Marcos R. Betancourt
Jeffrey Skolnick
Laboratory of Computational Genomics,
The Donald Danforth Plant Science Center,
893 N. Warson Rd.,
Creve Coeur, MO 63141

Received 24 January 2001;
accepted 22 March 2001

Universal Similarity Measure for Comparing Protein Structures

Abstract: *We introduce a new variant of the root mean square distance (RMSD) for comparing protein structures whose range of values is independent of protein size. This new dimensionless measure (relative RMSD, or RRMSD) is zero between identical structures and one between structures that are as globally dissimilar as an average pair of random polypeptides of respective sizes. The RRMSD probability distribution between random polypeptides converges to a universal curve as the chain length increases. The correlation coefficients between aligned random structures are computed as a function of polypeptide size showing two characteristic lengths of 4.7 and 37 residues. These lengths mark the separation between phases of different structural order between native protein fragments. The implications for threading are discussed.* Biopolymers 59: 305–309, 2001

Keywords: *protein structure correlations; relative root mean square distance; universal structure similarity measures; protein folding*

INTRODUCTION

Measures of structural similarity between proteins are a valuable tool for the analysis of protein structures and folding simulations.[1,2] One of the most commonly used measures is the coordinate root mean square distance, or RMSD[3–6] [see Eq. (1)], which describes the mean (RMS) distance per residue between two optimally aligned structures. The RMSD can be computed analytically and has the appealing property that it directly compares the real space coordinates between structures. The latter property allows the RMSD to be more sensitive to global changes in the structure and to differentiate between symmetry-related conformations such as mirror images. This is not necessarily the case for other popular similarity measures, such as the ones based on the comparison of local coordinates (e.g., torsion angles[7]), which are less sensitive to global changes, or based on interresidue distances (e.g., the fraction of native contacts[8]), which do not distinguish between mirror images. These and other measures are convenient for specific applications such as finding common substructures in proteins.[9] However, the RMSD is arguably one of the most discriminating and convenient measures for comparing the global structure of proteins.

While most similarity measures can identify two identical structures, the interpretation of how dissimilar two unequal structures are is strictly measure dependent. In particular, the correlation between the RMSD and other similarity measures is typically high

Correspondence to: Jeffrey Skolnick
Biopolymers, Vol. 59, 305–309 (2001)

as long as the structures are sufficiently similar. However, the correlation diminishes as the structures become significantly different from each other. Furthermore, the RMSD is dependent on the structure size and increases roughly proportionally to the protein's radius of gyration.[10] So, to assess the degree of global structural difference between two structures, one must consider the size of the structures.

Our goal is to define a universal RMSD-related measure that could better differentiate between similar and dissimilar topologies. This requires a benchmark that indicates when two structures are globally uncorrelated. First, we compute an average correlation coefficient by aligning random protein structures of a given length. This allows us to investigate the correlation properties of native structures and to establish the random structure benchmark. From this analysis, we define a universal similarity measure and compute the asymptotic probability distribution for random proteins.

PROTEINS STRUCTURAL CORRELATIONS

We define two structures to be globally uncorrelated when their RMSD value is the same as the average one of two randomly chosen structures of corresponding sizes. More specifically, given two conformations, α and β of N residues, let $\mathbf{r}_{\alpha i}$ and $\mathbf{r}_{\beta i}$ be the respective coordinates of their residues at position i, for 1, . . . , N. Without loss of generality, only one set of coordinates per residue (e.g., the C-α coordinates) is used. If $D_{\alpha\beta}$ is the RMSD between α and β, then

$$D_{\alpha\beta}^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_{\alpha,i} - Q\mathbf{r}_{\beta,i})^2 = R_{g\alpha}^2 + R_{g\beta}^2 - 2\left\langle \frac{\sum_{i=1}^{N}\mathbf{r}_{\alpha,i}\cdot Q\mathbf{r}_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r_{\alpha,i}^2 \sum_{i=1}^{N} r_{\beta,i}^2}} \right\rangle R_{g\alpha}R_{g\beta} \quad (1)$$

