J Mol Evol (1994) 39:206–209

Journal
of Molecular
Evolution

Evolution of Protein 3D Structures as Diffusion in Multidimensional Conformational Space

Alexander M. Gutin,* Azat Ya. Badretdinov

Institute of Protein Research, 142292, Pushchino, Moscow Region, Russia

Received: 8 March 1993 / Revised: 20 July 1993

Abstract. A theory of protein spatial-structure evolution in terms of random walks in multidimensional conformational space is proposed. It is shown that the spatial divergence in pairs of homologous proteins depends only on their sequence similarity and is independent of the protein size. X-ray data are reasonably well described in terms of the theory developed.

Key words: Three-dimensional structure—Tertiary structure—Multidimensional diffusion

Introduction

During biological evolution the primary structure of a naturally occurring protein is subjected to various point mutations. Some of them do not lead to frustrating changes in protein folding. Thus in the process of evolution groups of functionally analogous proteins with essential sequence similarity are formed.

We study analytically the effect of sequential point mutations on the protein 3D structure in terms of the root mean square (RMS) deviation from 3D structure of the original protein. We also reveal the relation between the RMS deviation and sequence similarity in pairs of functionally analogous proteins. The main idea of the approach is that the process of protein spatial-structure divergence is considered as diffusion in conformational space, the number of mutations representing the time.

Recently, Chothia et al. (Chothia and Lesk 1986; Lesk and Chothia 1986) have measured the divergence of structures in pairs of proteins vs their sequence homology in terms of RMS deviations. They took into account only those residues that form a "common core" of each pair of proteins. In another article (Hubbard and Blundell 1987) the authors consider the same relationship but establish other rules for the "common core" determination.

We compare these experimental data with the results of our analytical treatment. Our approach results in a reasonably good approximation of the experimental data.

Theory

Consider a protein of N residues. We represent its spatial structure by coordinates of its backbone atoms (x^i, y^i, z^i), where $1 \le i \le 4N$ (where 4 is the number of backbone atoms in each peptide unit). After one accepted point mutation, meaning an amino acid replacement at one position in the protein sequence that does not change dramatically the protein spatial structure, coordinates of protein atoms change. We express the difference between 3D structures of the mutant protein and the original one in terms of the RMS deviation Δ:

$$\Delta^2(1) = \frac{1}{4N} \sum_{i=1}^{4N} [(x_1^i - x_0^i)^2 + (y_1^i - y_0^i)^2 + (z_1^i - z_0^i)^2] \tag{1}$$

where (x_0^i, y_0^i, z_0^i) and (x_1^i, y_1^i, z_1^i) are coordinates of the i-th atom of the backbone of the original protein and those of the mutant protein after one mutation, respec-

* *Present address:* Department of Chemistry, Harvard University, Cambridge, MA 02138, USA
Correspondence to: A.M. Gutin

tively. Assume now that mutations occur randomly and the perturbations caused by different mutations are not correlated. Then the process of natural evolution is equivalent to diffusion in the multidimensional conformational space, the number of mutations, *t*, being equivalent to the time. According to the random walk theory, the following expression describes the RMS deviation between the original protein structure and that of protein after *t* mutations:

$$<\Delta^2(t)> = \Delta_1^2 \cdot t \tag{2}$$

where the effective diffusion coefficient is determined by

$$\Delta_1^2 \equiv <\Delta^2(1)> \tag{3}$$

and the $< \ldots >$ in (2) and (3) means averaging over all possible mutations.

We suppose that each point mutation effects only the positions of spatially nearest residues, their number being independent of the protein size. Thus the effective diffusion coefficient

$$\Delta_1^2 = D/N$$

where *D* is independent of the protein size, and expressions (2) and (3) are transformed into

$$<\Delta^2(t)> = (D/N) \cdot t \tag{2'}$$

and

$$D \equiv <\Delta^2(1)> \cdot N \tag{3'}$$

respectively.

