Proc. Natl. Acad. Sci. USA
Vol. 75, No. 5, pp. 2180–2184, May 1978
Biochemistry

A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme

(x-ray structure/structure comparison/evolution/protein folding)

STEPHEN J. REMINGTON AND BRIAN W. MATTHEWS*

Institute of Molecular Biology and Department of Physics, University of Oregon, Eugene, Oregon 97403

Communicated by George Streisinger, February 13, 1978

ABSTRACT **A method is proposed that permits the structural similarity between any pair of proteins to be analyzed in a completely general manner. In the proposed procedure, all possible structural segments of a given length from one protein are compared with all possible segments from the other protein. This set of comparisons reveals any structural similarities between the two proteins being compared, and also provides a basis for estimating the probability that a particular degree of structural homology could have occurred by chance. Application of the method to the comparison of T4 bacteriophage lysozyme and carp calcium-binding protein suggests that the previously reported structural similarity between parts of these two proteins [Tufty, R. M. & Kretsinger, R. H. (1975) *Science* 187, 167–169] is no better than would be expected by chance. On the other hand, the structural correspondence between phage lysozyme and hen egg-white lysozyme [Rossman, M. G. & Argos, P. (1976) *J. Mol. Biol.* 105, 75–96] does appear to be significant.**

There has been much discussion recently concerning similarities in protein folding. The reasons are twofold. (*i*) The recognition of frequently occurring patterns of folding ("super secondary structures") will help us to understand the types of interactions that are important in the folding of proteins. (*ii*) The identification of structurally related proteins is important in understanding protein evolution, especially in cases where the amino acid sequences of the proteins being compared have little, if any, detectable homology.

The methods that have been proposed for the comparison of protein folding fall into two categories. First, there are the *topological* comparisons, exemplified by the papers of Schulz and Schirmer (1), Sternberg and Thornton (2), Richardson (3), and Levitt and Chothia (4). In this case, the folding of a protein is represented in a simplified diagrammatic manner that facilitates the recognition and categorization of observed patterns of folding. Such approaches are clearly useful, but are limited in being qualitative rather than quantitative. The second category of structure correlation is one of *spatial* comparison, which has been extensively used for closely related structures (e.g., see refs. 5, 6), and extended by Rossman and coworkers (7–9) to compare structures which are less obviously similar. In Rossmann's procedure, the α-carbon positions of one protein are rotated and translated onto those of another and, by making "insertions" and "deletions" in the respective polypeptide chains, the number of positions "equivalenced" to a specified degree of accuracy is maximized. Again, the procedure is clearly valuable, but, because many insertions and deletions are allowed, and because nonequivalenced positions are ignored, it is difficult to assess the probability that a given structural agreement might have been obtained by chance.

We wish to propose a general method of protein structure comparison which is spatial in nature and which overcomes some of the above limitations. The basic idea is to apply to structural comparisons the statistical technique developed by Fitch (10, 11) for amino acid sequence comparisons. In the proposed method, the structural similarity between two backbone segments of length L residues from the respective proteins is determined by rotating and translating the α-carbon positions of one segment onto those of the other. The resultant minimized root mean square error between the α-carbon positions is used as a measure of the structural homology between the two segments. This is repeated for all possible segments of length L from the respective proteins, and the resultant distribution of the root mean square error is analyzed by statistical techniques to assess the probability that a given degree of structural homology could arise by chance. In effect, the agreement between all of the unrelated segments in the two structures is used as a statistical "control" to calibrate the "goodness" of the best agreement.

METHODS

A Fortran computer program was written to minimize the function $R_{C\alpha}{}^{ij}$ (α, β, γ) with respect to the angles $(\alpha, \beta,$ and $\gamma)$.