where Q is the unitary rotation matrix that optimally aligns the vectors, and $R_{g\alpha}$ and $R_{g\beta}$ are the radius of gyration for structures α and β, respectively. The optimal alignment between α and β requires that their average position coincides, i.e., $\Sigma_{i=1}^{N}\mathbf{r}_{\alpha i} = \Sigma_{i=1}^{N}\mathbf{r}_{\beta i}$. Because the common value of these averages is irrelevant to the problem at hand, the averages were set to zero. This choice of coordinate system allows expressing the radius of gyration for α simply as $R_{g\alpha}^2 = \Sigma_{i=1}^{N} r_{\alpha i}^2$ and similarly for β. The term in parentheses is the correlation coefficient between the aligned structures, which we now refer to as the aligned correlation coefficient (ACC). We are interested in obtaining the average of the ACC for a random ensemble of structures as a function of chain length. This quantity is defined as

$$C(N) \equiv \left\langle \frac{\sum_{i=1}^{N}\mathbf{r}_{\alpha,i}\cdot Q\mathbf{r}_{\beta,i}}{\sqrt{\sum_{i=1}^{N} r_{\alpha,i}^2 \sum_{i=1}^{N} r_{\beta,i}^2}} \right\rangle_{\alpha\beta} \quad (2)$$

where the angle brackets are an average between all pairs of structures of a given length in the random ensemble.

We calculated the average and standard deviation of the ACC from continuous N-residue fragments of almost 1300 nonhomologous random structures (with less than 30% sequence identity), obtained from the Protein Data Bank (PDB). Only one random fragment per structure was selected. The results are plotted in Figure 1. The curve for $C(N)$ asymptotically fits a multiexponential function given approximately by

$$C(N) \approx 0.42 - 0.05(N-1)e^{-(N-1)/4.7} + 0.63e^{-(N-1)/37}. \quad (3)$$

This curve is represented by two characteristic lengths, approximately given by 4.7 and 37 residues. The first one, 4.7, which approximately corresponds to 2.8 nm, is near the typical persistence length for globular proteins.[11] The second one, 37, indicates an aligned correlation length for intermediate size polypeptides. The interpretation of these lengths is

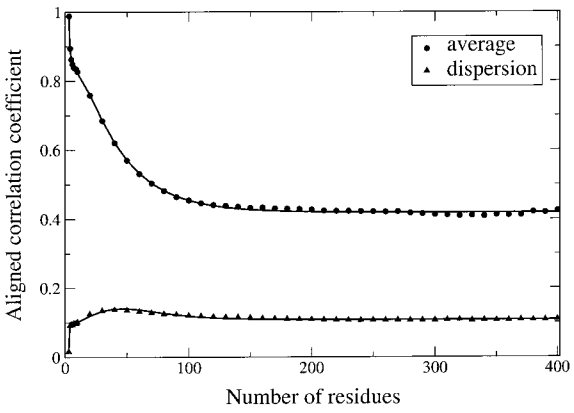

FIGURE 1 Average and dispersion of the ACC for random structures. The data were obtained from various fragments of almost 1300 nonhomologous protein structures.

that small random protein fragments (near or below 5 residues) have relatively stiff and restricted conformations, while intermediate ones (between 4.7 and 37) have significant correlations generated by the recurring types of secondary structure elements. It also suggests that polypeptide chains with more than 37 residues can achieve a wide range of conformations. In the long chain limit, $C(N \to \infty) = c \approx 0.42$. This value is rapidly attained after 100 residues. That this correlation does not decay to zero as the chain length increases is the result of the alignment process. In comparison, $C(N)$ for an ideal chain (random walk) quickly converges to 0.64, indicating that ideal chains are easier to overlap perhaps due to a lack of excluded volume effects. We also computed $C(N)$ for a randomly selected group of 1300 proteins from the PDB and the results were identical to the ones with the nonhomologous proteins.