It should be mentioned that relative fluctuations of $\Delta^2(t)$ at large *N* and *t* are $O(1/N) + O(1/t) << 1$.This means that in the case of protein evolution (i.e., diffusion in the multidimensional ($N >> 1$) conformational space) the effect of a large number of random mutations ($t >> 1$) is practically independent of the evolutionary path and depends only on *t*. Therefore we can omit averaging in (2′):

$$\Delta^2(t) \approx <\Delta^2(t)> \tag{4}$$

Since the number of mutations, *t*, is an unobservable quantity, we express it in terms of sequence similarity. Assuming that the subsequent mutations occur independently, we can obtain the fraction P_k of residues, mutated exactly *k* times, to obey Poisson's law:

$$P_k = \frac{\left(\frac{t}{N}\right)^k}{k!} \exp\left(-\frac{t}{N}\right) \tag{5}$$

Hence, the similarity *H*—that is, a fraction of residues, mutated exactly 0 times—is determined by

$$H \equiv P_0 = \exp\left(-\frac{t}{N}\right) \tag{5}$$

Substituting this expression in (2) and using (4), we finally obtain

$$\Delta^2(H) = -D \log H \tag{7}$$

Generally speaking, formula (7) was obtained for two proteins, one of which is a direct precursor of the other. But for each pair of homologous proteins we cannot surely say that one of the proteins is a direct precursor of the other. Let us generalize this formula to include all pairs of homologous proteins.

Consider two homologous proteins with a common precursor. Designate the number of mutations experienced by each of these proteins from their nearest precursor by t_1 and t_2. Then the RMS distance in the selected pair is

$$\Delta^2 = D(t_1 + t_2) = -D \cdot \log(H_1 H_2) \tag{7'}$$

where H_1 and H_2 are similarities between these proteins and their precursor. Since the similarity between proteins $H = H_1 H_2$, formula (7) is valid for these proteins, too.

Note that relation (7) does not contain the protein size parameter *N*. This allows us to compare our analytical conclusions with X-ray data on proteins of different sizes.

Comparison with X-ray Data

To compare the theoretical conclusions with X-ray data, it is necessary to take into account differences in X-ray structures of the same protein, refined from different crystals or from a crystallographic cell, containing several molecules of the protein. Therefore expression (7) is transformed into the following form:

$$\Delta^2(H) = -D \log H + \Delta_0^2. \tag{7''}$$

where Δ_0 is of the order of the RMS deviation between known different X-ray structures of the same protein.

The values of RMS deviations calculated for known X-ray structures are taken from Chothia and Lesk (1986) and Hubbard and Blundell (1987). We do not combine the data of these references, because they use different methods of structure-deviation calculations.

Chothia et al. (Chothia and Lesk 1986) calculated the RMS deviation for backbone nonhydrogen atoms of residues constituting a "common core" in each pair of proteins. This "core" of the structures comprises major elements of secondary structure and residues flanking

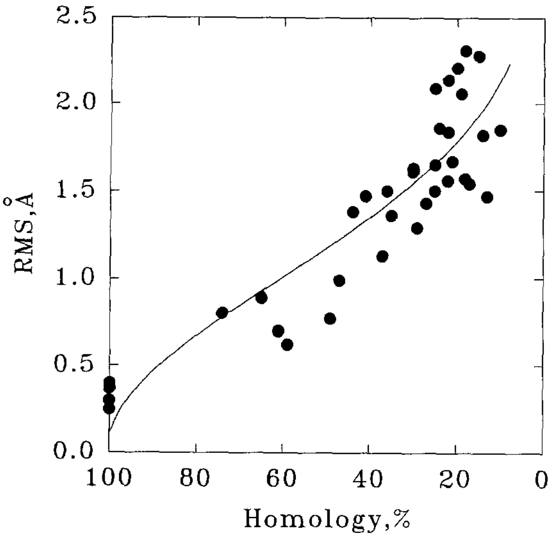

Fig. 1. The relation of primary structure similarity and the RMS deviation of backbone atoms of common cores (*filled circles*, data from Chothia and Lesk 1986; Lesk and Chothia 1986). Best-fitting curve (eq. 7″) shown.

