PROTEINS: Structure, Function, and Genetics 22:273–283 (1995)

Size-Independent Comparison of Protein Three-Dimensional Structures

Vladimir N. Maiorov and Gordon M. Crippen
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT **Protein structures are routinely compared by their root-mean-square deviation (RMSD) in atomic coordinates after optimal rigid body superposition. What is not so clear is the significance of different RMSD values, particularly above the customary arbitrary cutoff for obvious similarity of 2–3 Å. Our earlier work argued for an intrinsic cutoff for protein similarity that varied with the number of residues in the polypeptide chains being compared. Here we introduce a new measure, ρ, of structural similarity based on RMSD that is independent of the sizes of the molecules involved, or of any other special properties of molecules. When ρ is less than 0.4–0.5, protein structures are visually recognized to be obviously similar, but the mathematically pleasing intrinsic cutoff of $\rho<1.0$ corresponds to overall similarity in folding motif at a level not usually recognized until smoothing of the polypeptide chain path makes it striking. When the structures are scaled to unit radius of gyration and equal principle moments of inertia, the comparisons are even more universal, since they are no longer obscured by differences in overall size and ellipticity. With increasing chain length, the distribution of ρ for pairs of random structures is skewed to higher values, but the value for the best 1% of the comparisons rises only slowly with the number of residues. This level is close to an intrinsic cutoff between similar and dissimilar comparisons, namely the maximal scaled ρ possible for the two structures to be more similar to each other than one is to the other's mirror image. The intrinsic cutoff is independent of the number of residues or points being compared. For proteins having fewer than 100 residues, the 1% ρ falls below the intrinsic cutoff, so that for very small proteins, geometrically significant similarity can often occur by chance. We believe these ideas will be helpful in judging success in NMR structure determination and protein folding modeling.**

Key words: globular proteins; protein structure analysis; optimal rigid body superposition; three-dimensional structural motif

INTRODUCTION

Whenever the question of conformational similarity arises, particularly for globular proteins, the standard measure is the root-mean-square deviation (RMSD) in atomic coordinates after the two structures have been optimally superimposed by rigid body translation and rotation. Several clever algorithms have been devised to help decide which atoms of the one molecule are to be superimposed on which atoms of the other, but in all that follows we will concentrate on the simple case of comparing the C^{α} atoms of two n-residue polypeptide chains in the obvious way. While the superposition calculation itself is perfectly routine, the significance of the resulting number is not so clear. Most workers in the field simply adopt an arbitrary cutoff of 2–3 Å as the boundary between similar and dissimilar three-dimensional structures. Others have examined the distribution of RMSD in certain ensembles of comparisons and have thus attached statistical significance to the values.[2–4,12,14,16,17]

We, on the other hand, have sought an intrinsic measure of significance that is at least applicable to comparisons of globular proteins.[11] In that work we discovered two cutoffs: D_1 is the smallest RMSD such that a lower RMSD may sometimes be achieved by first mirror inverting one of the structures; and D_0 is the smallest RMSD that is observed between two segments of polypeptide chain coming from clearly unrelated proteins. Both of these cutoffs were found to be linear functions of $n^{1/3}$, apparently due to the constant density of amino acid residues packed into roughly spherical globules, implying that the protein's diameter is proportional to $n^{1/3}$. This highlights the often neglected fact that RMSD is affected by both the conformational similarity and the overall sizes of the proteins being compared. As we will show below, judging similarity by a single, fixed cutoff is not valid for pairs of proteins spanning a reasonable range in numbers of residues. Even our D_0 and D_1 cutoff functions of n lack generality in that they have been devised for globular proteins, and not general molecules or arbitrary configurations of points. In

Received December 28, 1994; revision accepted February 28, 1995.

Address reprint requests to Gordon M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109.

this paper we introduce a new measure of conformational similarity that automatically compensates for size effects and is applicable to the comparison of arbitrary sets of points.

METHODS

Conformational Similarity

Consider two arbitrary configurations, A and B, each consisting of n points in three-dimensional space. Assume we have numbered the points so that the given correspondence matches point i in A with point i in B for $i = 1, \ldots, n$. Denote the Cartesian coordinates of the points by $\mathbf{a}_i$ and $\mathbf{b}_i$. Our concern here is the optimal superposition of these matching pairs of points by rigid body translation and rotation of A and B. It is well known that such a superposition requires that the centroids of A and B must coincide,[9,19] so we will assume in all that follows that the centroids of both have already been translated to the origin, i.e., $\Sigma_i \mathbf{a}_i = \Sigma_i \mathbf{b}_i = 0$. Then there are many algorithms (see references in Maiorov and Crippen[11]) for finding the proper rotation matrix, R, where $\det(\mathbf{R}) = 1$, that minimizes

$$D^2(A, B) = n^{-1} \sum_i (\mathbf{R}\mathbf{a}_i - \mathbf{b}_i)^2 \tag{1}$$

so that D is the desired RMSD between the two configurations. Now the value of D reflects not only the similarity in relative placement of the points, but also the sizes of A and B and the disparity in their sizes. Let us take as our measure of size the radius of gyration, which can be calculated from the magnitudes of the center-of-mass coordinate vectors $\mathbf{r}_i$

$$R^2 = n^{-1} \sum_i r_i^2. \tag{2}$$

Suppose we start with center of mass coordinates of A and B where A has already been optimally rotated onto B so that in Eq. (1) the optimal rotation matrix is $\mathbf{R} = \mathbf{I}$, the identity matrix. Then Eq. (1) simplifies to

$$D^2(A, B) = n^{-1} \sum_i (\mathbf{a}_i - \mathbf{b}_i)^2 \tag{3}$$

which can be expanded and simplified by Eq. (2) to

$$D^2(A, B) = R^2(A) + R^2(B) - \frac{2}{n} \sum_i \mathbf{a}_i \cdot \mathbf{b}_i. \tag{4}$$

Depending on how similar A and B are, one can show the full range of D is[12,14]

$$0 \le D^2(A, B) \le R^2(A) + R^2(B). \tag{5}$$

Suppose we multiply all the coordinates of the points of both A and B by some scalar, f. Clearly from Eq. (2), $R(fA) = fR(A)$ and similarly for B. D is also proportional to f. In order to see this, note that optimal superposition implies[13]

$$\sum_i \mathbf{a}_i \times \mathbf{b}_i = 0. \tag{6}$$

Scaling one or both of the structures by f preserves this optimality condition, so no rotational readjustment is necessary. Thus

$$D^2(fA, fB) = n^{-1} \sum_i (f\mathbf{a}_i - f\mathbf{b}_i)^2 = f^2 D^2(A, B). \tag{7}$$

The behavior of D under scaling one of the structures is a little more complicated. The equivalent to Eq. (4) becomes

$$D^2(A, fB) = R^2(A) + f^2 R^2(B) - \frac{2f}{n} \sum_i \mathbf{a}_i \cdot \mathbf{b}_i \tag{8}$$

so that eliminating the summation between these equations gives

$$D^2(A, fB) = R^2(A) + f^2 R^2(B) - f[R^2(A) + R^2(B) - D^2(A, B)] \tag{9}$$

a quadratic in f having its minimal value when

$$f = \frac{R^2(A) + R^2(B) - D^2(A, B)}{2R^2(B)}. \tag{10}$$

In the special case of $A = B$, we have $D^2(fA, A) = (1 - f)^2 R^2(A)$ which has the pleasing property that $D^2(fA, A)$ is minimal when $f = 1$, that is, when the two structures have equal radii of gyration. However, this is not true in general for arbitrary A and B.

