J. Mol. Evol. 9, 339–342 (1977)

Journal of Molecular Evolution

Recognition of Phylogenetic Relationships from Polypeptide Chain Fold Similarities

Georg E. Schulz

Max-Planck-Institut für Medizinische Forschung, Abt. Biophysik
Jahnstraße 29, D-6900 Heidelberg, Federal Republic of Germany

Summary. Structural similarities between proteins with no amino acid sequence homology either indicate a phylogenetic relationship, or they are merely the expression of a physically preferred way of folding a polypeptide chain. It is shown that one can distinguish between these alternatives by evaluating the "significance of the similarity". Such significances have been derived for comparison between chain folds containing β-pleated sheets (Schulz and Schirmer, 1974; Richardson et al., 1976; Sternberg and Thornton, 1976). An extension of this method to comparisons between any two chain folds is outlined here.

Key words: Polypeptide chain fold – Phylogenetic relationship – Significance of a similarity – Mean C_α distance

During evolution the amino acid sequence of a protein changes much faster than its chain fold (Dayhoff, 1972), i.e. the three-dimensional layout of the chain as given by its C_α-coordinates. Consequently, a distant phylogenetic relationship might have been erased in the sequence but remained as chain fold similarity. However, chain folds have to obey stringent conditions: minimizing free energy and being able to fold spontaneously. Therefore, a similarity – as e.g. in an extreme case the similarity between two α-helices – may well be the expression of a favorable structure or the result of convergent evolution to a favorable structure. – In the following I suggest how to recognize those similarities which express a phylogenetic relationship.

The problem can be illustrated and the pertinent definitions introduced by amino acid comparison of two hypothetical proteins of 10 residues each. If we have 2 proteins with 5 residues at any 5 of the 10 positions in common, the "a priori probability" for such coincidence within the ensemble of proteins of 10 residues is

$$\left(\frac{1}{20}\right)^5 \cdot \left(\frac{19}{20}\right)^{10-5} \cdot \binom{10}{5} \simeq 0.00006.$$

Since certain amino acids occur more frequently than others and since not all sequences can form defined spatial structures, in fact a restricted subset of sequences is favored. For the sake of the argument let us assume that these preferences increase the

probability by a factor of 5. As the resulting probability of 0.0003 is corrected for all known contributions to non-randomness (except those caused by phylogenesis) it is called "standard-probability". With a standard-probability of 0.0003 it is very unlikely that the coincidence is random; there exists a strong residual deviation from randomness. In protein structures this points almost certainly to a phylogenetic relationship. However, if not 5 but only 3 residues are in common, the standard-probability increases to 0.05 and the coincidence may well be random.

It is convenient to work with significances instead of probabilities; therefore let us define the "a priori-significance of a structural similarity" as the inverse of the "a priori-probability" as given above, and likewise the "standard-significance". If the standard-significance exceeds a certain level it indicates a phylogenetic relationship. This level has to be established. In statistical tests a confidence level of 1 % is considered as a rather solid base for a hypothesis; hence I suggest to use the corresponding significance level of 100 as an initial estimate.

In the first example the standard significance exceeds this level by far, pointing rather clearly to a phylogenetic relationship. In the second example the standard significance is only 20 and below the level. This does not exclude that the proteins are phylogenetically related, but if such relation exists, it is too distant to be identified. Therefore, the similarity has to be taken as a random event.

However, at a low standard-significance it is also possible that the similarity has arisen by converging evolution of proteins of different origin. Since protein structures have to be stable, such convergence is the more likely the higher the physico-chemical favorization of the structure in question, i.e. the higher the reduction factor from the a priori-to the standard-significance. Consequently, a standard-significance below the level in conjunction with a high a priori significance points to convergent evolution.