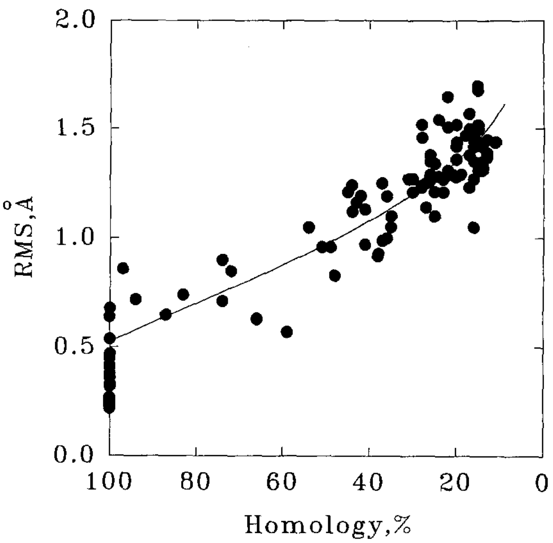

Fig. 2. The same as in Fig. 1 for the data from Hubbard and Blundell (1987).

them, including active-site peptides. For detailed information on the proteins under consideration see Chothia and Lesk (1986); 37 pairs of homologous X-ray structures with a resolution higher than 2.0 Å were investigated by Chothia et al. We unite their data with our best-fit theoretical curve in Fig. 1. Best-fitting parameters are: $D^{1/2}$ = 1.40 Å, Δ_0 = 0.11 Å.

Hubbard et al. (Hubbard and Blundell 1987) use all topologically equivalent residues as a core. They explore eight families of proteins with known X-ray structures, variously resolved. We exclude the data with a resolution lower than 2.0 Å following accepted standards for X-ray data analysis (see, e.g., Chothia and Lesk, 1986). The high-resolution data selected comprise 48 protein structures in 105 pairs. The data are shown in Fig. 2 with the best-fit theoretical curve.

We also calculate the best-fit values of $D^{1/2}$ and Δ_0 for each of the following protein families used by Hubbard et al. (Hubbard and Blundell 1987)—hemoglobins, cytochromes, immunoglobulins and serine proteinases, and miscellaneous β-proteins. The values are shown in Table 1.

As seen from Figs. 1 and 2, the RMS distances correlate reasonably well with the respective sequence similarities. This fact lends strong evidence for two theoretical conclusions: The RMS distances between two homologous proteins (1) do not depend on their size and (2) depend only on the similarity between the proteins. Point 1 is also confirmed in Table 1, where values of $D^{1/2}$ for different families are compared with mean protein sizes.

The differences in values of $D^{1/2}$ may reflect particular features of structures in each family of proteins and differences in RMS distances determination in different sources. The same may be said for the values of Δ_0.

It is extremely important to compare the calculated values of $D^{1/2}$ with those obtained from independent studies. We utilize X-ray data on various point mutants of bacteriophage T4 lysozyme (Matsumura et al. 1988, 1989; Gray and Matthews 1987; Faber and Matthews 1990; Nicholson et al. 1988; Alber et al. 1987) and the RMS deviation in coordinates for main-chain atoms between the wild-type and mutant proteins. Thus, the value of $D^{1/2}$ for mutant I3V (Matsumura et al. 1989, designation from the original article) is 1.3 Å; for I3Y (Matsumura et al. 1989) this quantity is 2.8 Å; for the Asp199 → Asn mutant of chloramphenicol acetyltransferase (Gibbs et al. 1990) this quantity is 5.9 Å. Note that in many cases (Gray and Matthews 1987; Faber and Matthews 1990; Nicholson et al. 1988; Alber et al. 1987; Matsumura et al. 1988) the RMS deviations are not published for the entire structure. In these cases, as a basis of evaluation, we assumed shifts of single atoms (Nicholson et al. 1988), residues (Alber et al. 1987), or elements of secondary structure (Matsumura et al. 1988) in the vicinity of a modified site and evaluated the overall RMS deviation, neglecting all other atoms shifts. Such a comparison shows that the $D^{1/2}$ values from homologous proteins and point mutants are equal in the order of magnitude, though the former values are a little smaller than the latter ones.