Figure 1 also shows that the dispersion of the ACC converges to a constant with a value near 0.11. The dispersion has a shallow maximum around 37 residues. This is reminiscent of a mild phase transition, separating a disordered phase for structures with more than 37 residues from a semidisorder phase for structures with less than 37 residues. Below four residues, the dispersion abruptly drops, marking a transition to ordered (or strongly correlated) structures. The probability distribution curve of the ACC converges very close to a Gaussian for large N, but becomes bimodal for small N. Some examples of probability distributions are shown in Figure 2. The lowest ACC value we have obtained is 0.04 for $N = 200$. That is, in contrast to the regular correlation coefficient in which the minimum value is -1 (anticorrelation), aligning

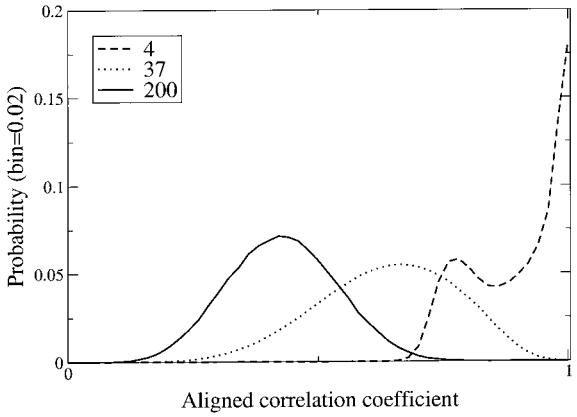

FIGURE 2 Examples of ACC distribution for random polypeptides with different chain lengths. The curves were obtained by comparing all pairs of structures for 4, 37, and 200 residues, separately. The probability was obtained by dividing the RRMSD in bins of 0.02.

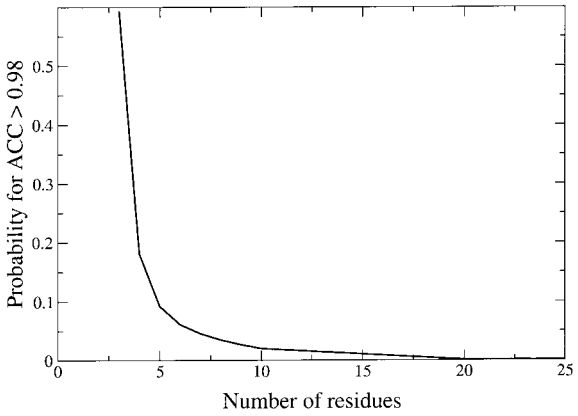

FIGURE 3 Probability that two random protein fragments have an aligned correlation coefficient of more than 0.98 as a function of fragment length.

structures prior to computing their correlation produces in our calculations only positive ACC values. This and similar results for the ideal chain suggest that the minimum value of the ACC (corresponding to the least alignable structures) is zero or near zero.

The structural correlations have implications in the problem of protein structure prediction by threading. Hu and Skolnick[12] investigated the use of sequence fragments to identify a native structure by threading. That is, given a sequence fragment of a known native structure and a library of varied protein structures including the native one, they studied the ability of the threading potential to recognize the native one as a function of fragment length. They found that as the fragment size decreases, the potential loses its ability of recognizing its native structure below 15–20 residues. One reason for this is that the structure of such protein fragments can be significantly stabilized by the interaction with other residues outside the fragment. Our results give an additional reason. In Figure 3 we have plotted the probability that two random fragments have an ACC of more than 0.98 as a function of fragment length. The probability drastically decreases above 4.7 and practically vanishes above 37 residues. For small fragments with less than 20 residues, the probability is greater than 0.1% and rapidly increases as the number of residues decreases. This implies that in a threading library of several thousand fragments of random protein structures there is a good chance of finding two fragments of different proteins with the same structure. Therefore, the ability of detecting the native fold using fragments with less than 20 residues is also hampered by the structural similarity of these fragments in different proteins. In other words, it is difficult to single out a particular structure by only looking at a small fragment, whose

geometry can be found in many structures. Another application of fragment threading is that of predicting the global fold by combining the predicted fragment structures. Our results indicate that the quality of the prediction for fragments with more than 37 residues will significantly diminish, unless the native protein or a related one is in the structure library. On the other hand, if the fragments are too small the potential loses its ability to select the correct structure and global structural information is lost. Therefore, one could argue that for combining structural fragments obtained by threading, the minimum fragment size should be somewhere near 37 residues.

The range of chain lengths in our calculations is limited to 400 residues because as the chain length increases further we begin to observe significant fluctuations in the results due to poor statistics. It is uncertain whether the limiting shape of the probability distribution [and consequently the limiting value of Eq. (3)] remains invariant far beyond 400 residues. Because proteins naturally arrange themselves in densely packed domains of perhaps 200 or 300 residues, it is possible that the limiting value of the ACC could decrease for proteins with many domains. This is because it should be generally more difficult to align multidomain proteins as compared, for example, to two domain proteins.