$$R_{C\alpha}{}^{ij} = \left[\frac{1}{L}\sum_{k=0}^{L-1}\left| X_{i+k} - A(\alpha, \beta, \gamma)\, X'_{j+k}\right|^2\right]^{1/2}$$

in which X_i and X'_j are, respectively, the coordinate vectors of the ith α-carbon atom and the jth α-carbon atom of the two proteins relative to the center of mass of the L atoms being compared. It is not difficult to show that the minimization of $R_{C\alpha}{}^{ij}$ requires that the centers of mass coincide. The rotation matrix A (α, β, γ) is defined by successive rotations of the angle γ about the Z axis of protein 2, the angle β about the new Y axis, and the angle α about the resultant X axis. Refinement is accomplished by a simple step search in the three angles, allowing each angle to increment only one step per cycle. Refinement is terminated when a single increment of either sign of each of the angles results in no further decrease in the residual. In the vicinity of a minimum, $R_{C\alpha}{}^{ij}$ (hereafter $R_{C\alpha}$) is a slowly varying function of the angles. For a deep minimum, a step size of 0.1 radian yields a final accuracy of about 0.1 Å. The use of the angles α, β, and γ rather than the conventional Eulerian angles results in a faster convergence and less susceptibility to false minima. The use of nonlinear function minimizers was explored [see, for example, Fletcher and Reeves (12)], but these were found to be very slow due to the extreme nonlinearity of the problem. Matrix methods, such as that of MacKay (13), minimize $R_{C\alpha}$ rapidly, but sometimes become ill conditioned when the minimum of $R_{C\alpha}$ is not well defined.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

* To whom reprint requests should be addressed.

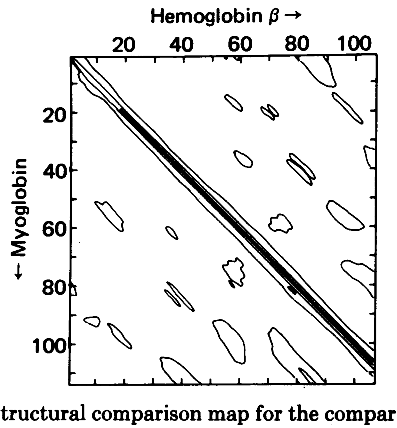

FIG. 1. Structural comparison map for the comparison of myoglobin and hemoglobin β chain with a probe length of 40 residues and step size of 0.1 radian. The contours represent $R_{C\alpha}$ in units of 1 σ (1.9 Å) below the mean value of 8.2 Å for all 40-residue segments.

The program calculates and stores the residual $R_{C\alpha}$ for all possible consecutive segments of length L in the two proteins. Because of overlap between consecutive segments, the angles α, β, and γ generally do not change rapidly, so that time is saved by commencing each refinement with the final angles from the previous comparison. The present method minimizes the residual for $L = 40$, using a step size of 0.2 radian, in about 0.7 sec on a Varian V76 minicomputer with a microcoded floating-point multiply time of about 30 μsec. For two proteins of 200 residues each, the total calculation takes about 5 hr, and is roughly independent of the probe length. This time could, for example, be divided by 4, with little loss of information, by calculating $R_{C\alpha}$ only for every second residue along the sequences of the two proteins.

Once $R_{C\alpha}$ has been calculated for all segments of the two proteins being compared, the results can be conveniently displayed in the form of a contour map, and the distribution of the $R_{C\alpha}$ values can be analyzed by standard statistical procedures. Specific examples are given in the following section.

Protein coordinates for the comparisons were obtained from the Protein Data Bank (14).

APPLICATIONS

In this section, three examples of the proposed structure comparison technique are given. The first example, myoglobin *vs* hemoglobin, compares two structures known to be very similar; the second example, bacteriophage T4 lysozyme *vs* calcium-binding protein, compares two structures that, as will be shown, are not similar; the third example, bacteriophage lysozyme *vs* hen egg-white lysozyme, compares two enzymes that are superfically different but have some structural resemblance.

Sperm Whale Myoglobin versus Hemoglobin β Chain. The myoglobin (15) and hemoglobin β-chain (16) backbones were compared using a probe length of 40 residues; a total of 12,198 40-residue segment alignments. The $R_{C\alpha}$ values are presented in contoured form in Fig. 1. It has been found convenient to calculate the average value and the standard deviation σ of $R_{C\alpha}$, (8.2 Å and 1.9 Å respectively, in this instance), and to contour the map in units of σ below the average value, i.e., with contour levels corresponding to structural agreement of 6.3 Å, 4.4 Å, and 2.5 Å.

Note that the residue numbers in Fig. 1 indicate the start of the 40-residue segments. For this reason the values of $R_{C\alpha}$ terminate 40 residues from the carboxy termini of the two proteins.

The similarity in folding of myoglobin and the hemoglobin β chain is obvious from the strong diagonal feature in Fig. 1, and the striking symmetry of the map about this diagonal.