It would be preferable to have some measure of dissimilarity that compensated for these simple size effects. It is helpful to introduce two artificial configurations: the "sum" or mean structure

$$\mathbf{s}_i = (\mathbf{a}_i + \mathbf{b}_i)/2, \quad i = 1, \ldots, n \tag{11}$$

and the "difference" structure

$$\mathbf{d}_i = (\mathbf{a}_i - \mathbf{b}_i)/2, \quad i = 1, \ldots, n. \tag{12}$$

Since $\mathbf{a}_i = \mathbf{s}_i + \mathbf{d}_i$ and $\mathbf{b}_i = \mathbf{s}_i - \mathbf{d}_i$, we can express $\mathbf{a}_i \cdot \mathbf{b}_i = s_i^2 - \mathbf{d}_i^2$, which simplifies Eq. (4) to

$$D^2(A, B) = R^2(A) + R^2(B) - 2R^2(s) + 2R^2(d) \tag{13}$$

where $R(s)$ and $R(d)$ are the respective radii of gyration of the sum and difference configurations. Since Eq. (3) expresses D in terms of the difference structure, Eq. (12), the definition of the radius of gyration, Eq. (2), results in

$$R^2(d) = D^2(A, B)/4 \tag{14}$$

so that Eq. (13) can be rearranged to give

$$2R^2(s) = R^2(A) + R^2(B) - D^2(A, B)/2. \tag{15}$$

As shown in Figures 1 and 2, as one progresses from similar to dissimilar configurations, the sum structure shrinks while the difference structure expands, and the radius of gyration of the difference structure

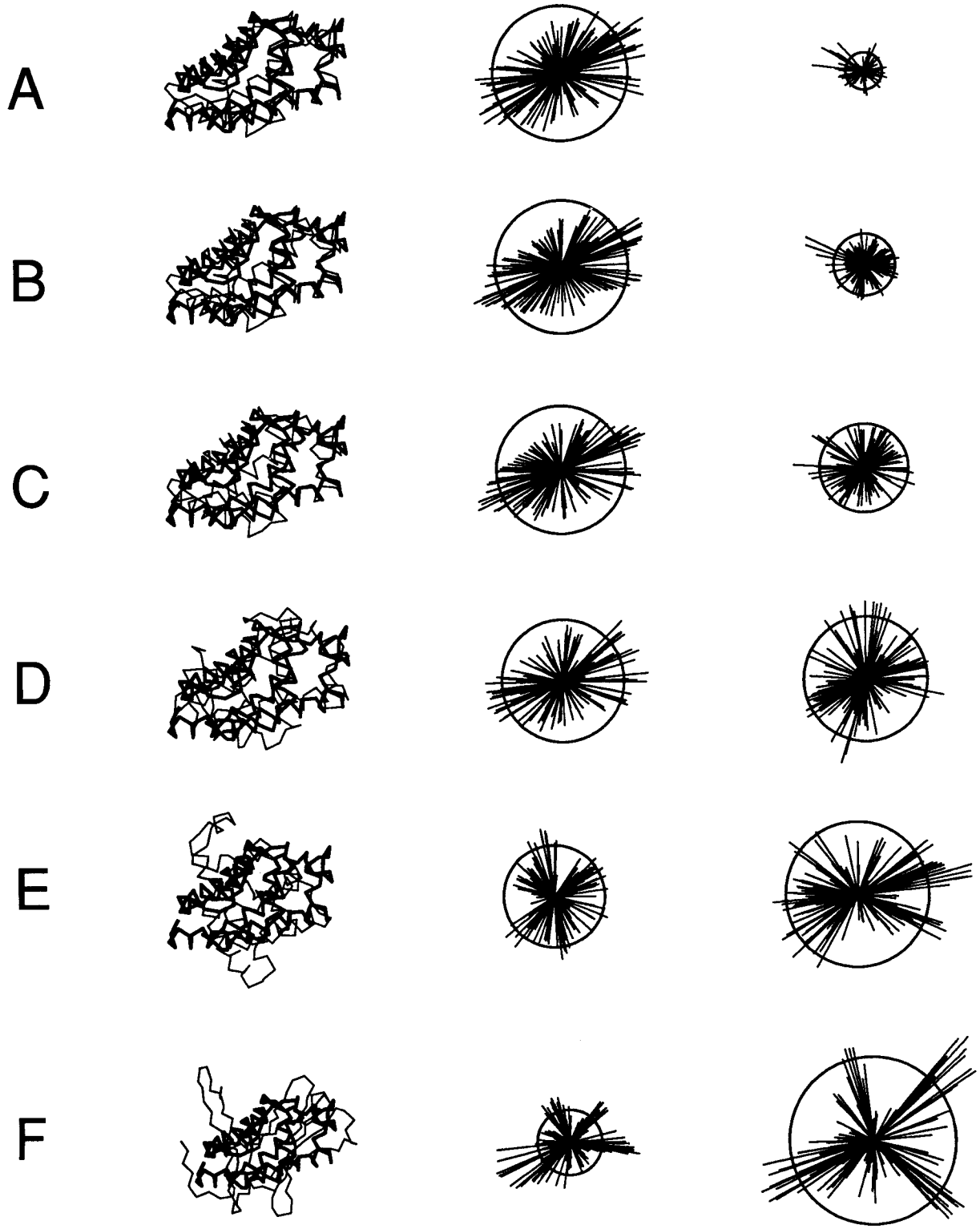

Fig. 1. Spatial similarity, sum, and difference structures. The left column shows examples of optimally superimposed polypeptide chains (heavy vs. light lines) which have increasing dissimilarity going from case **A** to **F.** The center column shows the corresponding sum structure vectors, $\mathbf{s}_i$ and a circle proportional to its radius of gyration, R_s. The right column shows the corresponding difference structure vectors and its radius of gyration. The radii are nearly equal in case **D,** corresponding to $\rho \approx 1$.

