J. Appl. Cryst. (1997). **30**, 547–549

Probabilistic evaluation of similarity between pairs of three-dimensional protein structures utilizing temperature factors

OLIVIERO CARUGO[a,b] AND FRANK EISENHABER[a,c]* *at* [a]*European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012, Heidelberg, Germany,* [b]*Dipartimento di Chimica Generale, Università di Pavia, via Taramelli 12, I-27100 Pavia, Italy, and* [c]*Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle Strasse 10, 13122 Berlin-Buch, Germany. E-mail: eisenhaber@embl-heidelberg.de*

(*Received 9 January 1997; accepted 28 February 1997*)

Abstract

A probabilistic measure of structural similarity is proposed which takes into account the degree of spatial localization of atoms expressed in atomic displacement parameters.

1. Introduction

Since the physical factors guiding the sequence-dependent formation of three-dimensional macromolecular structures are not fully understood, the structural similarity is generally evaluated by means of superposition techniques and root-mean-square (r.m.s.) distances between pairs of equivalenced atoms (Mizuguchi & Go, 1995). This approach is easy from a computational point of view, but r.m.s. values are often difficult to interpret since they do not have an upper limit and there are no objective thresholds for discriminating similarity and dissimilarity.

Here we present an alternative to r.m.s. which takes into account the degree of spatial localization expressed in the atomic displacement parameter (ADP, referred to hereafter in B units, as is usual in protein crystallography). Although protein ADPs are considerably approximate (Tronrud, 1996) since they are affected by functional restraints as well as by systematic errors in data collection and interpretation, there is a general confidence in them from comparisons between independently refined structures (Glusker, Lewis & Rossi, 1994) and from structure correlation studies (Ringe & Petsko, 1986).

2. Theory and methods

Superposition of two sets of coordinates can be achieved by maximizing the probability that pairs of equivalenced atoms (for example, the C_α atoms of proteins) occupy the same spatial position and similarity can be evaluated with the overall mean probability itself. The probability P_{ij} of identity of two equivalent atoms i and j can be evaluated as the overlap integral of their probability density functions which are derivable from crystallographic ADP B_i and B_j (see *Appendix*)

$$P_{ij} = \frac{32\pi^2}{3} L^3 \left[\frac{\pi}{(B_i + B_j)^3}\right]^{1/2} \exp\left(-\frac{4\pi^2 R^2}{B_i + B_j}\right).$$

Here, R denotes the distance between the two equilibrium positions of atoms i and j, L is the accuracy of crystallographic coordinate determination (~ 0.1 Å for main-chain atoms and about 0.1–0.5 Å for other atoms; Sheldrick, 1996). The probability $P_{\rm all}$ of identity of a set of n pairs of atoms is the weighted average of all P_{ij} values

$$P_{\rm all} = \sum P_{ij} / \sum P_{ij}^{\rm max}.$$

Thus, fragments of the structure with larger B values are down-weighted with respect to highly localized segments. The function $P_{\rm all}$ is nonlinear and can be iteratively maximized by full-matrix least squares after Taylor expansion (like a crystallographic refinement). Therefore, it is possible to refine an initial superposition matrix [for example, obtained by the methods of Kabsch (1978) and McLachlan (1979)] by maximizing $P_{\rm all}$.

3. Results and discussion

The difference between our similarity measure and the simple r.m.s. approach may be quite significant. The standard comparison of the equivalent C_α atoms between actin (1atn) and heat-shock protein 70 (3hsc), two distantly related proteins (Bork, Sander & Valencia, 1992), resulted in an r.m.s. of 3.5 Å and a $P_{\rm all}$ = 19%. Refinements of the initial Kabsch & McLachlan superposition matrices implied supplementary atom repositionings (r.m.s. increased to 3.8 Å) and $P_{\rm all}$ increased to 30%.