We next apply this concept to the comparison of chain folds. Here, a priori-significances for sheet topologies (i.e. the pathway of the chain as referred to a β-pleated sheet without reference to exact coordinates) have been derived for a number of proteins (Schulz and Schirmer, 1974). For instance, a value of 115 000 was found for the similarity between the nucleotide binding domains of any 2 of the 4 structurally known dehydrogenases, if the active site locations were taken into account. Allowing for the preferred handedness of β-strand $-$ α-helix $-$ β-strand units the a priori-significances have been converted to approximate standard significances (Sternberg and Thornton, 1976), yielding 4400 for the example quoted. A better approximation can be obtained if the observed neighbor-correlation in β-sheets (Richardson et al., 1976) is also considered. This results in a standard-significance of about 350 for the given example, which exceeds the level far enough to indicate a phylogenetic relationship between the dehydrogenases. In contrast, the a priori- and standard-significances for the similarity between the sheet topologies of two of Rossmann's *mono*nucleotide binding domains (Rossmann et al., 1974) consisting of 3 β-strands and 2 connections (usually α-helices), are only 12 and about 1.5, respectively. This is too low to indicate a phylogenetic relationship. More likely, such domain is a physico-chemically favored "super-secondary structure" as initially proposed by Rao and Rossmann (1973). A standard-significance has also been determined for the similarity of immunoglobulin and superoxide dismutase sheet topologies (Richardson et al., 1976). With a value of 3000 it clearly indicates a phylogenetic relationship.

Although significances for sheet topologies can be calculated rather easily, they restrict the comparisons to a small group of proteins and, furthermore, they neglect most of the information available in the exact chain geometry. Therefore, the more general method of comparing chain folds by computing mean distances between corresponding C_α-atoms: $<\Delta C_\alpha>$ (McLachlan, 1972) or by evaluating related indices (Rossmann and Argos, 1976) seems to be more appropriate. Such similarity indices, however, have the disadvantage that they (i) do not allow to derive a phylogenetic relationship directly and (ii) they do not take structural preferences (Chothia, 1973; Sternberg and Thornton, 1976; Richardson et al., 1976) into account.

These disadvantages are overcome if a relation between $<\Delta C_\alpha>$ and a priori-significances or even standard-significances can be obtained. In principle, this is possible because the a priori-significance that corresponds to $<\Delta C_\alpha> = d$ found in a given comparison is equal to $M(d)$, the number of geometrically possible chain folds that differ by more than d from each other. Since the conformational space is finite, also this number $M(d)$ is finite. $M(d)$ cannot be derived from the presently known structures, because they are too few. But it can be estimated by chain fold simulations.

For this purpose, the chain length should be restricted to about 120 residues which is the approximate size of a folding unit or domain (Wetlaufer, 1973). Furthermore, a computer program has to be devised which folds this chain to a globule, using a random number generator. After generating G such random globules, $<\Delta C_\alpha>$ values between any pair have to be determined and plotted as a frequency distribution shown schematically in Figure 1. G will be small as compared to $M(d)$, so that the $M(d)$ places of the conformational space will be sparsely occupied and the distribution will obey Poisson statistics. As derived from the frequency distribution there will be $Z(d)$ pairs with $<\Delta C_\alpha> = d$, or random coincidences. Since for Poisson statistics the rate of coincidences $Z(d)/G$ is equal to the average occupation $G/M(d)$, the number of distinct chain folds $M(d)$ can be estimated from $M(d) = G^2/Z(d)$. Thus, the general relation between a priori-significance and $<\Delta C_\alpha>$ can be derived from the frequency distribution. It allows to convert a $<\Delta C_\alpha> = d$ found in any chain fold comparison to the corresponding a priori-significance.