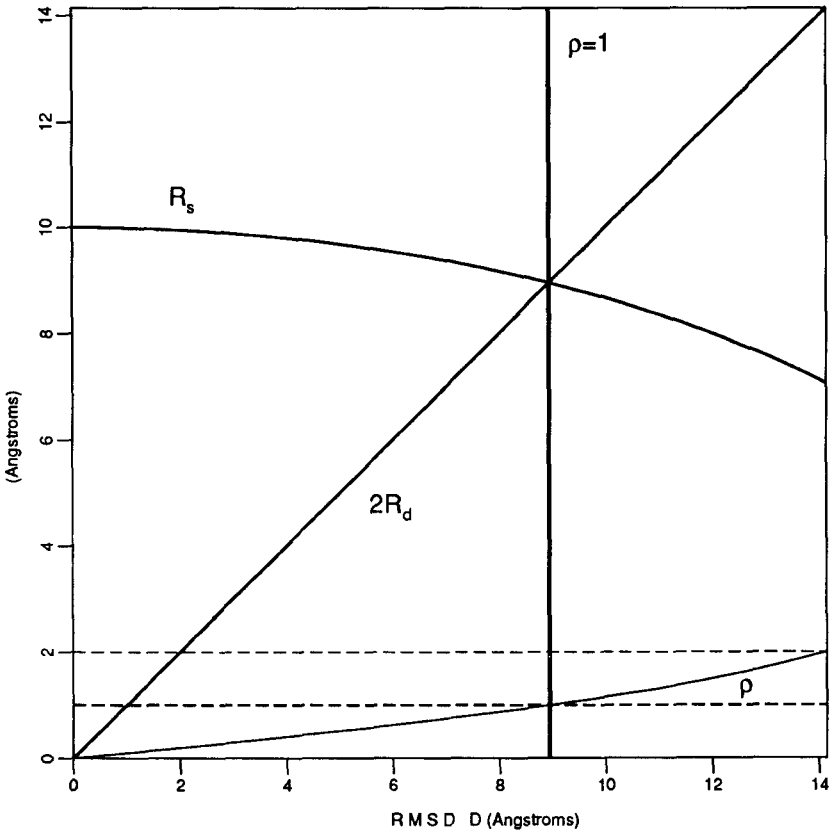

Fig. 2. A plot of $2R_d$, R_s, and ρ as a function of D for constant $R_A = R_B = 9.94$ Å.

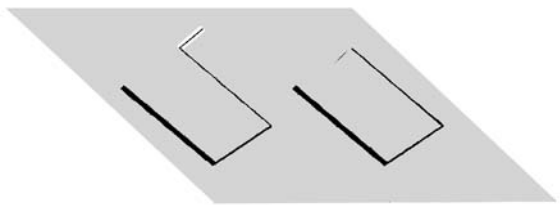

Fig. 3. An example of two antisimilar structures that are very similar. They both lie almost entirely in the shadowed plane except for the last segment in the background running up out of the plane in one and down in the other.

is just proportional to $D(A, B)$. As a measure of dissimilarity, we propose a ratio of the radii of gyration of the difference and sum structures, namely

$$\rho(A, B) = \frac{2R(d)}{R(s)} = \frac{2D(A, B)}{[2R^2(A) + R^2(B) - D^2(A, B)]^{1/2}}. \quad (16)$$

Remembering the range of D from Eq. (5), we see that $0 \leq \rho \leq 2$, independent of R_A and R_B (Fig. 2). Since D and R are both proportional to a common scaling factor, ρ is independent of scaling in the sense that $\rho(fA, fB) = \rho(A, B)$. If only one of the structures is caled, ρ has the nice property of being minimal when the two radii of gyration are equal, in general for any A and B. This can be shown by calculating $\rho^2(fA, B)$ from Eq. (9) and then solving $\partial\rho^2(fA, B)/\partial f = 0$ for f, which leads eventually to the conclusion that $f = R(B)/R(A)$. As a matter of mathematical esthetics, we suggest $\rho < 1$ for a size-independent criterion of similarity. Intuitively speaking, this is when the structural commonality dominates over the differences.

Spherically Scaled Conformational Similarity

While $\rho = 1$ is the point where the sum and difference structures have equal radii of gyration, one would like to choose a cutoff in the ρ scale having greater intrinsic mathematical significance. When A and B are very similar arrangements of points in general position, then $0 \approx D(A, B) < \hat{D}(A, B)$, where $\hat{D}(A, B)$ is the *anti*comparison RMSD calculated by first mirror inverting one of the structures and then optimally superimposing it on the other as usual. At the other extreme, if $B = \hat{A}$, the mirror image of A, then $0 = \hat{D}(A, B) < D(A, B)$. We will call arbitrary A and B an "antisimilar" pair of structures whenever the anticomparison RMSD is less than the ordinary RMSD. In between, there must be some minimal value of $D(A, B)$ or $\rho(A, B)$ below which an-

tisimilarity is impossible, and this value would be an intrinsic criterion for significant similarity.

First observe that $0 \approx D(A, B) \approx \hat{D}(A, B)$ when A and B are very similar and nearly planar, as in Figure 3. In order to establish a nontrivial minimal D for antisimilarity, we must treat only spherically scaled structures, defined as follows. Assume that the $n > 3$ points of A span three dimensions. If not, then $D(A, B) = \hat{D}(A, B)$ always, and this value can be 0 by choosing $B = A$. Next, translate A so that its centroid is at the origin, and rotate it to match its principle axes of inertia with the coordinate axes. Scale each axis independently so that each axis contributes equally to the radius of gyration, and $R(A) = 1$. Of course in general, B is transformed likewise but with different translation, rotation, and scaling. We derive in the appendix that the threshold value of the scaled $D^2_{sc}(A, B) = 2/3$ which corresponds to a scaled value of $\rho_{sc}(A, B) = \sqrt{4/5} \approx 0.894$. These thresholds are independent of the sizes, overall shapes, and numbers of points in the structures, as long as they span three dimensions.

It is worth mentioning one more landmark on the range of scaled comparisons in three dimensions, namely $D_{sc}(A, \hat{A}) = \hat{D}_{sc}(A, A)$, the similarity of A and its mirror image. Referring to Eq. (30) for scaled comparisons, we can show that $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$ and $S = 1$ for $D_{sc}(A, A) = 0$ because $U = I/3$, where I is the identity matrix. On the other hand, we noted earlier[11] that in general

$$D^2_{sc}(A, \hat{A}) = D^2_{sc}(A, A) + 4S\lambda_3 \quad (17)$$

implying that $D^2_{sc}(A, \hat{A}) = 4/3$. This converts to

$$\rho_{sc}(A, \hat{A}) = \sqrt{2} \quad (18)$$

for any three-dimensional structure A.