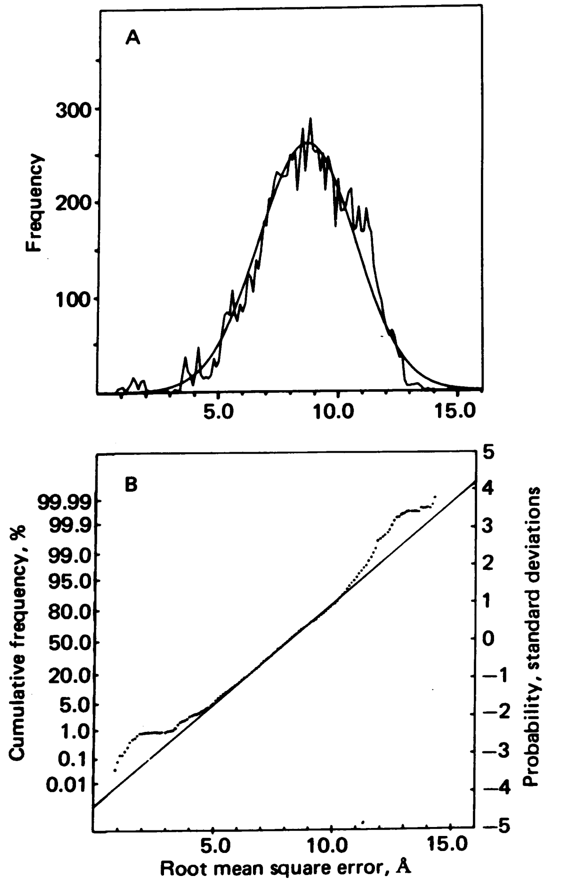

FIG. 2. (*A*) Frequency distribution of $R_{C\alpha}$ for the comparison of myoglobin and hemoglobin β chain, with $R_{C\alpha}$ tallied in cells of 0.1 Å. At a coarser interval the distribution would be much smoother. Probe length, 40 residues; step size, 0.1 radian. (*B*) Frequency distribution, from *A*, plotted as a cumulative distribution. The ordinate shows both the cumulative frequency of $R_{C\alpha}$, expressed as a percentage, and the corresponding probability in units of 1 σ. The straight line is a Gaussian distribution with the same mean and standard deviation as the observed distribution.

The frequency distribution of the $R_{C\alpha}$ values plotted in Fig. 1, tallied in cells of 0.1 Å, is presented in alternative forms in Fig. 2 *A* and *B*. Fig. 2*A* is a plot of the distribution superimposed upon a Gaussian with the same mean and standard deviation. The low values of $R_{C\alpha}$ that occur along the diagonal in Fig. 1 constitute the small peaks between 1 Å and 2 Å in Fig. 2*A*. This can be seen more clearly in a cumulative probability plot of the frequency distribution (Fig. 2*B*) in which the straight line corresponds to a Gaussian distribution. In this plot, the vertical coordinate of each point is calculated as $Q^{-1}\{F(R_{C\alpha})\}$, in which Q^{-1} is the inverse of the standard Gaussian probability integral (17), and $F(R_{C\alpha})$ is the cumulative frequency of $R_{C\alpha}$. The same result is obtained by plotting $F(R_{C\alpha})$ on standard "probability paper" (18). A Gaussian distribution will appear as a straight line with slope σ and zero intercept equal to the mean.

In Fig. 2*B*, the obvious departure from linearity at low $R_{C\alpha}$ values indicates that many more "good" structural agreements occur than would be expected by chance, confirming the known structural homology of myoglobin and the hemoglobin β-chain. The physical meaning of the departure from linearity at high values of $R_{C\alpha}$, seen in Fig. 2*A*, is not completely clear, and needs to be further explored. In the case of small, helical structures,