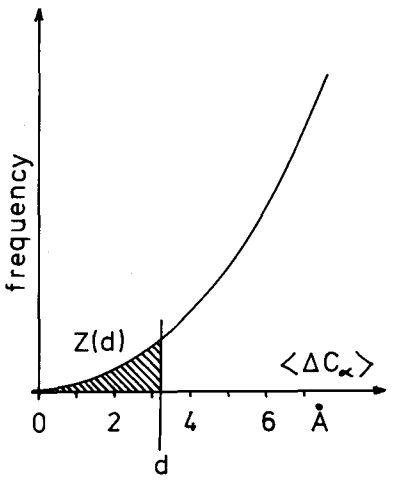

Fig. 1. Schematic frequency distribution of mean C_α-distances as to be derived from comparisons within a group of G randomly simulated chain folds. $Z(d)$ is the integral from zero to d. The a priori-significance $M(d)$ for a mean C_α-distance d found for a given chain fold comparison can be estimated by the formula $M(d) = G^2/Z(d)$

Clearly the necessary $\binom{G}{2} \simeq \frac{G^2}{2}$ comparisons between simulated chain folds will be the bottleneck of the proposed procedure, because for all pairs the best superposition has to be found (McLachlan, 1972; Rossmann and Argos, 1976). However, since no great accuracy is required, one could fix the 3 translations by superimposing the centers of mass and one could go through the 3 rotations, using a very coarse grid. A further gain in speed seems possible if one designs an algorithm that detects quickly whether $<\Delta C_\alpha>$ is larger than about 8 Å, because these cases are of no interest. In order to consider insertions and deletions appropriately one should define $<\Delta C_\alpha>$ as the minimal area stretched between both chains (Fig. 2) divided by the average chain length.

Furthermore, with this method it is possible to allow for structural preferences (Chothia, 1973; Sternberg and Thornton, 1976; Richardson et al., 1976). For this purpose, one can select only those simulated chain folds taking account of these preferences, or one can build these preferences into the chain fold simulation itself. In either case $Z(d)$ will increase for a given G, and $M(d)$ will decrease from the a priori-significance to the standard-significance. With the standard-significance of a given chain fold comparison at hand, one then would be able to recognize phylogenetic relationships in the quantitative manner described above.

Acknowledgement. I thank M.G. Rossmann and K.C. Holmes for discussions.

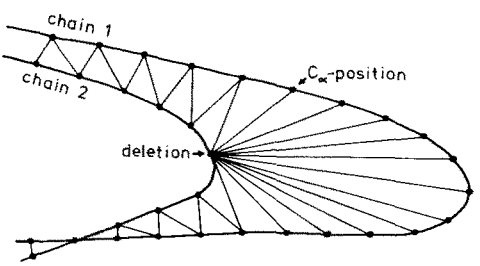

Fig. 2. Proposed handling of insertions and deletions by defining $<\Delta C_\alpha>$ as the minimal area stretched between both chains (corresponding to the area of an elastic film stretched between two wires) divided by the average chain length. The sum of the triangles between C_α-atoms, which are indicated in the sketch, is a good approximation to this minimal area, and easy to calculate

References

Chothia, C. (1973). J. Mol. Biol. **75**, 295–302
Dayhoff, M. (1972). Atlas of protein sequence and structure. Washington, D.C. Natl. Biomed. Res. Foundation
McLachlan, A.D. (1972). Nat. New Biol. **240**, 83–85
Rao, S.T., Rossmann, M.G. (1973). J. Mol. Biol. **76**, 241–256
Richardson, J.S., Richardson, D.C., Thomas, K.A., Silverton, E.W., Davies, D.R. (1976). J. Mol. Biol. **102**, 221–235
Rossmann, M.G., Argos, P., (1976). J. Mol. Biol. **105**, 75–96
Rossmann, M.G., Moras, D., Olsen, K.W. (1974). Nature **250**, 194–199
Schulz, G.E., Schirmer, R.H. (1974). Nature **250**, 142–144
Sternberg, M.J.E., Thornton, J.M. (1976). J. Mol. Biol. **105**, 367–382
Wetlaufer, D.B. (1973). Proc. Natl. Acad. Sci. USA **70**, 697–701

Received August 27, 1976; Revised Version January 10, 1977