RESULTS

ρ and Visual Comparison

Traditionally, $D < 2$ to 3 Å has been used as a criterion for spatial similarity of proteins. It turns out to be a substantial error to use such a fixed cutoff in D over the range of protein sizes commonly studied, particularly for small proteins. Earlier we empirically observed[10] that the most compact globular proteins have radii of gyration

$$R_{min}(n_r) = -1.26 + 2.79n_r^{1/3} \quad (19)$$

in Å, and a typical compact protein structure exceeds this by 10 to 15%. Consider for a moment only those compact globular protein structures satisfying $R = 1.10R_{min}$. Then Table I shows that if we compare two such compact 25-residue structures and find $D = 3$ Å, this corresponds to $\rho = 0.40$ according to Eq. (16). However, this same 3 Å cutoff in D when applied to longer chains lengths amounts to requiring as stringent a similarity as $\rho = 0.19$ for long ones. To put it the other way around, adopting $\rho < 0.40$ as a criterion for obvious spatial similarity implies $D < 2.98$ Å for short chains but only $D < 6.50$ Å for long ones. Figure 4 gives one a visual feeling for this effect. (A) shows two 25-residue fragments having $D = 3.34$ Å and $\rho = 0.463$. There is an obvious but modest similarity. The 200-residue comparison in (B) has the same ρ and a comparable appearance of conformational similarity, but now $D = 7.19$ Å. In order to return to a D comparable to that in (A) while staying with 200 residues, the visual resemblance must be much greater, as in (C) where $D = 3.77$ Å and $\rho = 0.217$.

TABLE I. Comparison of Compact Globular Proteins Having n_r Residues, Giving ρ for Fixed D = 3 Å, D for fixed ρ = 0.40, and $\rho_{sc,\%1}$

n_r	R (Å)*	ρ†	D (Å)‡	$\rho_{sc,1\%}$Σ
25	7.59	0.40	2.98	0.50
50	9.92	0.31	3.89	0.68
100	12.87	0.23	5.04	0.87
200	16.56	0.19	6.50	1.07

*Typical radius of gyration = $1.1R_{min}$ from Eq. (19).
†If the two protein structures differ by $D = 3$ Å, then this is the corresponding ρ.
‡If the two protein structures differ by $\rho = 0.40$, then this is the corresponding D.
ΣCalculated according to Eq. (20) with $a = 0.054$ and $b = 0.581$, the values for protein comparisons.

If $\rho < 0.3$ to 0.5 corresponds to the subjective consensus in the field for clear conformational similarity of globular proteins, then our proposed $\rho < 1$ test must imply only the most general level of similarity. Consider for example $n_r = 30$ and residues 1–30 of metallothionine [Brookhaven Protein Data Bank (PDB)[1] entry 2mhu] compared to residues 51–80 of Trp repressor (1wrp, chain "R"). These segments have very similar radii of gyration (8.18 and 7.86 Å, respectively), and the optimal superposition gives $\rho = 0.54$ ($D = 4.20$ Å, $\rho_{sc} = 0.52$), but the superposition matches α-helical segments with coil segments (see Fig. 5). The similarity between the two becomes obvious, however, when the C^α coordinates are averaged over a 7-residue window sliding up each chain, because this at least straightens out the helical segments.

As a more striking example, residues 1–75 of bovine calcium-binding protein (3icb) form a helical bundle, while residues 9–83 of the FAB immunoglobulin KOL (2fb4, light chain) are β-sheet strands (Fig. 6). Their respective radii of gyration happen to be similar (11.15 and 12.68 Å), and $\rho = 0.76$ ($D = 8.45$ Å, $\rho_{sc} = 0.71$). Figure 7 shows how their smoothed chain traces are obviously similar to the eye. We have found dozens of similar examples in PDB with such long chain segments, great differences in secondary structure, and yet $\rho < 1$. Our experience has been that visual similarity is detectable as long as ρ_{sc} is less than about 0.8, which is

