mdh.A

Table 3

Protein structure pairs between most of which significant structural similarities occur ("similar protein pairs")

	Protein 1	Protein 2		Protein 1	Protein 2		Protein 1	Protein 2
1	1ccr	2c2c	23	2act	3rp2.A	45	2sga	1cse.E
2	1ecd	1lh4	24	2act	4ape	46	2sga	2act
3	1ecd	1mba	25	2alp	2act	47	2sga	2alp
4	1ecd	1mbd	26	2alp	3rp2.A	48	2sga	3rp2.A
5	1ecd	2hhb.B	27	2alp	4tln	49	2sga	4tln
6	1ecd	2lhb	28	2alp	9pap	50	2sga	9pap
7	1fx1	1tim.A	29	2aza.A	2sod.O	51	351c	1cc5
8	1fx1	4mdh.A	30	2b5c	1ccr	52	3fab.L	2fb4.L
9	1fx1	6ldh	31	2b5c	5cyt.R	53	3fxn	1fx1
10	1lh4	1mbd	32	2c2c	155c	54	3fxn	1tim.A
11	1lyz	1cts	33	2fb4.L	3fab.H	55	3fxn	6ldh
12	1lyz	1lz1	34	2hhb.A	1lh4	56	3rp2.A	1cse.E
13	1mba	1lh4	35	2hhb.A	1mba	57	3rp2.A	1est
14	1mba	1mbd	36	2hhb.A	1mbd	58	3rp2.A	4tln
15	1mba	2hhb.B	37	2hhb.A	2lhb	59	3rp2.A	5cpa
16	1mba	2lhb	38	2hhb.B	1lh4	60	4dfr.A	8adh
17	1paz	2aza.A	39	2hhb.B	1mbd	61	4sbv.A	4rhv.1
18	1paz	3gpd.G	40	2hhb.B	2lhb	62	5cyt.R	155c
19	1paz	8adh	41	2pka.A	3rp2.A	63	5cyt.R	1ccr
20	1rei.A	2fb4.L	42	2pka.B	1est	64	5cyt.R	2c2c
21	1rei.A	2rhe	43	2rhe	2rb4.L	65	9pap	2act
22	1rei.A	3fab.H	44	2rhe	3fab.H	66	9pap	3rp2.A

different structural/functional (super)families, to avoid already known cases of spatial similarity between sequentially distant proteins (Holm *et al.*, 1992). Table 3, on the other hand, is a list of spatially similar pairs of structures. Here, we relied mainly on commonly accepted structural resemblance (in both sequence and 3D structure), which is certainly the case if the candidate proteins are either homologous or representatives of the same structural/functional (super)family. Sufficient arbitrarily chosen similar protein pairs were added to Table 3 to make it approximately as large as Table 2.

Figure 5 is a typical $\hat{D}/D$ diagram for $L = 60$, involving 133,676 comparisons between compact fragments of dissimilar protein pairs. $D_0 = 6$ Å is the smallest (positive CMD) observed D, which shows that the minimal rearrangement required to convert one compact conformation into what is generally viewed as a completely different one, is actually larger than is commonly believed. D_1, the smallest observed D having negative CMD, is only 1·5 Å larger. The analogous diagram for similar protein pairs, seen in Figure 6, shows a much smaller D_0, as expected, but a comparable D_1. From an analysis of similar diagrams for $L = 27, 40, 60, 80, \ldots, 180$ we fit the observed D_0 and D_1 values by linear regression to:

$$D_{\text{cutoff}} = a + b(N_{\text{res}})^{1/3}. \tag{7}$$

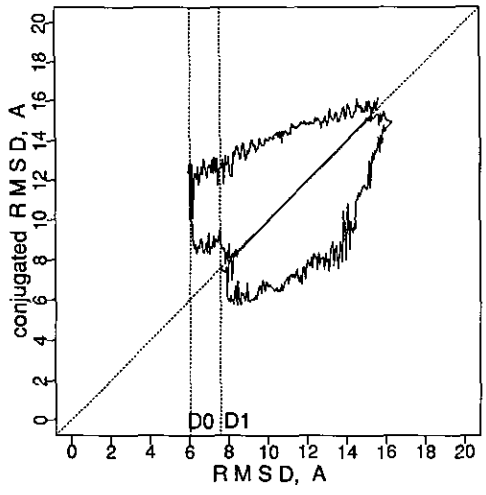

Figure 5. Typical $\hat{D}/D$ diagram for the comparison of 60 residue segments from dissimilar proteins. D_0 and D_1 cutoffs are marked.