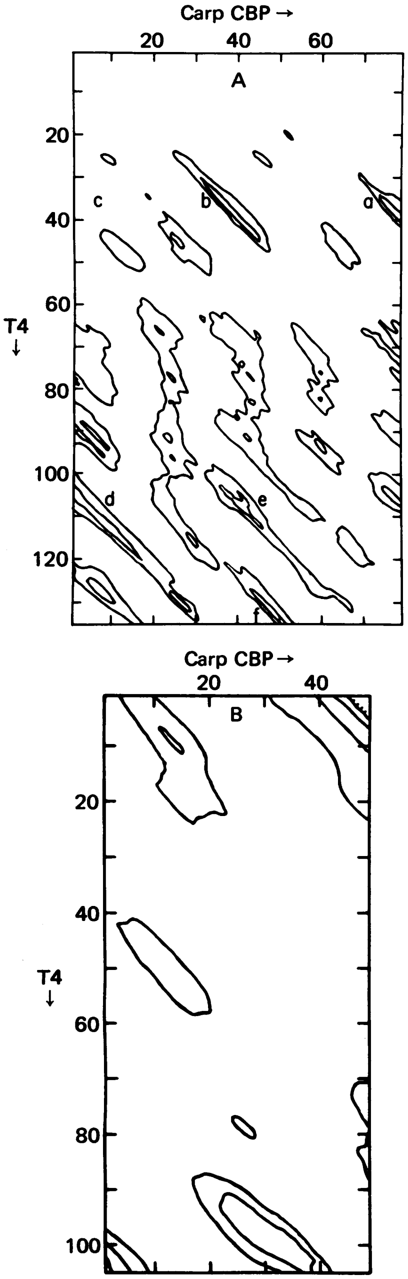

FIG. 3. Structural comparison maps for the T4 lysozyme *vs* carp calcium-binding protein (CBP) with probe lengths of 30 (*A*) and 60 (*B*) residues. Contour intervals are 1.3 Å (1 σ) below the mean of 6.9 Å in the 30-residue map and 1.5 Å below the mean of 10.7 Å in the 60-residue map. The peaks labeled a–f are discussed in the text.

such as hemoglobin and myoglobin, the folding may be more compact than for a random structure, resulting in unusually few "bad" agreements, as indicated in Fig. 2A.

T4 Phage Lysozyme versus Carp Calcium-Binding Protein. Shortly after we had determined the structure of the lysozyme from bacteriophage T4 (19), Tufty and Kretsinger (20) reported a similarity between a portion of the sequence and structure of the lysozyme and calcium-binding parvalbumin from carp muscle. An amino acid sequence comparison technique, based on a scoring scheme of their own design, was used to identify sequences homologous to a 29-residue calcium-binding region of the carp protein. A 29-residue segment of T4 lysozyme scored fairly high by this technique and, using a preliminary set of α-carbon coordinates supplied by us, Kretsinger reported an average (not root mean square) α-carbon–α-carbon distance of 1.9 Å for the 29 compared α-carbon positions in the two proteins. Tufty and Kretsinger went on to predict that T4 phage lysozyme binds calcium and suggested that T4 lysozyme derived from a calcium-binding protein in the host bacterium, which presumably had an origin in common with the carp protein.

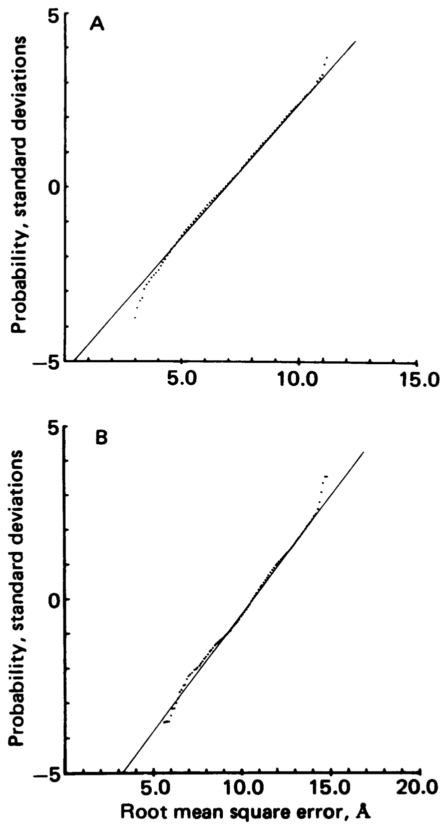

FIG. 4. Cumulative probability plots for the T4 lysozyme–calcium-binding protein comparisons, with $R_{C\alpha}$ tallied in cells of 0.1 Å. Probe lengths were 30 (*A*) and 60 (*B*) residues.