Our concept is especially useful for the objective quantification of identity for very similar structures (local rearrangements, comparison of domains). Fig. 1 shows the relationship

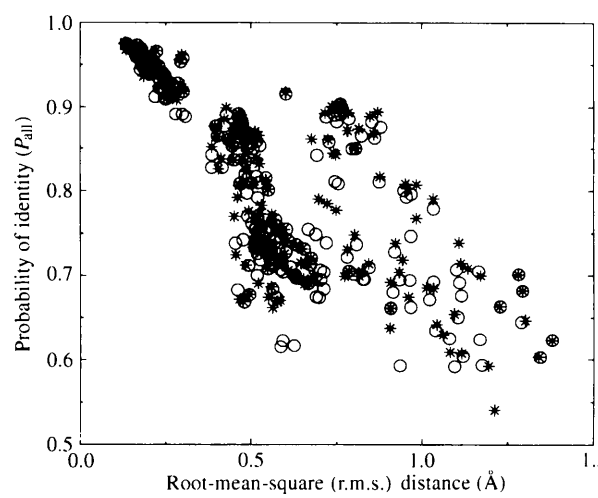

Fig. 1. The relationship between r.m.s. distance and $P_{\rm all}$ for Kabsch–McLachlan (○) and probabilistic (*) superpositions.

© 1997 International Union of Crystallography
Printed in Great Britain – all rights reserved

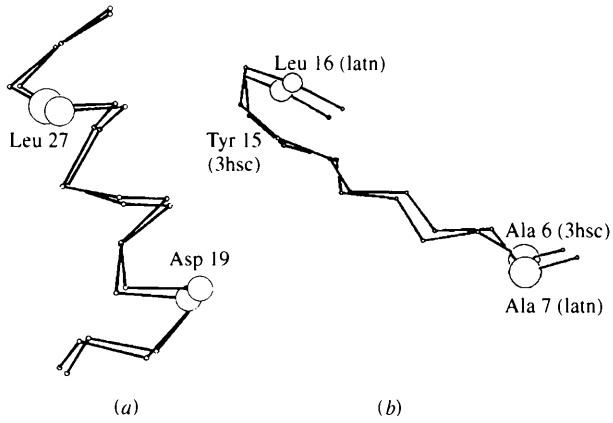

Fig. 2. (*a*) C_α traces of segments 15–30 after superposition of equivalenced pairs of C_α atoms of 2sec onto 1tec. C_α of Asp 19 and Leu 27 are shown at the 50% probability level. Residues are numbered according to 1tec. (*b*) C_α traces of segments 6–17 of 1atn and 5–16 of 3hsc after superposition of equivalenced pairs of C_α atoms. C_α of Ala 7 and Leu 16 of 1atn, and of Ala 6 and Tyr 15 of 3hsc are drawn at the 50% probability level. The figure was prepared using *ORTEP*III (Burnett & Johnson, 1996).

between r.m.s. and P_{all} calculated for all pairs of structures of eglin-C (1acb, 1cse, 1mee, 1tec, 2tec, 2sec, 3tec), ribonuclease A (1rca, 1rbx, 1rha, 1rob, 1rnq, 1rpg, 3rn3, 8rat) and superoxide dismutase (1cob, 1sdy, 1xso). The r.m.s. distances range from very low values (0.1–0.2 Å), usually associated with extreme similarity, to quite high values (1.3–1.4 Å), generally associated with moderate dissimilarity. The r.m.s. values of 0.1–0.2 Å correspond to a P_{all} higher than 90% and r.m.s. values of 1.3–1.4 Å to a P_{all} of only 50–70%.

Fig. 2 illustrates how a probabilistic evaluation of similarity can provide a better insight when comparing local features. C_α atoms of Asp 19 and Leu 27 from 1tec and 2sec are nearly at the same distance after refined superposition (0.8 and 0.7 Å, respectively) but the probabilities of identity are clearly different (40 and 65%, respectively) due to the fact that the ADPs of C_α from Asp 19 (14.2 and 14.6 Å^2 for 1tec and 2sec, respectively) are much smaller than those from Leu 27 (28.6 and 19.7 Å^2 for 1tec and 2sec, respectively). Similarly, the C_α atoms of the pairs Ala 7/Ala 6 and Leu 16/Tyr 15 in the phosphate 1 consensus region of the superposition latn/3hsc (Bork *et al.*, 1992) are nearly at the same distance (0.94 and 1.03 Å, respectively) but have very different probabilities of identity (62 and 28%, respectively) owing to the difference in ADPs [37.5 and 35.9 Å^2 for C_α atoms in Ala 7 (latn) and Ala 6 (3hsc), respectively, and 14.0 and 18.8 Å^2 for C_α atoms in Leu 16 (1atn) and Tyr 15 (3hsc), respectively]. Hence, our approach may also be applicable to the improvement of the three-dimensional alignment of distantly related proteins and the assessment of structural conservation of sequence motif regions.