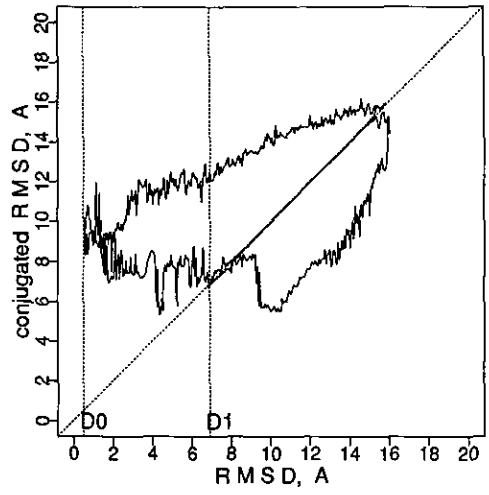

Figure 6. Same as in Fig. 5, but the list of similar protein pairs (Table 3) was used to collect 187,696 comparisons.

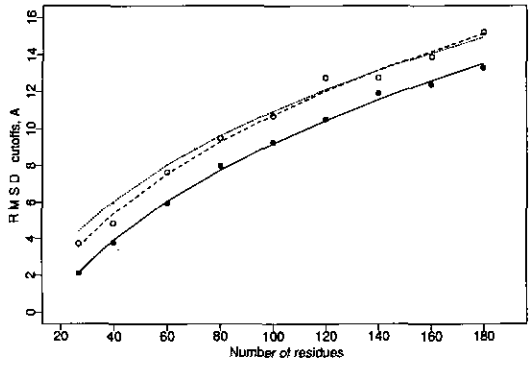

Figure 7. Plots of RMSD cutoffs *versus* number of residues in protein fragments compared. Dependencies of D_0, D_1 and that for the 1% level, *versus* probe length for dissimilar protein pairs (Table 2) fitted by linear regression to eqn (7). Filled circles and continuous curve represent the D_0 cutoff dependence. Open circles and dashed line represent D_1 cutoff data. The dotted line shows the curve for the RMSD cutoff from statistical estimation of similarity significance at the 1% level, as in McLachlan (1979) and Aleksandrov *et al.* (1992).

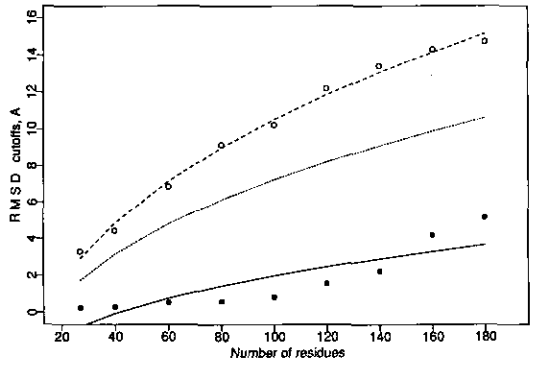

Figure 8. Same as Fig. 7, except for similar protein pairs (Table 3).

In addition, we also derived the 1% level RMSD cutoff dependence as a reference to compare with the results of other authors. The resulting curves are shown in Figures 7 and 8, and the regression parameters (intercepts, slopes and correlation coefficients) are listed in Table 4. The fitting for dissimilar proteins is particularly accurate. The D_1 curve runs consistently about 1 Å above D_0 and close to the 1% cutoff, which is, in turn, in good agreement with the results of Aleksandrov *et al.* (1992). For similar protein pairs, the D_0 curve, of course, runs much lower because it simply signifies minor perturbations in structure, rather than substantial rearrangements. The greater scatter only reflects exactly how close a match between segments could be found, given that the proteins are generally rather similar in structure. As can be seen from Table 4 and Figure 9, D_1 depends on comparing compact structures of a given chain length, independent of whether they arise from similar or dissimilar proteins. A fit of D_1 for the combined similar and dissimilar comparisons has a root-mean-square fitting deviation of 0·38 Å, which one could view as the accuracy. Correspondingly, D_0 for dissimilar proteins only, has an RMS fitting deviation of only 0·18 Å.