Discussion

In this paper we have developed an analytical theory explaining the natural evolution of protein 3D structures in terms of diffusion in multidimensional conformational space. According to the conclusions of our analytical treatment, the RMS distances in pairs of similarity proteins depend only on the degree of homology in pairs and do not depend on the size of the proteins

Table 1. Calculated parameters of diffusion (see (7″)) for different families of homologous proteins (RMS calculations from Hubbard and Blundell 1987)

	$\sqrt{D}$, Å	Δ_0, Å	N
Hemoglobins	1.01	0.58	~140
Immunoglobulins	0.87	0.72	~110
Serine proteinases	0.99	0.38	~210
Miscellaneous β-proteins	1.17	0.42	—
Total	0.99	0.52	—

involved. The available X-ray data lend strong evidence that naturally occurring homologous proteins obey this law reasonably well, in spite of great heterogeneity of protein families and the sizes of the proteins involved (Table 1). The calculated values of "diffusion" constants in the order of magnitude coincide with available RMS distances between proteins differing by one mutation.

Note that the following essential features of proteins are not included in the present consideration. First, all backbone atoms of a protein are linked into a chain, and therefore cannot diffuse independently. However, the effect of joined diffusion becomes significant at distances much greater than the length of backbone bonds not included in the data set studied. Second, another effect—namely, the existence of spatial boundaries of diffusion—arises from the finite size of a protein molecule. This means that there is an upper limit for an RMS distance between proteins. This upper limit reveals itself at low sequence similarities and can also contribute to the independence of the protein size in this region. As was mentioned, the data at low similarities are outside the scope of this article. Thus, our theoretical conclusion concerning the dependence on the protein size remains valid in the similarity region under consideration.

Third, we neglect reverse mutations. In other words, there is a nonzero minimal sequence similarity between *any* proteins. The conservation of some residues (and their mutual conformation) through the evolution (for example, in active sites of enzymes) enlarges the minimal sequence similarity. The minimal similarity value caused by reverse mutations cannot differ significantly from $1/L$, where $L = 20$ is the number of naturally occurring amino acids, and the one caused by conservation of functionally significant residues cannot be much greater. So, this effect may be revealed at low similarities which are not included in the experimental set of pairs. Therefore the effect of minimal similarity cannot be seen from the available experimental data.

Acknowledgments. We thank Prof. O.B. Ptitsyn, Dr. A.V. Finkelstein, and Dr. A.G. Murzin for fruitful discussions and useful comments.

References

Alber T, Dao-pin S, Wilson K, Wozniak JA, Cook SP, Matthews BW (1987) Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. Nature 330:41–46.

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826.

Gibbs MR, Moody PCE, Leslie AGW (1990) Crystal structure of the aspartic acid-199 → asparagine mutant of chloramphenicol acetyltransferase to 2.35-Å resolution: structural consequences of disruption of a buried salt bridge. Biochemistry 29:11261–11265.

Gray TM, Matthews BW (1987) Structural analysis of the temperature-sensitive mutant of bacteriophage T4 lysozyme, glycine 156 → aspartic acid. J Biol Chem 262:16858–16864.

Faber HR, Matthews BW (1990) A mutant T4 lysozyme displays five different crystal conformations. Nature 348:263–266.

Hubbard TJP, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling. Prot Eng 1:159–171.

Lesk AM, Chothia CH (1986) The response of protein structures to amino-acid sequence changes. Philos Trans R Soc Lond A 317: 345–356.

Matsumura MS, Becktel WJ, Matthews BW (1988) Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitution of Ile 3. Nature 334:406–410.

Matsumura M, Wozniak JA, Dao-pin S, Matthews BW (1989) Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. J Biol Chem 264:16059–16066.

Nicholson H, Becktel WJ, Matthews BW (1988) Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. Nature 336:651–656.