In conclusion, we think that P_{all} is an appropriate and useful criterion of similarity. In contrast to r.m.s., normalized P_{all} has both upper (1.0) and lower (0.0) limits. The physical meaning of a threshold value discriminating similar from dissimilar structures can be more easily appreciated with P_{all} since it is simply a probability.

APPENDIX

In the isotropic model, the probability density p of finding the atom i at location r_i is described by a Gaussian function

$$p(r_i) = [(2\pi\sigma_i^2)^{1/2}]^{-3} \exp[-(r_i - r_{io})^2/2\sigma_i^2]$$

where $B_i = 2\pi\sigma_i^2$ is the temperature factor and r_{io} is the equilibrium position of atom i. If atoms i and j are equivalent atoms of two structures, the probability P_{ij} of both atoms being in a given volume V is

$$P_{ij} = (2\pi)^{-3}\sigma_i^{-3}\sigma_j^{-3} \iiint_V dV_i \iiint_V dV_j \exp(-f) \quad (1)$$

with

$$f = [(r_i - r_{io})^2/2\sigma_i^2] + [(r_j - r_{jo})^2/2\sigma_j^2]. \quad (2)$$

We will derive a formula for the probability of coincidence of the two atoms, *i.e.* we calculate the probability P_{ij} for small distances l between atoms i and j. It is elementary geometry (but a very tedious task; major help is given below) to show that

$$f = \frac{R^2 + l^2 - 2lR\cos\phi}{2(\sigma_i^2 + \sigma_j^2)} + \frac{r^2}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right)^{-1} \quad (3)$$

where R is the distance between the two equilibrium positions r_{io} and r_{jo}, l is the distance between the two atoms r_i and r_j, and ϕ is the angle between $r_{io}-r_{jo}$ and r_i-r_j) (see Fig. 3 for the definitions of distances and angles). The value r is the distance between two points $O_{io,jo}$ and O_{ij} located on the lines through r_{io} and r_{jo} as well as through r_i and r_j, respectively. The introduction of these two points is a clue to proving the theory. The temperature factors B_i and B_j serve as barycentric coordinates for $O_{io,jo}$ and $O_{i,j}$ *i.e.* the distances $|r_{io}-r_{jo}|$ and $|r_i-r_j|$ are divided by these two points into parts x_{io} and x_{jo} as well as x_i and x_j, respectively, with

$$\begin{aligned} x_{io} &= B_i R/(B_i + B_j) \\ x_{jo} &= B_j R/(B_i + B_j) \\ x_i &= B_i l/(B_i + B_j) \\ x_j &= B_j l/(B_i + B_j). \end{aligned} \quad (4)$$

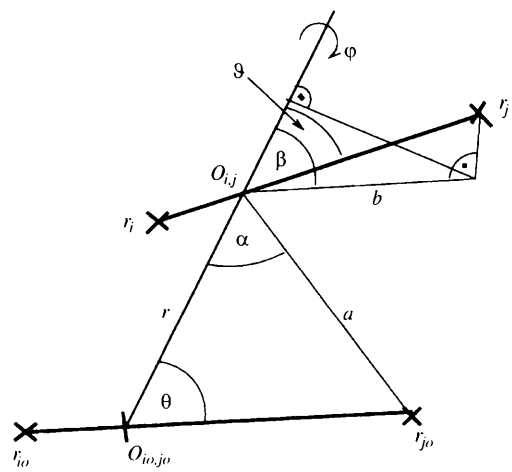

Fig. 3. Geometric definitions of atomic positions, angles and distances.

