Statistical validation of the root-mean-square-distance, a measure of protein structural proximity

Oliviero Carugo1,2,3
1Department of General Chemistry, University of Pavia, Pavia, Italy
2Department of Biomolecular Structural Chemistry, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria
3To whom correspondence should be addressed.
Email: oliviero.carugo@univie.ac.at

Despite its well-documented limitations, the root-mean-square-distance (rmsd) between pairs of equivalent atoms is routinely used to monitor the degree of similarity between two optimally superposed protein three-dimensional structures. A robust method for assessing the statistical significance of the difference between two rmsd values is presented here. It is based on the comparison of two protein structures through the correlation coefficient between equivalent inter-atomic distances and the subsequent application of the Fisher transformation that allows one to estimate the probability of identity between two correlation coefficient values. The relationship between the rmsd and Fisher correlation coefficient allows then to estimate the statistical significance of the difference between two rmsd values. Such a procedure is exemplified with the analysis of the possible classifications of the immunoglobulin-like domains of filamin and is compared to related estimations of structural similarity. The possibility to estimate the probability of the difference between two rmsd values can be used to optimize the protein structural classifications and comparisons, independent of the procedure used to derive the rmsds.

Keywords: filamin rod domain/protein structural similarity/protein structure classification/root-mean-square-distance

Introduction

The similarity between two protein three-dimensional structures is usually measured through the root-mean-square-distance (rmsd) between pairs of equivalent Cα atoms, computed after optimal superposition of the two structures. This is done either when two conformations of the same protein (bound/unbound, monomeric/oligomeric, etc.) are compared or when the comparison involves two different proteins that have different amino acid sequences, though compared or when the comparison involves two different protein (bound/unbound, monomeric/oligomeric, etc.) are compared. The use of the rmsd might seem quite peculiar, since many other similarity measures were proposed and used in various fields of structural and molecular biology (Carugo and Eisenhaber, 1997; Carugo and Pongor, 2002a, 2002b; Carugo, 2006). Such an abundance of alternatives to rmsd is largely due to the fact that the rmsd values are known to have several major drawbacks. One is obvious. Superpositions and the resulting rmsd values may not be the best way to compare two protein structures, especially if a small perturbation in just one part of a protein (e.g. a hinge between two domains) can create large rmsd values, suggesting that the two structures are very different, despite the fact that they are not. In other words, the local structural similarity may be much greater than the overall similarity. Such a problem may be solved, at least partially, by subdividing the structures into single domains.

Moreover, the rmsd does not behave as a metric, in the mathematical sense, unless the structures that are compared are very similar to each other (Maiorov and Crippen, 1995; Betancourt and Skolnick, 2001). For this reason, the rmsd_100 was proposed (Carugo and Pongor, 2001), which is the rmsd value that would be measured if the structures that are compared contained 100 residues. It was also observed that the rmsd values depend on the accuracy of the experimentally determined protein structures (Carugo, 2003). On average, smaller rmsd values are observed for protein structure pairs at better resolution and the rmsd values tend to increase if the two proteins that are compared were refined at different resolutions.

These drawbacks make it difficult to use rmsd values when several protein three-dimensional structures are compared in order to classify them through multivariate statistical techniques of machine learning algorithms. Nevertheless, it is still very common for structural biologists to use and publish rmsd values.

An open question about rmsd is the statistical significance of rmsd differences. In other words, if the two protein structures A and B (independent of their sequence similarity) are associated with an rmsd value equal to \( r_{AB} \) and the two protein structures C and D (independent of their sequence similarity) are associated with an rmsd value equal to \( r_{CD} \), and if \( r_{AB} < r_{CD} \), it is possible to affirm that the similarity between A and B is greater than the similarity between C and D. It is nevertheless impossible to estimate the probability with which the absolute value of the difference \( r_{AB} - r_{CD} \) is different from 0. In other words, it is impossible, on the basis of the rmsd values, to estimate the statistical significance of the value of \( |r_{AB} - r_{CD}| \) and the probability that this difference is different from the alternative hypothesis that \( r_{AB} = r_{CD} \).