There are two remarkable features of comparisons having $D < D_0$ that are revealed when one compact, native, "reference" protein structure is compared to many, not necessarily compact segments of the same length taken from larger, similar and dissimilar proteins. For example, in Figure 10 we see that below D_0, the radius of gyration of the segments is essentially equal (±5%) to that of the reference, while simultaneously the coordinate and distance RMSDs are linearly related. These relations hold for all eight reference proteins we considered (Table 5). In the eight fits of $D_{dis} = a + bD$, we find a ranges from 0·03 Å to 0·63 Å, and b ranges from 0·62 to 0·81, which is in reasonable agreement with figures reported by Levitt (1976) (0·00 Å, 0·82) and Cohen & Sternberg (1980) (0·19 Å, 0·75). Furthermore, the $D < D_0$ comparisons are always due to homologous proteins, as listed in the Table. The vast majority of the 51,482 comparisons for all eight reference structures fall above D_0, and are due to both similar and dissimilar proteins. The only clear regularities seen for large D are the tracks in Figure 10 due to the small and coherent changes in

Table 4

Linear regression parameters for dependencies of RMSD cutoffs against number of residues (probe length) for comparisons among dissimilar and similar pairs of proteins (eqn (7))

Type of input data	RMSD cutoff†	Intercept±error (Å)	Slope±error	Correlation coefficient
Dissimilar pairs	D_0	−10·82±0·37	4·31±0·08	0·997
	D_1	−9·61±0·72	4·38±0·16	0·996
	1%	−7·56±0·84	3·99±0·18	0·993
	D_{dis0}‡	−4·54±0·37	2·36±0·07	0·938
Similar pairs	D_0	−5·74±1·85	1·66±0·40	0·842
	D_1	−10·96±0·66	4·63±0·14	0·997
	1%	−8·32±3·22	3·35±0·70	0·874

† Rows marked by D_0 and D_1 represent the data for dependences of the lowest observed RMSD and those for comparisons with negative CMD, respectively. 1% marks the data for dependence of a statistically estimated RMSD cutoff (see the text).

‡ This entry is obtained for distance RMSD by linear regression to eqn (7) for comparisons of dissimilar protein structures.

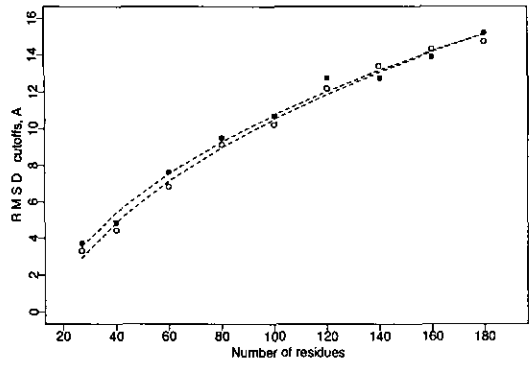

Figure 9. Comparison of the D_1 curves from dissimilar (Fig. 7, filled circles) and similar (Fig. 8, open circles) protein comparisons.

the plotted variables when the reference is compared to overlapping segments taken from larger protein structures. For very large D, the observed asymptotically linear dependence of R and D_{dis} arise from the overwhelming contribution of the larger of the two radii of gyration in equation (3).

Given the linear relationship between D_{dis} and D for low values of D, we can derive the distance RMSD equivalent to D_0 for dissimilar proteins as a function of polypeptide chain length:

$$D_{dis\,0} = -4{\cdot}54 + 2{\cdot}36(N_{res})^{1/3}, \qquad (8)$$

with a correlation coefficient of 0·938 (Table 4). Thus, in spite of the inability of distance RMSD to distinguish enantiomers, it can be applied to identify the rearrangement threshold marking significant similarity of protein structures.

As an example of a practical application of D_0, consider the problem of calculating an ensemble of structures for a protein, given interatomic distance constraints from NMR experiments. If there are many constraints affecting all parts of the chain, or if the procedure for generating conformations subject to these constraints fails to sample widely, then the ensemble will be a cluster of minor variations on a basic conformation. RMSD values exceeding D_0 constitute objective evidence that significantly different classes of conformations have been sampled. Alternatively, discovering that all RMSD values are below D_0 shows that only minor variations have been sampled but, of course, it does not reveal which of the two reasons is behind it. Note that according to equation (7) and Table 4, D_0 has a traditionally acceptable value of 2 Å for a 26 residue chain, but for 100 residues, RMSD values less than 9·2 Å indicate clustering.