For example, the distance $|r_{jo}-r_j|$ can now be represented as

$$(r_j - r_{jo})^2 = a^2 + b^2 + 2ab\cos(\alpha+\beta) + x_j^2 \sin^2\vartheta \sin^2\varphi.$$

If we use the following elementary geometric relationships

$$P_{ij}^{\max} = (32\pi^2/3)L^3\left[\pi/(B_i + B_j)^3\right]^{1/2}.$$

and apply the theorem of cosines to the respective triangles, we get the results

$$\begin{aligned} a^2 &= x_{jo}^2 + r^2 - 2rx_{jo}\cos\theta \\ b^2 &= x_j^2(\sin^2\vartheta\cos^2\varphi + \cos^2\vartheta) \\ 2ab\cos\alpha\cos\beta &= 2x_j(r - x_{jo}\cos\theta)\cos\vartheta \\ -2ab\sin\alpha\sin\beta &= 2x_{jo}x_j\sin\theta\sin\vartheta\cos\varphi. \end{aligned}$$

For the distance $|r_{jo}-r_j|$, the following equation can be obtained

$$\begin{aligned} (r_j - r_{jo})^2 = {} & x_{jo}^2 + x_j^2 + r^2 - 2rx_{jo}\cos\theta \\ & + 2rx_j\cos\vartheta - 2x_{jo}x_j\cos\phi \end{aligned} \tag{5}$$

with ϕ being the angle between $r_{io}-r_{jo}$ and r_i-r_j calculated *via*

$$\cos\phi = \sin\theta\sin\vartheta\cos\varphi + \cos\theta\cos\vartheta. \tag{6}$$

A formula similar to (5) is analogously derived for the distance $|r_{io}-r_i|$. After placing both distance expressions into (2), several summations become possible because of the definition of x_{io}, x_{jo}, x_i and x_j as barycentric coordinates. After a few rearrangements (easy for the reader who has followed the arguments to this point), equation (3) is obtained.

The execution of the double volume integration for all atomic positions r_i and r_j with a maximal mutual distance L yields the equation

$$P_{ij} = \left[\frac{1}{2\pi(\sigma_i^2+\sigma_j^2)^3}\right]^{1/2} \exp\left[-\frac{R^2}{2(\sigma_i^2+\sigma_j^2)}\right]$$

$$\times \int_0^L l^2\,\mathrm{d}l \int_{-1}^{l} \exp\left[-\frac{l^2 - 2lR\cos\phi}{2(\sigma_i^2+\sigma_j^2)}\right]\mathrm{d}(-\cos\phi).$$

The limit $L\to\infty$ yields $P_{ij} = 1$, consistent with expectations. For small L, we consider the Taylor series of the exponential function under the integral sign. If all terms containing L in the fourth or a higher degree are ignored, we get

$$P_{ij} = \frac{2}{3}L^3\left[\frac{1}{2\pi(\sigma_i^2+\sigma_j^2)^3}\right]^{1/2} \exp\left[-\frac{R^2}{2(\sigma_i^2+\sigma_j^2)}\right]$$

or the formula in the main text. L should correspond to the accuracy of atomic coordinate determination. In the case of identical structures ($R = 0$), the maximal possible probability is

$$P_{ij}^{\max} = (32\pi^2/3)L^3\left[\pi/(B_i + B_j)^3\right]^{1/2}.$$

This value may be used for normalizing P_{ij}.

The authors thank Shamil Sunyaev for valuable discussions.

References

Bork, P., Sander, C. & Valencia, A. (1992). *Proc. Natl Acad. Sci. USA*, **89**, 7290–7294.

Burnett, M. N. & Johnson, C. K. (1996). *ORTEP*III. Report ORNL-6895. Oak Ridge National Laboratory, Tennessee, USA.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal Structure Analysis for Chemists and Biologists*. New York: VCH Publishers.

Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.

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

Mizuguchi, K. & Go, N. (1995). *Curr. Opin. Struct. Biol.* **5**, 377–382.

Ringe, D. & Petsko, G. A. (1986). *Methods Enzymol.* **131**, 389–433.

Sheldrick, G. M. (1996). *Proceedings of the CCP4 Study Weekend*, pp. 47–58. CCLRC Daresbury Laboratory, Warrington, England.

Tronrud, D. E. (1996). *J. Appl. Cryst.* **29**, 100–104.