In the present article, we present a procedure that allows one to estimate the statistical significance of the differences between pairs of rmsd values. This is clearly of fundamental importance in all the circumstances in which protein three-dimensional structures are compared in order to extract any biological information, such as, for example, structure–function relationships or evolutionary pathways.

Methods

Rmsd values

Rigid body superpositions were performed on the Cα atoms with the method of Kabsch (1976, 1978). Rmsd values were
computed as
\[
\text{rmsd} = \sqrt{\frac{\sum d_i^2}{n}},
\]
(1)

where \(n\) is the number of pairs of equivalent Ca atoms and \(d_i\) is the Euclidean distance between the two Ca atoms of the \(i\)th pair. The rmsd values were standardized as rmsd_100 (Carugo and Pongor, 2001)
\[
\text{rmsd}_100 = \frac{\text{rmsd}}{1 + \ln \sqrt{n/100}},
\]
(2)

where \(n\) is the number of residues in the proteins that are compared. Rmsd_100 values are the rmsd that would be measured if the structures that are compared contained 100 residues.

**r_F values**

A protein structure containing \(N\) residues can be described with the \(n = N(N - 1)/2\) unique distances between Ca atoms and two three-dimensional models of the same protein can thus be described by two vectors \(X = (x_1, x_2, \ldots, x_n)\) and \(Y = (y_1, y_2, \ldots, y_n)\), containing \(n\) elements and where each \(i\)th element of \(X\) is equivalent to the \(i\)th element of \(Y\). The comparison between two protein three-dimensional structures can therefore be performed by comparing the \(X\) and \(Y\) vectors. This can be performed in a wide variety of ways (Theodoridis and Koutroubas, 2003), for example by computing the Pearson correlation coefficient \(r_p\), defined as
\[
\text{\(r_p\) = \frac{\sum (x - \langle x \rangle)(y - \langle y \rangle)}{\sqrt{\sum (x - \langle x \rangle)^2 \sum (y - \langle y \rangle)^2}},
\]
(3)

where \(\langle x \rangle\) (or \(\langle y \rangle\)) is defined as
\[
\langle x \rangle = \frac{\sum x_i}{n}.
\]
(4)

Robust statistical methods allow one to estimate the probability with which a certain \(r_p\) value is different from 0. In contrast, the comparison between two \(r_p\) values cannot be performed in a statistically robust way, since the \(r_p\) values are bound between \(-1\) and \(+1\) and therefore the \(r_p\) distribution can be symmetrical only around 0 (Dowdy et al., 2004). The comparison between two correlation coefficients is possible if the Pearson correlation coefficient is transformed into the Fisher one, defined as
\[
\text{\(r_F\) = \ln \sqrt{\frac{1 + r_p}{1 - r_p}}},
\]
(5)

In this way, in fact, the comparison between two values \(r_F\) and \(r_F'\) is possible, independent of their values, by using a \(Z\)-test defined as
\[
\text{\(Z\)-test = \frac{|r_F - r_F'|}{\sqrt{2/(n - 3)}}},
\]
(6)

which is normally distributed and can lead to the probability that \(r_F \neq r_F'\).

**Modeling computations**

Protein three-dimensional models were generated by homology modeling, a methodology that allows the prediction of the structure of a target protein on the basis of the similarity between its amino acidic sequence and the sequence of a template protein, the three-dimensional structure of which is known.

We used five target sequences, taken from the CATH database of protein structural domains (Orengo et al., 1997) and classified into different Homologous Superfamily clusters. Two hundred template proteins were randomly selected to generate 200 models of each target, among the CATH domains clustered in the same Homologous Superfamily group of the target.