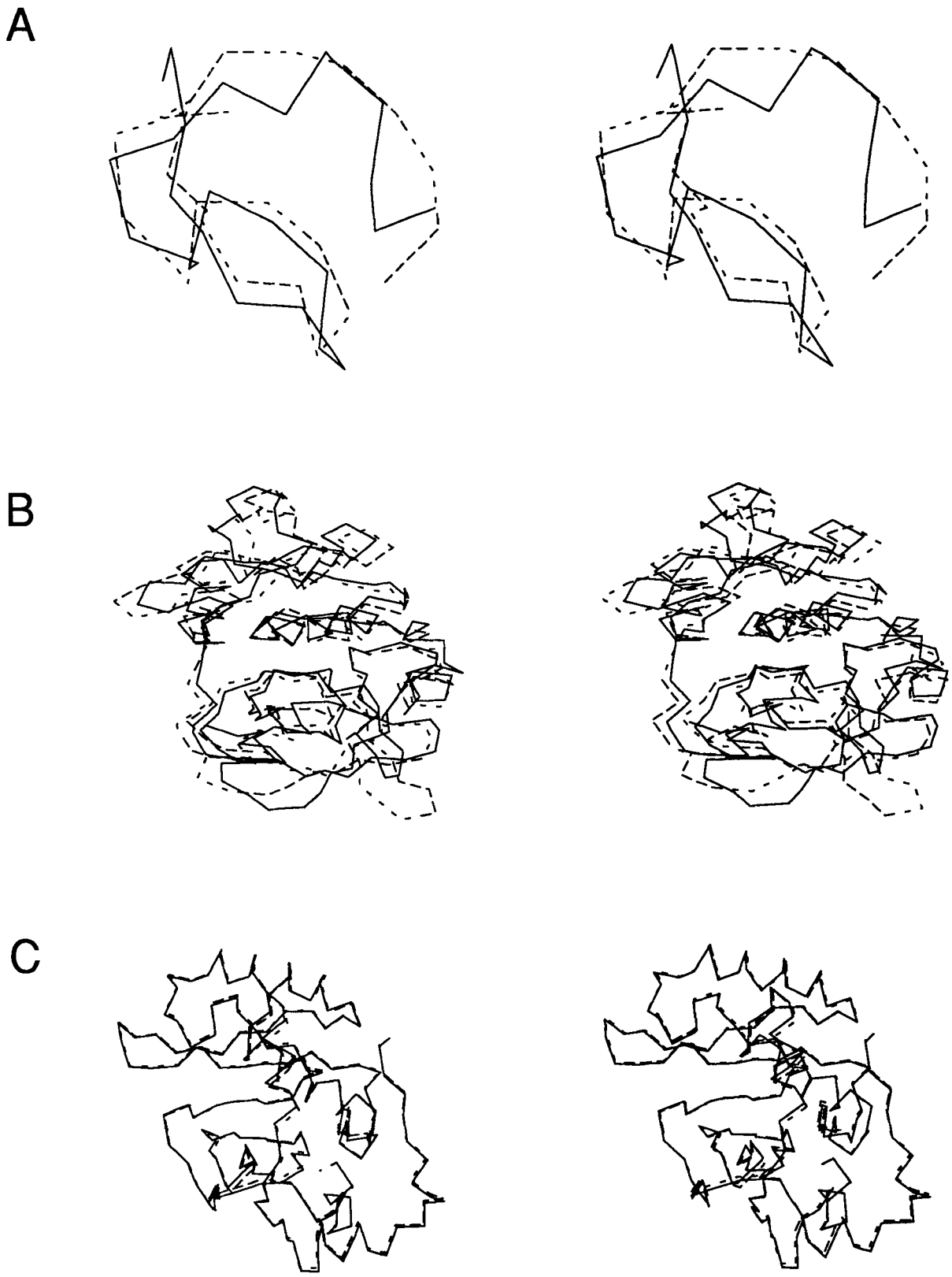

Fig. 4. Stereo pairs of the C^{α} traces for the optimal superposition of protein fragments having different lengths: (**A**) 25 residues from trypsin inhibitor II (PDB code 2eti) residues 1–25 and protein inhibitor 1cti residues 3–27; $D = 3.34$ Å, $\rho = 0.463$; (**B**) 200 residues from papain 9pap residues 5–204 and actinidin 2act residues 6–205; $D = 7.19$ Å, $\rho = 0.463$; (**C**) 200 residues from the same crystal structure of acetyltransferase 2cla: residues 1–200 and 2–201; $D = 3.77$ Å, $\rho = 0.217$.

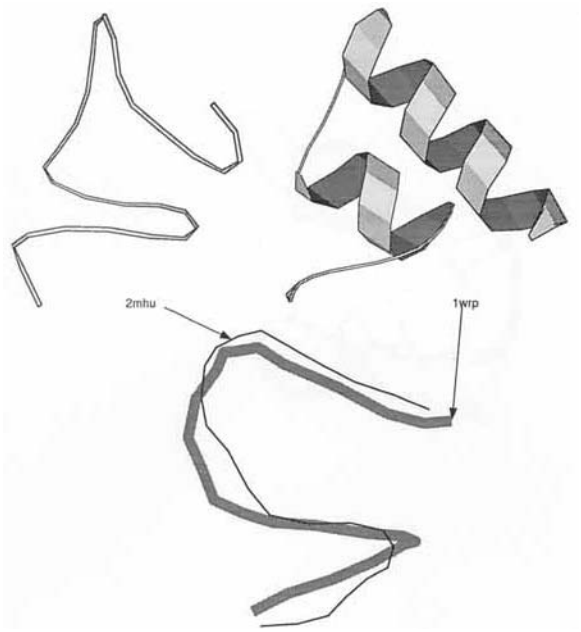

Fig. 5. Comparison of 2mhu residues 1–30 with 1 wrp R chain residues 54–83. Above are the two structures represented as MOLSCRIPT drawings[8] in relative orientations corresponding to their optimal superposition. Below is the same view of the optimal superposition with chain paths smoothed by averaging the C^{α} coordinates over a 7-residue sliding window.

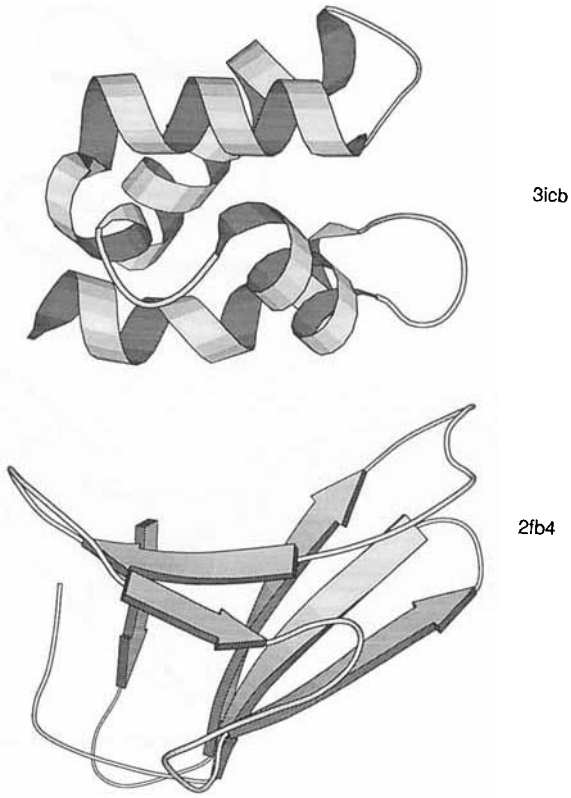

Fig. 6. Comparison of 3icb residues 1–75 with 2fb4 light chain residues 9–83 in their optimal superposition orientations, rendered by MOLSCRIPT.

surprisingly close to the absolute dissimilarity threshold of $\rho_{sc} = 0.894$. Clearly helical bundles and sheets are different folding motifs in the usually accepted sense, but from the broader perspective of $\rho < 1$, there may be a very limited number of different ways to pack a polypeptide chain up into an approximately spherical globule. Whether this is just a geometric restriction on curved lines in three-dimensional space or whether this says something special about protein folding, remains to be seen.

Statistical and Absolute Significance of ρ_{sc}

While absolute cutoffs are valuable to decide similarity vs. dissimilarity between structures, it is also helpful to compare to simple statistical models for a quantitative assessment of the degree of similarity. Consider as a model structure the N^3 points of an $N \times N \times N$ cubic lattice, where the points are labeled at random with some permutation of $1, \ldots, N^3$. Such structures have a cubical shape, a constant radius of gyration, and are self-avoiding in that all points are at least one step apart, but they are not Hamiltonian walks on the lattice as in polymer models. Figure 8 shows the distributions of ρ_{sc} for 100,000 comparisons between such pairs of random structures for $N = 3, 4, 5, 6$, and 7, corresponding to $n = 27, 64, 125, 216$, and 343 points each. Notice how the distributions become narrower and skewed to higher ρ_{sc} as n increases. We examined the left tails of these distributions by determining the cutoff for the first percentile of each, $\rho_{sc,\%1}$, as shown in the upper curve of Figure 9 ($n = 343$ is off scale). These points fit very well to the empirical function

$$\rho_{sc,1\%} = 2 - \frac{2}{1 + a(n-2)^b} \qquad (20)$$

which was chosen so that $\rho_{sc,1\%} \to 2$ as $n \to \infty$ and trivially all $\rho_{sc} = 0$ when $n = 2$. Nonlinear curve fitting gives $a = 0.246$ and $b = 0.593$.

Of course, other populations of structures can be expected to have different ρ_{sc} distributions. For example, comparisons of n-residue segments (counting C^{α} atoms only) taken from the same list of dissimilar protein crystal structures we used earlier[11] have distributions where the $\rho_{sc,\%1}$ fits the same empirical equation, but now with $a = 0.054$ and $b = 0.581$ (Fig. 9, lower curve). Note how the protein comparisons give much the same slope and curvature, but the whole curve is shifted down below that for scrambled cubic structures. Thus for fewer than about 100 residues, $\rho_{sc,\%1}$ is below the absolute dissimilarity cutoff of $\rho_{sc} = 0.894$, meaning that absolute geometric resemblance becomes statistically more likely for smaller protein structures.