Table 5

Eight reference protein structures having at least some $D < D_0$ *in comparisons with segments taken from (similar) counterpart protein structures. For these close comparisons, we give the parameters for linear regression of distance RMSD to coordinate RMSD,* $D_{dis} = a + bD$

PDB code of reference structure	Number of amino acid residues	Number of comparisons with RMSD below the D_0 cutoff	PDB codes of the respective comparison's counterparts†	Linear regression intercept a (Å)	Linear regression slope, b	Correlation coefficient
5cyt.R	103	16	2c2c 1ycc	0·23	0·62	0·995
1rei.A	107	8	2ig2.L 2rhe	0·06	0·80	0·999
2rhe	114	9	2ig2.L 3fab.H	0·03	0·81	0·997
1ecd	136	61	1mba 1mbd 2hhb.A 2hhb.B 2lhb 1lh4	0·57	0·63	0·993
2hhb.A	141	42	1mba 1mbd 2hhb.B 2lhb 1lh4	0·20	0·66	0·980
2hhb.B	146	18	1mba 1mbd 2lhb 1lh4	0·33	0·68	0·991
2sga	181	11	2alp	0·63	0·63	0·982
9pap	212	7	2act	0·19	0·72	0·993
pti‡	58	–		0·00	0·82	0·99
pti§	58	—	—	0·19	0·75	0·99

† Shown in this column are PDB codes and chain identifiers (if any) for those protein structures whose fragments have RMSD values below the D_0 threshold.

‡ According to Levitt (1976).

§ According to Cohen & Sternberg (1980).

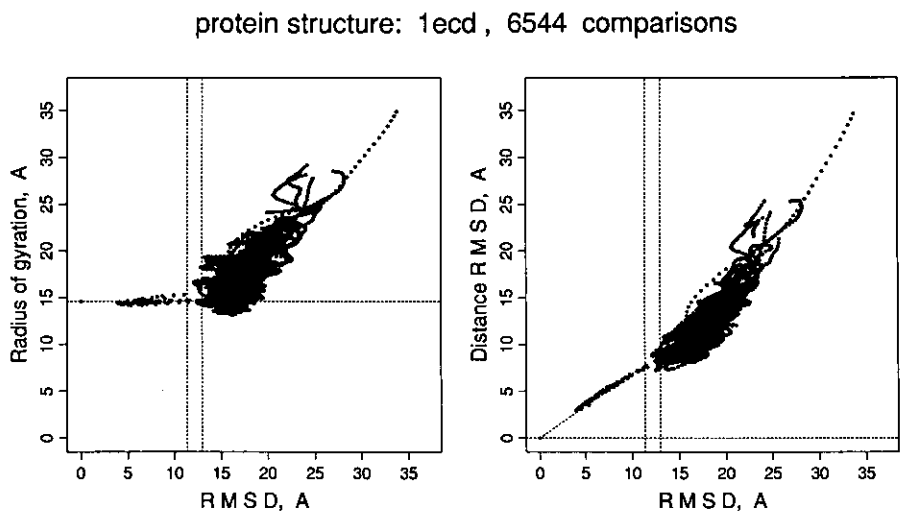

Figure 10. Plots of the radius of gyration (left) and distance RMSD (right) *versus* coordinate RMSD resulting from comparisons between reference protein 1ecd (deoxyerythrocruorin) and all segments of the same length taken from all larger proteins listed in Table 1. D_0 and D_1 are marked by vertical dotted lines. The positions of the reference structures on the ordinate axes are marked by dotted horizontal lines. On the plot of distance RMSD *versus* coordinate RMSD, the linear regression approximation (dotted line) for the range of coordinate RMSD below the threshold D_0 is shown. All counterparts to the reference structure with RMSD below D_0 belong to homologous proteins (Table 5).