The MODELLER suite of programs was used in the default mode. Although this might not be recommended to produce reliable computational models, especially when the sequence similarity between target and template is very low, this does not affect the results presented in the present article, since here models are necessary only to compute rmsd_100 and \(r_F\) values for a large number of protein pairs.

Since five targets were used and \(5 \times 200 = 1000\) proteins three-dimensional models were generated, \(5 \times 19,900 = 99,500\) unique pairs of structures were compared.

**Results**

The statistical validation of the difference between two rmsd_100 values (equation (2)) was performed by exploiting the relationship between the rmsd_100 values and the Fisher correlation coefficient values \(r_F\) (equation (5)), since the latter may be compared through standard and robust statistical techniques (see equation (6) in the Methods section for pertinent details). Rmsd_100 values were computed after optimal superposition of equivalent pairs of Ca atoms (external co-ordinates) and \(r_F\) values were computed on all pairs of equivalent Ca–Ca distances (internal co-ordinates). The use of different types of co-ordinates is justified by the fact that both are invariant relative to the orientation and position of the pair of protein structures that are compared.

The relationship between rmsd_100 and \(r_F\) values was determined by using a large set of pairs of protein three-dimensional structures. They were generated by homology modeling procedures with the computer program MODELLER (Marti-Renom et al., 2000) (see Methods section for pertinent details). Five protein structural domains were used as sequence targets and 200 computational models were generated for each of them by using 200 structural templates. The five targets were the domains 1a30A0, 1a3z00, 1a10G0, 1a5300 and 1a3h00 of the CATH database (Orengo et al., 1997), where they are classified in different fold types. Each structural template was randomly selected among the protein domains that are classified together with the template in CATH (at the classification level termed ‘Homologous Superfamily’). Since 200 models of each target protein were generated, 19,900 unique pairs of models were available. Since five different targets were used, this resulted in

34
Validation of rmsd

Fig. 1. Dependence of the $r_F$ on the rmsd_100 values.

99 500 pairs. The resulting rmsd_100 values were highly variable, with only 25% of the observations being associated with rmsd_100 < 2 Å.

Figure 1 shows the dependence of the $r_F$ values on the rmsd_100 values. It can be optimally described by the relationship

$$r_F = 0.148 + 4.233e^{-0.179 \text{ rmsd}_{100}}$$

(correlation coefficient = 0.995). It is known that two $r_F$ values can be compared with a Z-test, defined in equation (6) (see Methods section). Given the relationship between $r_F$ and rmsd_100, it is possible to re-write the Z-test as a function of the rmsd_100 values

$$Z = \frac{4.233(e^{-0.179 \text{ rmsd}_{100}} - e^{-0.179 \text{ rmsd}_{100}'})}{\sqrt{2/(n - 3)}}.$$  

(8)

Given that the probability $P$ can be associated with a Z-test value (Dowdy et al., 2004), it is therefore possible to determine the relationship between the $P$ values and the differences $\Delta \text{rmsd}_{100}$ defined as

$$\Delta \text{rmsd} = |\text{rmsd}_{100} - \text{rmsd}_{100}'|. \quad (9)$$

Figure 2 shows the relationships between $P$ and $\Delta \text{rmsd}_{100}$, which can be optimally fitted as

$$P = 1.270 \cdot \Delta \text{rmsd}_{100}^{-0.009} - 0.089$$

$$\cdot \Delta \text{rmsd}_{100}^{0.365} - 0.213 \cdot \Delta \text{rmsd}_{100}^{-0.524} \quad (10)$$

(correlation coefficient = 0.996). It can be seen that only if the two rmsd_100 values differ by at least 1.5 Å there is probability equal to 99% that the difference is not casual. The probability increases to 99.9% if the two rmsd_100 values reach 1.95 Å. Of course, the use of threshold $P$ values over which the difference $\Delta \text{rmsd}_{100}$ is considered statistically significant is rather arbitrary since various and different threshold values may be selected by different scientists.