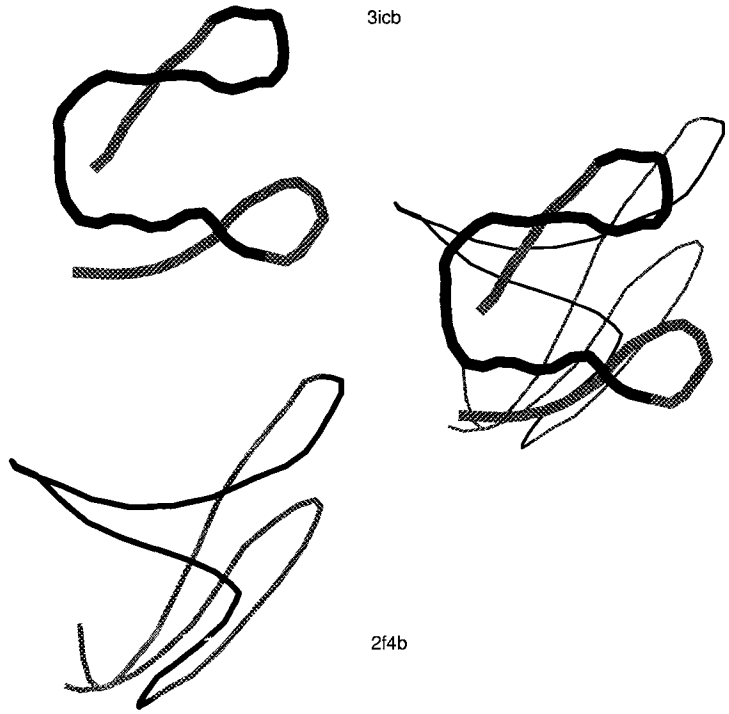

Fig. 7. Comparison of 3icb and 2fb4 in the same orientation as in Figure 6, but rendered as C^{α} paths smoothed over a 7-residue sliding window.

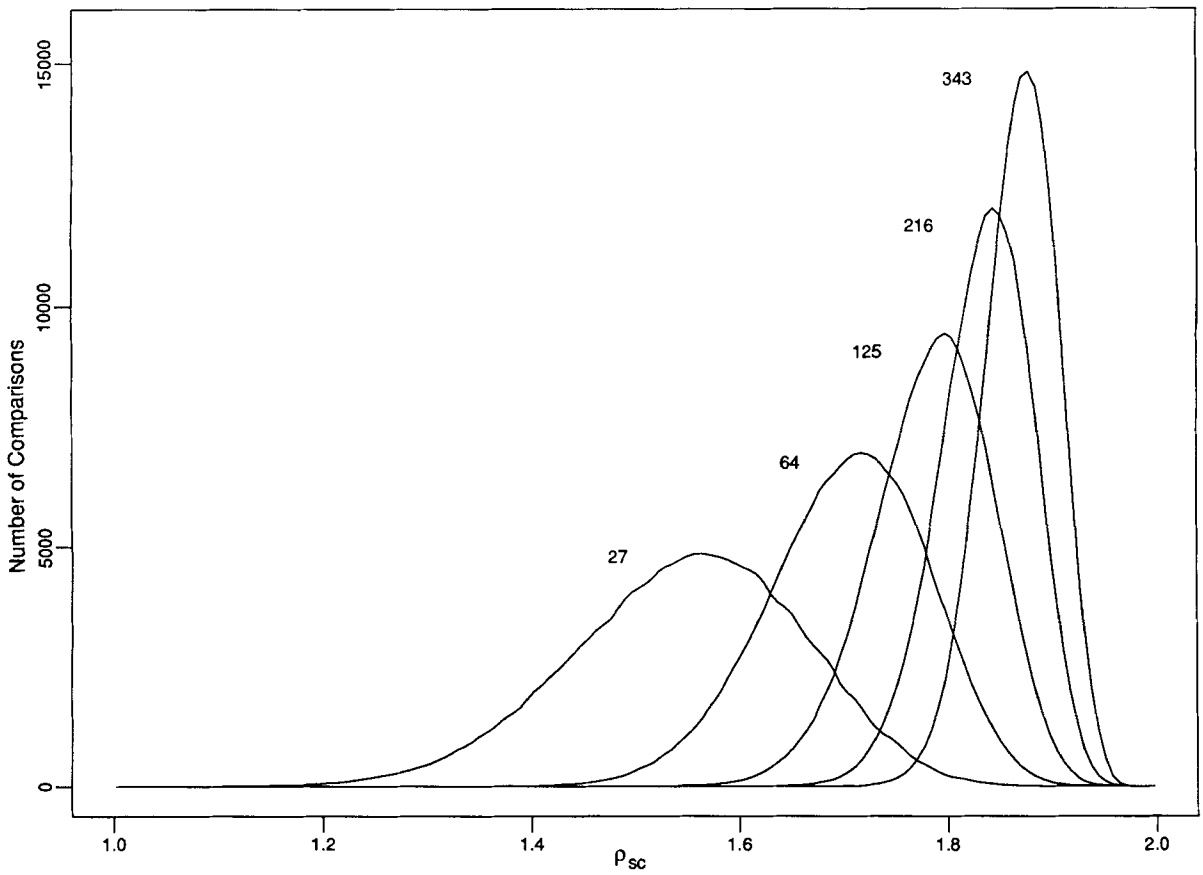

Fig. 8. Distributions of ρ_{sc} as a function of number of points in scrambled cubic lattice structure comparisons.