The structures of T4 lysozyme (21) and the calcium-binding protein (22) were compared by calculating comparison maps with probe lengths of 30 and 60 residues (Fig. 3 *A* and *B*). The peak labeled a in Fig. 3*A* corresponds to the structural homology reported by Tufty and Kretsinger (20). Positions b and c correspond to alignment of the same segment of phage lysozyme with the second calcium-binding loop of the carp protein, and with another loop of related structure (23). There are at least three other peaks, labeled d, e, and f, that indicate a higher degree of structural similarity than occurs at a and b. The distributions of the $R_{C\alpha}$ scores are shown in Fig. 4 *A* and *B*, and it can be seen that for both the 30-residue and 60-residue comparison there are, if anything, even fewer good structural agreements than would be expected on the basis of chance. The 60-residue map contains a single 3σ peak in the lower left corner which corresponds to $R_{C\alpha}$ = 5.6 Å for an alignment of the first 60 residues of the calcium-binding protein with the last 60 residues of the phage lysozyme. Although the respective segments align remarkably well, as illustrated in the stereo diagram (Fig. 5), this may be explained as a chance occurrence.

It is concluded that the structure correlation reported by Tufty and Kretsinger (20) for the 29-residue segments of T4 lysozyme and the calcium-binding protein is not unusual, and does not provide compelling support for the hypothesis that the two proteins share a common ancestor.

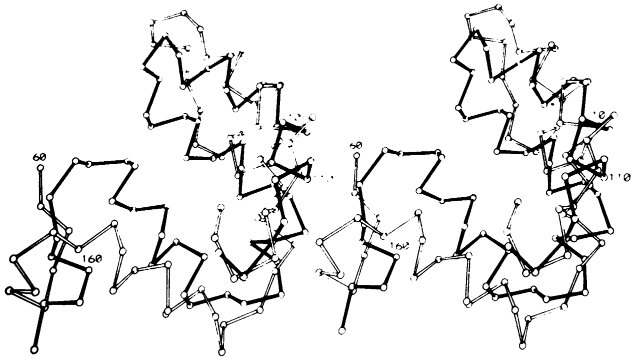

FIG. 5. Stereo diagram of 60-residue structural homology detected between T4 lysozyme and calcium-binding protein. Residues 104–163 of T4 lysozyme (dark bonds) are superimposed on residues 1–60 of the calcium-binding protein. As explained in the text, this structural homology is not statistically significant.

T4 Phage Lysozyme versus Hen Egg-White Lysozyme. The amino acid sequences of the lysozymes from T4 phage and hen egg-white lysozyme have no detectable homology. On first inspection, the three-dimensional structues of the respective enzymes were also thought to be different (19), but Rossman and Argos (8) have shown that there is some similarity in the two molecules.

Systematic structure comparisons of the two lysozymes (21, 24) have been carried out with a number of probe lengths, and plots for $L = 40$ and 80 residues are shown in Fig. 6 *A* and *B*. As one progresses to longer probe lengths the maps become noticeably cleaner and the peaks broader. Peaks due to chance structural similarities of short segments disappear, leaving broad peaks more representative of the overall folding topology. Comparison maps with short probe lengths, e.g., $L = 20$ residues, generally have many peaks due to the superposition of helices and other commonly occurring secondary structures. The general appearance of the maps indicates that at 40 residues or shorter it is not uncommon to find several regions of close structural similarity. At $L = 80$ the comparison plot for the two lysozymes (Fig. 6*B*) has a single peak of height 3.8 σ, indicating that the segments superimposed may have an unusually high degree of structural similarity. This alignment, of residues 1–80 of phage lysozyme with residues 27–106 of hen egg-white lysozyme, is essentially the same as that proposed by Rossmann and Argos (8), and aligns the active site clefts of the enzymes.

The normal probability plots corresponding to Fig. 6 *A* and *B* are shown in Fig. 7 *A* and B. At $L = 40$ residues, the distribution is approximately normal, indicating that structural similarities seen at this probe length are not statistically significant. On the other hand, the overall similarity in folding topology seen in the 80-residue map appears to be an unusual event. The distribution of $R_{C\alpha}$ scores in Fig. 7*B* shows that the observed structural agreement between the two lysozyme structural segments is very unlikely to have occurred by chance.