In the PDB there are several examples of multiple alternative NMR structures deposited for one protein, but only in the case of the 48 residue neurotoxin I from sea anemone (Fogh *et al.*, 1990; PDB entry 2sh1) do two of the structures differ by so much as 4·69 Å, a value close to $D_0 = 4{\cdot}84$ Å for this chain length. As can be seen in Figure 11, the two structures are very similar, except for the loop containing residues 8 to 16, which occupies two substantially different positions in order to maintain radii of gyration of 9·22 and 9·36 Å, which are close to 8·88 Å, the minimum value one can expect to find for such a polypeptide chain. This obviously resembles the minimal rearrangement threshold phenomenon for $3 \times 3 \times 3$ lattice walks, where all structures were confined to lattice points and had the same extremely low radius of gyration. However, because these protein structures are not confined to lattice points, some of the six others are only 2 Å from these extremes and are clearly minor variations. In order to fall between the two by positioning the loop straight out to the right in the Figure, the structure must violate the compactness limit otherwise observed and and expand its radius of gyration to 9·94 Å.

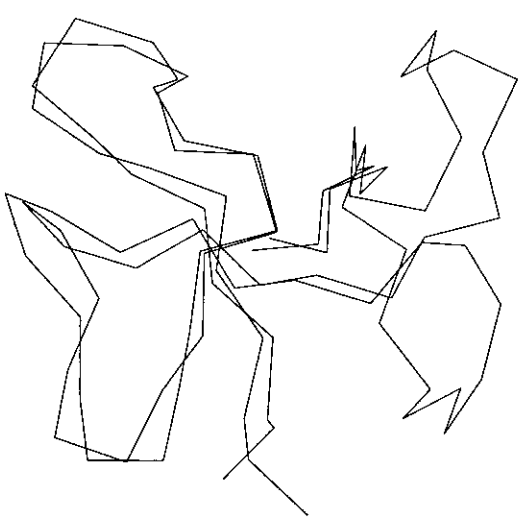

Figure 11. C^{α} tracings of models 2 and 6 of the 8 deposited in PDB file 2sh1, NMR structures for neurotoxin I from sea anemone (Fogh *et al.*, 1990). Most features of the fold are closely constrained except for the residues 8 to 16 loop on the right side.

5. Conclusions

The most important thing we have learnt from all this is that the comparison of protein structures has some general features arising from the restriction to compact, globular conformations of the polypeptide chain. These features are not special to proteins, since they can also be seen in compact walks on a cubic lattice, but they correspond closely to other, very different assessments of conformational similarity. (1) There is a coordinate RMSD cutoff, D_1, above which two conformations may or may not have $\hat{D} < D$, but below which one structure must more closely resemble the other structure than the other's mirror image. This is an intrinsic feature of globular conformations that can be seen in comparisons of simple lattice walks, compact fragments from sequentially dissimilar proteins, and compact fragments from highly homologous proteins. The value depends only on chain length and packing density. (2) D_1 corresponds to a level of similarity surpassed by only about 1% of all pairs of random, compact conformers. (3) There is a second cutoff, D_0, that falls 1 to 2 Å lower. For very compact globules, this is the least (cooperative) rearrange-

ment of the chain possible that still maintains the same very small radius of gyration. (4) When comparing conformations of contiguous segments taken from proteins generally considered as sequentially and structurally unrelated, $D > D_0$ in all cases. Segments from similar proteins may also fall in this range. The value of D_0 depends very simply and accurately on the number of residues in the chain. (5) Only when comparing conformations of segments taken from proteins generally viewed as sequentially or structurally very similar, can one find examples of $D < D_0$. (6) In this regime of extreme conformational similarity, the radii of gyration of the two globular structures must be nearly equal, and the distance RMSD is proportional to the coordinate RMSD.

In this work we have not addressed the difficult problem of comparing the structures of two different proteins while permitting chain insertions and deletions in the hope of detecting structural homology and distant evolutionary relations. However, our insights about intrinsic levels of similarity between pairs of compact conformations for the same protein may find application in many fields where optimal superposition is applied: experimental methods of spatial structure determination (X-ray crystallography and NMR protein spatial structure determination), protein structure analysis, protein database searches and computer modeling of protein folding.

This work was supported by grants from the National Institutes of Health (GM37123) and the National Institute on Drug Abuse (DA06746). We are indebted to all the crystallographers who deposited their protein structural data in the Protein Data Bank.