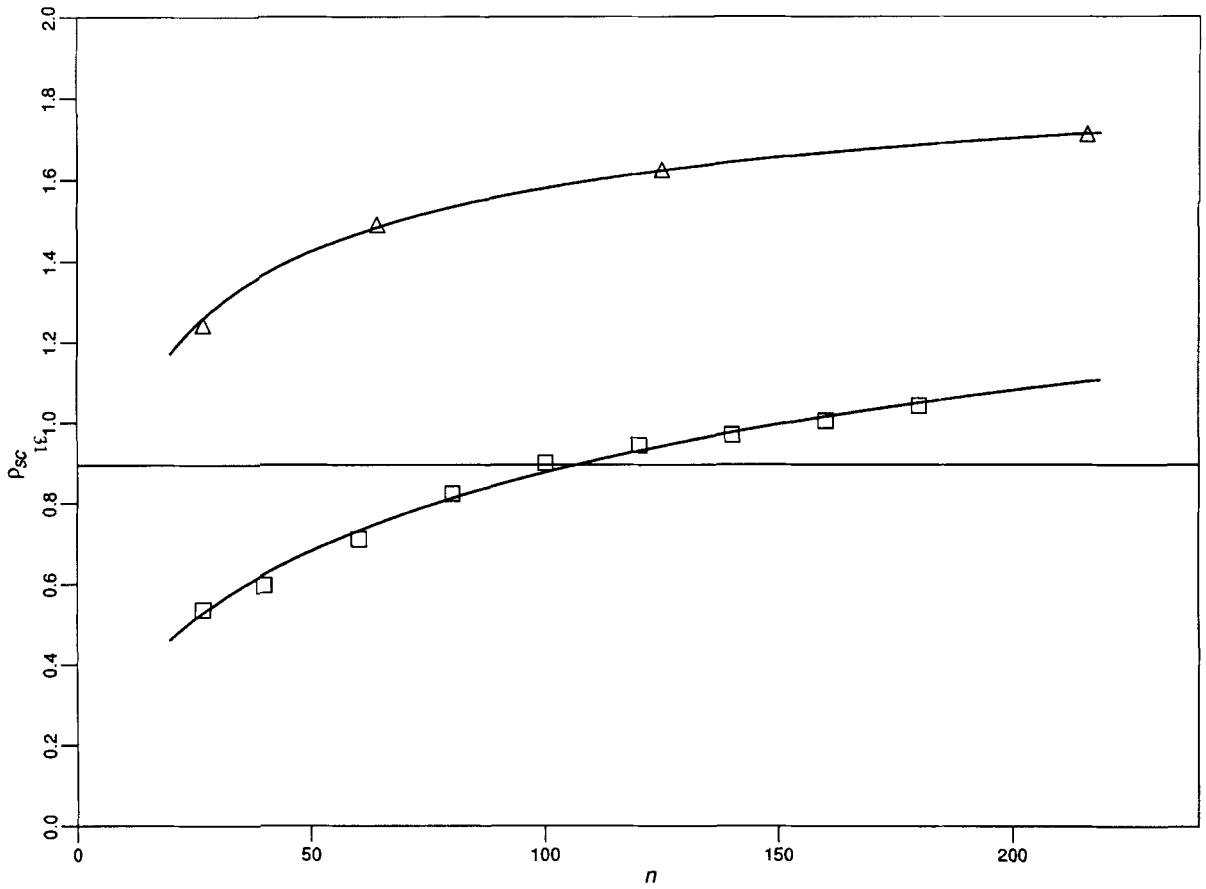

Fig. 9. $\rho_{sc,\%1}$ as a function of number of points, n, for scrambled cubic lattice structures (triangles, upper curve) and dissimilar protein segments (squares, lower curve). Horizontal line is the absolute anticomparison cutoff $\rho_{sc} = 0.894$.

CONCLUSIONS

We propose ρ_{sc} as a measure of similarity derived from the conventional RMSD, but being invariant under more kinds of scaling operations than RMSD. Table II summarizes the significant values of ρ_{sc} we have derived. This should be useful in protein conformational studies to resolve three main kinds of questions: (1) how well is a protein's tertiary structure determined by NMR, judging from the similarity of a set of structures derived from the experimental data[15]; (2) how good are protein structure predictions compared to the experimental conformations; and (3) are low-level conformational similarities between unrelated proteins in the Protein Data Bank significant?[18]

Although ρ_{sc} has some geometrically significant cutoff levels that are invariant under changes in radius of gyration, ellipticity, and numbers of points (or amino acid residues), the statistical significance is a function of n. Thus for small structures of less than about 40 residues, Figure 9 shows there may be a substantial probability of randomly choosing two conformations having geometrically significant similarity, whereas for more than 180 residues, it is very unlikely that geometrically significant similarity can be achieved by chance. This can also be seen in Table I, where even two 50-residue conformations chosen at random from contiguous segments of unrelated proteins have a 1% chance of having their mutual $\rho_{sc} < 0.68$. We believe this will be a useful test to apply when judging the quality of conforma-

TABLE II. Important Values of ρ_{sc} and Their Interpretation

ρ_{sc}	Meaning
0	Identical structures
< 0.3 to 0.5	Visually recognizable similarity
$<(4/5)^{1/2} = 0.894$	Antisimilarity impossible
<1	Structural commonality exceeds difference
$2^{1/2} = 1.414$	Similarity of a structure to its own mirror image
2	Maximally dissimilar

tions determined for small proteins by NMR and theoretical methods.

ACKNOWLEDGMENTS

This work was supported by a grant from the Office of the Vice President for Research of the University of Michigan. We are indebted to all those who deposited their structural data into the Protein Data Bank.

REFERENCES

1. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In "Crystallographic Databases—Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds). Bonn: Data Commission of the International Union of Crystallography, 1987:107–132.
2. Alexandrov, N.I., Takahashi, K., Go, N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. J. Mol. Biol. 225:5–9, 1992.
3. Alexandrov, N.I., Go, N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. Protein Sci. 3:866–875, 1994.
4. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: the significance of the root-mean-square deviation. J. Mol. Biol. 138:321–333, 1980.
5. Crippen, G.M., Havel, T.F. "Distance Geometry and Molecular Conformation." New York: Research Studies Press, Ltd. (Wiley), 1988.
6. Golub, G.H., Van Loan, C.F. "Matrix Computations," 2nd ed. Baltimore: Johns Hopkins University Press, 1989: 582.
7. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
8. Karulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
9. Lesk, A.M. A toolkit for computational molecular biology. II. On the optimal superposition of two sets of coordinates. Acta Cryst. A42:110–113, 1986.
10. Maiorov, V.N., Crippen, G.M. Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227:876–888, 1992.
11. Maiorov, V.N., Crippen, G.M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J. Mol. Biol. 235:625–634, 1994.
12. McLachlan, A.D. Gene duplications in the structural evolution of chymotrypsin. J. Mol. Biol. 128:49–79, 1979.
13. McLachlan, A.D. Rapid comparison of protein structures. Acta Cryst. A38:871–873, 1982.
14. McLachlan, A.D. How alike are the shapes of two random chains? Biopolymers 23:1325–1331, 1984.
15. Pastore, A., Atkinson, R.A., Saudek, V., Williams, R.J.P. Topological mirror images in protein structure computation: An underestimated problem. Proteins 10:22–32, 1991.
16. Remington, S.J., Matthews, B.W. A general method to assess similarity of protein structures with applications to T4 bacteriophage lysozyme. Proc. Nat. Acad. Sci. U.S.A. 75:2180–2184, 1978.
17. Remington, S.J., Matthews, B.W. A systematic approach to the comparison of protein structures. J. Mol. Biol. 140: 77–99, 1980.
18. Smith-Brown, M.J., Kominos, D., Levy, R. Global folding of proteins using a limited number of distance constraints. Prot. Eng. 6:605–614, 1993.
19. Zucker, M., Somorjai, R.L. The alignment of protein structures in three dimensions. Bull. Math. Biol. 51:55–78, 1989.