Thus, our results confirm that there is a significant structural similarity between the two lysozymes. In addition, there are similarities in the modes of binding of substrates to the respective enzymes that tend to strengthen the hypothesis that the enzymes have a common precursor. On the other hand, the constraints imposed by the necessity of binding a large substrate will restrict the protein folding in the region of the active site. Also, there are difficulties with the proposal that the catalytic mechanisms are similar. These considerations will be discussed in detail in a subsequent publication, which will include data from inhibitor-binding studies.

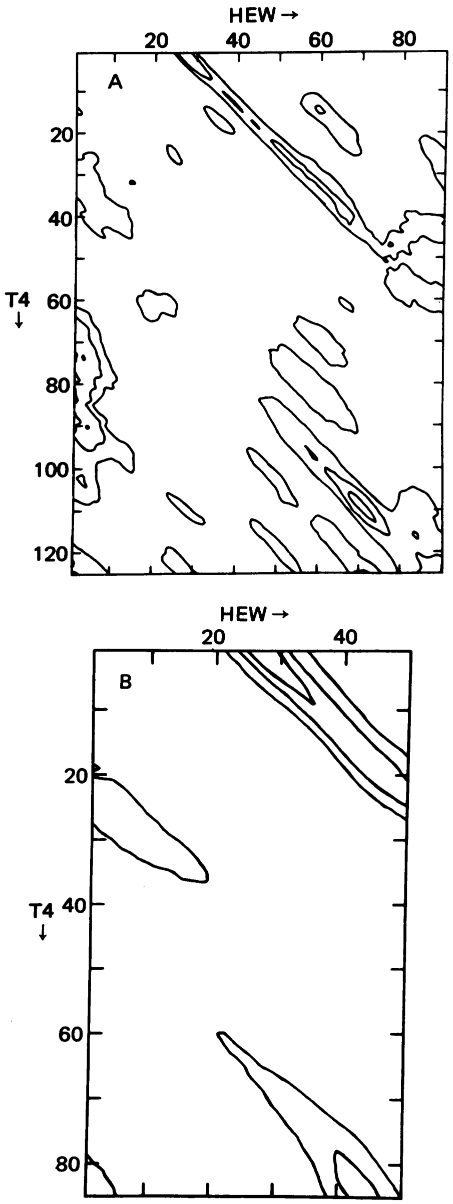

FIG. 6. Structural comparison maps for the T4 lysozyme–hen egg-white (HEW) lysozyme comparison with probe lengths of 40 (*A*) and 80 (*B*) residues. Contour intervals are 1.4 Å (1 σ) below the mean of 8.7 Å for the 40-residue map, and 1.6 Å below the mean of 12.3 Å for the 80-residue map.

DISCUSSION

The proposed method of structure comparison has the potential to help answer a number of questions about protein folding. In the first place, it can be used as an automatic procedure to detect structural similarity between any two proteins. Secondly, it provides a means of estimating the significance of a given structural correspondence. This will be useful not only in comparing structural segments of different proteins, but also in assessing the effectiveness of current attempts to predict the tertiary structures of proteins from their amino acid sequences.

It must be noted that the results we have presented are based on comparisons of a limited number of proteins, and it is not

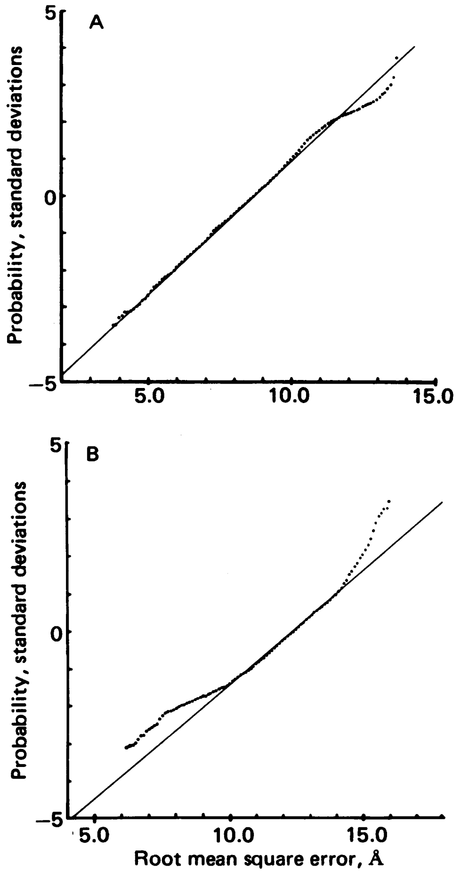

FIG. 7. Cumulative probability plots for the comparison of T4 lysozyme with hen egg-white lysozyme. $R_{C\alpha}$ values are tallied in cells of 0.1 Å. (*A*) Probe length, 40 residues; (*B*) probe length, 80 residues.

yet clear to what extent the method of internal calibration will be generally applicable. The results obtained to date suggest that, for a given probe length, the average value of $R_{C\alpha}$ and its standard deviation are roughly the same, irrespective of the proteins being compared. This also appears to be the case for proteins having a lower helix content than those described here, at least for longer probe lengths.