Nevertheless, the relationship above is a continuous function that allows the computation of any $P$ value given its corresponding $\Delta \text{rmsd}_{100}$ value.

Discussion

As an example of validation of the rmsd values, the classification of the filamin structural domains is reported here.

Human filamin, which is expressed in three very similar isoforms (a, b and c), contributes to the organization of the actin-based cytoskeleton. It contains two actin-binding calponin homology domains at the N-terminus followed by a long, flexible rod region made up of 24 immunoglobulin-like (Ig) domains. A similar rod containing only six domains is observed in Dictyostelium discoideum gelation factor. This rod region is responsible for several inter-molecular interactions with other cytoskeletal proteins and with various trans-membrane and cytoplasmic cell-signalling proteins (Gorlin et al., 1990; Stossel et al., 2001). The crystal structure of several of these domains was determined experimentally (Table I).

The classification of these Ig domains on the basis of their tertiary structures can be performed only if the proximity between each pair of domains is determined. This is possible, for example, by an all-against-all superposition. This was performed with combinatorial extension (CE) (Shindyalov and Bourne, 1998) and the resulting rmsd values were standardized to rmsd_100 (Carugo and Pongor, 2001) (see Table II).

A subsequent cluster analysis was performed with the 'neighbor' utility of the Phylip suite of programs (nearest neighbor criterion of similarity) and it resulted into the tree shown in Fig. 3. Since this is typical hierarchical agglomerative cluster analysis, the decision of which is the best number of partitions is rather ambiguous (Theodoridis and Koutrombas, 2003). Among the various possibilities, a reasonable partition (referred to as P_1) might be the following: domains 4 and 5 of gelation factor form a cluster (D1–D6), domains 6 of gelation factor form a second cluster (D12–D15) and a third cluster contains the domains 17, 21 and 24 of human filamin (D7–D10). Another reasonable partition (referred to as P_2) would move domain 24 of human filamin c (D7) into the cluster containing domains 4 and 5 of gelation factor (D1–D6). Both partitions contain three clusters, the only difference being the classification of domain
and the probability of the intra-cluster probability is 

since the probability of the inter-cluster distance is the same not a significant difference between partitions $P_1$ and $P_2$.

Different clusters. It can, therefore, be concluded that there is the same group and it is 0.88(1) if the proteins are grouped into $P$ probability, according to equation (10), which indicates its statistical significance. For partition $P_1$, the probability $P$ assumes the average value of 0.72(2) if the domains are clustered together and of 0.88(1) if they are segregated into different clusters. For partition $P_2$, the average value of the probability $P$ is 0.67(2) if the domains are clustered into the same group and it is 0.88(1) if the proteins are grouped into different clusters. It can, therefore, be concluded that there is not a significant difference between partitions $P_1$ and $P_2$, since the probability of the intra-cluster distance is the same [0.88(1)] and the probability of the intra-cluster probability is very similar (0.72(2) for $P_1$ and 0.67(2) for $P_2$). However, given that the statistical significance of the $\Delta$rmsd values computed for the domains that are classified into the same cluster is slightly higher in the case of partition $P_1$ (0.72(2)), it seems reasonable to prefer the other partition ($P_2$), where the probability that similarly classified domains are different is slightly lower (0.67(2)).

The example shown above is only a rather crude example of the use of the statistical validation of the rmsd values proposed in the present article. It must be remembered that in the case of a hierarchical agglomerative clustering other criteria for determining the optimal partition can be used (Theodoridis and Koutroumbas, 2003). The probability values $P$ associated with each pair of rmsd values can nevertheless be used in any other cluster analysis step and can thus provide a better strategy to compare and classify protein tertiary structures. Furthermore, the $P$ values computed with equation (10) were compared with similar measures of structural similarity provided by CE (Shindyalov and Bourne, 1998), a powerful and widely used algorithm for superposing pairs of protein three-dimensional structures. Among the results that CE offers,

D7, which is grouped with domains D8–D11 in partition $P_1$ and which is clustered with domains D1–D6 in partition $P_2$.