APPENDIX: THE ANTICOMPARISON THRESHOLD

We want to find the maximal value of D_{sc} over all pairs of structures such that $D_{sc} \leq \hat{D}_{sc}$. It helps to look at the two-dimensional case first, where the derivation is algebraically messy but straightforward to follow. It is convenient to think of the spherically scaled x and y coordinates of A and B and n-dimensional vectors, $\mathbf{a}_x$, $\mathbf{a}_y$, $\mathbf{b}_x$, and $\mathbf{b}_y$. The correlation matrix U between A and B has elements

$$u_{ij} = n^{-1}\mathbf{a}_i \cdot \mathbf{b}_j \qquad \text{for } i, j = x, y. \tag{21}$$

In two dimensions the rotation matrix

$$\mathbf{R} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{22}$$

involves only the single variable θ. Substituting this into Eq. (1), expanding, and solving $\partial D^2/\partial\theta = 0$, gives the optimal superposition value of

$$\theta^* = \arctan\left(\frac{u_{yx} - u_{xy}}{u_{xx} + u_{yy}}\right) \tag{23}$$

which results in an optimal value of

$$D^2_{sc}(A, B) = 2 - 2[(u_{xx} + u_{yy})^2 + (u_{yx} - u_{xy})^2]^{1/2}. \tag{24}$$

Now mirror inversion of A can be done by setting $\mathbf{a}_y = -\mathbf{a}_y$, which changes the sign of u_{yy} and u_{yx}, so that

$$\hat{D}^2_{sc}(A, B) = 2 - 2[(u_{xx} - u_{yy})^2 + (u_{yx} + u_{xy})^2]^{1/2}. \tag{25}$$

Setting $D^2_{sc} = \hat{D}^2_{sc}$ results in

$$0 = u_{xx}u_{yy} - u_{xy}u_{yx} = \det(U). \tag{26}$$

Minimizing D^2_{sc} subject to $\det(U) = 0$ by Lagrange multipliers yields four solutions all giving the same extremal value of $D^2_{sc} = 4u^2_{xx}$. In addition, the coordinates of A and B must be embeddable (viewed as four vectors in n-dimensional space, for $n > 4$), implying that the eigenvalues of the metric matrix

$$n^{-1}\begin{pmatrix} a_x \cdot a_x & a_x \cdot a_y & a_x \cdot b_x & a_x \cdot b_y \\ a_x \cdot a_y & a_y \cdot a_y & a_y \cdot b_x & a_y \cdot b_y \\ a_x \cdot b_x & a_y \cdot b_x & b_x \cdot b_x & b_x \cdot b_y \\ a_x \cdot b_y & a_y \cdot b_y & b_x \cdot b_y & b_y \cdot b_y \end{pmatrix} \tag{27}$$

$$= \begin{pmatrix} 1/2 & 0 & u_{xx} & u_{xy} \\ 0 & 1/2 & u_{yx} & u_{yy} \\ u_{xx} & u_{xy} & 1/2 & 0 \\ u_{yx} & u_{yy} & 0 & 1/2 \end{pmatrix}$$

must be nonnegative.[5] (The diagonal blocks of elements arise from the principal axes and unit radius of gyration conditions on the coordinates.) Combining this condition with each of the four solutions to the optimization, always results in the constraint that $|u_{xx}| < 8^{-1/2}$, so that the maximal antisimilar $D^2_{sc} = 1/2$.

Now in three dimensions, the spherical scaling of A in terms of its n-dimensional coordinate vectors, $\mathbf{a}_x$, $\mathbf{a}_y$, and $\mathbf{a}_z$, implies

$$\mathbf{a}_x \cdot 1 = \mathbf{a}_y \cdot 1 = \mathbf{a}_z \cdot 1 = 0$$

$$\mathbf{a}_x^2 = \mathbf{a}_y^2 = \mathbf{a}_z^2 = 1/3$$

$$\mathbf{a}_x \cdot \mathbf{a}_y = \mathbf{a}_x \cdot \mathbf{a}_z = \mathbf{a}_y \cdot \mathbf{a}_z = 0 \quad (28)$$

where $1 = (1, 1, \ldots)$. In other words, the inertial tensor of A is just 1/3 the unit matrix. The correlation matrix U between A and B is now 3×3, having elements $u_{ij} = n^{-1}\, \mathbf{a}_i \cdot \mathbf{b}_j$ for $i, j = x, y, z$. Following the presentation by McLachlan,[12] the symmetric matrix

$$W = \begin{pmatrix} 0 & U \\ U^T & 0 \end{pmatrix} \quad (29)$$

built out of the unsymmetric U has six eigenvalues in $\pm$ pairs, namely λ_1, λ_2, λ_3, $-\lambda_3$, $-\lambda_2$, and $-\lambda_1$, labeled in terms of the three nonnegative eigenvalues, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$. It can be shown[6,11,12] that

$$D^2(A, B) = R^2(A) + R^2(B) - 2(\lambda_1 + \lambda_2 + S\lambda_3) \quad (30)$$

where $S = \pm 1$ is the sign of det (U). $S = 1$ whenever $D(A, B) > \hat{D}(A, B)$, and $S = -1$ whenever $D(A, B) < \hat{D}(A, B)$, so the condition $D(A, B) = \hat{D}(A, B)$ implies $\lambda_3 = 0$. Furthermore, spherical scaling implies $R(A) = R(B) = 1$. Expressing any eigenvector of W as the concatenation of two three-dimensional vectors, $[\mathbf{e},\mathbf{f}]$, we see that

$$\lambda \begin{pmatrix} \mathbf{e} \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} 0 & U \\ U^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{e} \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} U\mathbf{f} \\ U^T\mathbf{e} \end{pmatrix} \quad (31)$$

so that $\lambda\mathbf{e} = U\mathbf{f}$ and $\lambda\mathbf{f} = U^T\boldsymbol{e}$. From this is follows that any eigenvector of W is also an eigenvector of the 6×6 metric matrix

$$\begin{pmatrix} I/3 & U \\ U^T & I/3 \end{pmatrix} \begin{pmatrix} \mathbf{e} \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} \mathbf{e}/3 + U\mathbf{f} \\ U^T\mathbf{e} + \mathbf{f}/3 \end{pmatrix} = \left(\frac{1}{3} + \lambda\right) \begin{pmatrix} \mathbf{e} \\ \mathbf{f} \end{pmatrix} \quad (32)$$

except that the eigenvalue is now $\lambda + 1/3$. Because the metric matrix must have nonnegative eigenvalues, we have $\lambda \geq -1/3$. When $[\mathbf{e}, \mathbf{f}]$ is the eigenvector corresponding to $\lambda = -\lambda_1$ or $\lambda = -\lambda_2$, at most $\lambda_1 = \lambda_2 = 1/3$. Then Eq. (30) reduces to the threshold value of the scaled $D^2_{sc}(A, B) = 2/3$. By the definition of ρ in Eq. (16), this corresponds to a scaled value of $\rho_{sc}(A, B) = \sqrt{4/5} \approx 0.894$.