It will also be necessary to investigate the effect of insertions and deletions in the sequences of proteins being compared. In contrast to methods of amino acid sequence comparison, where insertion of a single amino acid throws the alignment completely out of register, small insertions or deletions will not affect $R_{C\alpha}$ very much, especially for longer probe lengths and, therefore, will be of little consequence in a given structure comparison. On the other hand, larger insertions and deletions will have a more serious effect. While it is possible to generalize the proposed method to allow for insertions and deletions, this might require prohibitively long computing time.

All purely structural comparison methods, including the one proposed here, share the drawback that they cannot distinguish between structural similarity arising from either divergent evolution or convergent evolution. It seems reasonable to propose that close structural correspondence is an indicator of divergent evolution, and that marginal structural similarity is less likely due to divergent evolution, but the boundary between "close" and "marginal" is by no means clear. In principle, the structural comparison technique proposed here can be adapted to estimate evolutionary distance, in the same way that amino acid sequence homology is now used.

In summary, the proposed method of structure comparison appears to work well in a number of test cases. It is of general applicability, and provides an estimage of the significance of any structural similarity that may be detected.

This study was supported in part by the National Institutes of Health (GM20066; GM21967), by the National Science Foundation (PCM77-19310), and by a Career Development Award (GM70585) from the National Institutes of General Medical Science to B.W.M.

1. Schulz, G. E. & Schirmer, R. H. (1974) *Nature* **250**, 142–164.
2. Sternberg, M. J. E. & Thornton, J. M. (1977) *J. Mol. Biol.* **110**, 269–283.
3. Richardson, J. S. (1977) *Nature* **268**, 495–500.
4. Levitt, M. & Chothia, C. (1976) *Nature* **261**, 552–558.
5. Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, Ng. H. (1970) *Biochemistry* **9**, 1997–2009.
6. Huber, R., Epp, O., Steigemann, W. & Formanek, H. (1971) *Eur. J. Biochem.* **19**, 42–50.
7. Rao, S. T. & Rossmann, M. G. (1973) *J. Mol. Biol.* **76**, 241–256.
8. Rossman, M. G. & Argos, P. (1976) *J. Mol. Biol.* **105**, 75–96.
9. Rossmann, M. G. & Argos, P. (1977) *J. Mol. Biol.* **109**, 99–129.
10. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
11. Fitch, W. M. (1970) *J. Mol. Biol.* **49**, 1–14.
12. Fletcher, R. & Reeves, C. M. (1964) *Comput. J.* **7**, 149–154.
13. MacKay, A. L. (1977) *Acta Crystallogr. Sect. A* **33**, 212–215.
14. Bernstein, F. C. , Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M.D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
15. Watson, H. C. (1969) *Prog. Stereochem.* **4**, 299–333.
16. Fermi, G. (1975) *J. Mol. Biol.* **97**, 237–256.
17. Abramowitz, M. & Stegun, I. A., eds. (1965) *Handbook of Mathematical Functions* (Dover, New York).
18. Hald, A. (1952) *Statistical Theory with Engineering Applications* (Wiley, New York), p. 129.
19. Matthews, B. W. & Remington, S. J. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 4178–4182.
20. Tufty, R. M. & Kretsinger, R. H. (1975) *Science* **187**, 167–169.
21. Remington, S. J., Ten Eyck, L. F. & Matthews, B. W. (1977) *Biochem. Biophys. Res. Commun.* **75**, 265–269.
22. Moews, P. C. & Kretsinger, R. H. (1975) *J. Mol. Biol.* **91**, 201–228.
23. Kretsinger, R. H. (1972) *Nature New Biol.* **240**, 85–88.
24. Diamond, R. (1974) *J. Mol. Biol.* **82**, 371–391.