UNIVERSAL SIMILARITY MEASURE

With $C(N)$, we can estimate the average RMSD between two random polypeptides of size $R_{g\alpha}$ and $R_{g\beta}$ as

$$\bar{D}^2_{\alpha\beta} = R^2_{g\alpha} + R^2_{g\beta} - 2C(N)R_{g\alpha}R_{g\beta}. \qquad (4)$$

The ratio between the RMSD and Eq. (4) gives a measure with the desired universal properties. Therefore, we define the relative RMSD (RRMSD) as

$$d_{\alpha\beta} \equiv \frac{D_{\alpha\beta}}{\bar{D}_{\alpha\beta}} \qquad (5)$$

That is, $d_{\alpha\beta}$ is the RMSD relative to the approximate average RMSD between two random protein fragments, with sizes equivalent to the proteins being compared. We have verified that the average of this metric is highly independent of protein size when averaged over the random protein–fragment ensemble and that its value is approximately 1.00. In addition, the dispersion of RRMSD values for random structures approaches a value near 0.09. Assuming that the minimum ACC is 0.0, the RRMSD between two large structures ($N > 100$) ranges from 0.0 to a maximum of near 1.31, achieved for structures with an equal radius of gyration but a very different distribution of residues. Several RRMSD probability distributions, for chains with 200 residues or more, are shown in Figure 4. The limiting distribution is Gaussian-like, with a slower decaying tail for values smaller than one.

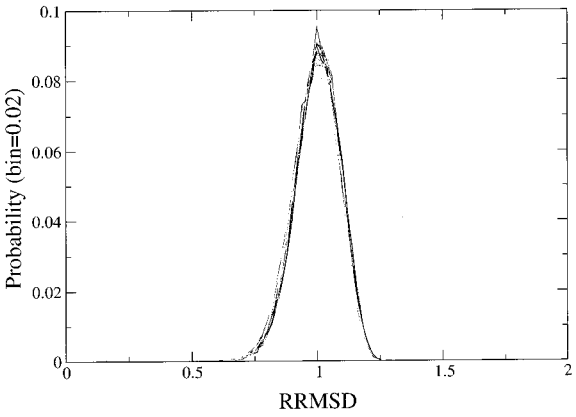

FIGURE 4 Probability distributions for the RRMSD of random structure fragments between 200 and 400 residues (in steps of 10). The curves converge to a universal Gaussian-like curve with an average of 1.0 and a standard deviation of 0.09.

A simpler version of the RRMSD is obtained by using the limiting value of $C(N)$ for large N, that is, c. This version varies significantly from Eq. (5) only for $N \leq 100$. In the large N limit, the correlation between $d_{\alpha\beta}$ and $D_{\alpha\beta}$ is strong but may decrease if the difference between the size of structures being compared is significant. To see this, let $R_{g\beta} = \gamma R_{g\alpha}$, for some arbitrary $\gamma \leq 1$. Also, to be able to compare the RMSD and the RRMSD, we define a dimensionless RMSD as $D_{\alpha\beta}/R_{g\alpha}$. The ratio between $(D_{\alpha\beta}/R_{g\alpha})^2$ and $d^2_{\alpha\beta}$, namely $(\bar{D}_{\alpha\beta}/R_{g\alpha})^2$, has a minimum at $\gamma = c = 0.42$, is one for $\gamma = 2c = 0.84$, and is maximum for $\gamma = 1$. This means that around $\gamma = 0.42$, 0.84, and 1.00, the RRMSD can be up to 10% higher, equal, or near 8% lower than the dimensionless RMSD, respectively. Compared to the dimensionless RMSD, the RRMSD differentiates more between structures with significant size differences and less between similar size structures. Therefore, there is some decorrelation between the RMSD and the RRMSD as their values increase.