References

Abagyan, R. A. & Maiorov, V. N. (1988). A simple quantitative representation of polypeptide chain folds: comparison of protein tertiary structures. *J. Biomol. Struct. Dynam.* **5**, 1267–1279.

Abagyan, R. A. & Maiorov, V. N. (1989). An automatic search for similar spatial arrangements of α-helices and β-strands in globular proteins. *J. Biomol. Struct. Dynam.* **6**, 1045–1060.

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein data bank. In *Crystallographic Databases - Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergerhoff, G. & Sievers, R., eds), pp. 107–132, Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.

Aleksandrov, N. I., Takahashi, K. & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.

Cohen, F. E. & Sternberg, M. J. E. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.* **138**, 321–333.

Diamond, R. (1976). On the comparison of conformations using linear and quadratic transformations. *Acta Crystallogr. sect. A*, **32**, 1–10.

Diamond, R. (1988). A note on the rotational superposition problem. *Acta Crystallogr. sect. A*, **44**, 211–216.

Diamond, R. (1990). Chiralty in rotational superposition. *Acta Crystallogr. sect. A*, **46**, 423–423.

Diamond, R. (1992). On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Sci.* **1**, 1279–1287.

Fogh, R. H., Kem, W. R. & Norton, R. S. (1990). Solution structure of neurotoxin I from the sea anemone *Stichodactyla helianthus*. A nuclear magnetic resonance, distance geometry and restrained molecular dynamics study. *J. Biol. Chem.* **265**, 13016–13028.

Gerber, P. R. & Muller, K. (1987). Superimposing several sets of atomic coordinates. *Acta Crystallogr. sect A*, **43**, 426–428.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. sect. A*, **32**, 922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. sect. A*, **34**, 827–828.

Kearsley, S. K. (1989). On the orthogonal transformation used for structural comparisons. *Acta Crystalllogr. sect. A*, **45**, 208–210.

Kearsley, S. K. (1990). An algorithm for the simultaneous superposition of a structural series. *J. Comput. Chem.* **11**, 1187–1192.

KenKnight, C. E. (1984). Comparison of methods of matching protein structures. *Acta Crystallogr. sect. A*, **40**, 708–712.

Lesk, A. M. (1986). A toolkit for computational molecular biology. II. On the optimal superposition of two sets of molecules. *Acta Crystallogr. sect. A*, **42**, 110–113.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Mackay, A. L. (1984). Quarternion transformation of molecular orientation. *Acta Crystallogr. sect. A*, **40**, 165–166.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

McLachlan, A. D. (1972). A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. sect. A*, **26**, 656–657.

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

McLachlan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallogr. sect. A*, **38**, 871–873.

McLachlan, A. D. (1984). How alike are the shapes of two random chains? *Biopolymers*, **23**, 1325–1331.

Nishikawa, K. & Ooi, T. (1972). Tertiary structure of protein. II. Freedom of dihedral angles and energy calculations. *J. Phys. Soc. Japan*, **32**, 1338–1347.

Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**, 241–256.

Remington, S. J. & Matthews, B. W. (1978). A general method to assess similarity of protein structures, with appliations to T4 bacteriophage lysozyme. *Proc. Nat. Acad. Sci., U.S.A.* **75**, 2180–2184.

Remington, S. J. & Matthews, B. W. (1980). A systematic approach to the comparison of protein structures. *J. Mol. Biol.* **140**, 77–99.

Rossmann, M. G. & Argos, P. A. (1976). Exploring structural homology of proteins. *J. Mol. Biol.* **105**, 75–95.

Rossmann, M. G. & Argos, P. A. (1977). The taxonomy of protein structure. *J. Mol. Biol.* **109**, 99–129.

Shakhnovich, E. I. & Gutin, A. M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature (London)*, **346**, 773–775.

Shapiro, A., Botha, J. D., Pastore, A. & Lesk, A. M. (1992). A method for multiple superposition of structures. *Acta Crystallogr. sect. A*, **48**, 11–14.

Sippl, M. J. (1982). On the problem of comparing protein structures. Development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J. Mol. Biol.* **156**, 359–388.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins: Struct. Funct. Genet.* **11**, 52–58.

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

Edited by F. Cohen

(Received 6 July 1993; accepted 17 August 1993)