To decide which of these alternative partitions is better than the other, it is possible to compute the $\Delta$rmsd values for two types of protein structures. On the one hand, it is possible to calculate the values of $|\text{rmsd}_{100}(i, j) − \text{rmsd}_{100}(i, k)|$ in the cases in which the structures $i, j$ and $k$ are classified into the same cluster and, on the other, the values of $\Delta$rmsd can be computed in the cases in which the protein domains are classified into different clusters. Both types of $\Delta$rmsd values can be associated with a probability, according to equation (10), which indicates its statistical significance. For partition $P_1$, the probability $P$ assumes the average value of 0.72(2) if the domains are clustered together and of 0.88(1) if they are segregated into different clusters. For partition $P_2$, the average value of the probability $P$ is 0.67(2) if the domains are clustered into the same group and it is 0.88(1) if the proteins are grouped into different clusters. It can, therefore, be concluded that there is not a significant difference between partitions $P_1$ and $P_2$, since the probability of the intra-cluster distance is the same [0.88(1)] and the probability of the intra-cluster probability is very similar (0.72(2) for $P_1$ and 0.67(2) for $P_2$). However, given that the statistical significance of the $\Delta$rmsd values computed for the domains that are classified into the same cluster is slightly higher in the case of partition $P_1$ (0.72(2)), it seems reasonable to prefer the other partition ($P_2$), where the probability that similarly classified domains are different is slightly lower (0.67(2)).

The example shown above is only a rather crude example of the use of the statistical validation of the rmsd values proposed in the present article. It must be remembered that in the case of a hierarchical agglomerative clustering other criteria for determining the optimal partition can be used (Theodoridis and Koutroumbas, 2003). The probability values $P$ associated with each pair of rmsd values can nevertheless be used in any other cluster analysis step and can thus provide a better strategy to compare and classify protein tertiary structures. Furthermore, the $P$ values computed with equation (10) were compared with similar measures of structural similarity provided by CE (Shindyalov and Bourne, 1998), a powerful and widely used algorithm for superposing pairs of protein three-dimensional structures. Among the results that CE offers,
three are important here, the number of aligned residues (n_ali), which is the number of equivalenced atoms, the rmsd, and the Z-score, which is a measure of probability that the rmsd values did not occur by chance. If the two compared protein structures are identical, n_ali is equal to the number of residues that they contain, rmsd is equal to zero, and Z reaches high values, close to 7. In contrast, n_ali decreases, rmsd increases and Z decreases if the two structures diverge. Z values lower than 3.5–3.8 suggest that there is no significant difference between two protein tertiary structures, though such values are often considered to be quite low and are therefore associated with some relationship between the protein structures that are compared.

It is nevertheless important to outline that the statistical appreciation of the rmsd variations is of fundamental importance in order to make a correct use of this very commonly used measure of structural similarity.

**Acknowledgement**

The Bioinformatics Integration Network II (GAN-AU, Austria) is gratefully acknowledged.

**References**


**Validation of rmsd**

A statistically robust method to estimate the probability with which two rmsd values are different is described in the present communication.

It must be observed that the probability values P (equation (10)) are strictly geometrical, in the sense that they monitor only geometrical features like the positional vectors of the protein atoms. They are thus totally independent on the fact that proteins are linear polymer with strong constraints due to the fact that the distances between adjacent residues cannot assume any real value. It is therefore not surprising that rmsd values of ~2 Å are associated with a statistically significant difference between two protein tertiary structures, though such values are often considered to be quite low and are therefore associated with some relationship between the protein structures that are compared.

Fig. 3. Classification of the immunoglobulin-like domains of filamin.

Received July 19, 2006; revised October 12, 2006; accepted November 03, 2006

Edited by Joel Sussman