The use of random protein fragments as the reference ensemble includes not so dense structures as well as dense ones. Is spite of this, a calculation of the RMS radius of gyration for this ensemble is fitted by

the power law $\langle R_g \rangle \approx 3.08N^{1/3}$, which is comparable to that of dense globular proteins.[10] For comparison, we computed a few RRMSD distributions for a random ensemble of whole protein structures, which on the average are somewhat denser than the protein fragments. For this ensemble the statistics are poorer due to a smaller number of nonhomologous proteins. The calculations show a small but noticeable narrowing and shift of the distributions toward higher RRMSD values, with the average increasing to near 1.05, in particular for short ($N < 100$) chains. This implies (as we have verified) that the structural correlation of a random ensemble of whole proteins is slightly smaller than that of a random ensemble of protein fragments with the same length. This loss of correlation may be because a chain that is a fragment of a larger protein can be in an extended conformation with a more regular structure. In contrast, when the same chain consists of the entire globular protein (not a fragment of it), its structure is frequently bent for compactness, leading to slightly more irregular shapes.

DISCUSSION

We have studied the structural correlations of diverse monomeric proteins after optimal alignment and defined a universal similarity measure for comparing protein structures. The aligned correlation coefficient of nonhomologous structures has revealed two characteristic lengths of effectively 4.7 and 37 residues. These quantities represent the boundaries between three phases of structural order. In particular, protein fragments of more than 37 residues are significantly decorrelated and reach almost full decorrelation between 100 to 150 residues. This suggests that to adopt a wide variety of native conformations, proteins should preferably consist of more than 37 residues. The number of conformations for proteins shorter than 37 residues is much more limited and a degeneracy of sequences for each native structure is likely.

The RRMSD is a convenient universal measure to determine when the similarity between two structures is more significant than random. In particular, a RRMSD value of 0.0 means identical structures and 1.0 means structures as different as random structures on average. This avoids ambiguities resulting from the dependence of the RMSD on protein size. For RRMSD values between zero and one, the standard deviation allows us to classify the similarity between structures. For proteins with more than 100 residues, the RRMSD probability distribution becomes Gaussian-like with a standard deviation near 0.11. In particular, the probability of finding two arbitrary structures with an RRMSD below 0.5 (corresponding to about five standard deviations below the mean) is approximately 10^{-6} for a Gaussian distribution. This probability is somewhat higher because the probability is not truly Gaussian. From the actual probability distribution the probability of finding two arbitrary structures with an RRMSD below 0.5 is closer to 10^{-5}. For smaller chain lengths, the RRMSD distribution deviates significantly from Gaussian, and the actual distribution or dispersion must be used to better estimate the significance of the RRMSD between two structures. The probability for two arbitrary chains with 20 residues having a RRMSD of less than 0.5 increases to near 10^{-3}.

The RRMSD helps to quantify the quality of a protein structure prediction by using a more uniform scale. In ab initio folding predictions, the RRMSD distribution of predicted structures can be examined to determine if the structures deviate significantly from random. It can also be used to define a general criterion for which a group of structures can be considered to belong to the same or different family of structures. The latter application is useful in structure clustering, where different structures are clustered in groups of significantly distinct topologies.

This research was supported in part by NIH grant No. GM-37408 of the Division of General Medical Sciences.

REFERENCES

1. Brown, N. P.; Orengo, C. A.; Taylor, W. R. Comput Chem 1996, 20, 359–380.
2. Lackner, P.; Koppensteiner, W. A.; Domingues, F. S.; Sippl, M. J. Proteins Suppl 1999, 37, 7–14.
3. McLachlan, A. D. Acta Cryst A 1972, 28, 656–657.
4. Kabasch, W. Acta Cryst A 1976, 32, 922–923.
5. Kabasch, W. Acta Cryst A 1978, 34, 827–828.
6. McLachlan, A. D. J Mol Biol 1979, 128, 49–79.
7. Taylor, W. R., Ed. Patterns in Protein Sequence and Structure; Springer Series in Biophysics; Springer Verlag, New York, 1992; Vol 7.
8. Holm, L.; Sandler, C. J Mol Biol 1993, 233, 123–138.
9. Kedem, K.; Chew, L. P.; Elber, R. Proteins 1999, 37, 554–564.
10. Reva, B. A.; Finkelstein, A. V.; Skolnick, J. Fold Des 1998, 3, 141–147.
11. Damschum, G.; Damschum, H.; Gast, K.; Zirwer, D. Biochemistry (Moscow) 1998, 63, 259–275.
12. Hu, W. P.; Skolnick, J. Protein Eng 1997, 10, 317–331.